Marie Davidian · Xihong Lin
Jeffrey S. Morris · Leonard A. Stefanski
Editors

The Work of Raymond J. Carroll

The Impact and Influence of a Statistician

Springer

The Work of Raymond J. Carroll

Marie Davidian • Xihong Lin
Jeffrey S. Morris • Leonard A. Stefanski
Editors

The Work of Raymond J. Carroll

The Impact and Influence of a Statistician

Springer

Editors

Marie Davidian
Department of Statistics
North Carolina State University
Raleigh, NC, USA

Xihong Lin
Department of Biostatistics
Harvard School of Public Health
Boston, MA, USA

Jeffrey S. Morris
Department of Biostatistics
The University of Texas
MD Anderson Cancer Center
Houston, TX, USA

Leonard A. Stefanski
Department of Statistics
North Carolina State University
Raleigh, NC, USA

ISBN 978-3-319-05800-9 ISBN 978-3-319-05801-6 (eBook)
DOI 10.1007/978-3-319-05801-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939953

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Ray

MICHIGAN

THE NORTH FACE

Preface

Raymond J. Carroll's impact on statistics and numerous other fields of science is far-reaching and substantial. His vast catalog of work spans the spectrum from fundamental contributions to statistical theory to innovative methodological development to new insights in a number of subject matter areas. From the outset of his career, rather than taking the "safe" route of pursuing incremental advances, Ray has focused on tackling the most important statistical research challenges of our time, and in doing so it is fair to say that he has literally shaped and defined a host of areas of statistics, including weighting and transformation in regression, measurement error modeling, quantitative methods for nutritional epidemiology, and non- and semi-parametric regression. It is indisputable that Ray is one of the giants of the field, and we are honored to have had the opportunity to prepare this volume, which highlights some of his most influential work.

The book is organized into seven main parts, each focused on a key area in which Ray has made significant contributions. The seven subject areas reviewed in this book were chosen by Ray himself, as were the articles representing each area. Each part is focused around these key papers, and, for each, we asked distinguished researchers in the area to provide a commentary giving insight into not only the significance of the featured papers but also on Ray's impact on the area more broadly. The commentaries not only review Ray's work, but they also are filled with history and anecdotes that reflect the fact that Ray is also a really nice guy! Indeed, as former students and collaborators of Ray, we are pleased that the personality, generosity, friendship, and enthusiasm we know so well emerge throughout all of the commentaries, whose authors have almost all had the pleasure of working with Ray firsthand as we have. We are deeply grateful to these contributors, whose thoughtful, insightful commentaries provide an inspiring roadmap to Ray's achievements. Due to their extraordinary efforts, this book is a fitting tribute to a scholar and educator whose influence on not only science but also on the individual students, postdocs, and junior colleagues he has mentored is legendary.

Our elation with the authors who contributed their insights into Ray's work and personality is tempered by the death of George Casella. George provides an entertaining overview of Ray's work in a hodgepodge of "Other" areas. He was both a close friend and colleague of Ray. We are grateful that George was able to contribute his personal reflections before his passing.

Putting together this volume was made even easier by Ray himself, and we cannot thank him enough. He provided us with extensive materials, including not only the list of articles around which the book is focused but also a detailed narrative of his own thoughts on his work, his biography, and other resources.

We would also like to acknowledge Jennifer Moy, a student at North Carolina State University, whose assistance in preparing Ray's complete bibliography was invaluable.

At the beginning of each commentary, the articles included in this volume that form the basis for the commentary are listed and are identified by acronyms in brackets; for example, "MEM" for Measurement Error Models." The second number in brackets is the number of citations reported by Google Scholar at the time Ray compiled the list (2011).

Of course, a book devoted to the contributions of Raymond Carroll cannot possibly provide a full accounting of his work. Despite approaching the start of his fifth decade as a researcher, Ray has not slowed his pace one bit, and he continues to produce and inspire and mentor students and postdocs unabated. We fully expect to be called upon to put together "Volume 2," featuring still other areas Ray has already influenced and forthcoming contributions in areas that have yet to be defined.

Raleigh, NC — Marie Davidian
Boston, MA — Xihong Lin
Houston, TX — Jeffrey S. Morris
Raleigh, NC — Leonard A. Stefanski
December 2012

Contents

Biography of Raymond J. Carroll

Raymond J. Carroll is Distinguished Professor of Statistics, Nutrition, and Toxicology at Texas A&M University, where he has been on the faculty since 1987. He was the first statistician ever given a Method to Extend Research In Time (MERIT) Award from the National Cancer Institute (NCI) of the National Institutes of Health (NIH), receiving this honor for his seminal contributions to statistical methodology and the impact of that methodology on public health. He is the principal investigator of an NCI-funded Bioinformatics training program and is the founding director of the Texas A&M Center for Statistical Bioinformatics. He is also the Director for the Texas A&M Institute of Applied Mathematics and Computational Science (http://iamcs.tamu.edu).

Raymond Carroll was born April 21, 1949 in Yokohama, Japan, into an Irish Catholic military family, and he is the eldest of five siblings. His father, who spent the Second World War in India and China, was transferred successively from Yokohama to Nagoya, Japan, Washington DC, Wichita Falls, Texas, Ramstein, Germany, Wichita Falls, Omaha, Nebraska, and Seoul, Korea, and finally retired from his last assignment in Wichita Falls. He is married to Marcia Ory. A memorial tree with a plaque honoring the memory of his parents Regina and Norman is situated in the heart of the central campus a few feet southwest of “Sully,” a bronze statue of the first president of Texas A&M University. Three other memorial trees are adjacent, two honoring the memories of his father-in-law, mother-in-law, and brother-in-law, and the other honoring the memory of Don Risner, a good friend and fishing guide from North Texas. Raymond attended high schools in Germany, Texas, and Nebraska. He graduated from the University of Texas at Austin in 1971 with a BA in mathematics and was especially influenced by courses in analysis and measure theory given by E. W. Cheney and G. W. Stewart, respectively. He received his PhD in Statistics from Purdue in 1974 under the direction of Shanti Gupta, with wonderful advice from Leon Gleser. He has held positions at the University of North Carolina at Chapel Hill and the University of Pennsylvania. He has published over 350 papers and given over 300 invited talks. The peripatetic nature of his childhood has made him an avid traveler, a characteristic not shared by his siblings. Since his first invitation to Australia in 1987, he has visited that country over 20 times, and he has visited Germany, the site of two of his sabbaticals, nearly yearly since 1980. He is in addition a bad golfer who takes mulligans liberally and a mediocre although enthusiastic fly fisherman.

Dr. Carroll is one of the world's foremost experts on problems of measurement error, data transformation, and nonconstant variation, and more generally on statistical regression modeling. His work has found application in a broad variety of fields, including marine biology, laboratory assay methods, econometrics, epidemiology, molecular biology, and many others. He has served as Editor of *Biometrics*, a journal of the International Biometric Society, and as Editor of the *Journal of the American Statistical Association* (*JASA*) Theory and Methods section. He has won many honors in the profession, including the two major research awards. The first is the 1988 Committee of Presidents of Statistical Societies (COPSS) Presidents' Award, given annually by five major statistical societies to the outstanding statistician under the age of 40. Secondly, he gave the COPSS Fisher Lecture at the 2002 Joint Statistical Meetings, an award given by these statistical societies in honor of a senior statistician "whose research has influenced the theory and practice of statistics."

Carroll's work is characterized by a combination of deep theoretical advances, innovative methodological development, and close contact with science. His first seminal contribution to statistical methodology was to create methods for the analysis of data with nonconstant variation; these methods being the transform-both-sides method for nonlinear regression (together with David Ruppert) and the variance function estimation approach (with Marie Davidian), both still in wide use. This work developed from two projects, one on marine fisheries where he worked with a team investigating how to model and manage the menhaden fishery in the Atlantic, and the other project involving immunoassays at Eli Lilly and Company. In the early 1990s, with the inspiration of his close friend Mitchell Gail, he developed a deep interest in epidemiologic case–control studies that led to his receiving the George W. Snedecor Award from COPSS in 1997 for work in this area (together with Bruce Lindsay and Katherine Roeder). The span of his scientific work is amazing, including among many others (a) modeling ozone exposure in Houston (the 1997 *JASA* Applications Editor's Invited Paper); (b) understanding the effects of diet on breast cancer; and (c) discovering interactions between genes and the environment (with Nilanjan Chatterjee and Yi-Hau Chen).

Carroll is no doubt most well known for his work in the area of nonlinear measurement error modeling, with applications to nutritional and radiation epidemiology. The body of seminal research is of such depth, and of such importance, that at the International Biometric Conference in 2000 in Berkeley, Scott Zeger described him as the "grandfather" of measurement error modeling. His 1995 book and 2006 second edition with David Ruppert, Len Stefanski, and Ciprian Crainiceanu is the standard reference in the field. This work began with his landmark 1984 paper in *Biometrika* on measurement error in the binary regression framework and has continued to the present. He was the first to suggest the use of likelihood methods in the nonlinear measurement error context. Along with Len Stefanski, he developed the theory for and coined the name for regression calibration, the most commonly used method in nutritional epidemiology. His 1987 paper with Stefanski developed the method of conditional score function. His 1990 paper with Stefanski and his 1988 paper with Peter Hall on deconvolution established the theoretical basis showing how difficult it really is to understand latent variable distributions: this result provides the theoretical underpinnings for the semi-parametric approaches in measurement error models that have become increasingly popular. The deconvolution area has become of great importance and interest, and even 20 years later the papers have led others into the area. Carroll continues to produce important ideas, and his work continues to influence others, in such important problems as mixed models, segmented regression, instrumental variables, and nonparametric regression. More recently, he has written papers on reanalysis of important radiation epidemiology studies to account for measurement error, both in *Biometrics*.

Carroll's work on measurement error modeling is also one of the landmark works in nutritional epidemiology. He helped design the NCI-AARP Diet and Health Study, the first study to confirm a link between fat in diet and breast cancer. He was the senior author on the first major biomarker study (the OPEN Study) to understand how well common instruments such as the food frequency questionnaire actually measure diet. This study was funded because of the methodological developments done together in what is now a long collaboration with Laurence Freedman, Victor Kipnis, and Douglas Midthune suggesting that the heart of the problem of null studies was the instruments themselves.

Dr. Carroll has worked with many researchers from around the world, but no doubt his closest collaboration has been with David Ruppert, now of Cornell University. They were next door office neighbors at the University of North Carolina from 1977 to 1987, where they started their original collaboration, and they have written over 45 papers in addition to 4 books. Other colleagues with whom he has written 10 or more papers include Mitchell Gail, Victor Kipnis, and Douglas Midthune of the National Cancer Institute; Peter Hall of the University of Melbourne; Len Stefanski of North Carolina State University; Laurence Freedman of the Gertner Institute in Israel; Naisyin Wang of the University of Michigan; Joanne Lupton, Nancy Turner and Robb Chapkin, nutritionists at Texas A&M; Xihong Lin of Harvard; and Bani Mallick of Texas A&M.

More recently, Dr. Carroll has developed a deep interest in basic molecular cell biology and how it relates to nutrition and colon carcinogenesis. His research grants include as co-investigator Dr. Joanne Lupton (the endowed Professor of Human Nutrition at Texas A&M) and Dr. Nancy Turner. This work includes papers both in biology journals and in *JASA* and *Biostatistics*, with many more papers under development. Carroll is involved to the point of generating his own biological hypotheses, suggesting new ways of measurement, and providing support so that novel measurements can be undertaking to understand molecular pathways. More recently, this close work with biologists and electrical engineers has led to the establishment of an NCI-funded training program in Biostatistics and Bioinformatics, for which Dr. Carroll has been the principal investigator since 2001, and the program has recently been renewed until 2016. The program is unique because it aims to train statisticians and electrical engineers in biology and includes mentors from biological fields.

Dr. Carroll is an inspirational teacher and a major innovator for the Department's teaching program. In the 1990s he introduced the use of the computer and class projects into STAT 302, an undergraduate course aimed at life science students. Similarly, since 2000, in STAT 651 he was the first non-distance education expert to create a distance course, something now routine in the department. Dr. Carroll has won a College of Science Teaching Award, and he has graduated 35 PhD students, many of whom are leading figures in academia and industry. He has also been the mentor to many faculty members around the USA, including many who are now full professors, and he is legendary for his willingness to give advice and technical assistance.

PhD Students of Raymond J. Carroll

Name	Institution	Year
Gordon Johnston	University of North Carolina at Chapel Hill	1979
Paul Gallo	University of North Carolina at Chapel Hill	1981
David Giltinan	University of North Carolina at Chapel Hill	1983
Len Stefanski	University of North Carolina at Chapel Hill	1983
Doug Simpson	University of North Carolina at Chapel Hill	1985
Marie Davidian	University of North Carolina at Chapel Hill	1986
Stena Kettl	University of North Carolina at Chapel Hill	1987
Yin Yin	University of North Carolina at Chapel Hill	1988
Lie Ju Hwang	Texas A&M University	1990
Jungsywan Sepanski	Texas A&M University	1991
Rick Landin	Texas A&M University	1992
C. Y. Wang	Texas A&M University	1993
Ron Knickerbocker	Texas A&M University	1993
Bobby Gutierrez	Texas A&M University	1995
Stephen Eckert	Texas A&M University	1995
Jeff Maca	Texas A&M University	1997
Christian Galindo	Texas A&M University	1998
Steve Iturria	Texas A&M University	1998
Jeffrey S. Morris	Texas A&M University	2000
Hua Liang	Texas A&M University	2001
Inyoung Kim	Texas A&M University	2002
Chan Hee Jo	Texas A&M University	2003
Tanya Apanasovich	Texas A&M University	2004
Gosia Leyk	Texas A&M University	2004
Christie Spinka	Texas A&M University	2004
Veera Baladandayuthapani	Texas A&M University	2005
Yehua Li	Texas A&M University	2006
Iryna Lobach	Texas A&M University	2006
Bo Li	Texas A&M University	2006

Name	**Institution**	**Year**
Lian Liu	Texas A&M University	2007
Arnab Maity	Texas A&M University	2008
Seokho Lee	Texas A&M University	2009
Andrew Redd	Texas A&M University	2010
Jiawei Wei	Texas A&M University	2010
Saijuan Zhang	Texas A&M University	2010
Trijya Singh	Texas A&M University	2011
Xiaolei Xun	Texas A&M University	2012

Contributors

John P. Buonaccorsi
University of Massachusetts, Amherst, MA, USA

George Casella
University of Florida, Gainesville, FL, USA
(deceased)

Aurore Delaigle
University of Melbourne, Melbourne, Australia

Laurence (Larry) Freedman
Gertner Institute for Epidemiology, Tel Hashomer, Israel

Mitchell H. Gail
National Cancer Institute, Bethesda, MD, USA

Roger Koenker
University of Illinois, Urbana-Champaign, IL, USA

Yehua Li
Iowa State University, Ames, IA, USA

Hua Liang
The George Washington University, Washington, DC, USA

Dale L. Preston
Hirosoft International, Eureka, CA, USA

David Ruppert
Cornell University, Ithaca, NY, USA

Douglas Simpson
University of Illinois, Urbana-Champaign, IL, USA

Naisyin Wang
University of Michigan, Ann Arbor, MI, USA

Chapter 1
Measurement Error

By John P. Buonaccorsi and Aurore Delaigle

About the Authors. John Buonaccorsi is Professor of Mathematics and Statistics at the University of Massachusetts, Amherst. He received his PhD from Colorado State University in Statistics in 1982. He has been at the University of Massachusetts ever since, including participation in the University's Statistical Consulting Center for twenty years. His original research interests were in optimal experimental design, estimation of ratios, and calibration, followed by a focus on measurement error over the last 25 years. He is the author of the recently published book, *Measurement Error: Models, Methods and Applications*. He also publishes extensively in ecology, with a recent emphasis on temporal data. John has a long-standing collaboration with colleagues in the Medical School at the University of Oslo, focusing on the use of measurement error methods in epidemiology. His relationship with Ray dates back to two early conferences dedicated to measurement error—the 1989 National Institutes of Health workshop, and the 1990 AMS-IMS-SIAM conference at Humboldt State University. John and Ray have been in regular contact ever since.

Aurore Delaigle is Professor and Queen Elizabeth II Fellow, Department of Mathematics and Statistics, University of Melbourne. She received her PhD from the Université Catholique de Louvain (UCL) in Belgium, on the topic on nonparametric measurement error problems. Ray heard about her thesis during a visit at UCL, and later invited Aurore to visit Texas A&M, when she was an Assistant Professor at the University of California, San Diego. Ray visits Melbourne every year, often resulting in a measurement error paper jointly written by him, Aurore Delaigle, and Peter Hall.

Selected Papers on Measurement Error

[MEM-1]-[161] Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71, 19–25.

[MEM-2]-[163] Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics*, 13, 1335–1351.

[MEM-3]-[26] Carroll, R. J., Gallo, P. P. and Gleser, L. J. (1985). Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association*, 80, 929–932.

[MEM-4]-[145] Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74, 703–716.

[MEM-5]-[303] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184–1186.

[MEM-6]-[239] Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 165–184.

[MEM-7]-[193] Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652–663.

[MEM-8]-[86] Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993). Case-control studies with errors in predictors. *Journal of the American Statistical Association*, 88, 185–199.

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_1,
© Springer International Publishing Switzerland 2014

[MEM-9]-[61] Wang, N., Lin, X., Gutierrez, R. G,. and Carroll, R. J. (1998). Generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93, 249–261.

[MEM-10]-[81] Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression with errors in covariates. *Biometrika*, 86, 541–554.

[MEM-11]-[85] Liang, H., Härdle, W., and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Annals of Statistics*, 27, 1519–1535.

[MEM-12]-[81] Berry, S. A., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97, 160–169.

It is both a privilege and a challenge to summarize Ray Carroll's contributions in measurement error. Ray literally wrote the book on the topic with coauthors David Ruppert, Len Stefanski, and Ciprian Crainiceanu (Carroll et al., 2006), and his fingerprints are present in a huge amount of published research on measurement error over the past 30 years. In addition to the book, Ray has authored or coauthored close to 100 papers involving measurement error alone, addressing a vast array of problems. His work covers models from the fairly simple to the very complex with an emphasis ranging from the relatively applied to the highly theoretical. Our detailed discussion of Ray's work concentrates heavily on the twelve papers appearing in this volume, although this only scratches the surface of his contributions. We first discuss parametric models ([MEM-1]-[MEM-4] and [MEM-7]-[MEM-9]), then turn to non-parametric and semi-parametric models including deconvolution problems ([MEM-5],[MEM-6],[MEM-10]-[MEM-11]).

Parametric Models

To put Ray's early work in context, it is worth setting the stage a bit. Prior to the early 1980s, measurement error (or "errors-in-variables" as it was known prior to the 1980s) had attracted a fair amount of attention in both the statistical and econometrics literature. The focus until then was heavily on linear problems under what is now referred to as classical measurement error (independent additive error with mean zero and constant variance). Further, the emphasis leaned towards identifiability issues and correction methods based on knowledge about the measurement error variances (either known or estimated via replication) or assumptions about functions of them. Carroll et al. (1984 [MEM-1]) and Stefanski and Carroll (1985 [MEM-2]) are among the earliest papers to expand the treatment of measurement error in a number of practically useful directions, primarily in: (i) moving away from linear models for the true values and (ii) handling more complex nonadditive measurement error models and accommodating different types of data for estimation of the measurement error model. Of particular interest at that time was the treatment of measurement error in predictors in nonlinear models in general, and binary regression in particular. A good sense of these new directions can be gleaned from the proceedings of the workshop on errors-in-variables held at the National Institutes of Health (Byar and Gail, 1989) and the AMS-IMS-SIAM conference held at Humboldt State University (Brown and Fuller, 1990). One of us [JB] had the pleasure of attending both of these conferences. Ray was an important presence at both. His early work during this period, along with coauthors, blazed a number of new trails.

While measurement error in linear models poses significant challenges (see Fuller, 1987) the treatment of nonlinear models calls for certain fundamentally different approaches. Carroll et al. (1984 [MEM-1]), Stefanski and Carroll (1985 [MEM-2]), and later, Carroll, Gail, and Lubin (1993 [MEM-8]) deal explicitly with binary regression models. By the 1980s these models were in widespread use, especially in epidemiology, and there was early recognition that mismeasurement in the predictors could lead to bias in estimated coefficients and associated probabilities.

Motivated by the analysis of data from the Framingham Heart Study, Carroll et al. (1984 [MEM-1]) address a number of fundamental issues in dealing with error in the predictors in binary regression, attacking both the question of what the effects of measurement error are on (so-called naive) methods that ignore it and looking at how to correct for bias induced by measurement error. The true predictors are assumed normally distributed as are the measurement errors, which also are assumed to be additive with constant covariance matrix. The paper [MEM-1] makes use of an "induced" model, i.e., a model for Y given the observed, rather than the true, predictors. With a logistic model, it is impossible to write down the exact induced model in a form that is useful. An important insight in [MEM-1] is that if the original model is probit (which often provides a good approximation to the logistic), then the induced model is also a probit model with explicit expressions for the parameters in the induced model in terms of the coefficients in the original model, the covariance matrix of the measurement error and the mean and covariance matrix of the true predictors. The yield from this was twofold: (i) an exact expression for the limiting values (and hence asymptotic bias) of the naive estimators that ignore the measurement error and (ii) an easy way to correct the naive estimators for bias. The bias results also show how measurement error in some predictors can induce bias in the coefficients of other perfectly measured predictors. The corrected estimators fall under the heading of pseudo-estimators that employ estimates of the measurement error covariance and the structural parameters obtained from replication. Inferences were based on the bootstrap, which at the time the paper was written, was a relatively new methodology.

The 1985 *Annals of Statistics* paper, Stefanski and Carroll (1985 [MEM-2]), joint with Len Stefanski and based on Len's PhD work as a student under Ray, also dealt with binary regression, but assuming a logistic model for the true values. This paper heads off into some fundamentally new directions, compared to Carroll et al. (1984 [MEM-1]), for a couple of reasons: (i) the lack of a clean expression for an induced model in the structural case posed significantly new challenges and (ii) the desire to address the so-called functional setting where the unobserved predictors are conditioned on and treated as fixed, an important consideration as it relaxes the assumption of an overall random sample. This was ground-breaking work, notable for both the high level of mathematical rigor and the novelty of the approaches taken in addressing both the behavior of the naive estimators and new strategies for correcting for measurement error. The approach here is in the context of the so-called small measurement error asymptotics, relying on very careful use of expansions for both the likelihood function (under normal measurement errors) and the corresponding naive estimating equations. After a rigorous treatment of asymptotics for small measurement error, leading to characterization of bias of naive estimators, three correction methods are explored. These were based on: (i) subtracting an estimate of the approximate bias of the naive estimator; (ii) using an approximate maximum likelihood estimator under normal measurement errors; and (iii) under normality, finding a sufficient "statistic" Δ (see the discussion of Stefanski and Carroll (1987 [MEM-4]) below) so that $Y|\Delta$ follows a logistic model. This leads to the development of unbiased (nonlinear) estimating equations; i.e., estimating equations calculated from the observed data that have expected value of zero at the parameters of the true-data model. With the use of certain approximations, all three of these approaches, coming from fairly different directions, lead to carrying out a logistic regression but with the observed/error prone predictors replaced by an updated/imputed value. These methods are predecessors of regression calibration, which also uses imputed values, but motivated from yet another perspective. Simulations provide additional evidence that these estimators greatly improve on the naive approaches in many situations. Finally, [MEM-2] also touches on the fact that naive tests for parameters involving perfectly measured predictors may not always be correct.

Following this initial work on binary regression are two landmark papers Stefanski and Carroll (1987 [MEM-4]) and Carroll and Stefanski (1990 [MEM-7], extending earlier work in major ways using methods that continue to impact both applications and methodological research. Stefanski and Carroll (1987 [MEM-4]) extends the sufficient-statistic method in Stefanski and Carroll (1985 [MEM-2]) to handle additive normal measurement error with a generalized linear model for the response in terms of the error-free predictors. It modifies the naive estimating equations to get unbiased estimating equations in three different ways covering both functional and structural models, and addressing questions of efficiency. As with Stefanski and Carroll (1985 [MEM-2]), accommodating the functional setting is a critical contribution. (For consistency of terminology with the rest of this chapter we refer to the true value as X and the error-prone measure as W, even though in the original paper U was the true value and X the error-prone measure.) There are three types of estimators here, all making use of $\Delta_i = W_i + Y_i \Omega\beta$, where Ω is the covariance matrix of the measurement errors. The quantity Δ_i is referred to as "parameter-dependent sufficient statistic" in that the distribution of $Y_i|\Delta_i$ does not depend on the unobserved X_i. The first part of [MEM-4] treats the functional case with two related sets of "unbiased" score equations. The first results in what is called a "sufficiency estimator" while the other, which garnered more attention, is based on the conditional score approach of Bruce Lindsay, developed for related problems. This leads to a so-called conditional estimator. The resulting two sets of estimating equations have common components but differ in one multiplicative factor. The conditional estimating equations involve an arbitrary function of Δ, $t(\Delta)$, where ideally $t(\cdot)$ is chosen to obtain efficient estimators. They then turn their attention to handling the structural model with an unspecified distribution $g(\cdot)$, for the true values. In this setting they are able to characterize unbiased score functions and lay out efficient corrected scores. The resulting estimating equations are of the same form as those leading to the conditional estimator but with $t(\Delta) = E(X|\Delta)$. Details are provided for the all-important linear, logistic, and Poisson models.

Carroll and Stefanski (1990 [MEM-7]) is among the most influential measurement error papers published. The scope is remarkably broad, providing a very general framework for attacking measurement error problems. It allows for a rich class of models for the true values, with $E(Y|X) = f_m(X,\beta)$ and $V(Y|X) = \sigma^2 v_m(X,\beta,\theta)$, where X contains the true predictors. It also allowed for non-additive measurement error or Berkson error. At the same time it accommodates various types of additional data, including combinations of reliability data (involving replication) and validation data (involving mismeasured and true values), with either type of data possibly being either internal or external to the main study data. This work carefully derives approximations for $E(Y|W)$ and $V(Y|W)$, where the conditional mean potentially depends on W, β and the measurement error parameters, while the variance may depend on these and the additional variance parameters (σ^2 and θ). The development of these approximations depends in turn on a model, exact or approximate, for $E(X|W)$ and $V(X|W)$. Using the model for $E(Y|W)$ and $V(Y|W)$ three general approaches to obtaining corrected estimators are given. The most enduring of these are based on the use of general quasi-likelihood methods for fitting the model for $E(Y|W)$ taking into account $V(Y|W)$, in conjunction with the use of the estimating equations for the measurement error parameters that arise from the reliability and/or validation data. As this method has evolved it is often implemented in a pseudo manner, first estimating the measurement error parameters, substituting them in the model for $Y|W$ and then estimating the remaining parameters. A special case of this approach is what is now known as regression calibration, developed at around the same time by Gleser (1990) and Rosner, Willett and Spiegelman (1989) in special settings. Regression calibration runs the naive analysis after substituting an estimate of $E(X|W)$ in place of the unobserved X. However, the estimated covariance matrix of the corrected estimates does not simply follow from the analysis on the imputed values. A general asymptotic theory is developed here, applicable to regression calibration as a special case. We agree with Ray's assessment that the more general versions of the quasi-likelihood approach are underused. [MEM-7] also presents two additional methods, one based on approximating the quasi-likelihood estimating equations, which generalized other earlier work by others, and a second method based on correcting the naive estimator by subtracting an estimator of the approximate bias.

The 1993 case–control paper by Carroll, Gail, and Lubin (1993 [MEM-8]) is connected to Ray's earlier work on logistic regression problems in that the model for $Y|X$ is still logistic. Here, however, the data arise from a case–control design involving independent samples from populations with outcome $Y = 0$ or $Y = 1$. It is well known that in the absence of measurement error, data from the case–control setting can be treated as if it is prospective, at least as far as estimation of the non-intercept coefficients are concerned. With measurement error, much more attention needs to be given to the distinctions between the case–control and prospective settings. Among other considerations, the nature of case–control studies often leads to differential measurement error as a result of recall bias, while many MEM methods are built on the assumption of non-differential measurement error. [MEM-8] is the first to provide a comprehensive look at measurement error in case–control studies, beyond those that had addressed the problem in highly parametric fashion based on the normal discriminant model and normal measurement errors (e.g., Armstrong, Whittemore, and Howe, 1989; Buonaccorsi, 1990). As with a number of Ray's other papers this one is notable for its scope. In [MEM-8] the authors take a likelihood approach based on the retrospective nature of case–control data while incorporating the prospective logistic model for the outcome given the covariates. It uses internal validation data in the form of subsamples from the cases and controls in which the true (X) as well as the mismeasured (W) covariates are obtained. This validation data allow for estimating the distribution of W given X and Y (modeled in a parametric manner) as well as the distribution of X within each outcome status. The latter are estimated nonparametrically using the empirical distribution function of the Xs from the validation data, computed separately for each outcome status. In this sense the model is semi-parametric. The use of these empirical distribution leads to the need for new theory in deriving the asymptotic behavior of the likelihood estimators. They also give attention to the use of external validation data as well as replication under additive non-differential measurement error. Finally they address the fact that methods designed for the prospective setting (such as regression calibration) may encounter trouble in the case–control scenario; see Guolo (2008) for a recent discussion.

Linear Problems. Some of Ray's earliest work was on measurement error in linear models. In addition to Carroll, Gallo, and Gleser (1985 [MEM-3]), related to earlier work by Carroll and Gallo (1982), his other important contributions include studying misuse of orthogonal least squares in "errors-in-variables" based on potentially invalid assumptions about the measurement error variances (Carroll and Ruppert, 1996); and investigating the somewhat under-discussed problem of model diagnostics in the presence of measurement error (Carroll and Spiegelman, 1992). The broader context of Carroll, Gallo, and Gleser (1985 [MEM-3] lies in the fact that measurement error in some predictors often leads to biases in estimated coefficients and in tests associated with other perfectly measured predictors. This is true of both linear and nonlinear models. However there are conditions under which inferences for certain linear combinations of coefficients of perfectly measured predictors are robust to the measurement error, depending on the correlation structure among the different predictors involved. One example discussed in the paper is the analysis of covariance with individuals randomized to treatments, and measurement error in the covariate; see Carroll (1989) for additional discussion. It is shown that the balance in covariates arising via the randomization leads to valid inference regarding the treatment effects. However, [MEM-3] goes quite a bit further. It asks the question of when the so-called naive estimator is better than an errors-in-variables maximum likelihood estimator obtained under the assumption that the measurement error is known up to a proportionality constant. This question is addressed via a comparison of the asymptotic properties of the two estimators. Further, and more important, it is shown that if the condition on the design matrix that leads to robustness of the naive estimators is incorporated into the model, then the MLE using this information along with the assumed structure of the measurement error covariance matrix is the same as the usual least squares estimator.

Mixed Models. Wang et al. (1998 [MEM-9]), with Wang, Lin, and Gutierrez, is another great example of Ray's foundational contributions studying measurement error in new settings. In 1998 there was a relatively slim literature on measurement error in mixed models, even for the linear case. The paper [MEM-9] jumped to the general problem with a generalized linear mixed model for the true values, with an emphasis on additive measurement error for predictors associated with the fixed effects part of the model. More specifically with i indexing a cluster and j an observation within a cluster, Y_{ij} denoting the response, b_i a vector of random effects, A_{ij} and Z_{ij} being known vectors, and X_{ij} a vector of predictors subject to measurement error, the conditional generalized linear model they consider has $g(E(Y_{ij}|X_{ij}, Z_{ij}, A_{ij}, b_i) = \beta_0 + X_{ij}^T\beta_X + Z_{ij}^T\beta_Z + A_{ij}^T b_i$. The Z are treated as fixed throughout, while the Xs are assumed random with a distribution that may depend on Z. The authors development allows for a general measurement error model via specification of the distribution of $(W|X, Z)$, where W is observed in lieu of X, but they later focus on additive errors. Additional variance covariance parameters enter the model via a parameter ϕ and $Cov(b_i) = D(\theta)$ and an important consideration in mixed models is that the variance/covariance parameters are themselves often of interest.

There are three general contributions in [MEM-9]. The most important one is the analysis of the induced model, with the twofold goal of bias assessment and suggesting correction methods. For standard regression problems, the induced model frequently is in the same form as the original model, either exactly or approximately, although there are a number of exceptions to this rule. The mixed model proved to be much more sensitive. As shown in [MEM-9] the measurement error often perturbs the structure of the fixed and random effects portions, leading to model misspecification in both. The key result in Wang et al. (1998 [MEM-9]) is manifested in their very general equation (6). They follow with a detailed analysis of a number of special cases illustrating how the general form of the fixed or random effects part of the model may, or may not, be altered by the measurement error. An important point here is that the nature of the biases depend on the structure of the distribution of $X|Z$, which itself can have a mixed model structure. They provide a detailed discussion of settings where the X_{ij} are assumed i.i.d., called the *homogeneous case*, or the X_{ij} follow a one-way random effects model with cluster specific means, called the *heterogeneous case*. The second contribution of [MEM-9] characterizes the biases that occur when (assuming the measurement error variance is known) maximum likelihood estimation is used assuming the homogeneous model, when in fact the heterogeneous model holds. Exact results are given for the linear mixed model, showing there is still bias in estimators of both the fixed effects coefficients and the variance estimates. Biases are assessed numerically for the logistic case because of the absence of a closed-form solution. The final contribution of [MEM-9] addresses correction methods. The authors first show that in most situations the use of regression calibration encounters problems because the induced model does not retain the same structure as the original model, the exception being in estimation of the fixed effects coefficients in certain linear mixed models. As an alternative to regression calibration, SIMEX estimators are proposed, evaluated via simulation, and illustrated using Framingham Heart Study data.

Non- and Semiparametric Models

Ray made outstanding contributions to nonparametric curve estimation in the presence of measurement error, in both the density and the regression contexts. Work by Ray and his coauthors initiated a huge and growing literature on nonparametric estimation in the presence of measurement error.

His first contributions (separately with Hall and Stefanski) study a nonparametric kernel density estimator that corrects for the contamination present in the data (Carroll and Hall, 1988 [MEM-5] and Stefanski and Carroll, 1990 [MEM-6], which together have attracted nearly 600 citations). In that problem, we observe data on W (the contaminated observations), but we are interested in estimating the density f_X of X, where $W = X + U$. Here, U is an unobserved measurement error, but its density f_U is known. Since the density f_W of the contaminated data is the convolution of f_X and f_U, the consistent kernel density estimator of f_X is usually referred to as the deconvolution kernel density estimator.

Although Carroll and Hall (1988 [MEM-5]) appeared in print first, Stefanski and Carroll (1990 [MEM-6]) was written first and contains many important results that have been used extensively by others. It derives the deconvolution kernel density estimator, calculates its bias, variance, and asymptotic mean integrated squared error in the Fourier domain, suggests a cross-validation bandwidth, and applies the method to data from a breast cancer study. Working in the Fourier domain makes the elegant theory possible and provides the framework for the now well-established distinction between the ordinary-smooth and the super-smooth errors. Another important result of [MEM-6] is that, conditionally on the unobserved X_is, the expectation of the deconvolution kernel density estimator of f_X is equal to the standard kernel density estimator of f_X constructed from the X_is. Together with the Fourier transform approach, this result turned out to be the key to solving the long-open problem of developing a local polynomial regression estimator with measurement errors (Delaigle, Fan, and Carroll, 2009).

Carroll and Hall (1988 [MEM-5]) also contains major results. It was the first to establish minimax convergence rates of the deconvolution kernel density estimator, and opened the way to a long series of influential papers about nonparametric density and regression estimators in the measurement error context. In particular, [MEM-5] showed that the very slow convergence rates of the deconvolution estimator in the case where the errors are normal, are not due to poor performance of the estimator itself, but are inherent to the difficulty of the problem. In other words, they showed that it is not possible to construct a nonparametric estimator that has faster convergence rates than the deconvolution kernel estimator.

Later Ray attacked complex nonparametric regression problems involving measurement errors. Fan and Truong (1993) were the first to extend the deconvolution kernel density estimator to the regression context, where the goal is to estimate a regression curve $m(X) = E(Y|X)$ from data on (W, Y), where $W = X + U$ with X and U independent and the measurement error density f_U is known. However, this estimator suffered from the same slow convergence rates as in the density case, and no data-driven bandwidth had yet been derived to calculate it in practice. The idea of Ray and his collaborators was that instead of trying to consistently estimate a curve that can only be estimated with much difficulty, why not target instead a curve that is only approximately equal to $m(\cdot)$, but which is more easily estimated? This lead to very innovative procedures and a lot of subsequent work by others.

One such approximation method is a nonparametric version of the SIMEX ideas of Cook and Stefanski (1994). In SIMEX, we learn how measurement error modifies a target by artificially adding more noise to the data and monitoring the effects. By learning about the relation between $E(Y|W)$ and $E(Y|W + \text{error})$, Carroll, Maca and Ruppert (1999 [MEM-10]) were able to extrapolate back that information to estimate the curve $E(Y|X)$ from a nonparametric estimator of $E(Y|W)$ (their method is actually more sophisticated than this). They showed that nonparametric SIMEX is only consistent if the variance of U tends to zero, but in practice it can give great results when this variance is not too large. This small variance idea was ingeniously exploited by Carroll and Hall (2004) to remove the bias of naive kernel estimators, and was later used by other authors in a variety of different contexts.

The structural regression splines in Carroll, Maca, and Ruppert (1999 [MEM-10]) improve upon SIMEX. Their idea is to approximate $m(\cdot)$ by a spline $\tilde{m}(\cdot)$ with a fixed number of knots, and estimate $\tilde{m}(\cdot)$ (instead of $m(\cdot)$) from the contaminated data. Since $E(Y|W)$ is well approximated by $E(\tilde{m}(X)|W)$, the spline coefficients of $\tilde{m}(\cdot)$ can be estimated by fitting $E(\tilde{m}(X)|W)$ to the (W_i, Y_i) data. In practice this requires estimators of moments of functions of X conditional on W. The distribution of $X|W$ could be estimated nonparametrically, but because it results in slow convergence rates they instead make the structural assumption that the distributions of X and W are both normal. Even though this assumption is usually not exactly satisfied in real data applications, it is often good enough to give reasonable approximations. There were two drawbacks to this method, though: (i) choosing the smoothing parameter was too difficult; and (ii) the near orthogonality of the conditional means of the spline basis functions caused numerical instability. To overcome these difficulties, a Bayesian version of smoothing and regression spline was suggested in Berry, Carroll, and Ruppert (2002 [MEM-12]).

Another influential work was Liang, Härdle, and Carroll (1999 [MEM-11]) wherein the authors show how to consistently estimate partially linear models when the explanatory variables in the linear part are measured with error. In the error-free case, Severini and Staniswalis (1994) suggest estimating the nonparametric part assuming the parametric part known, then estimate the parameters by least squares, plugging in the nonparametric estimator. Because this method is not consistent in case of measurement errors, Liang, Härdle, and Carroll (1999 [MEM-11]) add a penalty to the least-squares sum to overcome the attenuation effect of measurement errors.

Ray made so many other major contributions in extending non- and semiparametric estimation problems with measurement errors that it is impossible to list them all here. These include the use of instrumental variables in Carroll et al. (2004), the development of locally efficient estimators for semiparametric models in Ma and Carroll (2006), the combination of Berkson and classical errors in nonparametric regression in Carroll, Delaigle and Hall (2007), the provocative parametric rates in nonparametric prediction in measurement error models in Carroll, Delaigle and Hall (2009), and the development of methods for quantile regression in Wei and Carroll (2009), to cite just a few.

References

Other publications by Ray Carroll cited in this chapter.

Carroll, R. J. and Gallo, P. P. (1982). Some aspects of robustness in the functional errors-in-variables regression-model. *Communications in Statistics, Part A-Theory and Methods*, 11, 2573–2585.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8, 1075–1093.

Carroll, R. J. and Spiegelman, C. H. (1992). Diagnostics for nonlinearity and heteroscedasticity in errors-in-variables regression. *Technometrics*, 34, 186–196.

Carroll, R. J. and Ruppert, D. (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *American Statistician*, 50, 1–6.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D., and Karagas, M. R. (2004). Nonlinear and nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, 99, 736–750.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*, 2nd ed. London: Chapman & Hall.

Carroll, R. J., Delaigle, A., and Hall, P. (2007). Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *Journal of the Royal Statistical Society, Series B*, 69, 859–878.

Carroll, R. J., Delaigle, A., and Hall, P. (2009). Nonparametric Prediction in Measurement Error Models. *Journal of the American Statistical Association*, 104, 993–1003.

Delaigle, A., Fan, J. and Carroll, R.J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*, 104, 348–359.

Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, 101, 1465–1474.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association*, 104, 1129–1143.

Publications by other authors cited in this chapter.

Armstrong, B. G., Whittemore, A. S., and Howe, G. R. (1989). Analysis of case-control data with covariate measurement error: Application to diet and colon cancer. *Statistics in Medicine*, 8, 1151–1163.

Brown, P. J. and Fuller, W. A. (1990). *Statistical Analysis of Measurement Error Models and Applications: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference*, June 10–16, 1989. Providence: American Mathematical Society.

Buonaccorsi, J. P. (1990). Double sampling for exact values in the normal discriminant model with applications to binary regression. *Communications in Statistics, Theory and Methods*, 19, 4569–4586.

Byar, D. P. and Gail, M. (1989). Introduction. Errors-in-variables workshop. *Statistics in Medicine*, 8, 1027–1029.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314–1328.

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics*, 21, 1900–1925.

Fuller, W. A. (1987). Measurement error models. New York: John Wiley.

Gleser, L. J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables problems. *Contemporary Mathematics*, 112, 99–114.

Guolo, A. (2008). A flexible approach to measurement error correction in case-control studies. *Biometrics*, 64, 1207–1214.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8, 1051–1070.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89, 501–511.

On errors-in-variables for binary regression models

By RAYMOND J. CARROLL

Department of Statistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

CLIFFORD H. SPIEGELMAN

Statistical Engineering Division, National Bureau of Standards, Washington, D.C., U.S.A.

K. K. GORDON LAN, KENT T. BAILEY AND ROBERT D. ABBOTT

National Heart, Lung and Blood Institute, Bethesda, Maryland, U.S.A.

Summary

We consider binary regression models when some of the predictors are measured with error. For normal measurement errors, structural maximum likelihood estimates are considered. We show that if the measurement error is large, the usual estimate of the probability of the event in question can be substantially in error, especially for high risk groups. In the situation of large measurement error, we investigate a conditional maximum likelihood estimator and its properties.

Some key words: Functional model; Logistic regression; Measurement error; Probit regression; Structural model.

1. Introduction

The Framingham Heart Study (Gordon & Kannel, 1968) is a prospective study of the development of cardiovascular disease. This study has been the basis for a considerable amount of epidemiologic research. For example, there has been considerable emphasis on analysing the probability of developing coronary heart disease. Many of the analyses have attempted to relate baseline risk factors to the probability of developing heart disease; these risk factors include systolic blood pressure, serum cholesterol, etc. Often, in the analysis, logistic or probit binary regression is employed.

It is well known that many baseline risk factors are measured with error; systolic blood pressure is a good example (Armitage & Rose, 1966; Rosner & Polk, 1979). One of us was asked by a number of investigators whether such measurement errors could substantially affect the binary regression estimates and, if so, what could be done to correct for the measurement error. The present study is an outgrowth of these questions, although there are many important practical facets of the problem yet to be investigated.

Michalek & Tripathi (1980) show that ordinary logistic regression will not be too badly disturbed by measurement error as long as such error is moderate; see also Ahmed & Lachenbruch (1975). While our model is different, our methods provide alternatives to ordinary binary regression which will help the experimenter to get a more precise understanding of the effect of the measurement errors, especially if they are severe and the sample size is large.

The model is similar to that of Halperin, Wu & Gordon (1979). We have a sample of N persons from a particular population, e.g. males aged 45–54. The ith person in the sample is assumed to have a p-vector of baseline risk factors x_i, with the probability of developing disease given x_i taking the form

$$\mathrm{pr}(Y_i = 1 \mid x_i) = G(x_i'\beta) \quad (i = 1, \ldots, N), \tag{1·1}$$

where $G(.)$ is a known distribution function such as for logistic regression $G(a) = (1+e^{-a})^{-1}$, and for probit regression $G(a) = \Phi(a)$. Here $\Phi(.)$ is the standard normal distribution function. We return to probit regression later, but it is important to remember that probit and logistic regression often given similar results (Finney, 1964).

We partition the risk factors x_i into components observed without and with error, so that

$$x_i' = (w_i', z_i'), \quad \beta' = (\beta_1', \beta_2'). \tag{1·2}$$

In (1·2), the vectors $\{w_i\}$ can be observed at nearly exact levels, while the q-vectors $\{z_i\}$ are measured with nontrivial error and cannot be observed; rather, we observe

$$Z_i = z_i + u_i. \tag{1·3}$$

The $\{u_i\}$ are assumed independently and normally distributed with mean zero and nonsingular covariance Ω_M.

When the risk factors $\{z_i\}$ observed with error are unknown constants, we have a functional model (Kendall & Stuart, 1979, Chapter 29). In this instance, classical maximum likelihood theory does not apply. In fact, even in simple binary regression models, the functional maximum likelihood estimate of β is not consistent when Ω_M is known; details are available from the authors. This is in contrast to linear regression, where the functional maximum likelihood estimate is consistent if the ratio of the error variances is known or if there is finite replication of the predictors. Consistent and asymptotically normal estimates for the functional logistic regression model can be constructed when the measurement errors in (1·3) are normally distributed; this work will be reported elsewhere.

In § 2 we study the structural model, wherein the $\{z_i\}$ are themselves independent with common distribution function F, which we will suppose is that of a normal random vector with mean μ_z and covariance Ω_z. In effect, we condition on the observed values (w_i, Z_i) and replace (1·1) by $\mathrm{pr}(Y_i = 1 \mid w_i, Z_i)$, the probability of an event given the observed outcomes; see Armstrong & Oakes (1982) for a similar idea.

In this paper we present a small Monte Carlo study, as well as the analysis of actual data, in which we investigate the effect of measurement error on predicting the probability of heart disease on the basis of systolic blood pressure. The major purpose of these sections is to illustrate that for large sample sizes and realistically large measurement errors, the usual method of ignoring measurement error can be improved. This is not to suggest that the conditional estimator will emerge as the standard for practical use; this point is discussed in the concluding section.

As in linear regression, there are at least two good reasons to estimate the error-free regression. First, measurement processes may improve, making the errors-in-variables estimates more valuable. Second, it can be meaningful to investigate the true regression coefficient; see Wu, Ware & Feinleib (1980) for an example of linear regression where the errors-in-variables estimates are physically sensible but the least squares estimates are not.

2. Structural case: Normal distribution

The model is given by (1·1), (1·2) and (1·3), but in the structural case we eliminate the nuisance parameters $\{z_i\}$ by assuming they are independent and normally distributed with mean vector μ_z and covariance matrix Ω_z. The error vectors $\{u_i\}$ are also assumed to be normal random vectors with mean zero, covariance Ω_M, and independent of one another and of $\{z_i\}$. For the moment, assume μ_z, Ω_z and Ω_M are known. Then, except for a complicated constant of proportionality, the likelihood of Y_i, conditioned on $\{Z_i\}$ and as a function of G in (1·1) is given by

$$L(G,\beta_1,\beta_2,\Omega_M,\mu_z,\Omega_z) = \prod_{i=1}^{N} t_i^{Y_i}(1-t_i)^{1-Y_i}, \tag{2·1}$$

$$t_i = \int G\{w_i'\beta_1+(\beta_2' A\beta_2)^{\frac{1}{2}}v+d_i' A\beta_2\}\,\phi(v)\,dv, \tag{2·2}$$

$$A = (\Omega_M^{-1}+\Omega_z^{-1})^{-1}, \quad d_i = \Omega_z^{-1}\mu_z+\Omega_M^{-1}Z_i$$

and $\phi(\,.\,)$ is the standard normal density function.

In effect, the calculation of the likelihood (2·1) depends only on evaluating (2·2). This is no easy matter for the logistic function, although if the number of variables measured with error is small, (2·2) can in principle be evaluated by numerical integration. For probit regression, (2·2) can be evaluated explicitly:

$$t_i = \Phi\{(w_i'\beta_1+d_i' A\beta_2)(1+\beta_2' A\beta_2)^{-\frac{1}{2}}\}.$$

Since logistic and probit regression often give similar estimates of event probabilities, especially for our examples, in the rest of the paper we confine our discussion to probit regression.

In most instances, the nuisance parameters μ_z, Ω_z and Ω_M will be unknown, although one could conceive of setting $\mu_z = 0$ particularly to obtain some shrinkage. Joint estimation of these parameters and β through the full conditional likelihood may be computationally feasible, although this as well as existence of a sensible maximum for the conditional likelihood remains to be explored. An alternative is suggested by the work of Gong & Samaniego (1981), which is to find estimates of μ_z, Ω_z and Ω_M, substitute them in (2·1) and (2·2) and maximize. An obvious estimate for μ_z is the sample mean of the $\{Z_i\}$, while an estimate $\hat{\Omega}_{zM}$ for $(\Omega_z+\Omega_M)$ is the sample covariance of $\{Z_i\}$. A common device is to estimate Ω_M by replication. If, for example, each variable subject to error is measured twice (Z_{i1}, Z_{i2}), then an estimate $\hat{\Omega}_M$ of Ω_M is the sample covariance of $\frac{1}{2}(Z_{i1}-Z_{i2})$. This suggests the estimate

$$\hat{\Omega}_z = \hat{\Omega}_{zM}-\hat{\Omega}_M, \tag{2·3}$$

where $\hat{\Omega}_{zM}$ is the sample covariance of $Z_i = \frac{1}{2}(Z_{i1}+Z_{i2})$.

Two points relating to the above need to be emphasized. First, as we have proposed it, estimating Ω_M requires replication. In linear regression, if the ratio of the error variances is known, no such replication is necessary. It is not clear and remains to be seen whether direct maximization of the likelihood is computationally feasible and produces consistent and asymptotically normal estimates. A second point is that (2·3) is not necessarily positive-definite, a problem which is similar to that observed for moments estimates in variance components problems. In our applications and perhaps for most examples, N is large relative to q, so that $\hat{\Omega}_z$ will usually be positive-definite. This is no guarantee; further work is needed.

The covariance matrix of our estimates can be estimated by the bootstrap method (Efron, 1979); see §4. Alternatively, one could try to generalize equations (2·5) and (2·6) of Gong & Samaniego (1981) and take numerical derivatives. Finally, if maximum likelihood is used, one can in principle evaluate the sample information matrix.

3. RESULTS OF A SIMULATION STUDY

We performed a small Monte Carlo study for the probit model

$$\mathrm{pr}\,(Y_i = 1 \mid z_i) = \Phi(\beta_1 + \beta_2 z_i), \quad Z_{ij} = z_i + u_{ij} \quad (j = 1, 2;\ i = 1, \ldots, N).$$

The $\{z_i\}$ and $\{u_{ij}\}$ were generated as independent univariate normal random variables with means zero and variances σ_z^2, σ_m^2 respectively. We chose $\beta_1 = -1{\cdot}40$, $\beta_2 = 1{\cdot}34$ and $3\sigma_m^2 = \sigma_z^2$, with the two values of $\sigma_z^2 = 0{\cdot}0833, 0{\cdot}10$. The sample sizes were $N = 300, 600, 1200$. For each of the six combinations of (σ_z^2, N), we generated 400 simulated data sets.

In §4, we discuss an example which turns out to be very similar to one of our simulated cases, $\sigma_z^2 = 0{\cdot}0833$ and $N = 600$. The other cases were chosen to be both realistic and illustrative. We should emphasize that our experience has been with large data sets, and we would not recommend routinely correcting for measurement error for small sample sizes.

The results of the Monte Carlo study are reported in **Table 1**. From the reported results and our other simulations, we can make the following observations. First, usual probit regression is, as expected, more biased but less variable than the conditional likelihood estimate. Thus, in small samples where variance dominates, the usual probit regression will be preferred, while in large samples where bias dominates, the conditional likelihood approach will be preferred.

A second point which is not very clear from **Table 1** but which occurs consistently throughout our more extensive simulations is that, subject to fixed (β_1, β_2) and $3\sigma_m^2 = \sigma_z^2$, as the predictor variance σ_z^2 increases the conditional likelihood approach to correcting for measurement error improves. This phenomenon also occurs if we fix the variances (σ_m^2, σ_z^2) and increase the slope β_2.

A third point concerns estimating $\beta_1 + \beta_2$, which determines the disease probability for a very high risk event $z_i = 1$. Here, the conditional errors-in-variables method is about 10% more efficient than it is for estimating β_1 alone.

In §4 we discuss two simple methods which improve upon the conditional likelihood approach for errors-in-variables correction by about 10%. Hence for the case $N = 600$ and $\sigma_z^2 = 0{\cdot}0833$, which we know to be relevant, it is possible to get about 19% increase in efficiency by correcting for measurement error rather than using ordinary regression.

4. AN EXAMPLE

To get some idea of the possible effects of measurement error in a realistic context, we considered some of the data from the Framingham Study (Gordon & Kannel, 1968). Data used here were on 589 men aged 45–54. Individuals were called diseased cases if they developed coronary heart disease within the ten year interval after examination; 56 were eventually considered to be diseased. We used as our predictor variables not the actual systolic blood pressure but rather $\log\{(\text{systolic blood pressure} - 75)/25\}$, which was originally suggested by Cornfield (1962); these transformed observations appear reasonably normally distributed in our data set.

Table 1. *A Monte Carlo study for the probit regression model* $\mathrm{pr}(Y_i = 1 \mid z_i) = \Phi(1·34z_i - 1·40)$

		Sample size, $N = 300$				Sample size, $N = 600$			
		$\sigma_z^2 = 0·0833$		$\sigma_z^2 = 0·10$		$\sigma_z^2 = 0·0833$		$\sigma_z^2 = 0·10$	
		Probit	EIV	Probit	EIV	Probit	EIV	Probit	EIV
Bias	β_1	−0·00	−0·02	0·00	−0·02	0·01	−0·01	0·01	−0·01
	β_2	−0·19	0·03	−0·18	0·04	−0·20	0·01	−0·19	0·02
Standard deviation	β_1	0·12	0·13	0·12	0·13	0·08	0·09	0·08	0·09
	β_2	0·38	0·48	0·35	0·43	0·25	0·30	0·23	0·29
	$\beta_1+\beta_2$	0·35	0·43	0·31	0·38	0·22	0·27	0·20	0·25
10 × mean squared error	β_1	0·15	* 0·17	0·15	* 0·18	0·07	0·08	0·07	0·08
	β_2	1·82	* 2·29	1·52	* 1·88	0·98	* 0·91	0·90	* 0·82
	$\beta_1+\beta_2$	1·58	1·83	1·28	1·44	0·83	* 0·71	0·75	* 0·62
Eff.	β_1	86%		84%		88%		86%	
	β_2	80%		81%		108%		109%	
	$\beta_1+\beta_2$	86%		89%		118%		121%	

		Sample size, $N = 1200$			
		$\sigma_z^2 = 0·0833$		$\sigma_z^2 = 0·10$	
		Probit	EIV	Probit	EIV
Bias	β_1	0·01	−0·01	0·01	−0·01
	β_2	−0·20	0·01	−0·20	0·02
Standard deviation	β_1	0·06	0·06	0·06	0·06
	β_2	0·17	0·21	0·16	0·20
	$\beta_1+\beta_2$	0·16	0·19	0·15	0·18
10 × mean squared error	β_1	0·03	* 0·04	0·03	* 0·04
	β_2	0·67	* 0·43	0·65	* 0·41
	$\beta_1+\beta_2$	0·61	* 0·36	0·57	* 0·33
Eff.	β_1	86%		88%	
	β_2	155%		155%	
	$\beta_1+\beta_2$	170%		172%	

Values of $\{z_i\}$ normally distributed with mean zero and variance σ_z^2. Measurement error variance $\sigma_m^2 = \frac{1}{3}\sigma_z^2$. Probit, ordinary probit regression; EIV, estimates derived from conditional approach. Eff., mean squared error efficiency with respect to ordinary probit regression. * in mean squared error rows, difference significant at 1% level by signed rank test.

Table 2. *Framingham data: probit regression,* $\mathrm{pr}(Y_i = 1 \mid z_i) = \Phi(\beta_1 + \beta_2 z_i)$

	Usual probit	Probit EIV	SBP		Usual probit	Probit EIV
β_1	−2·13	−2·40	175	Probability	0·23	0·30
Mean	−2·13	−2·41		Mean	0·24	0·30
STD	0·22	0·31		STD	0·04	0·06
β_2	1·01	1·34	200	$\beta_1+\beta_2 z$	−0·51	−0·24
Mean	1·02	1·36		Mean	−0·49	−0·23
STD	0·24	0·34		STD	0·19	0·26
$\beta_1+\beta_2 z$, SPB = 175	−0·73	−0·54	200	Probability	0·31	0·40
Mean	−0·72	−0·53		Mean	0·32	0·41
STD	0·14	0·19		STD	0·07	0·10

Y_i indicates development of coronary heart disease; $z_i = \log\{(\mathrm{SBP} - 75)/25\}$, where SBP is true systolic blood pressure.

Mean, bootstrap; STD, bootstrap standard deviation; EIV, errors-in-variables.

Letting the $\{z_i\}$ and $\{Z_i\}$ of § 2 be these transformed observations, we estimated $\hat{\sigma}_z^2 = 0{\cdot}0833$, $\hat{\sigma}_M^2 = 0{\cdot}31\hat{\sigma}_z^2$; details of this estimation procedure, which uses method of moments and components of variance, are available from the authors. The estimates of the intercept and slope (β_1, β_2) for usual probit regression and the conditional likelihood errors-in-variables method are given in Table 2. Also given are the 'bootstrap' (Efron, 1979) means and standards deviations of these estimates, where in this case we bootstrap by randomly sampling with replacement 589 observations from the original data set, estimating σ_z^2 and σ_M^2 and then calculating the bootstrap estimates of (β_1, β_2). As expected, the usual probit regression estimates are attenuated, i.e. have larger intercept but smaller slope than the conditional likelihood method, and have smaller standard deviations.

Also of interest is the probability of disease and its inverse for those with very high blood pressure. We specifically focused on those whose true systolic blood pressure is 175 or 200. The estimates from the data, along with bootstrap means and standard deviations, are also given in Table 2. One of the basic consequences of the attenuation of (β_1, β_2) is that the usual probit estimates of the probability of developing disease will be lower than that of the errors-in-variables method, at least for individuals in the highest risk groups. Of course, since this is only one data set, we can make no claims that our errors-in-variables estimates are closer to the true values that are the usual probit estimates, but we believe our answers are physically meaningful.

5. CONCLUDING REMARKS

Correcting for measurement error will be worthwhile when the measurement error variance and sample size are such that the bias in the usual methods becomes large relative to the increased variance due to correction. For the situations we have investigated, this means that the sample size must be quite large. Simulations not reported here indicate the increased value of correction at a given sample size for increasing amounts of measurement error.

We view the conditional approach as a first step towards developing a useful method to correct for measurement error. We harbour no illusion that further work will show that the conditional approach is optimal. For example, the mean squared errors for the conditional approach given in Table 1 can be easily decreased by approximately 10% by one of two methods. The first method is based on a naive shrinkage idea and involves replacing d_i in (2·2) by $d_{i*} = (1-\alpha)\, d_i + \alpha Z_i$, where $\alpha = 25/M$ and M is the greater of the number of events, or N minus the number of events. In an unpublished 1982 North Carolina dissertation, R. Clark fixes μ_z, Ω_z and Ω_M, computes the Bayes estimate of z_i given Z_i, estimates $(\mu_z, \Omega_z, \Omega_M)$, and then uses ordinary logistic or probit regression on the result. For the realistic simulations reported here, the two alternative methods behave similarly. When σ_z^2 is increased by a factor of 3, Clark's alternative method was 10–15% more efficient for $N = 300$ and about 4% more efficient for $N \geqslant 600$; whether such a value of σ_z^2 for a given $\beta = 1{\cdot}34$ occurs routinely in practice remains to be seen. It is clear that shrinkage and Bayes ideas do improve upon the direct conditional likelihood approach, and these ideas should be pursued further.

Throughout, we have assumed normality. We have performed simulations where the predictors are highly skewed, e.g. chi-squared with one degree of freedom. The effect on all the errors-in-variables techniques tend to be markedly negative. This is a warning for practice and an area requiring further research.

The work of R. J. Carroll was supported by U.S. Air Force Office of Scientific Research. The authors thank the referee for a careful reading of the original manuscript and many helpful suggestions.

References

AHMED, S. & LACHENBRUCH, P. A. (1975). Discriminant analysis when one or both of the initial samples is contaminated: Large sample results. *EDV Med. Biol.* **6**, 35–42.

ARMITAGE, P. & ROSE, G. A. (1966). The variability of measurements of casual blood pressure, I. A laboratory study. *Clin. Sci.* **30**, 325–66.

ARMSTRONG, B. G. & OAKES, D. (1982). Effects of approximations in exposure assessments on estimates of exposure-response relationships. *Scand. J. Work Environ. Health*, **8**, 20–23.

CORNFIELD, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc.* **21**, 58–61.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

FINNEY, D. J. (1964). *Statistical Method in Biological Assay*, 2nd edition. London: Griffin.

GONG, G. & SAMANIEGO, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Ann. Statist.* **9**, 861–9.

GORDON, T. & KANNEL, W. E. (1968). *The Framingham Study*, introduction and general background in the Framingham study, §§ 1, 2. Bethesda, Maryland: National Heart, Lung, and Blood Institute.

HALPERIN, M., WU, M. & GORDON, T. (1979). Genesis and interpretation of differences in distribution of baseline characteristics between cases and non-cases in cohort studies. *J. Chronic Dis.* **32**, 483–91.

KENDALL, M. G. & STUART, A. (1979). *The Advanced Theory of Statistics,* **2**. London: Griffin.

MICHALEK, J. E. & TRIPATHI, R. C. (1980). The effect of errors of diagnosis and measurement on the estimation of the probability of an event. *J. Am. Statist. Assoc.* **75**, 713–21.

ROSNER, B. & POLK, B. F. (1979). The implications of blood pressure variability for clinical and screening purposes. *J. Chronic Dis.* **32**, 451–61.

WU, M., WARE, J. H. & FEINLEIB, M. (1980). On the relation between blood pressure change and initial value. *J. Chronic Dis.* **33**, 637–44.

[*Received July* 1982. *Revised July* 1983]

The Annals of Statistics
1985, Vol. 13, No. 4, 1335–1351

COVARIATE MEASUREMENT ERROR IN LOGISTIC REGRESSION*

BY LEONARD A. STEFANSKI AND RAYMOND J. CARROLL

Cornell University and University of North Carolina, Chapel Hill

In a logistic regression model when covariates are subject to measurement error the naive estimator, obtained by regressing on the observed covariates, is asymptotically biased. We introduce a bias-adjusted estimator and two estimators appropriate for normally distributed measurement errors —a functional maximum likelihood estimator and an estimator which exploits the consequences of sufficiency. The four proposals are studied asymptotically under conditions which are appropriate when the measurement error is small. A small Monte Carlo study illustrates the superiority of the measurement-error estimators in certain situations.

1. Introduction and motivation. Logistic regression is the most used form of binary regression [see Berkson (1951), Cox (1970), Efron (1975), and Pregibon (1981)]. Independent observations (y_i, x_i) are observed where (x_i) are fixed p-vector predictors and (y_i) are Bernoulli variates with

$$\Pr\{y_i = 1|x_i\} = F(x_i^T\beta_0) \triangleq \left(1 + \exp(-x_i^T\beta_0)\right)^{-1}, \qquad i = 1, \dots, n. \tag{1.1}$$

Subject to regularity conditions, the large-sample distribution of the maximum likelihood estimator of β_0 is approximately normal with mean zero and covariance matrix $(1/n)S_n^{-1}(\beta_0)$, where $S_n(\cdot)$ is defined for $\gamma \in R^p$ as

$$S_n(\gamma) = n^{-1}\sum_1^n F^{(1)}(x_i^T\gamma)x_i x_i^T. \tag{1.2}$$

Motivation for our paper comes from the Framingham Heart Study (Gordon and Kannel, 1968), a prospective study of the development of cardiovascular disease. This ongoing investigation has had an important impact on the epidemiology of heart disease. Much of the analysis is based on the logistic regression model with y an indicator of heart disease and x a vector of baseline risk factors such as systolic blood pressure, serum cholesterol, smoking, etc. It is well known that many of these baseline predictors are measured with substantial error, e.g., systolic blood pressure. When a person's "true" blood pressure is defined as a long-term average, then individual readings are subject to temporal as well as reader-machine variability. In one group of 45–54 year old Framingham males it was estimated that one fourth of the observed variability in blood pressure readings was due to within-subject variability. The second author was asked by some Framingham investigators to assess the impact of such substantial measure-

Received July 1983; revised July 1985.

*Supported by the Air Force Office of Scientific Research under grant AFOSR-F49620-82-C-0009.

AMS 1980 *subject classifications.* Primary 62J05; secondary 62H25.

Key words and phrases. Errors-in-variables, functional maximum likelihood, logistic regression, measurement error, sufficiency.

ment error and to suggest alternatives to usual logistic regression which account for this error. The present study is an outgrowth of these questions.

When covariates are measured with error the usual logistic regression estimator of β_0 is asymptotically biased, [see Clark (1982) and Michalik and Tripathi (1980)]. As a consequence of bias there is generally a tendency to underestimate the disease probability for high-risk cases and overestimate for low-risk cases; it will be said that measurement error attenuates predicted probabilities. Also, bias creates a problem with hypothesis testing; in Section 2 it is shown that the usual asymptotic tests for individual regression components can have levels different than expected. An example of this occurs in an unbalanced two-group analysis of covariance where interest lies in testing for treatment effect but the covariable is measured with error.

The severity of these problems depends, of course, on the magnitude of the measurement error. In some situations ordinary logistic regression might perform satisfactorily. However, when measurement error is substantial, alternative procedures are necessary. In addition, the availability of techniques which correct for measurement error can make clear the need for better measurement, e.g., more blood-pressure readings over a period of days.

In Section 2 our measurement-error model is defined and the asymptotic bias in the usual logistic-regression estimator is studied. Section 3 presents some alternative estimators; results of a Monte Carlo study are outlined in Section 4; proofs of the asymptotic results are given in Section 5.

Until recently the study of measurement-error models has focused primarily on linear models; see the review article by Madansky (1959) and the papers by Fuller (1980) and Gleser (1981). Interest in nonlinear models is increasing with recent contributions by Prentice (1982), Wolter and Fuller (1982a, 1982b), Carroll, Spiegelman, Lan, Bailey, and Abbott (1984), Armstrong (1984), Amemiya (1982), and Clark (1982). Of these articles Clark (1982) and Carroll et al. (1984) focus specifically on logistic regression. The asymptotic methods employed in this paper are similar to those used by Wolter and Fuller (1982a) and Amemiya (1982) in their studies of nonlinear functional relationships.

2. A measurement error model for logistic regression.

2.1. The model. Our measurement-error model starts with (1.1), but rather than observing the p-vector x_i we observe

$$X_i = x_i + \sigma v_i \quad \textit{where } v_i = \Sigma^{1/2}\varepsilon_i. \tag{2.1}$$

In (2.1) $\Sigma^{1/2}$ is the square root of a symmetric positive semidefinite matrix Σ scaled so that $\|\Sigma\| = 1$ and (ε_i) are independent and identically distributed random vectors with zero mean and identity covariance; also ε_i is independent of y_i, $i = 1, \dots, n$. The scale factor σ dictates the magnitude of the measurement error, e.g., if X_i is a mean of m independent replicate measurements of x_i then $\sigma \propto m^{-1/2}$. The asymptotic theory presented in this paper requires that $\sigma \to 0$ as $n \to \infty$, i.e., large sample, small measurement-error asymptotics. The asymptotics are relevant for two situations: (i) when X_i is an average of m-independent measurements of x_i, in which case the Central Limit Theorem suggests that (ε_i) should be viewed as normal random variates, and (ii) when measurement error is small but nonnegligible. In the latter case the moments of order greater than two of (ε_i) generally differ from those of a normal variate.

Our methods of correcting for bias require knowledge of the error covariance matrix $V \triangleq \sigma^2\Sigma$. Since this information is seldom available all asymptotic results are derived for the case in which V is replaced by an estimator $\hat{V}$ satisfying

$$n^{1/2}(\hat{V} - V) = O_p(\sigma^2). \tag{2.2}$$

Condition (2.2) is satisfied, for example, when V is estimated by replication. It is convenient to write $\hat{V} = \hat{\sigma}^2\hat{\Sigma}$ where $\hat{\sigma}^2 = \|\hat{V}\|$ and $\hat{\Sigma} = \hat{V}/\|\hat{V}\|$. Note that (2.2) then implies $n^{1/2}(1 - \hat{\sigma}^2/\sigma^2) = O_p(1)$.

2.2. *The effects of measurement error.* Our investigation starts with a study of the estimator obtained by regressing y_i on the observed X_i. This estimator, to be called $\hat{\beta}$, maximizes

$$L_n(\gamma) \triangleq n^{-1}\sum_1^n \{y_i \log F(c_i^T\gamma) + (1 - y_i)\log F(-c_i^T\gamma)\} \tag{2.3}$$

and satisfies

$$\sum_1^n (y_i - F(c_i^T\hat{\beta}))c_i = 0, \tag{2.4}$$

when $c_i = X_i$, $i = 1, \dots, n$. Our interest lies in the behavior of $\hat{\beta}$ as $\max(\sigma, n^{-1}) \to 0$. In addition to assumptions on the errors ε_i, some design conditions are necessary to insure weak consistency of $\hat{\beta}$. We shall work with the following assumptions where $\|\cdot\|$ denotes the Euclidean norm:

(C1) $G_n(\gamma)$ converges pointwise to a function $G(\gamma)$ possessing a unique maximum at β_0, where $G_n(\cdot)$ is defined as

$$G_n(\gamma) \triangleq n^{-1}\sum_1^n \{F(x_i^T\beta_0)\log F(x_i^T\gamma) + F(-x_i^T\beta_0)\log F(-x_i^T\gamma)\};$$

(C2) $\sum_1^n (\|x_i\|)^2 = o(n^2)$;

(C3) $E(\|\varepsilon_1\|) < \infty$.

The condition (C1) is an assumption of convenience since for each n, $G_n(\cdot)$ is concave with a maximum at β_0. Weaker conditions could thus be employed by studying subsequences of $G_n(\cdot)$ [see Theorem 10.9, Rockafellar (1970)].

Consistency of $\hat{\beta}$ is proved in Theorem 5.1. This result is necessary to establish the following asymptotic expansion which is crucial to our investigation. Theorem 1 gives conditions such that

$$\begin{aligned}\hat{\beta} &= \beta_0 + n^{-1/2}S_n^{-1}(\beta_0)Z_n \\ &\quad + \sigma^2 S_n^{-1}(\beta_0)(J_{n,1} + J_{n,2})\beta_0 + o_p(\max(\sigma^2, n^{-1/2})),\end{aligned} \tag{2.5}$$

where

$$Z_n = n^{-1/2}\sum_1^n \left(y_i - F(x_i^T\beta_0)\right)x_i$$

$$J_{n,1} = -(2n)^{-1}\sum_1^n F^{(2)}(x_i^T\beta_0)x_i\beta_0^T\Sigma$$

$$J_{n,2} = -n^{-1}\sum_1^n F^{(1)}(x_i^T\beta_0)\Sigma.$$

THEOREM 1. (Asymptotic expansion of $\hat{\beta}$). *Assume that $\hat{\beta}$ is a consistent estimator of β_0 satisfying* (2.4). *Also assume*:

(A1) *There exists a positive definite matrix M, $\delta > 0$ and $N_0 < \infty$, such that $S_n(\gamma) \geq M$ whenever $n \geq N_0$ and $\|\gamma - \beta_0\| \leq \delta$*;

(A2) $n^{-1}\Sigma_1^n\|x_i\|^2 \to x^2 < \infty$, $\max_{1\leq i\leq n}\|x_i\| = o(\sigma^{-2})$;

(A3) $E(\varepsilon_1) = 0$, $E(\varepsilon_1\varepsilon_1^T) = I$, $E(\|\varepsilon_1\|^{2+\alpha}) < \infty$ *for some* $\alpha > 0$.

Then $\hat{\beta}$ has the expansion given in (2.5).

Note that the first part of (A2) implies $\max_{1\leq i\leq n}\|x_i\| = o(n^{1/2})$. This fact is used repeatedly in the proofs in Section 5. Assumptions (A1) and (A2) are sufficient to prove asymptotic normality for $S_n^{-1/2}(\beta_0)Z_n$ by using the Cramér–Wold device (Billingsley, 1979, Theorem 29.4) and appealing to Proposition 5.3.2 of Laha and Rohatgi (1979). Thus Theorem 1 indicates that with $\lambda = n^{1/2}\sigma^2$, we can expect $n^{1/2}(\hat{\beta} - \beta_0)$ to be approximately normally distributed with mean $\lambda S_n^{-1}(\beta_0)(J_{n,1} + J_{n,2})\beta_0$ and covariance $S_n^{-1}(\beta_0)$, when n is large and σ is small. When X_i is a mean of m replicates, $\sigma^2 \propto m^{-1}$ and λ describes the relationship between the sample size and the rate of replication. The asymptotic bias obviously decreases with increasing replication.

We can use expansion (2.5) to construct a corrected estimator, $\hat{\beta}_c$, which has smaller asymptotic bias. Before doing so we comment on the problems with $\hat{\beta}$ alluded to in the introduction.

BIAS AND ATTENUATION. Consider simple logistic regression through the origin with $\beta_0 > 0$. One expects to see attenuation, i.e., a negative first-order bias term. For most designs this is true. Somewhat surprisingly and completely at odds with the linear regression case, $S_n^{-1}(\beta_0)(J_{n,1} + J_{n,2})\beta_0$ can be positive. One design in which this occurs arises when most cases have very high or very low risk, i.e., $|x_i^T\beta_0|$ is large for most i.

HYPOTHESIS TESTING. Consider a two-group analysis of covariance, $x_i^T = (1, (-1)^i, d_i)$, $\beta_0 = (\beta_0, \beta_1, \beta_2)$. The covariance d_i is measured with error having variance σ^2. Often interest lies in testing hypotheses about the treatment effect β_1. A standard method to test $\beta_1 = 0$ is to compute its logistic regression estimate compared to the usual estimate of its asymptotic standard error. When the asymptotics of Theorem 1 are relevant and $n^{1/2}\sigma \to \lambda > 0$, this test approaches its nominal level only if the second component of $S_n^{-1}(\beta_0)(J_{n,1} + J_{n,2})\beta_0$

approaches zero. Letting s_2 denote the second row of $S_n^{-1}(\beta_0)$ this is achieved only if

$$n^{-1}\sum_1^n s_2^T x_i F^{(2)}(x_i^T\beta_0)\sigma^2\beta_2^2 \to 0.$$

This will not hold in the common epidemiologic situation in which the true covariables are not balanced across the two treatments. Thus, when substantial measurement error occurs in a nonrandomized study, there will be bias in the asymptotic levels of the usual tests.

3. Accounting for measurement error. In this section three alternative approaches to estimatation are studied. The first is based on expansion (2.5) and is distribution free in the sense that only moment assumptions are made about the measurement errors. The second two methods are based on an assumption of normally distributed errors; their asymptotic properties are then studied under more general conditions.

3.1. Adjusting for bias in $\hat\beta$. Write $b_n = S_n^{-1}(\beta_0)(J_{n,1} + J_{n,2})\beta_0$ and $\hat b_n = \hat S_n^{-1}(\hat\beta)(\hat J_{n,1} + \hat J_{n,2})\hat\beta$, where

$$\begin{aligned} \hat S_n(\gamma) &= n^{-1}\sum_1^n F^{(1)}(X_i^T\gamma)X_iX_i^T \\ \hat J_{n,1} &= -(2n)^{-1}\sum_1^n F^{(2)}(X_i^T\hat\beta)X_i\hat\beta^T\hat\Sigma, \qquad (3.1) \\ \hat J_{n,2} &= -n^{-1}\sum_1^n F^{(1)}(X_i^T\hat\beta)\hat\Sigma; \end{aligned}$$

$\hat b_n$ depends only on the observed data and, under the conditions of Theorem 1 and (2.2), approximates b_n in the sense that $\hat b_n - b_n = o_p(1)$ as $\min(n, \sigma^{-1}) \to \infty$. This result suggests that the bias-corrected estimator $\hat\beta_c \triangleq \hat\beta - \hat\sigma^2\hat b_n$ should have smaller asymptotic bias for large n and small σ. We state these results as a theorem.

THEOREM 2. *Assume the conditions of Theorem* 1 *and* (2.2). *Then $\hat\beta_c$ is consistent and*

$$\hat\beta_c = \beta_0 + n^{-1/2}S_n^{-1}(\beta_0)Z_n + o_p(\max(\sigma^2, n^{-1/2})).$$

REMARKS. Theorem 2 follows from Theorems 5.1 and 5.2 which are proved using the following characterization of $\hat\beta_c$. Note that $\hat\beta_c = (I - \hat\sigma^2\hat B_n)\hat\beta$ where $\hat B_n = \hat S_n^{-1}(\hat\beta)(\hat J_{n,1} + \hat J_{n,2})$. Since $X_i^T\hat\beta = X_i^T(I - \hat\sigma^2\hat B_n)^{-1}\hat\beta_c$ it follows that $\hat\beta_c$ maximizes (2.3) when $c_i = \hat x_{i,c}$, defined as

$$\hat x_{i,c} = X_i + \hat\sigma^2(I - \hat\sigma^2\hat B_n^T)^{-1}\hat B_n^T X_i. \qquad (3.2)$$

In this sense $\hat\beta_c$ is a type of two-stage estimator obtained by doing logistic regression with $\hat x_{i,c}$ replacing X_i.

The estimator $\hat{\beta}_c$ is not unbiased just less biased. The Monte Carlo study of Section 4 shows that in realistic sampling situations the reduction in bias can be substantial.

3.2. Normal measurement error. When measurement error is present there is an added source of variability which is not accounted for by model (1.1). We now expand this model by assuming that (ε_i) are normally distributed, an assumption which is not unreasonable in some situations. The log-likelihood for estimating β_0 and $x_1, \dots, x_n$ is then

$$(3.3)\qquad \sum_1^n \Big\{ y_i \log\big(F(x_i^T\beta)\big) + (1 - y_i)\log\big(F(-x_i^T\beta)\big) - (2\sigma^2)^{-1}(X_i - x_i)^T\Sigma^{-1}(X_i - x_i)\Big\}.$$

The vectors $\tilde{\beta}_f$, $\tilde{c}_i$ maximizing (3.3) satisfy

$$\sum_1^n \big(y_i - F(\tilde{c}_i^T\tilde{\beta}_f)\big)\tilde{c}_i = 0$$

$$\tilde{c}_i = X_i + \big(y_i - F(\tilde{c}_i^T\tilde{\beta}_f)\big)\sigma^2\Sigma\tilde{\beta}_f, \qquad i = 1, \dots, n.$$

There are two problems with this estimator—it depends on the unknown matrix $\sigma^2\Sigma$ and solving for $\tilde{\beta}_f$ and $(\tilde{c}_i)$ is difficult. For these reasons we suggest an approximate version of $\tilde{\beta}_f$. Noting the form of $\tilde{c}_i$ we let

$$(3.4)\qquad \hat{x}_{i,f} = X_i + \big(y_i - F(X_i^T\hat{\beta})\big)\hat{\sigma}^2\hat{\Sigma}\hat{\beta}$$

and define $\hat{\beta}_f$ as the estimator obtained by maximizing (2.3) with $c_i = \hat{x}_{i,f}$; $\hat{\beta}_f$ is consistent and has an asymptotic expansion given in the next theorem. The assumption of normal errors is not necessary for Theorem 3.

THEOREM 3. *Assume the conditions of Theorem* 1 *and* (2.2). *Then* $\hat{\beta}_f$ *is consistent and*

$$(3.5)\quad \hat{\beta}_f = \beta_0 + n^{-1/2}S_n^{-1}(\beta_0)Z_n + \sigma^2 S_n^{-1}(\beta_0)J_{n,1}\beta_0 + o_p(\max(\sigma^2, n^{-1/2})).$$

REMARKS. A comparison of (2.5) and (3.5) indicates that our approximate functional maximum likelihood estimator, $\hat{\beta}_f$, and the uncorrected estimator, $\hat{\beta}$, have first-order biases of the same magnitude. It can be shown (Stefanski, 1983) that the bias term in $\hat{\beta}_f$ is not due to our one-step modification nor to use of $\hat{V}$ in place of V, i.e., when V is known the full functional maximum likelihood estimator, $\tilde{\beta}_f$, also has the expansion given in (3.5) even in the case of simple logistic regression. This is in contrast to linear regression where, if the ratio of error variances is known or if there is finite replication of the predictors, the functional maximum likelihood estimator is consistent.

Our final estimator starts with an assumption of normal errors and exploits the consequences of sufficiency. Given $\sigma^2\Sigma$ and β_0, a sufficient statistic for estimating x_i is $\bar{c}_i(\beta_0) = X_i + \sigma^2(y_i - \frac{1}{2})\Sigma\beta_0$. It follows that the distribution of

y_i given $\bar{c}_i(\beta_0)$ does not depend on x_i. The reason for using this particular sufficient statistic is that

$$P\{y_i = 1|\bar{c}_i(\beta_0)\} = F(\bar{c}_i^T(\beta_0)\beta_0) \tag{3.6}$$

and hence the score equation

$$\sum_1^n (y_i - F(\bar{c}_i^T(\beta)\beta))\bar{c}_i(\beta) = 0 \tag{3.7}$$

is unbiased for β_0. The conditional probability (3.6) also suggests another approach—replace c_i by $\bar{c}_i(\gamma)$ in (2.3) and maximize the resulting expression as a function of γ. However, a simple calculation indicates that the resulting score equation is not unbiased for β_0, thus we will confine our attention to (3.7).

Equation (3.7) can have multiple solutions not all which produce a consistent sequence of estimators. Since $\bar{c}_i(\beta)$ also depends on the unknown matrix $\sigma^2\Sigma$, we propose the following modification: Let

$$\hat{x}_{i,s} = X_i + \hat{\sigma}^2(y_i - \tfrac{1}{2})\hat{\Sigma}\hat{\beta} \tag{3.8}$$

and define $\hat{\beta}_s$, the sufficiency estimator, as the maximizer of (2.3) when c_i is replaced by $\hat{x}_{i,s}$. This estimator is consistent and has the expansion given in the next theorem.

THEOREM 4. *Assume the conditions of Theorem* 1 *and* (2.2). *Then* $\hat{\beta}_s$ *is consistent and*

$$\hat{\beta}_s = \beta_0 + n^{-1/2}S_n^{-1}(\beta_0)Z_n + o_p(\max(\sigma^2, n^{-1/2})). \tag{3.9}$$

REMARKS 1. Theorem 4 does not require the assumption of normal measurement error. Also, $\hat{\beta}$ can be replaced by any consistent estimator in the definition of $\hat{x}_{i,s}$. The effects of nonnormal measurement error and our particular choice of $\hat{x}_{i,s}$ become apparent only when $\hat{\beta}_s$ is expanded through terms of order $\max^2(\sigma^2, n^{-1/2})$. This analysis is lengthy and is not presented here [see Stefanski (1983)].

2. It is possible to define a sufficiency estimator for a large class of measurement-error models. In particular, we have in mind the generalized linear models with canonical link functions (McCullagh and Nelder, 1983). A complete exposition of this theory will appear elsewhere.

In the discussion following Theorem 1, it was noted that $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with nonzero mean provided $n^{1/2}\sigma^2 \to \lambda$. It follows from Theorems 2 and 4 that both $n^{1/2}(\hat{\beta}_c - \beta_0)$ and $n^{1/2}(\hat{\beta}_s - \beta_0)$ are asymptotically normal with zero means under the same conditions. Furthermore, it can be shown that for $\hat{\beta}_c$ and $\hat{\beta}_s$ asymptotic normality is obtained under the weaker condition $n^{1/2}\sigma^4 \to \lambda$ [see Stefanski (1983) for details].

In the next section results from a small Monte Carlo study are presented.

4. Monte Carlo.

We conducted a small simulation experiment to determine the relative merits of the four estimators $\hat{\beta}$, $\hat{\beta}_c$, $\hat{\beta}_f$, and $\hat{\beta}_s$. The model for the study was

$$\Pr\{y_i = 1|d_i\} = F(\alpha + \beta d_i), \qquad i = 1, \dots, n, \tag{4.1}$$

where $F(\cdot)$ is defined in (1.1).

As our estimators are derived for the functional case, one possible Monte Carlo study would have consisted of generating for fixed $(d_1, \dots, d_n)$ a sequence of response vectors $(y_1, \dots, y_n)$ according to (4.1), and a sequence of measurement-error vectors. This would allow evaluation of the estimators' performance for the particular design $(d_1, \dots, d_n)$. However, several different designs would have to be studied in order to obtain a useful overall measure of performance. We opted instead for a study in which at each step the design $(d_1, \dots, d_n)$ is generated at random and, in turn, a single response vector and measurement-error vector are generated. After a number of such steps are completed, the overall performance of the estimators is investigated [c.f. Olkin, Petkau, and Zidek (1981) and Dempster, Schatzoff, and Wermuth (1977)]. We believe this approach better indicates the estimators' performance in a wide variety of settings.

We considered these sampling situations where χ_1^2 denotes a chi-squared random variable with one degree of freedom:

$$\text{(I)} \qquad (\alpha, \beta) = (-1.4, 1.4), \qquad (d_i) \sim \text{Normal}(0, \sigma_d^2);$$

$$\text{(II)} \qquad (\alpha, \beta) = (-1.4, 1.4), \qquad (d_i) \sim \sigma_d(\chi_1^2 - 1)/\sqrt{2};$$

where,

$$\sigma_d^2 = 0.10, \qquad n = 300, 600.$$

For each case, we considered two sampling distributions for the measurement errors: (a) Normal$(0, \tau^2)$ and (b) a contaminated normal distribution, which is Normal$(0, \tau^2)$ with probability 0.90 and Normal$(0, 25\tau^2)$ with probability 0.10. For both cases, τ^2 was one third the variance of the true predictors $(\tau^2 = \sigma_d^2/3)$.

We believe these two sampling situations are realistic, but their representativeness is limited by the size of our study. The sample sizes $n = 300, 600$ may seem large, but our primary interest is in larger epidemiologic studies where such sample sizes are common. For example, Clark (1982) was motivated by a study with $n = 2580$, Hauck (1983) quotes a partially completed study with $n \geq 340$, and we have analyzed Framingham data for males aged 45–54 with $n = 589$. In addition, for the particular designs in our study, the unconditional probability of response $(y = 1)$ is only about 0.10. As in the case of Bernoulli trials, an estimator's variance decreases more like $1/np(1 - p)$ than $1/n$ and for this reason $np(1 - p)$ is sometimes called the effective sample size. In our study the effective sample sizes are only about 30 and 60 respectively. Furthermore, the results of the study suggest that correcting for measurement error when the effective sample size is small is unwarranted, except possibly when measurement error is larger than what we have studied.

The values of the predictor variance σ_d^2 and the normal measurement error variance τ^2 are similar to those found in the Framingham cohort mentioned in the previous paragraph when the predictor was $\log_e\{$(systolic blood pressure $-$ 75)/3$\}$, a standard transformation. The choice of (α, β) comes from Framingham data as well. All experiments were repeated 100 times.

In each experiment, we sampled two independent measurements $(D_{i,1}, D_{i,2})$ of each d_i; the observed covariate was $X_i = (1, \overline{D}_i)^T$, where $\overline{D}_i = (D_{i,1} + D_{i,2})/2$. Thus σ^2, the variance of $\overline{D}_i$, was equal to $(1/6)\sigma_d^2$ for the case of normal measurement error while for the contaminated normal error distribution $\sigma^2 = (3.4/6)\sigma_d^2$. The matrix $\sigma^2\Sigma$ has only one nonzero entry which was estimated by the sample variance of $(D_{i,1} - D_{i,2})/2$.

In addition to the four estimators presented in this paper, we included in the study a proposal due to Clark (1982). She suggests the estimator $\hat{\beta}_N$ obtained by maximizing (2.3) when c_i is replaced by $\hat{x}_{i,N} = X_i - \hat{\sigma}^2\hat{\Sigma}\hat{\Sigma}_X^{-1}(X_i - \hat{\mu})$ where $\hat{\mu}$ and $\hat{\Sigma}_X$ are the sample mean and covariance of the observed data. Motivation for this estimator derives from an assumption of normal errors and normal covariates x_i. In this case $E(x_i|X_i) = X_i - \sigma^2\Sigma\Sigma_X^{-1}(X_i - \mu)$ and hence $\hat{x}_{i,n}$ is a natural estimator of x_i. Theorems 5.1 and 5.2 can be used to prove consistency and derive an asymptotic expansion for this estimator. Like $\hat{\beta}$ and $\hat{\beta}_f$, $\hat{\beta}_N$ has a nonzero first-order bias although it is too lengthy to present here.

Sweeping conclusions cannot be made from such a small study. However, we can make the following qualitative suggestions. First $\hat{\beta}$ is less variable but more biased than the others. Sample sizes such as $n = 600$ as in the study or Clark's $n = 2580$ are such that bias dominates and hence are candidates for using corrected estimators. An opposite conclusion holds for small sample sizes where variance dominates. A second suggestion from **Table 1** is that when $\mathrm{Var}(\hat{\beta})$ is small relative to its bias [Case I(b), II(b), and when $n = 600$], the corrected estimators perform quite well.

Both $\hat{\beta}_s$ and $\hat{\beta}_f$ were defined via an assumption of normal errors yet they also performed well when the errors were contaminated normal [Cases I(b), II(b)]. Clark's estimator proved to be sensitive to the assumption of normal covariates; $\hat{\beta}_N$ performed very well in our study when the predictors were normally distributed, but it did have a noticeable drop in efficiency when the predictors were highly skewed (Case II). Finally, the corrected estimator $\hat{\beta}_c$, which was derived with no distributional assumptions for either the predictors or errors, performed well throughout the study.

In summary, the Monte Carlo results suggest that the estimators $\hat{\beta}_c$, $\hat{\beta}_f$, $\hat{\beta}_s$, and Clark's $\hat{\beta}_N$ are useful alternatives to $\hat{\beta}$ when covariates are measured with error. The pressing practical problem now appears to be how to delineate those situations in which ordinary logistic regression should be corrected for its bias. Studies of inference and more detailed comparisons of alternative estimators will be enhanced by the identification of those problems where measurement error severely affects the usual estimation and inference.

5. Proofs of theorems.

Consider the estimator $\tilde{\beta}$ obtained by maximizing (2.3) when c_i is replaced with $\tilde{x}_i$ where

$$\tilde{x}_i = x_i + \sigma v_i + \sigma^2 g_{in}. \tag{5.1}$$

In Theorem 5.1 we prove weak consistency of $\tilde{\beta}$ under conditions (C1), (C2), (C3),

TABLE 1

Results from our Monte Carlo study of the simple logistic regression model $\Pr\{y_i = 1 | d_i\} = F(\alpha + \beta d_i)$. *Observed covariates are* $X_i = (1, \bar{D}_i)^T$ *where* $\bar{D}_i$ *is the mean of two independent measurements of* d_i. *The normal measurement errors have variance* $\sigma_d^2 / 3$; *the contaminated normal errors have distribution function* $G(x) = 0.9\Phi(x/\tau) + 0.1\Phi(x/5\tau)$ *and variance* $(3.4/3)\sigma_d^2$. *"Efficiency" refers to mean-squared error efficiency with respect to ordinary logistic regression.*

	$\hat{\beta}$	$\hat{\beta}_c$	$\hat{\beta}_f$	$\hat{\beta}_N$	$\hat{\beta}_s$
CASE I(a). $(\alpha, \beta) = (-1.4, 1.4)$, $(d_i) \sim N(0, \sigma_d^2 = 0.1)$, *normal measurement error.*					
$n = 300$ Bias	−0.21	−0.04	−0.05	−0.02	−0.06
Std. Dev.	0.52	0.61	0.61	0.61	0.60
Efficiency	100%*	85%	85%	84%	88%
$n = 600$ Bias	−0.22	−0.05	−0.05	−0.02	−0.06
Std. Dev.	0.33	0.38	0.38	0.38	0.38
Efficiency	100%*	108%	106%	107%	108%
CASE I(b). Same as Case I(a) but measurement errors have the contaminated normal distribution.					
$n = 300$ Bias	−0.49	−0.16	−0.19	0.02	−0.20
Std. Dev.	0.34	0.48	0.48	0.54	0.46
Efficiency	100%*	143%	139%	121%	143%
$n = 600$ Bias	−0.53	−0.20	−0.21	−0.03	−0.22
Std. Dev.	0.24	0.33	0.34	0.38	0.33
Efficiency	100%*	223%	215%	234%	216%
CASE II(a). $(\alpha, \beta) = (-1.4, 1.4)$, $(d_i) \sim \sigma_d(\chi_1^2 - 1)/\sqrt{2}$, $\sigma_d^2 = 0.1$, *normal measurement error.*					
$n = 300$ Bias	−0.28	−0.05	−0.07	0.10	−0.08
Std. Dev.	0.47	0.58	0.57	0.66	0.56
Efficiency	100%*	90%	91%	69%	93%
$n = 600$ Bias	−0.27	−0.03	−0.04	0.11	−0.05
Std. Dev.	0.33	0.41	0.41	0.45	0.40
Efficiency	100%*	111%	110%	85%	112%
CASE II(b). Same as Case II(a) but measurement errors have the contaminated normal distribution.					
$n = 300$ Bias	−0.43	−0.13	−0.15	0.12	−0.17
Std. Dev.	0.33	0.44	0.45	0.53	0.43
Efficiency	100%*	141%	134%	103%	141%
$n = 600$ Bias	−0.46	−0.15	−0.16	0.10	−0.18
Std. Dev.	0.25	0.33	0.34	0.40	0.33
Efficiency	100%*	201%	190%	159%	194%

*By definition.

and

$$\text{(P1)} \qquad \sum_1^n \|g_{in}\|^2 = O_p(n).$$

In Theorem 5.2 an asymptotic expansion for $\tilde{\beta}$ is given. The consistency and asymptotic expansions of $\hat{\beta}$, $\hat{\beta}_c$, $\hat{\beta}_f$, and $\hat{\beta}_s$ follow from these general results by noting that X_i, $\hat{x}_{i,c}$, $\hat{x}_{i,f}$, and $\hat{x}_{i,s}$ all have the representation given in (5.1). We remind the reader that all the asymptotic expressions hold as $\max(\sigma, n^{-1}) \to 0$.

THEOREM 5.1 (Consistency). *Assume* (C1), (C2), (C3), *and* (P1), *then* $\tilde{\beta} - \beta_0 = o_p(1)$.

PROOF. Define $\tilde{L}_n(\gamma)$ to be the function obtained by taking $c_i = \tilde{x}_i$ in (2.3). The identity $\log(F(t)/(1 - F(t))) = t$ is used to show $\tilde{L}_n(\gamma) - G_n(\gamma) = R_{n,1} + R_{n,2}$, where

$$R_{n,1} = n^{-1}\sum_1^n \big(y_i - F(x_i^T\beta_0)\big)x_i^T\gamma$$

$$R_{n,2} = n^{-1}\sum_1^n \big\{y_i\big(\tilde{x}_i^T\gamma - x_i^T\gamma\big) + \log F\big(-\tilde{x}_i^T\gamma\big) - \log F\big(-x_i^T\gamma\big)\big\}.$$

Under (C2), $R_{n,1}$ has mean zero and asymptotically negligible variance, and by (C3) and (P1),

$$\|R_{n,2}\| \le 2\sigma\|\gamma\|n^{-1}\sum_1^n\|v_i + \sigma g_{in}\| = o_p(1).$$

Consequently (C1) implies that $\tilde{L}_n(\cdot)$ converges pointwise in probability to $G(\cdot)$. An appeal to Corollary II.2 of Andersen and Gill (1982) concludes the proof.

The consistency results follow by applying Theorem 5.1 first to $\hat{\beta}$, ($g_{in} = 0$) and then to $\hat{\beta}_c$, $\hat{\beta}_f$, and $\hat{\beta}_s$. Next we derive the asymptotic expansions for these estimators.

THEOREM 5.2 (Asymptotic expansion). *Assume* (P1) *and the conditions of Theorem* 1, *then*

$$\tilde{\beta} = \beta_0 + n^{-1/2}S_n^{-1}(\beta_0)Z_n + \sigma^2 S_n^{-1}(\beta_0)\{(J_{n,1} + J_{n,2})\beta_0 + b_{n,3} + b_{n,4}\}$$
$$+ o_p\big(\max(\sigma^2, n^{-1/2})\big),$$

where

$$b_{n,3} = n^{-1}\sum_1^n \big(y_i - F(x_i^T\beta_0)\big)g_{in},$$

$$b_{n,4} = -n^{-1}\sum_1^n F^{(1)}\big(x_i^T\beta_0\big)x_i g_{in}^T\beta_0,$$

where $S_n(\cdot)$ *is given in* (1.2), *and* Z_n, $J_{n,1}$, *and* $J_{n,2}$ *are defined in* (2.5).

Theorem 5.2 is proved with a series of lemmas. First we show how Theorems 1–4 follow as corollaries. Theorem 1 is immediate since $g_{in} \equiv 0$ for $\hat{\beta}$. For $\hat{\beta}_c$, $g_{in} = (\hat{\sigma}^2/\sigma^2)(I - \hat{\sigma}^2\hat{B}_n^T)^{-1}\hat{B}_n^T X_i$ where $\hat{B}_n = \hat{S}_n^{-1}(\hat{\beta})(\hat{J}_{n,1} + \hat{J}_{n,2})$. Assumptions (A2), (A3), Lemma 5.1, and (2.2) imply $b_{n,3} = o_p(1)$, and

$$-b_{n,4} = n^{-1}\sum_1^n F^{(1)}\big(x_i^T\beta_0\big)x_i X_i^T\hat{B}_n\big(I - \hat{\sigma}^2\hat{B}_n\big)^{-1}\beta_0$$
$$= S_n\big(\beta_0\big)\hat{B}_n\beta_0 + o_p(1)$$
$$= \big(J_{n,1} + J_{n,2}\big)\beta_0 + o_p(1),$$

thus proving Theorem 2.

For $\hat{\beta}_f$, $g_{in} = (\hat{\sigma}^2/\sigma^2)(y_i - F(X_i^T\hat{\beta}))\hat{\Sigma}\hat{\beta}$ and (A2), (A3), Lemma 5.1, and (2.2) imply $b_{n,4} = o_p(1)$, and

$$\begin{aligned} b_{n,3} &= n^{-1}\sum_1^n \big(y_i - F(x_i^T\beta_0)\big)^2\Sigma\beta_0 + o_p(1) \\ &= -J_{n,2}\beta_o + o_p(1). \end{aligned}$$

Theorem 3 follows. Finally for $\hat{\beta}_s$, $g_{in} = (\hat{\sigma}^2/\sigma^2)(y_i - \frac{1}{2})\hat{\Sigma}\hat{\beta}$. (A2), (A3), Lemma 5.1, and (2.2) imply

$$\begin{aligned} b_{n,3} &= n^{-1}\sum_1^n \big(y_i - F(x_i^T\beta_0)\big)\big(y_i - \tfrac{1}{2}\big)\Sigma\beta_0 + o_p(1) \\ &= -J_{n,2}\beta_0 + o_p(1) \\ b_{n,4} &= -n^{-1}\sum_1^n F^{(1)}(x_i^T\beta_0)\big(y_i - \tfrac{1}{2}\big)x_i\beta_0^T\Sigma\beta_0 + o_p(1) \\ &= -n^{-1}\sum_1^n F^{(1)}(x_i^T\beta_0)\big(F(x_i^T\beta_0) - \tfrac{1}{2}\big)x_i\beta_0^T\Sigma\beta_0 + o_p(1) \\ &= -J_{n,1}\beta_0 + o_p(1). \end{aligned}$$

In the last step we use the identity $F^{(2)}(t) = F^{(1)}(t)(1 - 2F(t))$. This proves Theorem 4. Notice that in deriving these results we used only the fact that $\hat{\beta} - \beta_0 = o_p(1)$. Thus the conclusions of theorems 3 and 4 remain unchanged if $\hat{\beta}$ is replaced by any other *consistent* estimator in the definitions of $\hat{x}_{i,f}$ and $\hat{x}_{i,s}$. In particular, this can be shown to imply that the fully iterated versions of the functional and sufficiency estimators (provided consistent versions are chosen) also satisfy Theorems 3 and 4, respectively (Stefanski, 1983).

The proof of Theorem 5.2 starts with the following weak law.

LEMMA 5.1. *Let $u_1, u_2, \ldots$ be independent random vectors such that $E(u_i) = 0$ for all i, and $E(|u_{ij}|^{1+\alpha}) \le B$ for all i and j, and some $\alpha > 0$ and $B < \infty$, where u_{ij} is the j*th *element of u_i. If $\Sigma_i^n|a_i| = O(n)$ and* $\max_{1\le i\le n}(|a_i|/n) = o(1)$ *then $n^{-1}\Sigma_1^n a_i u_i = o_p(1)$.*

PROOF. The proof of the lemma entails a routine verification of the assumptions of Theorem 5.2.3 (Chung, 1974) and is not given here.

LEMMA 5.2. *Under the conditions of Theorem* 1,

$$n^{-1}\sum_1^n \big(y_i - F(X_i^T\beta_0)\big)X_i = n^{-1/2}Z_n + \sigma^2(J_{n,1} + J_{n,2})\beta_0 + o_p\big(\max(\sigma^2, n^{-1/2})\big).$$

PROOF. $n^{-1}\Sigma_1^n(y_i - F(X_i^T\beta_0))X_i = T_{n,1} + T_{n,2}$, where

$$\begin{aligned} T_{n,1} &= n^{-1}\sum_1^n \big(y_i - F(X_i^T\beta_0)\big)x_i, \\ T_{n,2} &= \sigma n^{-1}\sum_1^n \big(y_i - F(X_i^T\beta_0)\big)v_i. \end{aligned}$$

A Taylor series expansion of $F(\cdot)$ shows that

$$T_{n,1} = n^{-1/2}Z_n + \sigma^2 J_{n,1}\beta_0 + n^{-1/2}Q_{n,1,\sigma} + \sigma^2(D_{n,1} + R_{n,1}),$$

where

$$Q_{n,1,\sigma} = -\sigma n^{-1/2}\sum_1^n F^{(1)}(x_i^T\beta_0)v_i^T\beta_0 x_i$$

$$D_{n,1} = -(2n)^{-1}\sum_1^n \Big\{F^{(2)}(x_i^T\beta_0)\big((v_i^T\beta_0)^2 - \beta_0^T\Sigma\beta_0\big)x_i\Big\}$$

$$R_{n,1} = -(2n)^{-1}\sum_1^n \big(F^{(2)}(\tilde{X}_i^T\beta_0) - F^{(2)}(x_i^T\beta_0)\big)(v_i^T\beta_0)^2 x_i,$$

and $\tilde{X}_i$ is on the line segment joining x_i to X_i. $Q_{n,1,\sigma}$ has mean zero and asymptotically negligible variance thus $n^{-1/2}Q_{n,1,\sigma} = o_p(n^{-1/2})$. Assumptions (A2) and (A3) and Lemma 5.1 are used to show $D_{n,1} = o_p(1)$. Also note that

$$\|R_{n,1}\| \le (2n)^{-1}\sum_1^n \|x_i\|(v_i^T\beta_0)^2\min(1, \sigma|v_i^T\beta_0|) \le A_n A_n^*,$$

where

$$A_n = \left(n^{-1}\sum_1^n \|x_i\|^2(v_i^T\beta_0)^2\right)^{1/2},$$

$$A_n^* = \left(n^{-1}\sum_1^n (v_i^T\beta_0)^2\min^2(1, \sigma|v_i^T\beta_0|)\right)^{1/2}.$$

Assumptions (A2) and (A3) and Lemma 5.1 imply $A_n = O_p(1)$ while (A3), the fact that $\max(n^{-1}, \sigma) \to 0$, and the Dominated Convergence Theorem imply $A_n^* = o_p(1)$. It follows that $\sigma^2(D_{n,1} + R_{n,1}) = o_p(\sigma^2)$. Combining these results we get

$$T_{n,1} = n^{-1/2}Z_n + \sigma^2 J_{n,1}\beta_0 + o_p\big(\max(\sigma^2, n^{-1/2})\big). \tag{5.2}$$

Another Taylor series expansion of $F(\cdot)$ shows that

$$T_{n,2} = \sigma^2 J_{n,2}\beta_0 + n^{-1/2}Q_{n,2,\sigma} + \sigma^2(D_{n,2} + R_{n,2}),$$

where

$$Q_{n,2,\sigma} = \sigma n^{-1/2}\sum_1^n \big(y_i - F(x_i^T\beta_0)\big)v_i$$

$$D_{n,2} = -n^{-1}\sum_1^n F^{(1)}(x_i^T\beta_0)(v_i v_i^T - \Sigma)\beta_0$$

$$R_{n,2} = -n^{-1}\sum_1^n \big(F^{(1)}(\tilde{X}_i^T\beta_0) - F^{(1)}(x_i^T\beta_0)\big)v_i v_i^T\beta_0,$$

and $\tilde{X}_i$ lies on the line segment joining x_i to X_i. $Q_{n,2,\sigma}$, $D_{n,2}$ and $R_{n,2}$ are all $o_p(1)$, and the proofs are analogous to those for $Q_{n,1,\sigma}$, $D_{n,1}$, and $R_{n,1}$, respectively. Consequently,

$$T_{n,2} = \sigma^2 J_{n,2}\beta_0 + o_p\big(\max(\sigma^2, n^{-1/2})\big). \tag{5.3}$$

Combining (5.2) and (5.3) completes the proof of the lemma.

LEMMA 5.3. *Assume the conditions of Theorem* 1 *and* (P1) *and define* $\tilde{H}_n(\gamma) = n^{-1}\Sigma_1^n(y_i - F(\tilde{x}_i^T\gamma))\tilde{x}_i$. *Then*

$$\tilde{H}_n(\beta_0) = n^{-1/2}Z_n + \sigma^2((J_{n,1} + J_{n,2})\beta_0 + b_{n,3} + b_{n,4}) + o_p(\max(\sigma^2, n^{-1/2})).$$

PROOF. $\tilde{H}_n(\beta_0) = W_{n,1} + W_{n,2} + W_{n,3} + W_{n,4}$, where

$$W_{n,1} = n^{-1}\sum_1^n (y_i - F(X_i^T\beta_0))X_i,$$

$$W_{n,2} = \sigma n^{-1}\sum_1^n (F(X_i^T\beta_0) - F(\tilde{x}_i^T\beta_0))(v_i + \sigma g_{in}),$$

$$W_{n,3} = \sigma^2 n^{-1}\sum_1^n (y_i - F(X_i^T\beta_0))g_{in},$$

$$W_{n,4} = n^{-1}\sum_1^n (F(X_i^T\beta_0) - F(\tilde{x}_i^T\beta_0))x_i.$$

Note that in light of (A2) and (P1)

$$\|W_{n,2}\| \le \sigma^2 n^{-1}\sum_1^n \|g_{in}\|(\|v_i\| + \sigma\|g_{in}\|) = o_p(\sigma^2).$$

Also,

$$\begin{aligned}\|W_{n,3} - \sigma^2 b_{n,3}\| &\le \sigma^2 n^{-1}\sum_1^n \left|F(x_i^T\beta_0) - F(X_i^T\beta_0)\right|\|g_{in}\| \\ &\le \|\beta_0\|\sigma^3 n^{-1}\sum_1^n \|v_i\|\,\|g_{in}\| \\ &\le \|\beta_o\|\sigma^3\left(n^{-1}\sum_1^n\|v_i\|^2\right)^{1/2}\left(n^{-1}\sum_1^n\|g_{in}\|^2\right)^{1/2} \\ &= o_p(\sigma^2),\end{aligned}$$

using (A3) and (P1). One term in a Taylor series expansion of $F(\cdot)$ and Lemma 5.1, (A2), and (P1) show that

$$\begin{aligned}\|W_{n,4} - \sigma^2 b_{n,4}\| &\le \sigma^2\|\beta_0\|^2 n^{-1}\sum_1^n (\sigma\|v_i\| + \sigma^2\|g_{in}\|)\|x_i\|\,\|g_{in}\| \\ &\le \sigma^2\|\beta_0\|^2\left\{\sigma n^{-1}\sum_1^n\|v_i\|\,\|x_i\|\,\|g_{in}\| + \sigma^2 n^{-1}\sum_1^n\|x_i\|\,\|g_{in}\|^2\right\} \\ &\le \sigma^2\|\beta_0\|^2\left\{\sigma\left(n^{-1}\sum_1^n\|v_i\|^2\|x_i\|^2\right)^{1/2}\left(n^{-1}\sum_1^n\|g_{in}\|^2\right)^{1/2}\right. \\ &\qquad\left. +\sigma^2\left(\max_{1\le i\le n}\|x_i\|\right)n^{-1}\sum_1^n\|g_{in}\|^2\right\} \\ &= o_p(\sigma^2).\end{aligned}$$

An expansion for $W_{n,1}$ is given in Lemma 5.2. Combining the above results proves the lemma.

Define

$$\tilde{S}_n(\gamma) = n^{-1}\sum_1^n F^{(1)}(\tilde{x}_i^T\gamma)\tilde{x}_i\tilde{x}_i^T$$

and note that

$$\tilde{S}_n(\gamma) = (\partial/\partial\gamma)\tilde{H}_n(\gamma), \tag{5.4}$$

where $\tilde{H}_n(\cdot)$ is defined in Lemma 5.3.

LEMMA 5.4. *Conditions* (*A*2), (*A*3), *and* (*p*1) *imply* $\tilde{S}_n(\bar{\beta}) - S_n(\bar{\beta}) = o_p(1)$ *for any* $\bar{\beta}$ *on the line segment joining* β_0 *and* $\tilde{\beta}$.

PROOF. $\tilde{S}_n(\bar{\beta}) - S_n(\bar{\beta}) = H_{n,1} + H_{n,2}$, where

$$H_{n,1} = n^{-1}\sum_1^n F^{(1)}(\tilde{x}_i^T\bar{\beta})(\tilde{x}_i\tilde{x}_i^T - x_ix_i^T),$$

$$H_{n,2} = n^{-1}\sum_1^n \{F^{(1)}(\tilde{x}_i^T\bar{\beta}) - F^{(1)}(x_i^T\bar{\beta})\}x_ix_i^T.$$

The boundedness of $F^{(1)}(\cdot)$ and some elementary inequalities are used to show

$$\|H_{n,1}\| \le n^{-1}\sum_1^n \left(2\|x_i\|\,\|\sigma v_i + \sigma^2 g_{in}\| + \|\sigma v_i + \sigma^2 g_{in}\|^2\right)$$
$$\le 2\left(n^{-1}\sum_1^n\|x_i\|^2\right)^{1/2}\left(n^{-1}\sum_1^n\|\sigma v_i + \sigma^2 g_{in}\|^2\right)^{1/2} + n^{-1}\sum_1^n\|\sigma v_i + \sigma^2 g_{in}\|^2.$$

Assumption (A2) implies $n^{-1}\Sigma_1^n\|x_i\|^2 = O_p(1)$ and (A3) and (P1) imply $n^{-1}\Sigma_1^n\|\sigma v_i + \sigma^2 g_{in}\|^2 = o_p(1)$. Thus $\|H_{n,1}\| = o_p(1)$ as $\min(\sigma^{-1}, n) \to \infty$. A Taylor series expansion of $F^{(1)}(\cdot)$ and the boundedness of $F^{(2)}(\cdot)$ are used to show

$$\|H_{n,2}\| \le \|\bar{\beta}\|n^{-1}\sum_1^n\|\sigma v_i + \sigma^2 g_{in}\|\,\|x_i\|^2$$
$$\le \|\bar{\beta}\|\left\{\sigma n^{-1}\sum_1^n\|v_i\|\,\|x_i\|^2 + \sigma^2\left(\max_{1\le i\le n}\|x_i\|\right)\left(n^{-1}\sum_1^n g_{in}^2\right)^{1/2}\left(n^{-1}\sum_1^n\|x_i\|^2\right)^{1/2}\right\}. \tag{5.5}$$

Assumption (A2) and Lemma 5.1 imply $n^{-1}\Sigma_1^n\|v_i\|\,\|x_i\|^2 = O_p(1)$, and (A2) and (P1) imply that the second term in (5.5) is $o_p(1)$. Thus $\|H_{n,2}\| = o_p(1)$ as $\min(\sigma^{-1}, n) \to \infty$ and the proof is complete.

LEMMA 5.5. *Assume* (P1) *and the conditions of Theorem* 1, *then*

$$\tilde{\beta} - \beta_0 = O_p(\max(\sigma^2, n^{-1/2})).$$

PROOF. Let $\tilde{H}_n(\cdot)$ be the function defined in Lemma 5.3. Consider the real-valued function of γ defined as $\tilde{J}_n(\gamma) = \tilde{H}_n^T(\gamma)(\tilde{\beta} - \beta_0)$. The Mean Value Theorem proves the existence of some $\bar{\beta}$ on the line segment joining $\tilde{\beta}$ to β_0 such that

$$\tilde{H}_n^T(\beta_0)(\tilde{\beta} - \beta_0) = (\tilde{\beta} - \beta_0)^T \tilde{S}_n(\bar{\beta})(\tilde{\beta} - \beta_0),$$

where $\tilde{S}_n(\cdot)$ is defined in (5.4).

It follows that $\|\tilde{\beta} - \beta_0\| \leq \|\tilde{H}_n(\beta_0)\| \lambda_{\min}^{-1}(\tilde{S}_n(\bar{\beta}))$ where $\lambda_{\min}(A) =$ minimum eigenvalue of A. By Lemma 5.4, $\tilde{S}_n(\bar{\beta}) - S_n(\bar{\beta}) = o_p(1)$ hence by (A1), $P\{\lambda_{\min}^{-1}(\tilde{S}_n(\bar{\beta})) \leq 2\lambda_{\min}^{-1}(M)\} \to 1$. Thus $\|\tilde{\beta} - \beta_0\|$ and $\|\tilde{H}_n(\beta_0)\|$ have the same order which, from Lemma 5.3, is $O_p(\max(\sigma^2, n^{-1/2}))$.

We are now in a position to prove Theorem 5.2.

Proof of Theorem 5.2. By definition $n^{-1}\Sigma_1^n(y_i - F(\tilde{x}_i^T\tilde{\beta}))\tilde{x}_i = 0$; expanding $F(\cdot)$ in a Taylor series shows that $\tilde{S}(\tilde{\beta} - \beta_0) = \tilde{H}_n(\beta_0)$, where

$$\tilde{S} = n^{-1}\sum_1^n F^{(1)}\big(x_i^T\bar{\beta}_i\big)\tilde{x}_i\tilde{x}_i^T$$

and for each i, $\|\bar{\beta}_i - \beta_o\| \leq \|\tilde{\beta} - \beta_0\|$. (A2), (A3), (P1), and the conclusion of Lemma 5.5 are used to show $\tilde{S} - S_n(\beta_0) = o_p(1)$. The theorem follows from Lemma 5.5.

Acknowledgment. This research was supported by the Air Force Office of Scientific Research under grant AFOSR-49620-82-C-0009. We thank Rob Abbott of the National Heart, Lung, and Blood Institute for suggesting the problem, David Ruppert for helpful discussions, and the referees for many useful comments on an earlier draft of this paper.

REFERENCES

Amemiya, Y. (1982). Estimators for the errors-in-variables model. Unpublished Ph.D. thesis, Iowa State University, Ames.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120.

Armstrong, B. (1984). Measurement error in the generalized linear model. To appear in *Comm. Statist.*

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics* **7** 327–339.

Billingsley, P. (1979). *Probability and Measure.* Wiley, New York.

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bailey, K. T. and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika* **71** 19–25.

Chung, K. L. (1974). *A Course in Probability Theory.* Academic, New York,

Clark, R. R. (1982). The errors-in-variables problem in the logistic regression model. Unpublished Ph.D. thesis, University of North Carolina, Chapel Hill.

Cox, D. R. (1970). *Analysis of Binary Data.* Chapman and Hall, London.

Dempster, A. P., Schatzoff, M. and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Assoc.* **72** 77–91.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70** 892–898.

FULLER, W. A. (1980). Properties of some estimators for the errors-in-variables model. *Ann. Statist.* **8** 407–422.

GLESER, L. R. (1981). Estimation in a multivariate "errors-in-variables" regression model: Large sample results. *Ann. Statist.* **9** 24–44.

GORDON, T. and KANNEL, W. E. (1968). *Introduction and General Background in the Framingham Study—The Framingham Study, Sections 1 and 2.* National Heart, Lung, and Blood Institute, Bethesda, Maryland.

HAUCK, W. W. (1983). A note on confidence bands for the logistic response curve. *Amer. Statist.* **37** 158–160.

LAHA, R. G. and ROHATGI, V. K. (1979). *Probability Theory.* Wiley, New York.

MADANSKY, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Amer. Statist. Assoc.* **54** 173–205.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MICHALIK, J. E. and TRIPATHI, R. C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *J. Amer. Statist. Assoc.* **75** 713–721.

OLKIN, I. A., PETKAU, J. A. and ZIDEK, J. A. (1981). A comparison of *n* estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **76** 637–642.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705–724.

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331–42.

ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton University Press, Princeton.

STEFANSKI, L. A. (1983). Influence and measurement error in logistic regression. Institute of Statistics Mimeo Series No. 1548, University of North Carolina, Chapel Hill.

WOLTER, K. M. and FULLER, W. A. (1982a). Estimation of nonlinear errors-in-variables models. *Ann. Statist.* **10** 539–548.

WOLTER, K. M. and FULLER, W. A. (1982b). Estimation of the quadratic errors-in-variables model. *Biometrika* **69** 175–182.

DEPARTMENT OF ECONOMIC AND
SOCIAL STATISTICS
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514

Comparison of Least Squares and Errors-in-Variables Regression, With Special Reference to Randomized Analysis of Covariance

RAYMOND J. CARROLL, PAUL GALLO, and LEON JAY GLESER*

In an errors-in-variables regression model, the least squares estimate is generally inconsistent for the complete regression parameter but can be consistent for certain linear combinations of this parameter. We explore the conjecture that, when the least squares estimate is consistent for a linear combination of the regression parameter, it will be preferred to an errors-in-variables estimate, at least asymptotically. The conjecture is false, in general, but it is true for some important classes of problems. One such problem is a randomized two-group analysis of covariance, upon which we focus.

KEY WORDS: Measurement error; Randomized studies; Functional models; Structural models; Asymptotic theory.

1. INTRODUCTION

The literature on the problem of linear regression when some of the predictors are measured with error is substantial (for example, see Reilly and Patino-Leal 1981). Recent work includes the theoretical study of Gleser (1981) and the important practical shrinkage suggestions of Fuller (1980). Also see Anderson (1984) and Healy (1980).

A subarea of this literature concerns two-group analysis of covariance when some of the predictors are measured with error (for example, see Lord 1960, Cochran 1968, DeGracie and Fuller 1972, and Cronbach 1976).

Lord (1960) discussed the case of one covariate measured with error. He noted that it may "happen. . .that the usual covariance analysis (least squares) will fail to detect a statistically significant difference between groups. . .when such a difference actually exists and can be detected by proper statistical procedures" (p. 309). He also gave a numerical example of this phenomenon.

Cochran (1968) and DeGracie and Fuller (1972) discussed two-group analysis of covariance, providing in particular some discussion of the case in which the true values of covariates are themselves random variables; this is usually called a "structural" model in the literature. They showed that if the covariables are unbalanced, as might happen in an observational study, then the meaurement error will cause least squares to inconsistently estimate the true treatment difference. In the sense of asymptotics, when the covariables are unbalanced one should then correct for measurement error if it is substantial; a global small-sample statement of this type cannot be made.

Now consider a completely randomized study, where the covariables will be balanced on average across the two treatments. In this case, Cochran (1968) and DeGracie and Fuller (1972) indicated that least squares will consistently estimate the treatment difference. The question that remains to be answered is: Should we correct for measurement error when the least squares estimate consistently estimates the treatment effect? It is the purpose of this article to partially answer this question. Using large-sample distribution theory, we show that in a balanced, completely randomized study with measurement error in the covariables, the least squares estimate of the treatment difference will be generally preferred when compared to a particular errors-in-variables regression estimator. This result can be generalized, so in a large class of problems, when least squares is consistent for a linear combination of the regression parameter, it will be preferred, at least asymptotically. Further, for a smaller but not insubstantial class of problems, when least squares is consistent for a linear combination of the regression parameter, it is the maximum likelihood estimate of this linear combination, taking the consistency into account.

2. THE NORMAL CASE WITH NO REPLICATION: TECHNICAL BACKGROUND

A special case of considerable interest occurs when all errors are normally distributed and no replicates of the variables measured with error are available. The general model considered here, which includes the analysis of covariance as a special case, is given by

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\ C &= X_2 + U \\ \beta &= [\beta_1^T, \beta_2^T]^T. \end{aligned} \tag{2.1}$$

Here, Y and ε are $(N \times 1)$ vectors, X_1 is an $(N \times p)$ matrix observed without error, and X_2 is an $(N \times q)$ matrix of true values that we cannot observe exactly. Rather, we observe C. The rows of the matrix (U, ε) will be assumed to be mutually independent normally distributed random vectors with mean zero and covariance matrix Σ.

In comparing least squares and errors-in-variables methods, we must pick a representative member of the latter class. In the main, we will do this by following Gleser (1981) for the case that no replicated estimates of X_2 are available; the replicated case will be discussed at the end of the article. Gleser studied the functional model in which $X_1 = (1, 1, \ldots, 1)^T$ and X_2 are considered fixed constants. A special case of his

* Raymond J. Carroll is Professor of Statistics, University of North Carolina, Chapel Hill, NC 27514. Paul Gallo is Group Leader, Preclinical Statistics, Lederle Laboratories, Pearl River, NY 10954. Leon Jay Gleser is Professor of Statistics, Purdue University, West Lafayette, IN 47907. Carroll's work was supported by Air Force Office of Scientific Research Contract AFOSR-F-49620-82-C-0009 and by United States Army Contract DAAG29-80-C-0041. Gleser's work was supported by National Science Foundation Grant DMS-8121948. The authors thank Robert D. Abbott and Leonard A. Stefanski for helpful discussions.

Journal of the American Statistical Association
December 1985, Vol. 80, No. 392, Theory and Methods

model assumes that there is a known matrix Σ_0 and an unknown constant σ^2 for which

$$\Sigma = \sigma^2 \Sigma_0 = \sigma^2 \begin{pmatrix} \Sigma_{uo} & 0 \\ 0 & 1 \end{pmatrix} . \qquad (2.2)$$

If Σ_u is the covariance matrix of the rows of U, then in (2.2) we are assuming that we know the ratio of the elements of Σ_u to σ_ε^2, the variance of the elements of ε. Gallo (1982) exhibited the maximum likelihood estimate of β, which is given in Appendix A.

He also proved the following.

Theorem 1 (from Gallo 1982). Suppose that

$$\Delta = \lim_{N\to\infty} N^{-1}(X_1, X_2)^T(X_1, X_2)$$

exists and is positive definite. Then if $\hat\beta_M$ is the functional maximum likelihood estimate, $N^{1/2}(\hat\beta_M - \beta)$ is asymptotically normally distributed with zero mean and covariance matrix

$$\operatorname{cov}(\hat\beta_M) = d\left\{\Delta^{-1} + \Delta^{-1}\begin{pmatrix} 0 & 0 \\ 0 & Q \end{pmatrix}\Delta^{-1}\right\},$$

where $d = [\beta_2^T, -1]\ \Sigma\ [\beta_2^T, -1]^T$ and $Q^{-1} = [I, \beta_2]\ \Sigma^{-1}\ [I, \beta_2]^T$.

3. MAIN RESULTS

There are instances other than randomized two-group analysis of covariance in which certain linear combinations of the least squares estimate are consistent for the same linear combinations of the parameter. Consider the model (2.1) with $\beta^T = (\beta_1^T, \beta_2^T)$ in which it is desired to estimate the parameter $\gamma^T\beta$, where $\gamma^T = (\gamma_1^T, \gamma_2^T)$. Partitioning Δ in (2.3) into components Δ_{ij}, informally the least squares estimate

$$\hat\beta_L = ((X_1, C)^T(X_1, C))^{-1}(X_1, C)^TY$$

converges in probability to

$$\begin{bmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} + \Sigma_u \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ -\Sigma_u\beta_2 \end{pmatrix} + \beta. \qquad (3.1)$$

This leads to the following result, which was proved formally by Gallo (1982).

Theorem 2. The least squares estimate $\gamma^T\hat\beta_L$ is consistent for $\gamma^T\beta$; that is, it converges in probability to $\gamma^T\beta$ for all β, σ^2, Σ_u, if and only if

$$\gamma_2^T = \gamma_1^T\Delta_{11}^{-1}\Delta_{12}. \qquad (3.2)$$

Computing the asymptotic distribution of least squares is fairly complicated. Recall that X_1 is observed exactly, and we will assume it has a column of ones; X_2 is measured with error as in (2.1). Suppose we are interested in estimating a linear combination $\gamma^T\beta$ for which least squares is known to be consistent—that is, (3.2) holds. Then the next result gives a description of an important case for which least squares will be asymptotically preferred to the functional maximum likelihood estimate.

Theorem 3. Make the following assumption.

Given X_1, the rows of $R = X_2 - X_1\Delta_{11}^{-1}\Delta_{12}$ are independent and identically distributed with mean zero and covariance

$$\Delta_{22.1} = \Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}. \qquad (3.3)$$

Further, suppose that R is distributed independently of ε and U. Define

$$\Lambda = (\Delta_{22.1} + \Sigma_u)^{-1}.$$

Then the least squares and functional maximum likelihood estimates $a^T\hat\beta_L$ and $a^T\hat\beta_M$ are asymptotically normally distributed with mean $a^T\beta$ and variances $\sigma^2(L)/N$, $\sigma^2(M)/N$, respectively, where $\sigma^2(L) \le \sigma^2(M)$. In fact,

$$\sigma^2(L) = \sigma^2(M) - (\gamma_1^T\Delta_{11}^{-1}\gamma_1)\beta_2^T\ \Sigma_u\ \Lambda\ \Sigma_u\ \beta_2$$

$$\sigma^2(M) = (\gamma_1^T\Delta_{11}^{-1}\gamma_1)(\sigma^2 + \beta_2^T\ \Sigma_u\ \beta_2).$$

The proof of Theorem 3 is sketched in Appendix B. The asymptotic distribution of least squares when (3.3) does not hold has been computed by Gleser, Carroll, and Gallo (1985), but here one need not necessarily prefer least squares. This issue is discussed in the next section. Note that assumption (3.3) holds if X_1 and X_2 are independent random matrices.

It may be considered a bit unfair to compare least squares to a "maximum likelihood estimator" that does not take into account the consistency of least squares. It turns out that, under normality assumptions, the maximum likelihood estimate of $\gamma^T\beta$ when it is known that least squares is consistent for $\gamma^T\beta$ is simply the least squares estimate of $\gamma^T\beta$. Specifically, we have the following.

Theorem 4. Suppose that the assumptions of Theorem 3 hold and that the rows of R are normally distributed independently of ε and U. Then the maximum likelihood estimate of $\gamma^T\beta$ given X_1 and subject to (3.2) is simply the least squares estimate of $\gamma^T\beta$.

4. EXAMPLES AND EXTENSIONS

Consider a completely randomized two-group analysis of covariance, with covariables subject to error. Formally, this problem can be subsumed into the more general structure (2.1) by letting X_2 be the covariables and

$$X_1^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ s_1 & s_2 & \cdots & s_N \end{pmatrix}, \qquad \beta_1 = \begin{pmatrix} \mu \\ a \end{pmatrix},$$

$$X_2^T = (x_{21}, x_{22}, \ldots, x_{2N}). \qquad (4.1)$$

Here the $\{s_i\}$ are zero–one variates representing the assignment to the two groups. The parameter of interest is a, the treatment or group effect. In the notation of Section 3, we wish to estimate $\gamma^T\beta$, where $\gamma_2 = 0$, $\gamma_1^T = (0, 1)$. By using Theorem 2, it is easy to show that least squares is consistent for the group effect a only when the limiting means of the covariables are the same for the two groups. If treatment assignment is random, then the least squares estimate is consistent, assumption (3.3) holds, and by Theorem 3 the least squares estimate of group effect has a smaller limiting variance than the maximum likelihood estimate.

It is reasonable to conjecture that complete randomization is not necessary for least squares to be preferred in the context of an analysis of covariance. For example, one might randomize in blocks or use alternative balancing schemes (see Wei 1978). The details of the proof of Theorem 3 might prove helpful in studying this conjecture.

It should be noted that in a balanced randomized study, the usual t test for treatment effect has correct nominal level asymptotically. Thus, from both an estimation and inferential standpoint, for large samples least squares will be preferred over the functional estimate.

When assumption (3.3) is violated, difficulties arise. For instance, consider analysis of covariance for the pure functional case in which the group assignment variables $\{s_i\}$ occur in the fixed sequence $\{-1, +1, -1, +1, \ldots\}$. Let the covariables $\{x_i\}$ be fixed. In a variety of circumstances, Gleser et al. (1985) have shown that the least squares estimate $\hat{a}_L$ of the treatment effect a satisfies

$$N^{1/2}(\hat{a}_L - a) = V + AN^{-1/2}\sum_{i=1}^{n} s_i x_i, \tag{4.2}$$

where A is a constant and V is asymptotically normally distributed. To obtain an asymptotic distribution for $N^{1/2}$ $(\hat{a}_L - a)$, one must make assumptions about the behavior of the second term on the right side of (4.2). Generalization of (4.2) and further discussion are given by Gleser et al. (1985).

Even if X_1 and X_2 are random variables, violation of (3.3) can cause failure of the main conclusion of Theorem 3. The reason is that in this case the limit distribution of least squares can depend on the fourth moment of X_1 whereas that of the functional estimate depends only on the second moment of X_1.

Finally, in some instances an assumption such as (2.2) will not be tenable, so a functional estimate cannot be computed. There are many ways out of this dilemma. One is to take independent replicates of C_1, C_2 of X_2 in (2.1). One can compute the normal theory functional estimate in this case and obtain a result similar to Theorem 1 but more general in the sense that the underlying random variables need not actually be normally distributed. The computation of this functional estimate and its asymptotic distribution theory are available in, for example, Gallo (1982).

5. CONCLUSION

In a particular errors-in-variables regression model, we have shown that least squares will often be asymptotically more efficient than a particular functional regression estimate, when the former is known to be consistent. This happens in particular when those variables X_2 subject to error are distributed independently of those variables X_1 measured without error, or more generally when X_2 follows a linear regression in X_1. An important special case of this least squares preference phenomenon is a randomized analysis of covariance in which one wants to estimate the treatment effect. Finally, if the linear regression of X_2 on X_1 follows a multinormal distribution, and if it is known that the least squares estimate is consistent for the linear combination $\gamma^T\beta$, then the least squares estimate is the maximum likelihood estimate for $\gamma^T\beta$.

APPENDIX A: THE MAXIMUM LIKELIHOOD ESTIMATOR FOR MODEL (2.1)

Define $L = I - X_1(X_1^TX_1)^{-1}X_1^T$ and $W = [C, Y]^TL[C, Y]$. Let θ be the smallest eigenvalue of Σ_0^{-1} W, where Σ_0 is given in (2.2). Define

$$C_* = [X_1, C] = [X_1, X_2 + U],$$

$$D = C_*^TC_* - \theta\begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{uo}\end{pmatrix}.$$

The matrix D is nonsingular with probability one, and the functional estimate is $\hat{\beta}_M = D^{-1}C_*^TY$. The formula for $\hat{\beta}_M$ is derived by Gallo (1982) and relies on similar work of Gleser (1981) and Healy (1980).

APPENDIX B: THE ASYMPTOTIC DISTRIBUTION OF LEAST SQUARES

The following general result can be justified formally and is at the heart of the analysis of covariance calculations. We sketch herein a proof without stating all the necessary regularity conditions. Let $e_*^T = (1, 1, \ldots, 1)$.

Lemma 1. Define

$$\Delta_{22.1} = \Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}, \qquad \wedge = (\Delta_{22.1} + \Sigma_u)^{-1},$$

and suppose that γ satisfies (3.2) as well as

$$N^{-1/2}X_1^T(R + U) = O_p(1), \tag{B.1}$$

where $R = X_2 - X_1\Delta_{11}^{-1}\Delta_{12}$. Then the least squares estimate satisfies

$$N^{1/2}\gamma^T(\hat{\beta}_L - \beta) = N^{-1/2}\gamma_1^T\Delta_{11}^{-1}X_1^T(\varepsilon - U\beta_2) + N^{-1/2}\gamma_1^T\Delta_{11}^{-1}X_1^T(R + U) + o_p(1), \tag{B.2}$$

where $\xi = \wedge\,\Sigma_u\,\beta_2$.

Proof (Sketch). Define $C_* = [X_1, X_2 + U]$. Then

$$\frac{(C_*^TC_*)}{N}(\hat{\beta}_L - \beta) + \begin{pmatrix} 0 \\ \Sigma_u\beta_2\end{pmatrix} = C_*^T(\varepsilon - U\beta_2)/N + \begin{pmatrix} 0 \\ \Sigma_u\beta_2\end{pmatrix}. \tag{B.3}$$

Multiply both sides of (B.3) by $N^{1/2}\gamma^T(C_*^TC_*/N)^{-1}$ to get

$$N^{1/2}\gamma^T(\hat{\beta}_L - \beta) = N^{1/2}\gamma^T(C_*^TC_*/N)^{-1}\left(C_*^T(\varepsilon - U\beta_2)/N + \begin{pmatrix} 0 \\ \Sigma_u\beta_2\end{pmatrix}\right) - N^{1/2}\gamma^T(C_*^TC_*/N)^{-1}\begin{pmatrix} 0 \\ \Sigma_u\beta_2\end{pmatrix}. \tag{B.4}$$

By Slutsky's theorem, the first term on the right-hand side of (B.4) equals

$$N^{1/2}\gamma^T\Delta + \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_u\end{pmatrix}^{-1} C_*^T(\varepsilon - U\beta_2)/N + \begin{pmatrix} 0 \\ \Sigma_u\beta_2\end{pmatrix} + o_p(1) = N^{-1/2}\gamma_1^T\Delta_{11}^{-1}X_1^T(\varepsilon - U\beta_2) + o_p(1), \tag{B.5}$$

which is the same as the first term on the right-hand side of (B.2). The second term in (B.4) is

$$N^{1/2}\gamma_1^T(X_1^TX_1)^{-1}X_1^T(R + U)W\,\Sigma_u\,\beta_2, \tag{B.6}$$

where

$$W \xrightarrow{P} (\Delta_{22.1} + \Sigma_u)^{-1}, \qquad (X_1^TX_1/N) \longrightarrow \Delta_{11}.$$

By (B.1), this completes the proof.

One should note that (B.1) is satisfied in the randomized two-group analysis of covariance.

Using Lemma 1 and writing for the analysis of covariance

$$X_2^T = (x_{21}\, x_{22} \cdots x_{2N})$$

$$U^T = (u_1\, u_2 \cdots u_N)$$

$$m_2 = N^{-1}\sum_{i=1}^{N} x_{2i},$$

we see that

$$N^{1/2}(\hat{a}_L - a) = N^{-1/2}\sum_{i=1}^{N} s_i\{\varepsilon_i + (\eta - \beta_2)^T\mathrm{u}_i + \eta^T(x_{2i} - m_2)\}, \tag{B.7}$$

$$\eta = (o^{-2}\Delta_{22.1} + \Sigma_u)^{-1}\,\Sigma_u\,\beta_2.$$

Expression (B.7) shows why Theorem 3 may apply to alternative randomization schemes.

Proof of Theorem 3. The form of $\sigma^2(M)$ follows directly from Theorem 1. The form of $\sigma^2(L)$ follows from (B.2) and the assumptions of the theorem.

Proof of Theorem 4. First assume that $\Delta_{11}^{-1}\Delta_{12}$ is known. Define

$$\pi = (I, \Delta_{11}^{-1}\Delta_{12})\beta$$

$$\pounds = \sigma^2 + \beta_2^T \Sigma_u \beta_2 - \beta_2^T \Sigma_u \wedge \Sigma_u \beta_2.$$

Given (X_1, C), we have

$$(Y \mid X_1, C) = X_1\pi + S\wedge\Delta_{22.1}\beta_2 + F, \tag{B.8}$$

where $S = C - X_1\Delta_{11}^{-1}\Delta_{12}$ and the rows of F are independent normal random variables with mean zero and variances $\pounds^2$. If we define

$$\Sigma_x = \sigma^{-2}\Delta_{22.1}, \quad \xi = \wedge\Delta_{22.1}\beta_2, \quad L = \wedge^{-1},$$

then it is fairly direct to show that the mapping of $\beta_1, \beta_2, \sigma^2, \Delta_{22.1}$ to π, ξ, σ^2, L is one to one from the space $\{\sigma^2 > 0, \Delta_{22.1} > 0\}$ to the space $\{\sigma^2 > 0, L - \sigma^2 \Sigma_{uo} > 0\}$.

One can show next that the map π, ξ, L, σ^2 to $\pi, \xi, L, \pounds^2$ is also one to one onto the space $\{\pounds^2 > 0, L > 0\}$. To see this, note that

$$\begin{aligned}\pounds^2 &= \sigma^2 + \beta_2^T\Delta_{22.1}\beta_2 - \beta_2^T\Delta_{22.1}\wedge\Delta_{22.1}\beta_2 \\ &= \sigma^2 + \xi^T L(L - \sigma^2 \Sigma_{uo})^{-1} L\xi - \xi^T L = H(\sigma^2).\end{aligned}$$

Thus $\pounds^2$ is a function of π, ξ, L, σ^2. For the converse we must show that given $\pi, \xi, L, \pounds^2$ there exists $\sigma^2 > 0$ such that $\pounds^2 = H(\sigma^2)$ and $L - \sigma^2 \Sigma_{uo} > 0$. Write

$$\Sigma_{uo}^{-1/2} L \Sigma_{uo}^{-1/2} = \Gamma D \Gamma^T,$$

where Γ is an orthogonal matrix and D is diagonal with elements $d_1 \geq d_2 \geq \cdots \geq d_p$. Then

$$H(\sigma^2) = \sigma^2 - \xi^T L\xi + \sum_{K=1}^{p} (\Gamma^T \Sigma_{uo}^{-1/2} L\xi)_K^2/(d_k - \sigma^2),$$

where $(\Gamma^T \Sigma_{uo}^{-1/2} L\xi)_K$ is the Kth element of the matrix. Moreover, $L - \sigma^2 \Sigma_{uo} > 0$ if and only if $D - \sigma^2 I_p > 0$ if and only if $0 < \sigma^2 < d_p$. Hence we must show that there exists σ^2 such that $0 < \sigma^2 < d_p$ and $H(\sigma^2) = \pounds^2$. However, $H(\sigma^2)$ is continuous and increasing in σ^2. Further,

$$\lim_{\sigma \to 0} H(\sigma^2) = 0 < \pounds^2 \quad \text{and} \quad \lim_{\sigma^2 \to d_p} H(\sigma^2) = \infty > \pounds^2.$$

Hence a solution exists, and the map is one to one. We next complete the proof of Theorem 4.

Now, the maximum likelihood estimates of π and ξ are seen from (B.8) to be

$$\{(X_1, S)^T(X_1, S)\}^{-1}(X_1, S)^T Y.$$

Since the column space of (X_1, C) is the same as the column space of (X_1, S), it follows that, given $(X_1, S, \Delta_{11}^{-1}\Delta_{22})$, the maximum likelihood and least squares estimates of π coincide; that is,

$$\pi(\text{MLE}) = (I, \Delta_{11}^{-1}\Delta_{12})\hat{\beta}_L.$$

This means that $\gamma^T\hat{\beta}_L$ is the maximum likelihood estimate (MLE) of $\gamma^T\pi$, given X_1, S and $\Delta_{11}^{-1}\Delta_{12}$. Since, under (3.2), $\gamma^T\pi = \gamma^T\beta$, the proof is complete.

[*Received September 1982. Revised April 1985.*]

REFERENCES

Anderson, T. W. (1984), "Estimating Linear Statistical Relationships," *The Annals of Statistics*, 12, 1–45.

Cochran, W. G. (1968), "Errors of Measurement in Statistics," *Technometrics*, 10, 637–666.

Cronbach, L. J. (1967), "On the Design of Educational Measures," in *Advancement in Psychological and Educational Measurement*, eds. D. N. M. DeGruijter and L. J. Th. van der Kamp, New York: John Wiley.

DeGracie, J. S., and Fuller, W. A. (1972), "Estimation of the Slope and Analysis of Covariance When the Concomitant Variable Is Measured With Error," *Journal of the American Statistical Association*, 67, 930–937.

Fuller, W. A. (1980), "Properties of Some Estimators for the Errors-in-Variables Model," *The Annals of Statistics*, 8, 407–422.

Gallo, P. P. (1982), "Properties of Estimators in Errors-in-Variables Regression Models," unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill.

Gleser, L. J. (1981), "Estimation in a Multivariate 'Errors in Variables' Regression Models: Large Sample Results," *The Annals of Statistics*, 9, 24–44.

Gleser, L. J., Carroll, R. J., and Gallo, P. P. (1985), "The Limiting Distribution of Least Squares in an Errors-in-Variables Linear Regression Model," Mimeo Series No. 1577, University of North Carolina.

Healy, J. D. (1980), "Maximum Likelihood Estimation of a Multivariate Linear Functional Relationship," *Journal of Multivariate Analysis*, 10, 243–251.

Lord, F. M. (1960), "Large-Sample Covariance Analysis When the Control Variable Is Fallible," *Journal of the American Statistical Association*, 55, 307–321.

Reilly, P. M., and Patino-Leal, H. (1981), "A Bayesian Study of the Errors-in-Variables Model," *Technometrics*, 23, 221–231.

Wei, L. J. (1978), "Adaptive Biased Coin Designs," *The Annals of Statistics*, 6, 92–100.

Biometrika (1987), **74**, 4, *pp.* 703-16
Printed in Great Britain

Conditional scores and optimal scores for generalized linear measurement-error models

BY LEONARD A. STEFANSKI

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.

AND RAYMOND J. CARROLL

Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27514, U.S.A.

SUMMARY

This paper studies estimation in generalized linear models in canonical form when the explanatory vector is measured with independent normal error. For the functional case, i.e. when the explanatory vectors are fixed constants, unbiased score functions are obtained by conditioning on certain sufficient statistics. This work generalizes results obtained by the authors (Stefanski & Carroll, 1985) for logistic regression. In the case that the explanatory vectors are independent and identically distributed with unknown distribution, efficient score functions are identified. Related results for the structural case are given by Bickel & Ritov (1987).

Some key words: Conditional score function; Efficient score function; Functional model; Generalized linear model; Measurement error; Structural model.

1. INTRODUCTION

Given a covariate p-vector $U = u$, assume that Y has the density

$$h_Y(y;\, \theta, u) = \exp\left\{\frac{y(\alpha+\beta^{\mathrm{T}}u) - b(\alpha+\beta^{\mathrm{T}}u)}{a(\phi)} + c(y, \phi)\right\} \qquad (1·1)$$

with respect to a σ-finite measure $m(.)$. In (1·1), $\theta^{\mathrm{T}} = (\alpha, \beta^{\mathrm{T}}, \phi)$; $a(.)$, $b(.)$ and $c(.,.)$ are known functions; and the dominating measure $m(.)$ does not depend on θ or u. The density (1·1) is that of a generalized linear model in canonical form (McCullagh & Nelder, 1983, Ch. 2). Suppose that u cannot be observed but that k independent measurements, $X = (X_1, \ldots, X_k)$, of u are available. When measurement error is normally distributed the matrix X has the density

$$h_X(x;\, \theta, u) = \prod_{j=1}^{k} \frac{(2\pi)^{-\frac{1}{2}p}}{|\bar{\Omega}|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(x_j - u)^{\mathrm{T}}\bar{\Omega}^{-1}(x_j - u)\}, \qquad (1·2)$$

where $\bar{\Omega}$ is the covariance matrix of the measurement-error vector. Together (1·1) and (1·2) define a generalized linear measurement-error model with normal measurement error. If for a sample (Y_i, X_i) $(i = 1, \ldots, n)$ the covariables $\{u_i\}$ are unknown constants, a functional model is obtained; if $\{u_i\}$ are independent and identically distributed random vectors from some unknown distribution, a structural model is obtained (Kendall &

Stuart, 1979, Ch. 29). In the present paper the problem of deriving unbiased scores for θ in both functional and structural models is studied.

There is a vast literature on this problem when (1·1) is a normal density. This dates back to Adcock (1878) and has been reviewed by Anderson (1976); see also Moran (1971). Recently there has been considerble interest in nonlinear measurement-error models; see Prentice (1982); Wolter & Fuller (1982a, 1982b), Carroll et al. (1984), Stefanski (1985) and Stefanski & Carroll (1985).

The density (1·1) includes normal, Poisson, logistic and gamma regression models. The key feature these models have in common is a natural sufficient statistic for u when all other parameters are fixed. The same is true of the normal density in (1·2). In fact (1·2) could be replaced with any density possessing a natural sufficient statistic for u when other parameters are fixed and much of the theory in this paper holds with little or no modification.

In § 2 functional models are studied. This work generalizes results of Stefanski & Carroll (1985) on logistic regression. Structural models are studied in § 3 and efficient score functions for estimating θ are identified. Related work includes that of Bickel & Ritov (1987).

If the covariates $u_1, \ldots, u_n$ are observed without error the maximum likelihood estimator of θ maximizes $\Sigma \log h_Y(Y_i; \theta, u_i)$. Let $\bar{X}_i$ be the mean of the k measurements of u_i; that value of θ which maximizes $\Sigma \log h_Y(Y_i; \theta, \bar{X}_i)$ will be called the naive estimator. This estimator is usually inconsistent (Stefanski, 1985) although when $\bar{\Omega}/k$ is small its bias will be small.

2. FUNCTIONAL MODELS

2·1. *The functional likelihood*

Consider the functional version of (1·1) and (1·2) when $k=1$ and

$$\bar{\Omega}/a(\phi)=\Omega, \tag{2·1}$$

where Ω is known. In simple linear regression, (2·1) reduces to the common identifiability assumption that the ratio of the measurement-error variance to the equation-error variance is known. Similarly, (2·1) ensures identifiability of the parameters in the general model (1·1) and (1·2).

The random variables (Y_i, X_i) $(i=1, \ldots, n)$ are independent but not identically distributed since their distributions depend on the true regressors u_i, which vary with i. However, for notational convenience the subscript i will be dropped when referring to (Y_i, X_i) in situations where it causes no confusion. Under (1·1), (1·2) and (2·1) the joint density of (Y, X) is

$$h_{Y,X}(y, x; \theta, u)=h_Y(y; \theta, u)h_X(x; \theta, u). \tag{2·2}$$

For a set of n observations the log likelihood is

$$L(\theta, u_1, \ldots, u_n)=\sum_{i=1}^{n} \log\{h_{Y,X}(Y_i, X_i; \theta, u_i)\}. \tag{2·3}$$

When Y is normally distributed it is known that under (2·1) maximizing (2·3) with respect to $(\alpha, \beta^{\mathrm{T}}, \phi, u_1, \ldots, u_n)$ results in consistent estimators of the regression coefficients α and β (Gleser, 1981). For models other than the normal, the task of maximizing (2·3) with respect to its $n+p+2$ parameters is formidable. More importantly it is not generally true that maximizing (2·3) produces consistent estimators. In logistic regression the functional maximum likelihood estimator of (α, β) is not consistent under assumption (2·1) (Stefanski & Carroll, 1985). The problem is due to the large number of nuisance parameters (Neyman & Scott, 1948). The unwieldy functional likelihood, and its failure to produce consistent estimators underscore the need for an alternative approach to estimation.

2·2. *Unbiased score functions*

The literature on estimating a structural parameter in the presence of a large number of nuisance parameters dates back to the paper by Neyman & Scott (1948). They were the first to show that maximum likelihood estimation is not always a viable alternative. Andersen (1970) proved consistency of a conditional maximum likelihood estimator in a special class of models. Later we show a fundamental difference between our model (2·2), and those considered by Andersen (1970). Recent work on the problem includes that of Lindsay (1980, 1982, 1983, 1985) and Kumon & Amari (1984).

In this section unbiased score functions are obtained by conditioning on certain parameter-dependent sufficient statistics. It is shown how these scores relate to the conditional scores of Lindsay (1982) and some problems associated with their application are discussed.

Consider the density in (2·2). If u is viewed as a parameter and α, β and ϕ as fixed, then the statistic

$$\Delta = \Delta(Y, X, \theta) = X + Y\Omega\beta \tag{2·4}$$

is complete and sufficient for u. Consequently the distribution of $Y|\Delta$ depends only on Y, X and θ, but not on u. From this conditional distribution it is possible to derive unbiased estimating equations for θ which are independent of u.

Let $h_{Y|\Delta}(y\,|\,\delta;\,\theta)$ denote the conditional distribution of Y given $\Delta = \delta$. In the calculations that follow, δ is treated as a fixed conditioning argument until the final step of the analysis, equation (2·8), wherein δ is set equal to $\delta(y, x, \theta) = x + y\Omega\beta$; see equation (2·4). The Jacobian of the transformation which takes (Y, X) into $(Y, X + Y\Omega\beta)$ has determinant one. Thus $\text{pr}\,(Y = y, \Delta = \delta)dm(y)d\delta = \text{pr}\,(Y = y, X = \delta - y\Omega\beta)dm(y)d\delta$ and one finds

$$h_{Y|\Delta}(y\,|\,\delta;\,\theta) = \exp\,[y\eta - \tfrac{1}{2}y^2\beta^{\mathrm{T}}\Omega\beta/a(\phi) + c(y, \phi) - \log\{S(\eta, \beta, \phi)\}], \tag{2·5}$$

where $\eta = (\alpha + \delta^{\mathrm{T}}\beta)/a(\phi)$ and $S(.,.,.)$ is defined as

$$S(\eta, \beta, \phi) = \int \exp\{y\eta - \tfrac{1}{2}y^2\beta^{\mathrm{T}}\Omega\beta/a(\phi) + c(y, \phi)\}dm(y).$$

Note that (2·5) is an exponential family density in η with Y as the natural sufficient statistic. Thus moments of Y given $\Delta = \delta$ can be computed from the partial derivatives of $S(\eta, \beta, \phi)$ with respect to η; for example

$$E_\theta(Y\,|\,\Delta = \delta) = [(\partial/\partial\eta)\log\{S(\eta, \beta, \phi)\}]_{\eta = (\alpha + \beta^{\mathrm{T}}\delta)/a(\phi)}. \tag{2·6}$$

Using the exponential family representation in (2·5) and the fact that $m(.)$ does not depend on θ, it can be shown that

$$\int \dot{h}_{Y|\Delta}(y\,|\,\delta;\,\theta)dm(y) = 0, \tag{2·7}$$

where

$$\dot{h}_{Y|\Delta}(y\,|\,\delta,\,\theta)=(\partial/\partial\theta)h_{Y|\Delta}(y\,|\,\delta;\,\theta).$$

Thus defining $\psi_S(y, x, \theta)=(\partial/\partial\theta)\log h_{Y|\Delta}(y\,|\,\delta;\,\theta)$ evaluated at $\delta=x+y\Omega\beta$, we have that

$$\psi_S(y,x,\theta)=\left[\begin{array}{c}\{y-E(Y\,|\,\Delta=\delta)\}/a(\phi)\\ \{y-E(Y\,|\,\Delta=\delta)\}\delta/a(\phi)-\{y^2-E(Y^2\,|\,\Delta=\delta)\}\Omega\beta/a(\phi)\\ r(y,x,\theta)-E\{r(Y,X,\theta)\,|\,\Delta=\delta\}\end{array}\right]_{\delta=x+y\Omega\beta}, \tag{2·8}$$

where

$$r(y,x,\theta)=\frac{\partial c(y,\phi)}{\partial\phi}-y\frac{\alpha+\delta^{\mathrm{T}}\beta}{a^2(\phi)}a'(\phi)+y^2\frac{\beta^{\mathrm{T}}\Omega\beta}{2a^2(\phi)}a'(\phi).$$

Also $\psi_s(.,.,.)$ is unbiased for θ; that is $E_\theta\{\psi_S(Y,X,\theta)\}=E_\theta[E_\theta\{\psi_S(Y,X,\theta)\,|\,\Delta\}]=0$. The inner conditional expectation is zero by (2·7).

Any estimator, $\hat{\theta}_S$, satisfying

$$\sum_{i=1}^{n}\psi_S(Y_1,X_i,\hat{\theta}_S)=0 \tag{2·9}$$

will be called a sufficiency estimator. It is worth emphasizing that $\hat{\theta}_S$ does not maximize the conditional likelihood, $\Sigma\log\{h_{Y|\Delta}(Y_i\,|\,\Delta_i(\theta);\,\theta)\}$, where $\Delta_i(\theta)=X_i+Y_i\Omega\beta$. The estimator which does maximize this likelihood is generally not consistent, a consequence of the fact that the resulting score is not unbiased.

In the conditional likelihood the conditioning statistic depends on θ and it is here that our model differs from those studied by Andersen (1970). He studies models in which the sufficient statsitistics for the nuisance parammeters are independent of the structural parameter. The derivation of ψ_S exploited the fact that $h_{Y,X}$ factorizes into the product of $h_{Y,\Delta}$ and h_Δ. This factorization is similar to one used by Kalbfleisch & Sprott (1970). Other uses of conditional likelihoods like (2·5) arise in hypothesis testing problems (Cox & Hinkley, 1974, p.146).

Consider the density in (2·2) and let $\dot{h}_{Y,X}=(\partial/\partial\theta)h_{Y,X}$. Note that

$$\frac{\dot{h}_{Y,X}(y,x;\,\theta,u)}{h_{Y,X}(y,x;\,\theta,u)}-E\left\{\left[\frac{\dot{h}_{Y,X}(Y,X;\,\theta,u)}{h_{Y,X}(Y,X;\,\theta,u)}\right]_{\Delta=x+y\Omega\beta}\right\}$$

$$=\left[\begin{array}{c}\{y-E(Y\,|\,\Delta=\delta)\}/a(\phi)\\ \{y-E(Y\,|\,\Delta=\delta)\}u/a(\phi)\\ r(y,x,\theta)-E\{r(Y,X,\theta)\,|\,\Delta=\delta\}\end{array}\right]_{\delta=x+y\Omega\beta},$$

where $r(y, x, \theta)$ is defined in (2·8). As the expression in brackets above depends on the unknown covariate u only as a vector of weights this suggests the class of score functions

$$\psi_C(y,x,\theta)=\left[\begin{array}{c}\{y-E(Y\,|\,\Delta=\delta)\}/a(\phi)\\ \{y-E(Y\,|\,\Delta=\delta)\}t(\delta)/a(\phi)\\ r(y,x,\theta)-E\{r(Y,X,\theta)\,|\,\Delta=\delta\}\end{array}\right]_{\delta=x+y\Omega\beta} \tag{2·10}$$

indexed by the vector-valued function $t(.)$. When $t(\Delta)$ depends on (Y, X) only through Δ, we have $E[\{Y-E(Y\,|\,\Delta)\}t(\Delta)]=E(t(\Delta)E[\{Y-E(Y\,|\,\Delta)\}\,|\,\Delta])=0$ and thus ψ_C is unbiased. The score (2·10) is motivated by the work of Lindsay (1980, 1982, 1983) and will be called a conditional score.

Any estimator $\hat{\theta}_C$ satisfying

$$\sum_{i=1}^{n} \psi_C(Y_i, X_i, \hat{\theta}_C) = 0 \tag{2·11}$$

will be called a conditional estimator.

Ideally, $t(.)$ in (2·10) would be chosen to minimize the asymptotic variance of $\hat{\theta}_C$. However, for a functional model the optimal choice of $t(.)$ will depend on the particular sequence of covariates and thus no globally optimal choice exists. A related problem is noted by Cox & Hinkley (1974, p. 146) in a discussion of locally most powerful similar tests. The information unbiasedness criterion suggested by Lindsay (1982) leads to complex choices for $t(.)$ even in simple versions of the models studied in this paper. For example, in simple logistic regression through the origin ($\alpha = 0$) the function $t(.)$ for which ψ_C satisfies $E(\psi_C^2 + \psi'_C \,|\, \Delta) = 0$ takes the form

$$t(\delta) = \frac{-\{1+\exp(\delta\beta - \frac{1}{2}\beta^2)\}^{1/\beta} \exp\{\frac{1}{2}(\delta-\beta)^2\}}{d + \int \{1+\exp(\delta\beta - \frac{1}{2}\beta^2)\}^{1/\beta} \exp\{\frac{1}{2}(\delta-\beta)^2\} d\delta},$$

where d is a constant which may depend on β.

Some choices for $t(.)$ are now given which seem reasonable in the absence of any theory producing tractable alternatives. In terms of asymptotic efficiency no choice of $t(.)$ can outperform $t(\Delta_i) = u_i$; but of course this not available to the statistician. However, the fact that $t(\Delta_i) = u_i$ is optimal suggests that $t(\Delta_i)$ should be a good estimator of u_i. Further support for this argument is given in § 3 where it is shown that for structural models the optimal choice is $t(\Delta_i) = E(U_i \,|\, \Delta_i)$.

Consider simply taking $t(\Delta) = \Delta$. Since $E(\Delta) = E(X + Y\Omega\beta) = u + E(Y)\Omega\beta$, $t(\Delta)$ is a biased estimator of u, but if $|\Omega|$ is small then the bias is small. For logistic regression the choice $t(\Delta) = \Delta$ results in equivalence of ψ_S and ψ_C as shown in § 2·3. Another estimator of u is obtained by noting that X is unbiased for u and Δ is sufficient for u, thus $t(\Delta) = E(X \,|\, \Delta)$ is a uniformly minimum variance unbiased estimator of u. Also, since

$$E(X \,|\, \Delta) = \Delta - E(Y \,|\, \Delta)\Omega\beta, \tag{2·12}$$

only the conditional moment of $Y \,|\, \Delta$, given by (2·6), is needed to find $E(X \,|\, \Delta)$.

Since (2·8) and (2·10) are unbiased, regularity conditions will ensure the existence of consistent sequences of estimators $\hat{\theta}_S$ and $\hat{\theta}_C$ satisfying (2·9) and (2·11) respectively. It is not generally true that (2·9) and (2·11) define $\hat{\theta}_S$ and $\hat{\theta}_C$ uniquely. More importantly, there can exist sequences of solutions which are not consistent, and thus care must be taken when defining $\hat{\theta}_S$ and $\hat{\theta}_C$. Although we know of no definitive solutions to this problem in practice, certain solutions seem to work reasonably well. Our discussion focuses on $\hat{\theta}_C$ although it applies equally well to $\hat{\theta}_S$. In the case that X, u and β are scalars and the family of solutions to (2·11) has only one member with the same sign as the naive estimator, then choose $\hat{\theta}_C$ to be that solution. This selection rule is known to work in simple linear regression as detailed in § 2·3.

For multiple regression models ($p \geq 1$) two solutions are suggested. In the first $\hat{\theta}_C$ is defined as the solution to (2·11) which is closest to the naive estimator defined in § 1. This rule is justifiable when measurement error is small, however it can break down when measurement error is large. This is discussed in greater detail for the normal model in § 2·3. The second solution entails one or two steps of a Newton–Raphson iteration of (2·11) starting from the naive estimator. Again, this is generally appropriate only when the measurement is small. However, in some realistic sampling situations, Stefanski &

Carroll (1985) show that such an approach substantially improves upon the naive estimator in their study of measurement error in logistic regression.

When consistent sequences of solutions to (2·9) and (2·11) are obtained the asymptotic distributions of $\hat\theta_S$ and $\hat\theta_C$ are easily derived since both are M-estimators (Huber, 1967).

2·3. *Normal, logistic and Poisson regression*

In this section the strengths and limitations of the estimation theory are illustrated by studying it in three particular generalized linear models.

Suppose first that Y has a normal distribution with mean $\alpha+\beta^{\mathrm T}u$ and variance σ^2. For this model $\phi=\sigma^2$, $a(\phi)=\phi$ and $m(.)$ is Lebesgue measure. Using (2·5) one finds that the distribution of Y given $\Delta=\delta$ is normal with variance $\sigma^2/(1+\beta^{\mathrm T}\Omega\beta)$ and mean μ, where

$$\mu=(\alpha+\beta^{\mathrm T}\delta)/(1+\beta^{\mathrm T}\Omega\beta). \tag{2·13}$$

Corresponding to (2·8) one finds

$$\psi_S(y,x,\theta)=\left[\begin{array}{c} +\dfrac{1}{\sigma^2}(y-\mu) \\ \dfrac{\Omega\beta}{1+\beta^{\mathrm T}\Omega\beta}-\dfrac{1}{\sigma^2}\{(y-\mu)^2\Omega\beta-(y-\mu)(\delta-2\mu\Omega\beta)\} \\ \dfrac{-1}{2\sigma^2}+\dfrac{(y-\mu)^2(1+\beta^{\mathrm T}\Omega\beta)}{2\sigma^4} \end{array}\right]_{\delta=x+y\Omega\beta},$$

where μ is defined in (2·13). Define $\Delta_i^*=(I+\Omega\beta\beta^{\mathrm T})^{-1}\{\Delta(Y_i,X_i,\theta)-\alpha\Omega\beta\}$, where $\Delta(.,.,.)$ is given by (2·4), and consider the equations

$$\sum_{i=1}^n(Y_i-\alpha-\beta^{\mathrm T}\Delta_i^*)\begin{pmatrix}1\\ \Delta_i^*\end{pmatrix}=0,\quad \sigma^2=\frac{1+\beta^{\mathrm T}\Omega\beta}{n}\sum_{i=1}^n(Y_i-\mu_i)^2. \tag{2·14}$$

Every solution to (2·14) is also a solution to $\Sigma\psi_S(Y_i,X_i,\theta)=0$; that is any solution to (2·14) is a sufficiency estimator. The similarity of (2·14) to the usual normal equations is readily apparent. However, Δ_i^* depends on α and β and thus (2·14) is nonlinear in the parameters.

Note that Δ_i^* is also sufficient for u_i and that, given Δ_i^*, Y_i is normal with mean $\alpha+\beta^{\mathrm T}\Delta_i^*$. Since Δ_i^* is the functional maximum likelihood estimator for u_i in this model (Gleser, 1981), equation (2·14) shows that the functional maximum likelihood estimator is a sufficiency estimator.

From (2·14) it follows that $\hat\alpha_S=\bar Y-\hat\beta_S^{\mathrm T}\bar X$ and using this it is possible to deduce that $\hat\beta_S$ satisfies

$$-\hat\beta_S^{1}\left(\sum_{i=1}^n Y_i^*X_i^*\right)\Omega\hat\beta_S+\sum_{i=1}^n(Y_i^{*2}\Omega-X_i^*X_i^{*\mathrm T})\hat\beta_S+\sum_{i=1}^n Y_i^*X_i^*=0, \tag{2·15}$$

where $Y_i^*=Y_i-\bar Y$, $X_i^*=X_i-\bar X$.

Consider (2·15) for the case $p=1$; that is $\hat\beta_S$ is a scalar. This quadratic equation has two real roots (Kendall & Stuart, 1979, Ch. 29); unfortunately our derivation of (2·15) via the argument in § 2·2 does not indicate which root is appropriate. Had the equations (2·14) been derived as the gradient of the functional log likelihood, the appropriate root would have been dictated by the maximizing principle.

In § 2·2 it was suggested that in the case of multiple solutions to (2·9) and (2·11) we pick that solution closest to the naive estimator. In this particular case the two roots of (2·15) converge to β_0 and $-\sigma^2/(\beta_0\tau^2)$, where $\tau^2=\Omega\sigma^2$ is the measurement-error variance. The naive estimator converges to $\sigma_u^2\beta_0/(\sigma_u^2+\tau^2)$, where σ_u^2 is the limiting value of the sample variance of the true u_i's. Thus the suggested selection rule asymptotically chooses the right root whenever

$$|\beta_0|\tau^2/(\sigma_u^2+\tau^2)<\{|\beta_0|\sigma_u^2/(\sigma_u^2+\tau^2)+\sigma^2/(\tau^2|\beta_0|)\}.$$

This inequality holds if and only if $2\tau^2<[\sigma_u^2+\sigma^2/\beta_0^2+\{(\sigma_u^2+\sigma^2/\beta_0^2)^2+4\sigma^2\sigma_u^2/\beta_0^2\}^{\frac{1}{2}}]$. The infimum of the right-hand side above with respect to the ratio σ^2/β_0^2 is $2\sigma_u^2$. Thus whenever $\tau^2<\sigma_u^2$ the selection rule works no matter what the values of σ^2 and β_0^2; however, if $\tau^2>\sigma_u^2$ and σ^2/β_0^2 is sufficiently small then the selection rule chooses the wrong root. This is encouraging, for it is unusual to have measurement error so large that $\tau^2\geqslant\sigma_u^2$.

Finally for the normal model $E_\theta(Y_i|\Delta_i)=\alpha+\beta^{\mathrm{T}}\Delta_i^*$ and, from (2·12), $E_\theta(X_i|\Delta_i)=\Delta_i^*$. Thus ψ_S and ψ_C define the same estimators, that is $\hat{\theta}_S=\hat{\theta}_C$, when $t(\delta)=E_\theta(X|\Delta=\delta)$, which is linear in δ in this case.

Now consider logistic regression in which $\mathrm{pr}_\theta\,(Y=1|u)=F(\alpha+\beta^{\mathrm{T}}u)$, where $F(t)=1/(1+e^{-t})$. For this model $a(\phi)\equiv 1$ and $m(.)$ is counting measure on $\{0, 1\}$. Using (2·5) one obtains

$$\mathrm{pr}_\theta\,(Y=1|\Delta=\delta)=F\{a+(\delta-\tfrac{1}{2}\Omega\beta)^{\mathrm{T}}\beta\}, \tag{2·16}$$

and corresponding to (2·8) is the logistic sufficiency score

$$\psi_S(y, x, \theta)=[y-F\{\alpha+(\delta-\tfrac{1}{2}\Omega\beta)^{\mathrm{T}}\beta\}]\begin{pmatrix}1\\ \delta-\Omega\beta\end{pmatrix}\Bigg|_{\delta=x+y\Omega\beta}; \tag{2·17}$$

and setting $\Sigma\psi_S(Y_i, X_i, \theta)=0$ results in the equivalent equations

$$\sum_{i=1}^{n}\{Y_i-F(\alpha+\beta^{\mathrm{T}}\Delta_i^*)\}\begin{pmatrix}1\\ \Delta_i^*\end{pmatrix}=0, \tag{2·18}$$

where $\Delta_i^*=\Delta_i-\frac{1}{2}\Omega\beta$. Conditioned on Δ_i^*, Y_i is a Bernoulli variate with mean $F(\alpha+\beta^{\mathrm{T}}\Delta_i^*)$. Stefanski & Carroll (1985) show in a Monte Carlo study that, in spite of the possibility of multiple solutions to (2·18), a modified one-step version of $(\hat{\alpha}_S, \hat{\beta}_S^{\mathrm{T}})^{\mathrm{T}}$, starting from the naive estimator, performed well in some realistic sampling situations. Unlike the normal model the logistic sufficiency estimator does not correspond to the functional-maximum likelihood estimator, which in this case is not consistent (Stefanski & Carroll, 1985). In § 3·3 it is shown that the logistic sufficiency score is optimal for a particular structural model.

For logistic regression ψ_S and ψ_C are equivalent when $t(\delta)$ is linear but not when $t(\delta)=E(X|\Delta=\delta)$. Under linearity of $t(.)$ the equivalence follows by comparing (2·17) to (2·10) when $t(.)$ is linear. Howeve- 'vith $E_\theta(Y|\Delta=\delta)$ given by (2·16), $E(X|\Delta=\delta)=\delta-F\{\alpha+(\delta-\frac{1}{2}\Omega\beta)^{\mathrm{T}}\beta\}\Omega\beta$ and corresponding to (2·10) with $t(\delta)=E(X|\Delta=\delta)$ are the equations

$$\sum_{i=1}^{n}\{Y_i-F(\alpha+\beta^{\mathrm{T}}\Delta_i^*)\}\begin{bmatrix}1\\ \Delta_i^*+\{\frac{1}{2}-F(\alpha+\beta^{\mathrm{T}}\Delta_i^*)\}\Omega\beta\end{bmatrix}=0. \tag{2·19}$$

Although (2·18) and (2·19) clearly differ, the practical significance of this difference has not yet been investigated.

The final model considered is that of Poisson regression in which

$$\text{pr}_\theta\,(Y=k\,|\,u)=(k!)^{-1}\exp\{k(\alpha+\beta^{\mathrm{T}}u)-\exp(\alpha+\beta^{\mathrm{T}}u)\}.$$

For this model $a(\phi)\equiv 1$ and $m(.)$ is counting measure on $\{0, 1, \ldots\}$.

From (2·5) it follows that

$$\text{pr}_\theta\,(Y=k\,|\,\Delta=\delta)=\frac{(k!)^{-1}\exp\{k(\alpha+\beta^{\mathrm{T}}\delta)-\frac{1}{2}k^2\beta^{\mathrm{T}}\Omega\beta\}}{\Sigma(j!)^{-1}\exp\{j(\alpha+\beta^{\mathrm{T}}\delta)-\frac{1}{2}j^2\beta^{\mathrm{T}}\Omega\beta\}},\qquad (2\cdot20)$$

where the sum is over $j=0,\ldots,\infty$. Since (2·20) has no closed form the scores ψ_S and ψ_C are quite messy.

In the normal and logistic models it transpires that ψ_C and ψ_S are equivalent provided $t(.)$ is linear. These examples appear to be the exceptions rather than the rule; i.e. in general there is no choice of $t(.)$ which makes ψ_C equivalent to ψ_S. To see why, consider models in which $a(\phi)\equiv 1$. In these cases (2·8) and (2·10) are respectively

$$\psi_C=\begin{bmatrix} y-E(Y\,|\,\Delta=\delta) \\ \{y-E(Y\,|\,\Delta=\delta)\}t(\delta)\end{bmatrix},\quad \psi_S=\begin{bmatrix} y-E(Y\,|\,\Delta=\delta) \\ \{y-E(Y\,|\,\Delta=\delta)\}\delta-\{y^2-E(Y^2\,|\,\Delta=\delta)\}\Omega\beta\end{bmatrix}.$$

Because ψ_S involves the second-order conditional moments of Y there is in general no choice of $t(\delta)$ such that ψ_C and ψ_S are equivalent. In the special case of logistic regression, $Y=Y^2$ and taking $t(\delta)=\delta-\Omega\beta$ results in equality.

3. STRUCTURAL MODELS

3·1. *The structural likelihood*

In this section the model studied is the structural version of (1·1) and (1·2) wherein $u_1,\ldots,u_n$ are independent and identically distributed observations with unknown density $g(u)$. The density g is an element of $\mathscr{G}$, a family of densities with respect to Lebesgue measure, denoted $\nu(.)$. As in § 2, it is assumed that $k=1$ along with the identifiability condition (2·1). Under these conditions the joint density of (Y, X) is

$$f_{Y,X}(y, x;\, \theta, g)=\int h_{Y,X}(y, x;\, \theta, u)g(u)\,d\nu(u),\qquad (3\cdot1)$$

where $h_{Y,X}$ is defined in (2·2). Let $\dot f_{Y,X}(y, x;\, \theta, g)=(\partial/\partial\theta)f_{Y,X}(y, x;\, \theta, g)$ and assume that differentiation and integration can be interchanged in (3·1). If $g(.)$ were known then the efficient score for θ would be

$$\dot l(y, x, \theta, g)=(\partial/\partial\theta)l(y, x, \theta, g),$$

where $l(y, x, \theta, g)=\log f_{Y,X}(y, x;\, \theta, g)$ and the information available in (Y, X) for estimating θ would be $\mathscr{I}=E(\dot l\dot l^{\mathrm{T}})$.

A useful expression for $\dot l$ is obtained by noting that upon differentiating the logarithm of (3·1) we get

$$\dot l(y, x, \theta, g)=\frac{\int(\partial/\partial\theta)\log(h)hg\,d\nu}{\int hg\,d\nu}=E\left\{\frac{\partial}{\partial\theta}\log h(y, x;\, \theta, U)\,|\,Y=y, X=x\right\}.$$

Thus if $g(.)$ is viewed as a 'prior' for u then $\dot{l}$ has the interpretation as the posterior expectation of the functional maximum likelihood θ-score. Furthermore, since $\Delta = X + Y\Omega\beta$ is sufficient for u in the conditional model, $h_{Y,X}(y, x;\, \theta, u)$, the conditional distribution of $U|(Y=y, X=x)$ is the same as the conditional distribution of $U|\Delta = x + y\Delta\beta$. Thus

$$\dot{l}(y, x, \theta, g) = E\{(\partial/\partial\theta) \log h(y, x;\, \theta, U)\,|\,\Delta = x + y\Omega\beta\}. \qquad (3{\cdot}2)$$

3·2. *Efficient score functions and information bounds*

Efficient score functions for estimating $\theta = (\alpha, \beta^{\mathrm{T}}, \phi)^{\mathrm{T}}$ in the presence of the nuisance function $g(.)$ are now derived. As in § 2 the existence of certain sufficient statistics plays a key role here. The structural model studied is a generalization of a model considered by Bickel & Ritov (1987). Whereas they study simple linear regression under a number of conditions, including that of replicated measurements and our assumption (2·1), we consider the more general model only under the latter assumption.

Our structural model with unknown $g(.)$ is included in the class of semi-parametric models studied by Begun et al. (1983); see also Pfanzagl (1982, Ch. 14). Thus it is possible to apply their results to our model to find the efficient score for θ. However, our model has a good deal more structure than those considered by Begun et al. (1983) and this can be exploited to obtain a simple derivation of the efficient θ-score which uses only classical results. Furthermore, our derivation clearly illustrates the importance of the 'statistic' $\Delta = X + Y\Omega\beta$ encountered earlier in the discussion of functional models.

In § 2 the primary goal was to find unbiased scores for θ and it was of no consequence that we failed to give a precise statement of the relevant parameter space. The same is not true for the discussion of efficiency in structural models. The problem stems from the fact that for certain generalized linear models (1·1) there are linear restrictions on the quantity $\alpha + \beta^{\mathrm{T}}u$. These usually specify the sign of $\alpha + \beta^{\mathrm{T}}u$, for example for the gamma and inverse gamma models. For the normal, logistic and Poisson models there are no restrictions.

We now assume that the family of densities $\{h_Y(y;\, \eta)\}$, obtained by setting $\eta = \alpha + \beta^{\mathrm{T}}u$ and fixing ϕ in the right-hand side of (1·1), is a regular exponential family for $\eta \in H$ where H is one of the three open intervals $(-\infty, 0)$, $(0, \infty)$, $(-\infty, \infty)$. Let $\theta = (\alpha, \beta^{\mathrm{T}}, \phi)$ be an element in $\Theta = \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^+$ and g an element of $\mathscr{G}$. Write $\tau = (\theta, g)$ and, with supp(g) = support of g, define $T = \{\tau\colon \alpha + \beta^{\mathrm{T}}u \in H$ for $u \in \operatorname{supp}(g)\}$. The appropriate parameter space for our structural model is T.

In searching for an optimal score function for θ we restrict attention to only those functions ψ satisfying the following three conditions for all $\tau = (\theta, g)$ in T:

(i) $E_\tau\{\psi(Y, X, \theta)\} = 0$,

(ii) $E_\tau\{(\partial/\partial\theta)\psi(Y, X, \theta)\} = -E_\tau\{\psi(Y, X, \theta)\dot{l}^{\mathrm{T}}(T, X, \theta)\}$,

(iii) $E_\tau\{\|\psi(Y, X, \theta)\|^2\} < \infty$.

These conditions are fairly standard. The set of score functions satisfying (i)-(iii) is denoted by $\mathscr{S}$.

We now show that under an assumption concerning the richness of $\mathscr{G}$ that every score in $\mathscr{S}$ must be conditionally unbiased with respect to $\Delta = X + Y\Omega\beta$; that is if ψ is in $\mathscr{S}$ then, for all $\tau \in T$,

$$E_\tau\{\psi(Y, X, \theta)\,|\,\Delta\} = 0. \qquad (3{\cdot}3)$$

Before stating our assumption on $\mathscr{G}$ we make a definition.

Definition. A collection of functions, $\mathscr{H}$, is said to be complete with respect to a measure μ if a necessary condition for

$$\int t(s)h(s)\,d\mu(s)=0,$$

for all $h\in\mathscr{H}$, is $t(.)=0$ μ-almost surely.

For a fixed $\theta\in\Theta$ let $\mathscr{G}_\theta=\{g\in\mathscr{G}\colon(\theta,g)\in T\}$ and let ν_θ be Lebesgue measure on $\{u\in\mathbb{R}^p\colon \alpha+\beta^{\mathrm{T}}u\in H\}$. We now make the following assumption.

Assumption 1. For each fixed $\theta\in\Theta$, $\mathscr{G}_\theta$ is complete with respect to ν_θ.

Assumption 1 plays a role similar to the convexity condition (C) of Bickel (1982), and to assumption (S) of Begun et al. (1983). Note that if $H=(-\infty,\infty)$ then $T=\Theta\times\mathscr{G}$ and $\mathscr{G}_\theta=\mathscr{G}$ for all θ; and $\mathscr{G}$ will be complete provided it contains a complete parametric family of densities whose support is all of $\mathbb{R}^p$. For example, if $\mathscr{G}$ contains all of the p-dimensional normal densities then $\mathscr{G}$ is complete.

LEMMA 3·1. *With* $\Delta=X+Y\Omega\beta$, *Assumption* 1 *implies that any* ψ *in* $\mathscr{S}$ *must satisfy* (3·3).

Proof. Fix θ. For any ψ in $\mathscr{S}$ we know that $E_{\theta,g}\{\psi(Y,X,\theta)\}=0$ for all $g\in\mathscr{G}_\theta$. Conditioning first on Δ means that $E_{\theta,g}\{Q(\Delta)\}=0$ for all $g\in\mathscr{G}_\theta$ where $Q(\Delta)=E_{\theta,g}\{\psi(Y,X,\theta)\,|\,\Delta\}$. We now show that $E_{\theta,g}\{Q(\Delta)\}=0$ for all $g\in\mathscr{G}_\theta$ implies that $Q(\Delta)=0$ almost surely.

First note that $E\{Q(\Delta)\}=E[E\{Q(\Delta)\,|\,U\}]$. Now, for our structural model the conditional distributions of $Y\,|\,U$ and $X\,|\,U$, and hence that of $\Delta\,|\,U$, do not depend on g. Therefore, $E\{Q(\Delta)\,|\,U=u\}$ also does not depend on g. The fact that $E_{\theta,g}\{Q(\Delta)\}=0$ for all $g\in\mathscr{G}_\theta$ implies that, $0=\int E\{Q(\Delta)\,|\,U=u\}g(u)\,d\nu_\theta(u)$ for all $g\in\mathscr{G}_\theta$. Completeness of $\mathscr{G}_\theta$ with respect to ν_θ means that $E_\theta\{Q(\Delta)\,|\,U\}=0$, ν_θ-almost surely. Using properties of the bilateral Laplace transform it can be shown that the function $E\{Q(\Delta)\,|\,U=u\}$ is continuous over its domain and thus the fact that $E_\theta\{Q(\Delta)\,|\,U\}=0$, ν_θ-almost surely implies that $E\{Q(\Delta)\,|\,U=u\}$ is identically equal to zero for all u satisfying $\alpha+\beta^{\mathrm{T}}u\in H$. The density of $\Delta\,|\,U=u$ forms an exponential family in u. Since the parameter space for this family, $\{u\colon\alpha+\beta^{\mathrm{T}}u\in H\}$, is an open subset of $\mathbb{R}^p$, the family is complete. Note also that $\Delta=X+Y\Omega\beta$ has a Gaussian component and thus its distribution is absolutely continuous with respect to Lebesgue measure. Thus $E\{Q(\Delta)\,|\,U=u\}=0$ whenever $\alpha+\beta^{\mathrm{T}}u\in H$ implies $Q(\Delta)=0$ ν-almost surely. Since $Q(\Delta)=E\{\psi(Y,X,\theta)\,|\,\Delta\}$ the lemma follows. □

The implication of Lemma 3·1 is important; the only score functions which are unbiased for all $\tau\in T$ are those which satisfy (3·3). This fact enables us to deduce, quite simply, the form of the efficient score for θ.

As a measure of efficiency of a score ψ we use the positive-definite matrix $V_\psi=\{E(\psi\dot{l}^{\mathrm{T}})\}^{-1}E(\psi\psi^{\mathrm{T}})\{E(\dot{l}\psi^{\mathrm{T}})\}^{-1}$. Under sufficient regularity conditions V_ψ is the asymptotic covariance matrix of $n^{\frac{1}{2}}(\hat{\theta}-\theta)$, when $\hat{\theta}$ solves $\Sigma\psi(Y_i,X_i,\hat{\theta})=0$.

We now show that for every $\psi\in\mathscr{S}$, V_ψ is bounded below, in the sense of positive-definiteness, by V_{ψ^*}, where

$$\psi^*=\dot{l}(y,x,\theta,g)-E\{\dot{l}(Y,X,\theta,g)\,|\,\Delta=x+y\Omega\beta\};\qquad(3{\cdot}4)$$

that is ψ^* is the efficient θ-score.

THEOREM 3·1. *Assumption* 1 *implies* $V_\psi\geqslant V_{\psi^*}$ *for all* $\psi\in\mathscr{S}$.

Proof. Define IC_ψ to be the influence functions for ψ, that is $\mathrm{IC}_\psi = \{E(\psi \dot{l}^{\mathrm{T}})\}^{-1}\psi$; IC_{ψ^*} is the influence function for ψ^*.

Pick any $\psi \in \mathscr{S}$. Now, since $\psi \dot{l}^{\mathrm{T}} = \psi\psi^{*\mathrm{T}} + \psi E(\dot{l}^{\mathrm{T}} | \Delta)$, Lemma 3·1 implies, after first conditioning on Δ, that $E(\psi \dot{l}^{\mathrm{T}}) = E(\psi\psi^*)$. From this fact it follows that $V_{\psi^*}^{-1} = E(\psi^*\psi^{*\mathrm{T}})$ and $E\{\mathrm{IC}_\psi \mathrm{IC}_{\psi^*}^{\mathrm{T}}\} = E\{\mathrm{IC}_{\psi^*}\mathrm{IC}_\psi^{\mathrm{T}}\} = V_{\psi^*}$. Using these relationships we find

$$\begin{aligned} E\{(\mathrm{IC}_\psi - \mathrm{IC}_{\psi^*})(\mathrm{IC}_\psi - \mathrm{IC}_{\psi^*})^{\mathrm{T}}\} &= E(\mathrm{IC}_\psi \mathrm{IC}_\psi^{\mathrm{T}}) - E(\mathrm{IC}_{\psi^*}\mathrm{IC}_\psi^{\mathrm{T}}) - E(\mathrm{IC}_\psi \mathrm{IC}_{\psi^*}^{\mathrm{T}}) + E(\mathrm{IC}_{\psi^*}\mathrm{IC}_{\psi^*}^{\mathrm{T}}) \\ &= V_\psi - V_{\psi^*}. \end{aligned}$$

Thus for any $\psi \in \mathscr{S}$, V_ψ can be represented as the sum of V_{ψ^*} and a nonnegative-definite matrix which vanishes if and only if $\mathrm{IC}_\psi = \mathrm{IC}_{\psi^*}$ almost surely. This establishes the optimality of ψ^*. □

To find ψ^* note that Y and U are conditionally uncorrelated given Δ. This follows from the fact that the σ-field generated by Δ is a sub-σ-field of that generated by (Y, X). Thus $E(YU|\Delta) = E\{E(YU|Y, X)|\Delta\} = E\{YE(U|\Delta)|\Delta\} = E(Y|\Delta)E(U|\Delta)$. Using this and (3·2) we get

$$\psi^* = \dot{l} - E(\dot{l}|\Delta) = E(\dot{h}/h | Y, X) - E\{E(\dot{h}/h | Y, X)|\Delta\} = E(\dot{h}/h | Y, X) - E(\dot{h}/h|\Delta),$$

from which it follows that

$$\psi^*(y, x, \theta) = \left[\begin{matrix} \{y - E(Y|\Delta = \delta)\}/a(\phi) \\ \{y - E(Y|\Delta = \delta)\}E(U|\Delta = \delta)/a(\phi) \\ r(y, x, \theta) - E\{r(Y, X, \theta)|\Delta = \delta\} \end{matrix}\right]_{\delta = x + y\Omega\beta} . \tag{3·5}$$

Comparison with (2·10) shows that ψ^* is a conditional score with $t(\delta) = E(U|\Delta = \delta)$, hence the rationale for choosing $t(\Delta)$ as an estimator of U in the functional model.

3·3. *Some models in which $E(U|\Delta = \delta)$ is linear*

Since the optimal score in (3·5) depends on the unknown $g(.)$, it is not readily apparent how to construct asymptotically efficient estimators. The theoretical feasibility of doing so is suggested by the work of Bickel (1982), and Bickel & Ritov (1987) have successfully carried out such a construction for a model similar to ours. Our preliminary work on constructing efficient estimators for the class of models considered in the present paper is promising; however, the efficient estimators are necessarily complex and this makes them less than fully acceptable. For a simpler but related class of models Lindsay (1985) proposes modelling $E(U|\Delta = \delta)$ with a parametric family and estimating the additional parameters of this regression model. Such a procedure can result in fully efficient estimators only if the true regression of U on Δ is a member of the parametric family chosen. Furthermore, Lindsay considers only scalar models and obtains estimates of the model for $E(U|\Delta = \delta)$ by maximizing an empirical version of the information. An analogous approach in our higher-dimensional model would require maximizing an information matrix, a more difficult task.

In light of the difficulties involved in finding efficient estimators, an argument can be made for using a simple estimator provided it is optimal in some cases of interest. We now study the conditional score (2·10) when $t(\delta) = \delta$. For this choice of $t(.)$, ψ_{CL} is used to denote the conditional score. Since two scores are equivalent if and only if one is a nonsingular matrix multiple of the other, ψ_{CL} is equivalent to any conditional score for which $t(\delta)$ is a one-to-one linear function of δ. Using the same argument it follows that

ψ_{CL} is equivalent to the optimal score (3·5) for any model in which $E(U\,|\,\Delta=\delta)$ is linear. Thus conditions ensuring linearity of $E(U\,|\,\Delta=\delta)$ also ensure that ψ_{CL} is optimal.

A straightforward calculation indicates that

$$E(U\,|\,\Delta=\delta)=\frac{\int u\exp\{u^{\mathrm{T}}\Omega^{-1}\delta/a(\phi)\}g^*(u)\,d\nu(u)}{\int\exp\{u^{\mathrm{T}}\Omega^{-1}\delta/a(\phi)\}g^*(u)\,d\nu(u)}, \tag{3·6}$$

where

$$g^*(u)=\exp\left\{-\frac{u^{\mathrm{T}}\Omega^{-1}u+2b(\alpha+\beta^{\mathrm{T}}u)}{2a(\phi)}\right\}g(u).$$

Consider (3·6) when $E(U\,|\,\Delta=\delta)=M\delta+V$ and δ is replaced by $a(\phi)\Omega\gamma$. The resulting equation, which must hold for all $\gamma\in\mathbb{R}^p$, can be written in the form

$$a(\phi)M\Omega\gamma+V=(\partial/\partial\gamma)\log\{w(\gamma)\}, \tag{3·7}$$

where $w(\gamma)=w^*(\gamma)/w^*(0)$ and $w^*(\gamma)=\int\exp(u^{\mathrm{T}}\gamma)g^*(u)\,d\nu(u)$. Note that $w(\gamma)$ is the multivariate moment-generating function of the probability density $g^*(u)/w^*(0)$. Log convexity of $w(.)$ implies that $M\Omega$ is nonnegative which in turn implies that the differential equation (3·7) has the solution $\log\{w(\gamma)\}=\frac{1}{2}a(\phi)\gamma^{\mathrm{T}}M\Omega\gamma+\gamma^{\mathrm{T}}V$, and thus $g^*(u)/w^*(0)$ must be a $N(V, a(\phi)M\Omega)$ density. In terms of $g(.)$ this means that $E(U\,|\,\Delta=\delta)=M\delta+V$ if and only if

$$g(u)\propto\exp\left\{\frac{u^{\mathrm{T}}\Omega^{-1}u+2b(\alpha+\beta^{\mathrm{T}}u)-(u-V)^{\mathrm{T}}\Omega^{-1}M^{-1}(u-V)}{2a(\phi)}\right\}. \tag{3·8}$$

Finally we conclude that ψ_{CL} is optimal when (3·8) holds.

At first glance it seems that ψ_{CL} has a very weak claim to optimality in that it appears that there is only one $g(.)$ for which it is optimal, and furthermore that this $g(.)$ depends on θ. However, recall that ψ_{CL} is optimal no matter what M and V are, thus we are free to vary M and V in (3·8). This creates a family of distributions for which ψ_{CL} is optimal.

If in the normal regression model $g(.)$ is a normal density, then U and Δ are jointly normally distributed and $E(U\,|\,\Delta=\delta)$ is linear in δ. Thus for the normal regression model, ψ_{CL} is equivalent to ψ^* whenever the covariate U is normally distributed. This fact could also have been deduced by noting that for the normal regression model $b(t)=\frac{1}{2}t^2$ and thus the exponent in (3·8) is a quadratic form in u. Here $g(u)$ must be a normal density in order for $E(U\,|\,\Delta)$ to be linear. Furthermore, M and V can be chosen to obtain any mean and covariance matrix and thus $g(.)$ can be chosen independently of α, β in this case.

Now consider logistic regression. Much of the motivation for logistic regression is derived from its connection to normal theory discriminant analysis. Suppose that (Y, U) have the distributional properties

$$\operatorname{pr}(Y=0)=\Pi_0,\quad \operatorname{pr}(Y=1)=\Pi_1=1-\Pi_0,\quad U\,|\,Y=y\sim N(\mu_y,\Psi)\quad (y=0,1). \tag{3·9}$$

It then transpires that Y given U follows the logistic model with parameters depending on Π_0, μ_0, μ_1 and Ψ, that is $\operatorname{pr}(Y=1\,|\,U=u)=1/\{1+\exp(\alpha+\beta^{\mathrm{T}}u)\}$, where

$$\alpha=\log(\Pi_1/\Pi_0)+\tfrac{1}{2}(\mu_0^{\mathrm{T}}\Psi^{-1}\mu_0-\mu_1^{\mathrm{T}}\Psi^{-1}\mu_1),\quad \beta=\Psi^{-1}(\mu_1-\mu_0). \tag{3·10}$$

A routine calculation shows that, under (3·9),

$$E(U|\Delta=\delta)=(\Omega^{-1}+\Psi^{-1})^{-1}(\Omega^{-1}\delta+\Psi^{-1}\mu_0),$$

that is $E(U|\Delta=\delta)$ is linear, and thus for logistic regression ψ_{CL} is equivalent to ψ^* whenever (3·9) holds for any choice of Π_0, μ_0, μ_1 and Ψ. The covariate distribution $g(.)$ cannot be chosen entirely independently of $(\alpha, \beta^{\mathrm{T}})$ in this model. We are free to choose any Π_0, μ_0, μ_1 and Ψ which in turn determine both (a, β^{T}) and $g(.)$. However, since the $p+1$ components of (α, β) are functions of the $\frac{1}{2}p(p+1)+2p+1$ components of $(\Psi, \mu_0, \mu_1, \Pi_0)$, that is the relationship is not one-to-one, there are several different $g(.)$, all corresponding to the same (α, β). That is, even for a fixed (α, β) there are several different $g(.)$ for which ψ_{CL} is optimal. This family of $g(.)$ corresponds to all normal mixtures of the form (3·9) where Ψ, μ_0, μ_1 and Π_0 are constrainted to satisfy (3·10). Of course if the normal discriminant assumptions were known to hold a priori, then the linear discriminant, $\alpha+\beta^{\mathrm{T}}u$, would preferably be estimated using a full parametric likelihood (Efron, 1975; Michalik & Tripathi, 1980).

Finally note that for both the normal and logistic models the sufficiency score (2·8) is equivalent to ψ_{CL}. Thus for these two models the sufficiency score is optimal in the situations described above.

4. Concluding remarks

The assumption of normal errors, (1·2), is not crucial to the theory developed herein, the existence of a complete sufficient statistic for u when the other parameters are fixed is. The situation in which (1·2) is replaced with an assumption of replicated measurements, that is $k>1$ in (1·2), is conceptually no different than when (1·2) is assumed with the exception that both Ω and ϕ can now be estimated; thus there will be an additional $\frac{1}{2}p(p+2)$-dimensional component to all the scores.

Although no distributional assumption on the measurement errors is more natural than that of normality, it is still an unverifiable assumption unless replicate measurements are made. The sufficiency, conditional and efficient score lose their unbiasedness when the assumption of normal errors is erroneous. Thus when measurement error is nonnormal, estimates derived from these scores will be inconsistent and the asymptotic bias will generally not be computable. Approximations to the bias can be obtained using the small-measurement asymptotics employed by Stefanski (1985) although we have not attempted these calculations.

Acknowledgements

The work of R. J. Carroll and L. A. Stefanski was supported by the U.S. Air Force office of Scientific Research and the National Science Foundation repectively. We thank a referee for many helpful suggestions on an earlier draft of this paper.

References

Adcock, R. J. (1878). A problem in least squares. *Analyst* **5**, 53-4.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *J. R. Statist. Soc.* B **32**, 283-301.

Anderson, T. W. (1976). Estimation of linear functional relationships (with discussion). *J.R. Statist. Soc.* B **38**, 1-36.

Begun, J. M., Hall, W. J., Huang, W. M. & Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432-52.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-71.

BICKEL, P. J. & RITOV, Y. (1987). Efficient estimation in the errors-in-variables model. *Ann. Statist.* **15**, 513-40.

CARROLL, R. J., SPIEGELMAN, C. H., LAN, K. K., BAILEY, K. T. & ABBOTT, R. D. (1984). On errors-in-variables for binary models. *Biometrika* **71**, 19-25.

COX, D. R. & HINKELY, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Statist. Assoc.* **70**, 892-8.

GLESER, L. J. (1981). Estimation in a multivariate 'errors-in-variables' regression model: large sample results. *Ann. Statist.* **9**, 24-44.

HUBER, P. J. (1967). The behavior of maximum likelihood estimators under nonstandard conditions. *Proc. 5th Berkeley Symp.* **1**, 221-33.

KENDALL, M. G. & STUART, A. (1979). *The Advanced Theory of Statistics*, **2**. London: Griffin.

KALBFLEISCH, J. D. & SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. R. Statist. Soc.* B **32**, 175-208.

KUMON, M. & AMARI, S. (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameters. *Biometrika* **71**, 445-59.

LINDSAY, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Phil. Trans. R. Soc. Lond.* A **296**, 639-65.

LINDSAY, B. G. (1982). Conditional score functions: some optimality results. *Biometrika* **69**, 503-12.

LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11**, 486-97.

LINDSAY, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13**, 914-32.

MCCULLAGH, P. & NELDER, J. A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

MICHALIK, J. E. & TRIPATHI, R. C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *J. Am. Statist. Assoc.* **75**, 713-21.

MORAN, P. (1971). Estimating structural and functional relationships. *J. Mult. Anal.* **1**, 232-55.

NEYMAN, J. & SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory.* New York: Springer-Verlag.

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-42.

STEFANSKI, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72**, 583-92.

STEFANSKI, L. A. & CARROLL, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Statist.* **13**, 1335-51.

WOLTER, J. M. & FULLER, W. A. (1982a). Estimation of nonlinear errors-in-variables models. *Ann. Statist.* **10**, 539-48.

WOLTER, J. M. & FULLER, W. A. (1982b). Estimation of the quadratic errors-in-variables model. *Biometrika* **69**, 174-82.

[*Received October* 1985. *Revised January* 1987]

Optimal Rates of Convergence for Deconvolving a Density

RAYMOND J. CARROLL and PETER HALL*

Suppose that the sum of two independent random variables X and Z is observed, where Z denotes measurement error and has a known distribution, and where the unknown density f of X is to be estimated. One application is the estimation of a prior density for a sequence of location parameters. A second application arises in the errors-in-variables problem for nonlinear and generalized linear models, when one attempts to model the distribution of the true but unobservable covariates. This article shows that if Z is normally distributed and f has k bounded derivatives, then the fastest attainable convergence rate of any nonparametric estimator of f is only $(\log n)^{-k/2}$. Therefore, deconvolution with normal errors may not be a practical proposition. Other error distributions are also treated. Stefanski-Carroll (1987a) estimators achieve the optimal rates. The results given have versions for multiplicative errors, where they imply that even optimal rates are exceptionally slow.

KEY WORDS: Deconvolution; Density estimation; Errors in variables; Measurement error; Rates of convergence.

1. INTRODUCTION

Suppose that we wish to gain information about the density f of a random variable X, but because of measurement error can only observe $Y = X + Z$, where the measurement error Z is independent of X. Assume Z has a known density function f_Z with characteristic function ϕ_Z. We address the following question: From a sample $Y_1, \ldots, Y_n$, how well can f be estimated?

Applied problems in which knowledge of f is required were discussed by Mendelsohn and Rice (1982) (see also Medgyessy 1977). Nonparametric estimates of f were discussed by Stefanski and Carroll (1987a).

An application of our results is to the nonparametric empirical Bayes problem (see Berger 1980; Maritz 1980). Here f represents the prior distribution for a sequence of location parameters $X_1, \ldots, X_n$. The idea is to estimate the prior nonparametrically, as opposed to the alternative method of specifying a parametric form for the prior with parameters to be estimated. We consider how well a prior can be estimated nonparametrically.

Another application is to the problem of measurement error models (errors in variables) for nonlinear regression and generalized linear models (see Stefanski and Carroll 1987b). Other recent articles include Carroll, Spiegelman, Lan, Bailey, and Abbott (1984), Stefanski and Carroll (1985), Stefanski (1985), and Schafer (1987). In this problem, X is the true predictor, but because of measurement error Z we can observe only $Y = X + Z$. Although Stefanski and Carroll (1985) and Stefanski (1985) use a sensitivity analysis approach, Carroll et al. (1984) and Schafer (1987) assume a specific distributional form for f. This article addresses how well the data can be used in a nonparametric way to suggest a parametric form for f. Schafer (1987) shows that in generalized linear models, the EM algorithm for maximum likelihood requires knowledge of the first two conditional moments of X, given Y and the response variable in the generalized linear model. Other problems require the conditional moments of X, given Y. In either case, how well these conditional moments can be estimated from data depends on how well f can be estimated from data.

The case of normal measurement error is particularly important. We show that if f has k bounded derivatives and errors are normal, then the fastest rate of convergence of any estimator of f is only $(\log n)^{-k/2}$, and that this rate is achieved by a kernel estimator of Stefanski–Carroll (1987a) type. This very slow rate suggests that deconvolution to get precise point estimates of f may not be a practical procedure with normal errors, even if optimal estimators are employed. With $k = 2$, it also follows that the best achievable rate for estimating the distribution function of X can be no faster than $(\log n)^{-3/2}$. Thus even estimating probabilities for X is difficult.

We emphasize that our results pertain to precise point estimation of the density f and its distribution function. It is likely that other quantities may be estimated much more precisely, such as conditional moments of X, given Y or the number of modes of f.

We also show that Stefanski–Carroll estimators attain optimal convergence rates for many other error distributions, such as gamma, exponential, and double-exponential. For example, the optimal achievable rate in the double-exponential case is $n^{-k/(2k+5)}$. Our results indicate that if the error density is compactly supported and infinitely differentiable then the optimal convergence rate is slower than n^{-a} for any $a > 0$. Deconvolving a density with smooth measurement error is intrinsically difficult, with convergence rates much slower than those usually encountered in density estimation.

These results have obvious implications for models with multiplicative error, $Y = XZ$, that may be expressed additively by taking logs. The density of $\log Z$ is infinitely differentiable in many important cases, such as when Z is gamma or lognormal, so convergence rates are extremely

* Raymond J. Carroll is Professor and Head, Department of Statistics, Texas A&M University, College Station, TX 77843. Peter Hall is Professor, Department of Statistics, Australian National University, Canberra, Australian Capital Territory 2601, Australia. Carroll's work was supported by the U.S. Air Force Office of Scientific Research and undertaken during a visit to the Australian National University. The authors thank Len Stefanski and Cliff Spiegelman for helpful conversations.

Journal of the American Statistical Association
December 1988, Vol. 83, No. 404, Theory and Methods

slow. Hence deconvolution is difficult when errors are multiplicative.

Of course, our lower bounds to convergence rates continue to apply when error distributions are known imperfectly—for example, when errors are normal with unknown variance. In such cases where the error distribution is specified up to estimable parameters, the distribution can often be estimated $n^{1/2}$ consistently by replication. Since estimators of the X-density f converge at rates considerably slower than $n^{-1/2}$, replacing the true error distribution by its estimated version does not measurably affect convergence rates of Stefanski–Carroll estimators. Hence both our lower and upper bounds to convergence rates apply when error distributions are imperfectly specified, up to a parametric form.

Section 2 gives details of our calculations in the case of normal measurement errors. In Section 3 we briefly discuss other error distributions.

2. DECONVOLUTION WHEN ERRORS ARE NORMAL

Write $C_k(B)$ for the class of k-times differentiable densities f having $\sup f \leq B$ and $\sup|f^{(k)}| \leq B$. Let X have density f, Z be normal $N(0, 1)$ independent of X, and $Y = X + Z$. The following theorem provides bounds to the accuracy with which $f \in C_k(B)$ can be estimated from an n sample of Y's.

Let x_0 be any real number, and $\hat{f}(x_0)$ be any nonparametric estimator of $f(x_0)$, based on an n sample of Y's.

Theorem 1. Assume that the error distribution is normal $N(0, 1)$. If for some sequence of positive constants $\{a_n, n \geq 1\}$ we have

$$\liminf_{n\to\infty} \inf_{f\in C_k(B)} P_f\{|\hat{f}(x_0) - f(x_0)| \leq a_n\} = 1 \tag{2.1}$$

for each $B > 0$, then

$$\lim_{n\to\infty} (\log n)^{k/2} a_n = \infty. \tag{2.2}$$

Theorem 1 (see the Appendix for proof) declares that the rate of convergence of $\hat{f}$ to f cannot be faster than $(\log n)^{-k/2}$, over densities in $C_k(B)$. Kernel estimators attaining this rate of convergence were constructed by Stefanski and Carroll (1987a) and are given as follows. Let G be a symmetric function vanishing outside $(-1, 1)$, having $k + 2$ bounded derivatives on $(-\infty, \infty)$ and satisfying $G(t) = 1 + O(|t|^k)$ as $t \to 0$. Put $h \equiv (2/\log n)^{1/2}$, $G(w, h) \equiv (2\pi)^{-1} \int \cos(tw/h)G(t)\exp\{(t/h)^2/2\}\, dt$, and $\hat{f}(x) \equiv (nh)^{-1} \sum_j G(Y_j - x, h)$, where $\{Y_1, \ldots, Y_n\}$ is a random sample from the distribution of Y. The following result is an easy generalization to $k > 2$ of a result of Stefanski and Carroll (1987a).

Theorem 2. Assume that the error distribution is normal $N(0, 1)$. If the constants a_n satisfy (2.2) and $\hat{f}$ is the kernel estimator just defined, then (2.1) holds for each real number x_0 and each $B > 0$.

A referee has commented that the minimax nature of Theorem 1 may be unduly pessimistic. Further work with the Stefanski–Carroll estimator will be the judge of this concern. Cliff Spiegelman has conjectured that much better rates of convergence can be obtained if we limit consideration to smaller classes of densities, such as those confined to a known interval. Len Stefanski has also suggested that the small error approach of Stefanski and Carroll (1985) and Stefanski (1985) could be used to good effect in small samples.

The basic method of proof of Theorem 1 (see the Appendix) can be used to show that if $k = 2$, the distribution function of X can be estimated at a rate no faster than $(\log n)^{-3/2}$. Let F_n and F_0 be the distribution functions for f_n and f_0 in the proof of Theorem 1, and evaluate them at εx_0, where $x_0 > 0$. The calculations rely on an approximation to $H_{2l-1}(x_0)$ given by Magnus, Oberhettinger, and Soni (1966, p. 254), and various integral identities (p. 251). We omit the details. Using slightly different techniques, the same result has been obtained independently by Y. Ritov in an as-yet unpublished paper.

3. DECONVOLUTION FOR GENERAL ERRORS

There are versions of Theorems 1 and 2 for a variety of different types of error distributions. The general principle is: the smoother the residual distribution, the slower the optimal achievable rate of convergence. It is convenient to consider this principle in the Fourier domain, bearing in mind that smoother distributions have characteristic functions with thinner tails. If X, Y, and Z have respective characteristic functions ϕ_X, ϕ_Y, and ϕ_Z, and if $Y = X + Z$ where X and Z are independent, then the characteristic function of X is recoverable from that of Y via the formula $\phi_X = \phi_Y/\phi_Z$. Any data-based form of this inversion becomes increasingly difficult as the tails of ϕ_Z become thinner. For example, if Z has a gamma distribution with shape parameter α, then the tails of $\phi_Z(t)$ decrease like $|t|^{-\alpha}$ as $|t| \to \infty$, so deconvolution is difficult for large α. In fact, the fastest achievable rate of convergence over densities in $C_k(B)$ is $n^{-k/(2k+2\alpha+1)}$. This is made clear by the following analog of Theorem 1. Again, $\hat{f}(x_0)$ is a nonparametric estimator of $f(x_0)$.

Theorem 3. Assume that the error distribution is gamma with shape parameter $\alpha > 0$. If for some sequence of positive constants $\{a_n, n \geq 1\}$ we have $\liminf_{n\to\infty} \inf_{f\in C_k(B)} P_f\{|\hat{f}(x_0) - f(x_0)| \leq a_n\} = 1$ for each $B > 0$, then

$$\lim_{n\to\infty} n^{k/(2k+2\alpha+1)} a_n = +\infty. \tag{3.1}$$

The "double gamma" case, where Z is symmetric and $|Z|$ is gamma(α), is similar. There, Theorem 3 continues to hold for integer α, provided 2α in (3.1) is changed to $4(\alpha - [\alpha/2])$, where $[\alpha/2]$ denotes the largest integer not exceeding $\alpha/2$. In particular, the optimal rate of convergence when errors have a double-exponential distribution is $n^{-k/(2k+5)}$.

Proofs of results such as Theorem 3, where algebraic rates are available, run as follows. Let $\varepsilon \to 0$ as $n \to \infty$, and fix a k-times differentiable density f_0 that is bounded away from 0 in a neighborhood of the origin. Let H be a bounded, compactly supported function with at least k bounded derivatives, satisfying $H(0) \neq 0$ and $\int x^j H(x)\,dx = 0$ for $0 \leq j < \alpha + 1$. Put $f_n(x) \equiv f_0(x) + \varepsilon^k H(x/\varepsilon)$, and let g_n and g_0 be the convolution densities for f_0 and f_n, respectively. It may be shown that if $\varepsilon = n^{-1/(2k+2\alpha+1)}$, then I, defined at (A.6) (see the Appendix), satisfies $I = O(n^{-1})$. Then, arguing much as in the proof of Theorem 1, the best attainable rate of convergence emerges as being no faster than ε^k. Similar techniques show that for smooth, infinitely differentiable error densities such as the Cauchy, the optimal convergence rate is slower than n^{-a} for any $a > 0$.

Stefanski–Carroll–type kernel estimators achieve optimal rates in the normal, gamma, and double-gamma cases.

4. DISCUSSION

Deconvolution problems are important in their own right, as well as in nonparametric estimation of priors. In measurement error models, deconvolution arises if one wishes to use data either to suggest models for the unobservable predictors or to estimate conditional moments useful in likelihood calculations. When the measurement errors are normally distributed, our results are pessimistic, suggesting that it is difficult to deconvolve effectively over a wide class of distributions for X if one is interested in precise estimates of the true density f. Other functions of f may be estimated better, such as the conditional moments of X given Y or the general shape of f.

APPENDIX: PROOF OF THEOREM 1

To simplify notation, we relocate so that $x_0 = 0$ and rescale so that Z is normal $N(0, \frac{1}{2})$, with density $\psi(z) \equiv \pi^{-1/2}e^{-z^2}$. Let $\sigma \geq 1$, and write f_0 for the $N(0, \sigma^2)$ density; l for the integer part of $\log n$; $b_j \equiv 2^{-j}\{(2j)!\}^{-1/2}j^{1/4}$; $\eta \equiv l^{-k/2}\varepsilon^k\delta B$, where ε and $\delta \in (0, \frac{1}{2}]$ are fixed; and $H_0, H_1, \ldots$ for Hermite polynomials orthogonal with respect to ψ. The following properties are obtainable from Magnus et al. (1966, p. 252) and Sansone (1959, p. 324): $H_j(-x) = (-1)^j H_j(x)$;

$$\exp\{2x\varepsilon y - (\varepsilon y)^2\} = \sum_{j=0}^{\infty} H_j(x)(\varepsilon y)^j/j!; \qquad \text{(A.1)}$$

$$\int H_i(x)H_j(x)e^{-x^2}\,dx = \pi^{1/2}2^i i! \quad \text{if } i = j,\ 0 \text{ otherwise}; \qquad \text{(A.2)}$$

$$\int H_{2i}(x)x^{2j}\psi(x)\,dx = (2j)!/\{4^{j-i}(j-i)!\}; \qquad \text{(A.3)}$$

$$|b_j H_{2j}(x)\psi(x)| \leq C(1 + |x|^{5/2})e^{-x^2/2}; \qquad \text{(A.4)}$$

$$\eta \sup|(d/dx)^k b_l H_{2l}(x/\varepsilon)\psi(x/\varepsilon)| \leq C\,\delta B, \qquad \text{(A.5)}$$

where C depends only on k.

Put $f_n(x) \equiv f_0(x) + \eta b_l H_{2l}(x/\varepsilon)\psi(x/\varepsilon)$. By (A.4), and since $\eta(n) \to 0$ and $\varepsilon < 1 \leq \sigma$, f_n is a density for large n. If X has density f_0 or f_n then $Y = X + Z$ has density g_0 or g_n, respectively, where g_0 is the $N(0, \sigma^2 + \frac{1}{2})$ density, $g_n(x) \equiv g_0(x) + \eta b_l h_l(x)$, and

$$\begin{aligned} h_l(x) &\equiv \int H_{2l}(y/\varepsilon)\psi(y/\varepsilon)\psi(x-y)\,dy \\ &= \varepsilon\psi(x) \sum_{j=l}^{\infty} H_{2j}(x)\varepsilon^{2j}\{4^{j-l}(j-l)!\}^{-1}, \end{aligned}$$

using (A.1) and (A.3). Since $\psi(x)^2/g_0(x) \leq C_1 e^{-x^2}$,

$$\begin{aligned} I \equiv \int (g_n - g_0)^2(g_0)^{-1} &\leq C_1(\eta b_l)^2 \\ &\quad \times \int h_l(x)^2 e^{x^2}\,dx \\ &= C_2\varepsilon^{2k+2}\,\delta^2 2^{2l}\{(2l)!\}^{-1}l^{1/2-k} \\ &\quad \times \sum_{j=l}^{\infty} (\varepsilon^4/4)^j(2j)!\{(j-l)!\}^{-2}, \end{aligned} \qquad \text{(A.6)}$$

using (A.2). But $\{(2l)!\}^{-1} \leq C_3(l!)^{-2}2^{-2l}l^{1/2}$, $(2j)! \leq C_3(j!)^2 2^{2j}j^{-1/2}$, and $j!/(j-l)! = \binom{j}{l}l! \leq 2^j l!$. Hence, remembering that $\varepsilon \leq \frac{1}{2}$,

$$\begin{aligned} I \leq C_4\varepsilon^{2k+2}\,\delta^2 l^{1-k} &\sum_{j=l}^{\infty} (4\varepsilon^4)^j j^{-1/2} \\ &\leq C_5(\varepsilon, \delta)l^{1/2-k}(4\varepsilon^4)^l = o(n^{-1}). \end{aligned} \qquad \text{(A.7)}$$

Given $B > 0$, we see from (A.4) and (A.5) that by choosing σ large and δ small, not depending on B, we may ensure that f_0 and $f_n \in C_k(B)$ for large n. For an event A, let $P_n(A)$ and $P_0(A)$ denote the probability of A under f_n and f_0, respectively. If $\{a_n\}$ satisfies (2.1), then by (A.7) and the Cauchy–Schwarz inequality

$$\begin{aligned} &[P_n\{|\hat{f}(0) - f_n(0)| \leq a_n\}]^2 \\ &\leq P_0\{|\hat{f}(0) - f_n(0)| \leq a_n\}(1 + I)^n \\ &= \{1 + o(1)\}P_0\{|\hat{f}(0) - f_n(0)| \leq a_n\}, \end{aligned}$$

so both $P_0\{|\hat{f}(0) - f_n(0)| \leq a_n\}$ and $P_0\{|\hat{f}(0) - f_0(0)| \leq a_n\}$ converge to 1 as $n \to \infty$. Hence $|f_n(0) - f_0(0)| \leq 2a_n$ for large n. But $|f_n(0) - f_0(0)| = \eta b_l(2l)!/l!\pi^{1/2} \geq 2CB(\log n)^{-k/2}$, where C does not depend on B. Therefore, $a_n \geq CB(\log n)^{-k/2}$ for large n. Since this is true for each $B > 0$, then $(\log n)^{k/2}a_n \to \infty$, completing the proof.

[*Received July 1987. Revised March 1988.*]

REFERENCES

Berger, J. O. (1980), *Statistical Decision Theory,* New York: Springer-Verlag.

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bailey, K. R., and Abbott, R. D. (1984), "On Errors-in-Variables for Binary Regression Models," *Biometrika,* 71, 19–25.

Magnus, W., Oberhettinger, F., and Soni, R. P. (1966), *Formulas and Theorems for the Special Functions of Mathematical Physics,* Berlin: Springer-Verlag.

Maritz, J. S. (1980), *Empirical Bayes Methods,* London: Methuen.

Medgyessy, P. (1977), *Decomposition of Superpositions of Density Functions and Discrete Distributions,* New York: John Wiley.

Mendelsohn, J., and Rice, R. (1982), "Deconvolution of Microfluorometric Histograms With *B* Splines," *Journal of the American Statistical Association,* 77, 748–753.

Sansone, G. (1959), *Orthogonal Functions,* London: Wiley Interscience.

Schafer, D. W. (1987), "Covariate Measurement Error in Generalized Linear Models," *Biometrika,* 74, 385–391.

Stefanski, L. A. (1985), "The Effects of Measurement Error on Parameter Estimation," *Biometrika,* 72, 583–592.

Stefanski, L. A., and Carroll, R. J. (1985), "Covariate Measurement Error in Logistic Regression," *The Annals of Statistics,* 13, 1335–1351.

——— (1987a), "Deconvoluting Kernel Density Estimators," preprint.

——— (1987b), "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models," *Biometrika,* 74, 703–716.

statistics **21** (1990) 2, 169–184 Akademie-Verlag Berlin

Deconvoluting Kernel Density Estimators

LEONARD STEFANSKI and RAYMOND J. CARROLL

North Carolina State University and Texas A & M University

Summary. This paper considers estimation of a continuous bounded probability density when observations from the density are contaminated by additive measurement errors having a known distribution. Properties of the estimator obtained by deconvolving a kernel estimator of the observed data are investigated. When the kernel used is sufficiently smooth the deconvolved estimator is shown to be pointwise consistent and bounds on its integrated mean squared error are derived. Very weak assumptions are made on the measurement-error density thereby permitting a comparison of the effects of different types of measurement error on the deconvolved estimator.

AMS 1980 *subject classifications:* Primary 62J05; secondary 62H25, 62G05.

Key words: Convolution, deconvolution, density estimation, errors-in-variables, kernel, measurement error models.

1. Introduction

1.1. The deconvolution problem

Let U and Z be independent random variables with probability density functions g and h respectively. Then the random variable $X = U + Z$ has the density $f = g * h$ where $*$ denotes convolution. Assuming h is known we consider estimating g from a set of independent observations $\{X_j\}_{j=1}^n$ having the common density f.

The problem arises whenever data are measured with nonnegligible error and knowledge of g is desired. In this case U represents a true value, Z is an error of measurement and X is an observed value. An application is discussed in MENDELSOHN and RICE (1983); other applications and related work can be found in EDDINGTON (1913), TRUMPLER and WEAVER (1953), KAHN (1955), GAFFEY (1959), WISE et. al. (1977), and DEVROYE and WISE (1979). Our interest in the problem arises from its potential for application in measurement-error modelling. For example, work is in progress at the Radiation Effects Research Foundation, Hiroshima, Japan, to assess the health effects of radiation exposure. Measured exposures are known to contain substantial measurement errors. Some of the statistical models proposed for the data require estimation of the distribution (density) of true radiation exposures given only data on measured exposures and reasonable assumptions concerning the error distribution. Since the sample sizes involved are very large, nonparametric density estimation with deconvolution seems feasible. This paper establishes the elementary asymptotic theory associated with deconvolving a kernel density estimator and presents results from a simulation study demonstrating the difficulty of nonparametric deconvolution in moderate to large samples.

The deconvolution problem can also be cast in the format of an empirical Bayes problem wherein g represents the prior distribution for a sequence of location parameters. Estimation of prior distributions or mixing distributions have been studied by several authors, for example, Blum and Susarla (1977), Choi and Bulgren (1968), Deely and Kruse (1968), Maritz (1967) and Preston (1971). The estimator proposed in Section 1.2 is a transformation of a kernel density estimator of f. Some papers in which a transformation of a density estimator is of primary interest include Taylor (1982) and Hall and Smith (1986). Recent contributions to the literature on deconvolution include Carroll and Hall (1988), Fan (1988), Liu and Taylor (1988a, b) and Stefanski (1989).

Although many of the estimators proposed in the literature have been shown to be consistent less is known about their rates of convergence. The estimator we propose has the advantage of being analytically and, in some cases, computationally no more complex than an ordinary kernel density estimator, thus facilitating a discussion of its convergence properties. However, a price is paid for the reduction in complexity in that the resulting estimator is not range preserving, i.e., it assumes negative values with positive probability.

Throughout we make very weak assumptions on h and this allows us to assess the effects of different types of measurement error. A conclusion indicated by the analysis is that the difficulty of the deconvolution problem varies inversely with the smoothness of the measurement-error density. Thus deconvolution is particularly difficult under the common assumption of normally distributed errors.

Some conditions on h are necessary to insure that g is identifiable. We assume that h has a nonvanishing characteristic function, Φ_h, i.e.,

$$|\Phi_h(t)| > 0, \quad \text{for all real } t. \tag{1.1}$$

Although (1.1) is not the weakest assumption insuring identifiability of g it holds in many cases of interest, and in particular at the normal model.

1.2. The estimator

Let K be a bounded even probability density function whose characteristic function, Φ_K, satisfies, for each fixed $\lambda > 0$,

$$\sup_t |\Phi_K(t)/\Phi_h(t/\lambda)| < \infty; \qquad \int |\Phi_K(t)/\Phi_h(t/\lambda)| \, dt < \infty. \tag{1.2}$$

Implied by (1.2) are the facts that $\Phi_K^2/|\Phi_h(\bullet/\lambda)|^2$, $|\Phi_K|$ and Φ_K^2 are all integrable, which in turn implies that Φ_K is invertible, i.e.,

$$K(x) = (2\pi)^{-1} \int e^{-itx} \Phi_K(t) \, dt. \tag{1.3}$$

Let $\hat{f}$ be an ordinary kernel density estimator of f based on the kernel K,

$$\hat{f}(x) = (n\lambda)^{-1} \sum_{j=1}^{n} K\left((X_j - x)/\lambda\right). \tag{1.4}$$

The characteristic function of $\hat{f}$ is denoted $\Phi_{\hat{f}}$ and satisfies $\Phi_{\hat{f}}(t) = \hat{\Phi}(t)\, \Phi_K(\lambda t)$ where $\hat{\Phi}(t) = n^{-1} \sum_{j=1}^{n} \mathrm{e}^{itX_j}$ is the empirical characteristic function of $\{X_j\}_{j=1}^{n}$. Under (1.2) $\Phi_{\hat{f}}/\Phi_h$ is an integrable function and therefore possesses a FOURIER transform. The estimator we propose is $\hat{g}$ given by

$$\hat{g}(x) = (2\pi)^{-1} \int \mathrm{e}^{-itx}\{\Phi_{\hat{f}}(t)/\Phi_h(t)\}\, \mathrm{d}t. \tag{1.5}$$

Note that it is not possible to replace $\Phi_{\hat{f}}$ with $\hat{\Phi}$ in (1.5) since the resulting integral does not exist. By using $\Phi_{\hat{f}}$ in place of $\hat{\Phi}$ we are able to force integrability of the integrand in (1.5) by suitable choice of Φ_K.

Define the function K_λ^* as

$$K_\lambda^*(t) = (2\pi)^{-1} \int \mathrm{e}^{ity}\, \{\Phi_K(y)/\Phi_h(y/\lambda)\}\, \mathrm{d}y. \tag{1.6}$$

Then $\hat{g}$ has the representation

$$\hat{g}(x) = (n\lambda)^{-1} \sum_{j=1}^{n} K_\lambda^*\left((X_j - x)/\lambda\right). \tag{1.7}$$

Properties of $\hat{g}$ are best understood in terms of the properties of K_λ^* and the latent variables $\{U_j, Z_j\}_{j=1}^{n}$.

Equation (1.2) implies that $|K_\lambda^*|$ is bounded, thus $|\hat{g}|$ is also bounded and its expectation necessarily exists. Furthermore, an interchange of expectation and integration, justified by FUBINI's Theorem and (1.2), shows that

$$\mathsf{E}\, \{K_\lambda^*\left((X - x)/\lambda\right) \mid U\} = K\left((U - x)/\lambda\right). \tag{1.8}$$

From (1.8) it follows that

$$\mathsf{E}\{\hat{g}(x)\} = \lambda^{-1}\mathsf{E}\, \{K\left((U - x)/\lambda\right)\} = \int \lambda^{-1} K\left((u - x)/\lambda\right) g(u)\, \mathrm{d}u. \tag{1.9}$$

Thus $\hat{g}$ has the same bias as an ordinary kernel density estimator. Formally, this is a consequence of the fact that the linear operations of expectation and FOURIER transformation commute. Furthermore if $\hat{g}^*$ is the kernel estimator

$$\hat{g}^*(x) = (n\lambda)^{-1} \sum_{j=1}^{n} K\left((U_j - x)/\lambda\right)$$

then it follows from (1.8) that $\mathsf{E}\{\hat{g}(x) \mid U_1, \ldots, U_n\} = \hat{g}^*(x)$. Thus, conditionally $\hat{g}$ can be viewed as an unbiased estimator of $\hat{g}^*$.

The fact that Φ_K is even and real implies that K_λ^* is real and thus $\hat{g}$ is also. When h is even, K_λ^* is even. If $\Phi_K/\Phi_h(\bullet/\lambda)$ possesses m continuous integrable derivatives, then it follows from the RIEMANN-LEBESGUE Lemma that $K_\lambda^*(t) = o(|t|^{-m})$ as $|t| \to \infty$ and for $m \geqq 2$ this means that K_λ^* and hence $\hat{g}$ are integrable. Furthermore in this case the FOURIER Inversion Formula indicates that

$$\Phi_K(\lambda y)/\Phi_h(-y) = \lambda^{-1} \int \mathrm{e}^{ity} K_\lambda^*(t/\lambda)\, \mathrm{d}t. \tag{1.10}$$

Evaluating (1.10) at $y=0$ shows that $\int K_\lambda^*(t)\,\mathrm{d}t=1$ implying that $\int \hat{g}(x)\,\mathrm{d}x=1$. Since K_λ^* has many properties of an ordinary kernel we call it a *deconvoluting* kernel. The one property it lacks is nonnegativity; the left hand side of (1.10) is not positive definite thus K_λ^* cannot be a true probability density. Since, conditioned on $\{U_j\}_{j=1}^n$, $\hat{g}$ is an unbiased estimator of $\hat{g}^*$, this problem can be viewed as a failure of unbiased estimators to be range preserving.

In summary, provided $\Phi_K/\Phi_h(\bullet/\lambda)$ is smooth and (1.2) holds, $\hat{g}$ is continuous, real-valued and integrable with $\int \hat{g}(x)\,\mathrm{d}x=1$. Although the severity of (1.2) depends on Φ_h, it is always possible to satisfy these conditions when (1.1) holds by choosing Φ_K so that it vanishes outside a finite interval. For example, we can take Φ_K proportional to $U^{(2m)}$ where $U^{(2m)}$ is the $2m$-fold convolution of the uniform density, $\chi(|x|\leqq 1)/2$, with itself. The corresponding density is proportional to $\{\sin(t)/t\}^{2m}$. When $m\geqq 2$, $U^{(2m)}$ has two continuous integrable derivatives and the smoothness conditions on $\Phi_K/\Phi_h(\bullet/\lambda)$ are obtained provided Φ_h is sufficiently smooth. If Φ_h is not smooth then $\hat{g}$ need not be integrable although it will still be bounded and square integrable.

For certain measurement-error distributions (1.6) has a closed-form expression. For example, when $h(x)=(1/2)\,\mathrm{e}^{-|x|}$, $K_\lambda^*(t)=K(t)-\lambda^{-2}K''(t)$. In fact the integral in (1.6) can be evaluated analytically whenever $1/\Phi_h$ is a polynomial. Unfortunately, it does not seem possible to obtain K_λ^* in closed form for the normal measurement-error model.

2. Asymptotic results

In this section we establish the point-wise consistency of $\hat{g}$ and derive an approximation to its integrated mean square error. Throughout we work under the assumptions that g is continuous and bounded and hence square integrable.

Theorem 2.1. *If Φ_K and Φ_h are such that* (1.1) *and* (1.2) *hold and g is continuous and bounded then $\hat{g}(x)$ defined by* (1.5) *is a consistent estimator of $g(x)$ provided $n\to\infty$, $\lambda\to 0$ and $(n\lambda)^{-1}\int \Phi_K^2(t)\,|\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t\to 0$.*

Proof. Since $|\Phi_K/\Phi_h(\bullet/\lambda)|$ is square integrable we have from (1.6) and Parseval's Identity

$$\int \{K_\lambda^*(x)\}^2\,\mathrm{d}x=(2\pi)^{-1}\int \Phi_K^2(t)\,|\,\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t\,. \tag{2.1}$$

In light of (1.9) we can appeal to known results on kernel density to claim that the bias of $\hat{g}(x)$ converges to zero as $\lambda\to 0$.

Define

$$A(\lambda,a)=\int \{K_\lambda^*(x)\}^2\,g\,(a+\lambda x)\,\mathrm{d}x\,[\int \{K_\lambda^*(x)\}^2\,\mathrm{d}x]^{-1}$$

and note that $A(\lambda,a)$ is bounded by $B_g=\sup_x g(x)$. Now note that

$$\mathsf{E}\,\{K_\lambda^*\,((X-x)/\lambda)\}^2=\int\int \{K_\lambda^*\,((z+u-x)/\lambda)\}^2\,g(u)\,\mathrm{d}u h(z)\,\mathrm{d}z$$

and after the change of variables $t=(z+u-x)/\lambda$ in the inner integral we get

$$\mathsf{E}\,\{K_\lambda^*\,((X-x)/\lambda)\}^2=\lambda\int A\,(\lambda,x-z)\,h(z)\,\mathrm{d}z\int\{K_\lambda^*(t)\}^2\,\mathrm{d}t \qquad (2.2)$$
$$\leqq\lambda B_g\int\{K_\lambda^*(t)\}^2\,\mathrm{d}t\,.$$

Now since $n\lambda^2\,\mathsf{Var}\,\{\hat g(x)\}$ is bounded by the left hand side of (2.2) we find that

$$\mathsf{Var}\,\{\hat g(x)\}\leqq(n\lambda)^{-1}\,B_g\int\{K_\lambda^*(t)\}^2\,\mathrm{d}t \qquad (2.3)$$
$$=(2\pi n\lambda)^{-1}\,B_g\int\Phi_K^2(t)\,|\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t\,,$$

upon appealing to (2.1). Under the assumptions of the theorem, (1.9) and (2.3) show that $\mathsf{E}\{g(x)\}\to\hat g(x)$ and $\mathsf{Var}\,\{\hat g(x)\}\to 0$ thus concluding the proof. ■

Now we derive an approximation to the integrated mean squared error of $\hat g$. Using PARSEVAL's Identity, (2.1) and the change of variables $t=(X-y)/\lambda$ we have that as $n\to\infty$ and $\lambda\to 0$,

$$\int\mathsf{Var}\,\{\hat g(y)\}\,\mathrm{d}y \qquad (2.4)$$
$$=n^{-1}\int\lambda^{-2}\,\mathsf{E}\,\{K_\lambda^*\,((X-y)/\lambda)\}^2\,\mathrm{d}y-n^{-1}\int[\mathsf{E}\,\{\lambda^{-1}\,K_\lambda^*\,((X-y)/\lambda)\}]^2\,\mathrm{d}y$$
$$=n^{-1}\,\mathsf{E}\int\lambda^{-2}\,\{K_\lambda^*\,((X-y)/\lambda)\}^2\,\mathrm{d}y-(2n\pi)^{-1}\int|\Phi_g(t)|^2\,\Phi_K^2(\lambda t)\,\mathrm{d}t$$
$$=(\lambda n)^{-1}\int\{K_\lambda^*(t)\}^2\,\mathrm{d}t-(2n\pi)^{-1}\int|\Phi_g(t)|^2\,\Phi_K^2(\lambda t)\,\mathrm{d}t$$
$$=(2\pi n\lambda)^{-1}\int\Phi_K^2(t)\,|\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t+o\{(n\lambda)^{-1}\}\sim(2\pi n\lambda)^{-1}\int\Phi_K^2(t)\,|\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t\,.$$

If in addition to previous assumptions, g possesses two bounded integrable derivatives then as $\lambda\to 0$,

$$\int[\mathsf{E}\{g(x)\}-g(x)]^2\,\mathrm{d}x\sim(\lambda^4/4)\,\mu_{K,2}^2\int\{g''(x)\}^2\,\mathrm{d}x \qquad (2.5)$$

when $\mu_{K,2}=\int y^2K(y)\,\mathrm{d}y<\infty$. Combining (2.4) and (2.5) we have that to a first-order approximation

$$\int\mathsf{E}\,\{\hat g(x)-g(x)\}^2\,\mathrm{d}x\sim(2\pi n\lambda)^{-1}\int\Phi_K^2(t)\,|\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t \qquad (2.6)$$
$$+(\lambda^4/4)\,\mu_{K,2}^2\int\{g''(x)\}^2\,\mathrm{d}x\,.$$

The first term in (2.6) can be much larger than the variance component of the integrated mean squared error of an ordinary kernel density estimator. This is the price paid for not measuring $\{U_j\}_{j=1}^n$ precisely. The rate at which

$$V_{K,h}(\lambda)=\int\Phi_K^2(t)\,|\Phi_h(t/\lambda)|^{-2}\,\mathrm{d}t \qquad (2.7)$$

diverges as λ decreases is dictated by the tail behavior of $|\Phi_h|$, which in turn is related to the smoothness of h. Suppose that Φ_K is strictly positive on $(-B,B$ and vanishes off this interval. Then considering (2.7) when h is standard normal, CAUCHY and double-exponential we have respectively that for each $0<\varepsilon<B$ there exist positive constants $c_1,\ldots,c_5$ such that

$$c_1\mathrm{e}^{(B-\varepsilon)^2/\lambda^2}\leqq V_{K,h}(\lambda)\leqq c_2\mathrm{e}^{B^2/\lambda^2}\quad\text{(normal)}$$
$$c_3\mathrm{e}^{(B-\varepsilon)/\lambda}\leqq V_{K,h}(\lambda)\leqq c_4\mathrm{e}^{2B/\lambda}\quad\text{(CAUCHY)}$$
$$V_{K,h}(\lambda)\sim c_5\lambda^{-4}\quad\text{(double exponential)}\,.$$

Thus in these cases in order for (2.4) to converge to zero, λ must approach zero at a rate no faster than $\{\log(n)\}^{-1/2}$ for the normal model, no faster than $\{\log(n)\}^{-1}$

for the Cauchy model, and no faster than $n^{-1/5}$ for the double-exponential model. In other words, these are necessary conditions on λ if the first term on the right hand side of (2.6) is to be asymptotically negligible. By considering these rates in the right hand side of (2.5), it follows that the best possible rates on the integrated mean squared errors of $\hat{g}$ are $\{\log(n)\}^{-2}$, $\{\log(n)\}^{-1}$ and $n^{-4/9}$ for the normal, Cauchy and double-exponential cases respectively. Here we have considered only the order of the bandwidth. A more complete discussion of bandwidth selection can be found in Stefanski (1989).

The normal, Cauchy and double-exponential densities share the same ordering with respect to their peakedness at the origin; the normal is least peaked and the double-exponential is most peaked. The relationship between the variance of $\hat{g}$ and the peakedness of h is intuitive if we think of peakedness as a measure of the closeness of h to a delta function for which measurement error is identically zero and the deconvolution problem disappears. This analogy can be pushed a little further by considering a model wherein only $100\,p\%$ $(0<p<1)$ of the data are measured with error and the remaining data are error free. In this case we have $X=U+Z^*$ where $\mathsf{P}(Z^*=0)=1-p$ and $\mathsf{P}(Z^*=Z)=p$. The characteristic function, Φ^* of Z^* is $\Phi^*(t)=(1-p)+p\Phi_h(t)$. Nothing in the previous analysis required Z to have an absolutely-continuous distribution. If Z^* and not Z is the measurement-error variable then the variance term in (2.7) becomes

$$(2\pi n\lambda)^{-1}\int\{\Phi_K^2(t)\,|1-p+p\Phi_h(t/\lambda)|^{-2}\}\,\mathrm{d}t\sim(2\pi n\lambda)^{-1}\int\Phi_K^2(t)\,\mathrm{d}t/(1-p)^2$$

which is the same order of magnitude as the variance term for ordinary kernel density estimation. Thus for the model in which some data are error free we get convergence of the integrated mean-squared error at the usual rates. A similar model for discrete data was studied by Devroye and Wise (1979). We do not know of any instances in which this model has been studied for continuous variates.

3. Normal measurement error

We now argue that the poor performance of $\hat{g}$ at the normal measurement error model is intrinsic to the deconvolution problem and that convergence rates like $\{\log\,(n)\}^{-p}$, $p>0$, are to be expected.

Let $\bar{g}$ be any estimator of g which is continuous, bounded and integrable. Then $\bar{g}$ determines an estimator of f, namely $\bar{f}=h*\hat{g}$ where h is the standard normal density. It follows that $\bar{f}$ and all of its derivatives are continuous, bounded and integrable.

Let $\mathfrak{C}$ be the convolution operator corresponding to standard normal measurement error and let $\mathfrak{D}$ be the differential operator. Expanding $\mathrm{e}^{t^2/2}$ in a Maclaurin series we get that formally $\mathfrak{C}^{-1}=\sum_{j=1}^{\infty}(-1/2)^j\,\mathfrak{D}^{2j}/j!$. Thus when $f=h*\bar{g}$, $\bar{g}(x)=\sum_{j=1}^{\infty}(-1/2)^j\,\bar{f}^{(2j)}\,(x)/j!$ provided the series is convergent. Therefore we cannot expect to estimate g any better than we can estimate arbitrarily high derivatives of f. In contrast, when h is double-exponential, $\Phi_h(t)=1/(1+t^2)$, and deconvolution corresponds to a differential operator of order 2. Thus the rate of convergence for deconvolving double-exponential errors is the same as for estimating a second derivative.

Theorem 3.1. below shows that under certain regularity conditions, if the performance of $\bar{f}$ deteriorates sufficiently upon *one* differentiation, then the rate of convergence of the integrated mean squared error of $\bar{g}$ can be no better than inverse powers of log (n) for a large class of estimators.

In the theorem it is assumed that $\bar{g}$, $\bar{f}$ and $\bar{f}'$ are bounded, integrable estimators of g, f and f' respectively. For an integrable random function, $\bar{s}(\bullet)$, with an integrable expectation, $\mathsf{E}\{s(\bullet)\}$, let $\Phi_{\bar{s}}(t)$ and $\Phi_{\mathsf{E}\{\bar{s}\}}(t)$ denote $\int e^{itx}\bar{s}(x)\,dx$ and $\int e^{itx}$ $\mathsf{E}\{\bar{s}(x)\}\,dx$ respectively. With this notation we have that $\Phi_{\bar{f}}=\Phi_h\Phi_{\bar{g}}$ and $\Phi_{\bar{f}'}(t)=$ $=-it\Phi_{\bar{f}}(t)$ under the conditions stated above.

Theorem 3.1. *If for $\bar{s}=\bar{g}$, $\bar{f}$ and $\bar{f}'$,*

$$\mathsf{E}\{\Phi_{\bar{s}}(t)\}=\Phi_{\mathsf{E}\{\bar{s}\}}(t); \tag{3.1}$$

$$\int \mathsf{Var}\,\{\bar{s}(x)\}\,dx=(2\pi)^{-1}\int \mathsf{E}\,\{|\Phi_{\bar{s}}(t)-\Phi_{\mathsf{E}\{\bar{s}\}}(t)|^2\}\,dt; \tag{3.2}$$

$$\int \mathsf{Var}\,\{\bar{f}(x)\}\,dx=n^{-1}c_{1,n}a_n; \tag{3.3}$$

$$\int [\mathsf{E}\{\bar{f}(x)\}-f(x)]^2\,dx=c_{2,n}a_n^{-r}; \tag{3.4}$$

$$\int \mathsf{Var}\,\{\bar{f}'(x)\}\,dx=n^{-1}c_{3,n}a_n^{1+\varepsilon}; \tag{3.5}$$

where r and ε are positive constants; the sequence $\{a_n\}\to\infty$; and $\{c_{1,n}\}$, $\{c_{2,n}\}$ and $\{c_{3,n}\}$ are convergent sequences with positive limits; then the integrated mean squared error of $\bar{g}$ exceeds $c_{2,n}\,\{(c_{1,n}/c_{3,n})\log(n)\}^{-r/\varepsilon}$ for large n.

Proof. For convenience drop the subscript n on a_n, $c_{1,n}$, ..., $c_{3,n}$. The relationship $\Phi_{\bar{f}}=\Phi_{\bar{g}}\Phi_h$, (3.1), (3.2), JENSEN's inequality, (3.3) and (3.5) are used to show that

$$\begin{aligned}\int \mathsf{Var}\,\{\bar{g}(x)\}\,dx&=(2\pi)^{-1}\int e^{t^2}\mathsf{E}\,\{|\Phi_{\bar{f}}(t)-\Phi_{\mathsf{E}\{\bar{f}\}}(t)|^2\}\,dt\\ &\geqq\int \mathsf{Var}\,\{\bar{f}(x)\}\,dx\exp\left(\left[\int \mathsf{Var}\,\{\bar{f}'(x)\}\,dx\right]\left[\int \mathsf{Var}\,\{\bar{f}(x)\}\,dx\right]^{-1}\right)\\ &=(n^{-1}ac_1)\exp\{(c_3/c_1)\,a^{\varepsilon}\}\,.\end{aligned}$$

If $(c_3/c_1)\,a^{\varepsilon}\geqq\log(n)$, then $(c_1a/n)\exp\{(c_3/c_1)\,a^{\varepsilon}\}$ diverges and thus the integrated mean squared error of $\bar{g}$ diverges unless $a<\{(c_1/c_3)\log(n)\}^{1/\varepsilon}$. But (3.4), (3.1) and two simple inequalities together imply that

$$\begin{aligned}c_2a^{-r}&=\int [\mathsf{E}\,\{\bar{f}(x)\}-f(x)]^2\,dx=(2\pi)^{-1}\int e^{-t^2}\,|\Phi_{\mathsf{E}\{\bar{g}\}}(t)-\Phi_g(t)|^2\,dt\\ &\leqq(2\pi)^{-1}\int|\Phi_{\mathsf{E}\{\bar{g}\}}(t)-\Phi_g(t)|^2\,dt\\ &\leqq\mathsf{E}\left[\int\{\bar{g}(x)-g(x)\}^2\,dx\right].\end{aligned}$$

Thus when $a<\{(c_1/c_3)\log(n)\}^{1/\varepsilon}$ the integrated mean squared error of $\bar{g}$ exceeds

$$c_2\,\{(c_1/c_3)\log(n)\}^{-r/\varepsilon}\,,$$

which concludes the proof. ■

When $\hat{f}$ is a kernel density estimator with nonnegative kernel, a^{-1} is the bandwidth, $r=4$, $\varepsilon=2$ and the theorem shows that the mean integrated squared error of $\bar{g}$ can converge at a rate no faster than $\{\log(n)\}^{-2}$.

The significance of Theorem 3.1 lies in the fact that most nonparametric density estimators used in applications show a significant drop in performance upon differentiation. The theorem indicates that if such an estimator is deconvolved to remove a normal component, then the resulting estimator will have a mean integrated squared error that converges no faster than negative powers of log (n).

In a paper that has appeared since submission of this paper, (Carroll and Hall, 1988), it is established that the best *point-wise* convergence rates for deconvolving normal measurement error are proportional to $\{\log(n)\}^{-d/2}$ when g has d bounded derivatives. This result and Theorem 3.1 suggest that deconvolving normal measurement error is generally not likely to be very successful except in very large samples or when the density being estimated is extremely smooth and its smoothness is exploited. This often means resorting to estimators of f and g which are neither positive or integrable or both. Smoothness assumptions on g are not unreasonable in some applications and if we are interested primarily in determining the gross structure of g, e.g., presence and location of modes (Silverman, 1981), nonnegativity and integrability are not crucial.

4. A procedure for normal errors

Due to the difficulties inherent in deconvolving normal errors we investigated use of the so-called sinc kernel, $K(x)=(\pi x)^{-1}\sin(x)$, (Davis, 1975; Tapia and Thompson, 1978) with characteristic function $\Phi_K(t)=\chi(|t|\leqq 1)$. This kernel takes full advantage of smoothness properties of g by allowing bias to decrease at rates dictated by the tail behavior of $|\Phi_g|$. Lighter tails of $|\Phi_g|$ correspond to better convergence rates for $\hat{g}$. The improved asymptotic performance is obtained at the expense of nonnegativity of $\hat{f}$ and integrability of both $\hat{f}$ and $\hat{g}$. These are properties which might reasonably be sacrificed for the sake of determining shape.

The estimator proposed in Section 1 requires a bandwidth-selection rule for implementation. We now show that a cross-validation approximation to the integrated squared error of $\hat{g}$, $\widetilde{AISE}_g(\lambda)$, is given by

$$AISE_g(\lambda)=\int_0^{1/\lambda}\frac{2-(n+1)\,|\hat{\Phi}(t)|^2}{(n-1)\,\pi\,|\Phi_h(t)|^2}\,\mathrm{d}t\,. \tag{4.1}$$

Let $\{X\}_{(j)}$ denote the sequence $\{X_j\}$ with X_j removed and let $\hat{\Phi}_{(j)}$ be the empirical characteristic function of $\{X\}_{(j)}$. Then the fact that

$$\mathsf{E}\,[\mathrm{e}^{-\mathrm{i}tX_j}\hat{\Phi}_{(j)}(t)\mid U_j,\{X\}_j]=\hat{\Phi}_{(j)}(t)\,\mathrm{e}^{-\mathrm{i}tU_j}\Phi_h(-t)\,,$$

motivates the approximation via Parseval's relation

$$\int\hat{g}g\mathrm{d}x\sim(2\pi n)^{-1}\int\frac{\sum\limits_{j=1}^{n}\mathrm{e}^{-\mathrm{i}tX_j}\hat{\Phi}_{(j)}(t)\;\Phi_K(\lambda t)}{|\Phi_h(t)|^2}\,\mathrm{d}t\,.$$

Thus for the purpose of minimization

$$\int (\hat{g}-g)^2 \, \mathrm{d}x \sim \int \frac{n \, |\Phi_K(\lambda t)|^2 \, |\hat{\Phi}(t)|^2 - 2 \sum_{j=1}^{n} \mathrm{e}^{-\mathrm{i}tX_j} \hat{\Phi}_{(j)} \Phi_K(\lambda t)}{(2\pi n) \, |\Phi_h(t)|^2} \, \mathrm{d}t = AISE_g(\lambda) \, .$$

The last equality follows from (4.1) upon invoking the relationship

$$\sum_{j=1}^{n} \mathrm{e}^{-\mathrm{i}tX_j} \hat{\Phi}_{(j)}(t) = (n-1)^{-1} \, \{n^2 \, |\hat{\Phi}(t)|^2 - n\} \, ,$$

substituting $\chi(|\lambda t| \leqq 1)$ for $\Phi_K(\lambda t)$ and noting that $|\hat{\Phi}(\bullet)|^2$ is even.

The cross-validation approximation to the integrated squared error of $\hat{f}$, $AISE_f(\lambda)$, is given by the right side of (4.1) upon setting $\Phi_h(t) \equiv 1$. Differentiating (4.1) with respect to λ shows that extreme points of $AISE_g$ and $AISE_f$ both satisfy $\hat{I}_n(\lambda) = 0$, where

$$\hat{I}_n(\lambda) = n \, (n-1)^{-1} \, \{2 - (n+1) \, |\hat{\Phi}(1/\lambda)|^2\} \, . \tag{4.2}$$

Equation (4.2) indicates that the optimal cross-validation bandwidths for estimating g and f respectively are identical except possibly when (4.2) has multiple solutions. This is a consequence of using the sinc kernel and the fact that f is infinitely differentiable. A sufficient condition for the *mean* integrated squared errors of $\hat{g}$ and $\hat{f}$ to have identical minima is given in the Appendix.

5. Simulation results

We conducted a Monte Carlo study to determine if $\hat{g}$ is capable of revealing features of g which are masked by convolution with h. In particular we took g to be a 50—50 mixture of normal densities with means $\pm(2/3)^{1/2}$ and common variances 1/3. Normal measurement error with variance 1/3 was added to g so that f is a 50—50 normal mixture with means, $\pm(2/3)^{1/2}$, and common variances, 2/3. For this parameterization, g is bimodal, f is unimodal and the measurement error variance is 1/3 the variance of g, although it equals the variance of each normal component of g.

Observations were generated from f, and $\hat{g}$ was computed according to (1.5) with $\Phi_K(t) = \chi \, (|t| \leqq 1)$. This required numerical integration of

$$\hat{g}(x) = \pi^{-1} \int_0^{1/\lambda} q_{n,\sigma}(t) \, \mathrm{d}t$$

where $q_{n,\sigma}(t) = n^{-1} \sum_{j=1}^{n} \cos\{t \, (X_j - x)\} \, \mathrm{e}^{t^2\sigma^2/2}$. The integral was evaluated using SIMPSON's rule with sequential doubling of the grid size until successive iterations of $\hat{g}(x)$ differed by less than 10^{-5}. The computational procedure is accurate although it is slow.

In our study estimates were calculated over a 65-point grid spanning $[-3.5,$ $3.5]$. Sample size was set at $n = 2500$. With samples this large $AISE_g(\lambda)$ and $AISE_f(\lambda)$ are well-behaved and estimated bandwidths can be reliably computed by solving (4.2). Optimal bandwidths were also determined by minimizing the integrated squared error of g. Due to the large number of computations involved only 25 repetitions were performed. Figure 1 summarizes the findings of the simulation. Of the 25 density estimators, two were of significantly poorer quality than the rest due to estimated bandwidths which were much too small. Figure 1a contains an overlap plot of the remaining 23 estimates and gives a good idea of the variability inherent to the estimators.

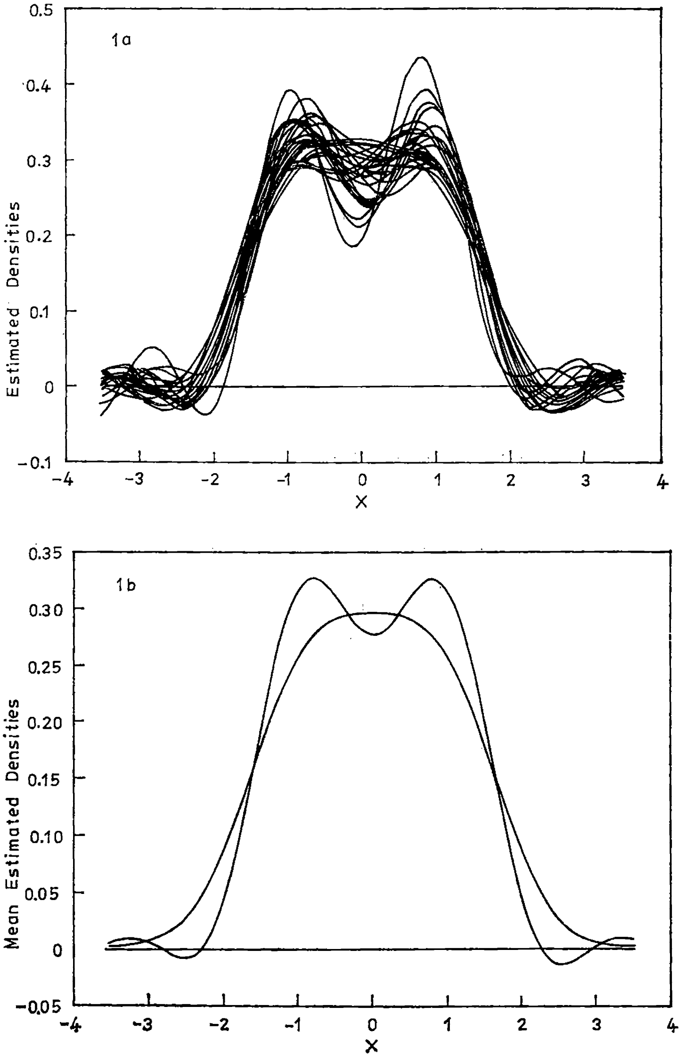


Figure 1 b displays the mean of the density estimates $\hat{f}_t$ and $\hat{g}_t$ over the same 23 observations alluded to above. The mean density estimates are very similar to their population counterparts, apart from the negativity in $\bar{g}$.

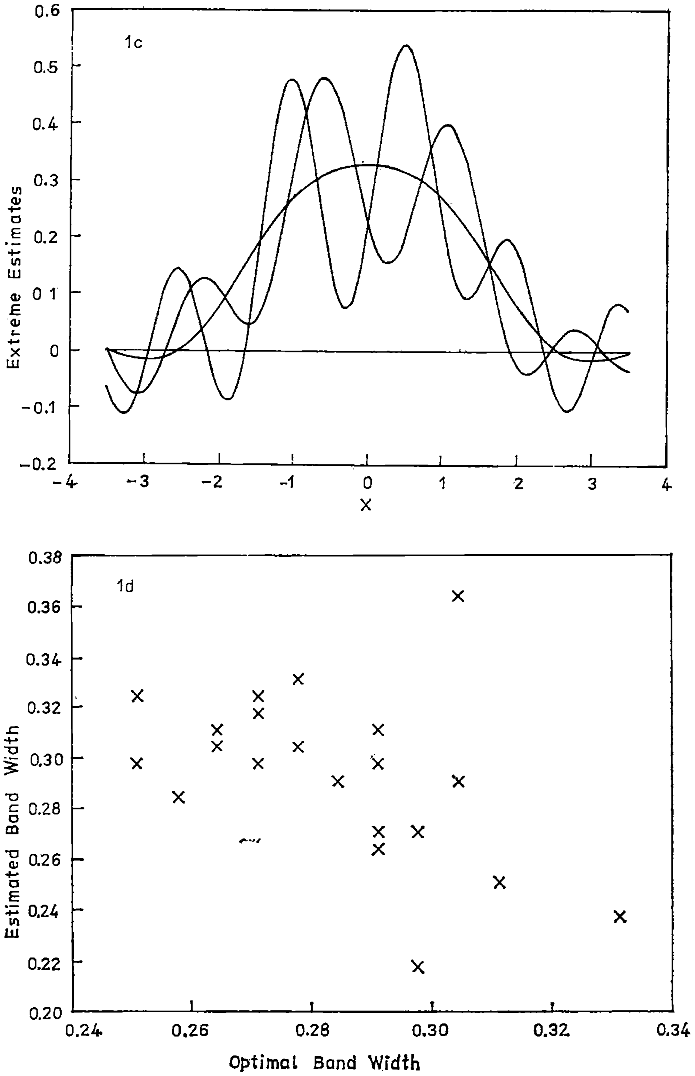


Fig. 1. Simulation results: 1a. Twenty-three estimates $\hat{g}$; 1b. Means of twenty-three estimates of $\hat{f}$ (unimodal) and $\hat{g}$ (bimodal); 1c. Three worst estimates $\hat{g}$; 1d. Scatterplot of estimated bandwidths versus optimal bandwidths.

Figure 1c graphs the three most extreme estimated densities. These densities correspond to the largest and two smallest estimated bandwidths.

Figure 1d contains a scatter plot of estimated versus optimal bandwidths. The latter were determined by minimizing $\int (\hat{g}-g)^2 \, dx$. The discrete nature of the data is an artifact of the optimization procedures employed, which searched over the grid

$$(1.5,\ 1.475,\ \ldots,\ 0.75)\ \sigma/\{\log\,(2\pi\sigma^2 n^{1/2})\}^{1/2}, \quad \sigma^2 = 1/3, \quad n = 2500\ .$$

All bandwidths, estimated and optimal, fell within the boundaries of this grid. The correlation coefficient for these data is -0.458. Removing the maximum and minimum estimated bandwidths changes the correlation to -0.670. The negative correlation between estimated and optimal bandwidths is typical of bandwidth selection procedures.

Of the 25 estimates, 13 showed clear evidence of bimodality, 4 showed questionable evidence and the remaining 8 gave little or no evidence of bimodality. The marginal performance of $\hat{g}$ is consistent with the asymptotic results of Sections 2 and 3 even though the latter pertain specifically to nonnegative kernel estimators. The model is artificial only with regards to the loss of bimodality in the presence of measurement error. The signal-to-noise ratio is not extreme. Thus a reasonable conclusion is that deconvolution is generally going to be a viable technique only with very large sample sizes. And in these cases computational efficiency may dictate the choice of estimator, at least to some extent.

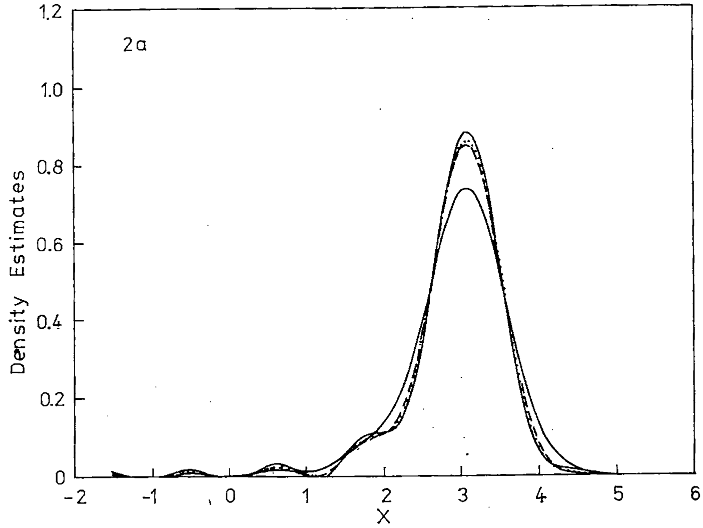
6. Applications

The theoretical results and simulation evidence in the previous sections indicate that deconvolution with normal errors will generally be feasible only with very large sample sizes and this limits its applicability. Furthermore, the need to specify the error density and the fact that our asymptotic results indicate widely varying performance under different error models, suggest that deconvolution may not be robust to choice of error model. We now examine the extent of these limitations in a particular application.

We consider estimating the density of long-term log daily saturated fat intake in women, using data from a study on the relationship of breast cancer incidence and dietary fat, see Jones, et al. (1987). We use the same 2888 observations on women of age less than 50 employed by Stefanski and Carroll (1989). Estimates of the error variance for these data suggest that as much as 50–75 % of the variability in the observed data may be due to measurement error. The simulation results suggest that for a sample of size 2888, deconvolving this much noise is problematic. However, it is possible to gain some insight into the data by deconvolving lesser amounts of noise.

In the example we used the sinc kernel and bandwidth selection procedure described in Section 4. Recall that for the sinc kernel, bandwidth selection is independent of the error density asymptotically, see Section 4 and the appendix. The deconvolved density estimator was computed under three different assumptions on the error density, normal (N), double exponential (DE) and hyperbolic cosine (HC) $\left((2/\pi)\,(e^t+e^{-t})^{-1}\right)$. These densities were chosen for their qualitatively different behaviour at the origin and in the tails, and because of their analytical tractability. In each case the densities were scaled to have common variance.

Because the estimators are not range preserving, in applications we suggest employing the positive projections of the estimators renormalized to integrate to one over the range of the data. Figures 2a and 2b display the resulting estimators



$\hat f$, $\hat g_N$, $\hat g_{DE}$, and $\hat g_{HC}$ assuming that $\sigma_Z^2=(1/5)\,\sigma_X^2$ and $(1/3)\,\sigma_X^2$ respectively. The density estimates are graphed over the range of the observed data. For the case $\sigma_Z^2=(1/5)\,\sigma_X^2$, the three deconvolved densities are nearly identical. For the case of larger measurement error, distinctions between the three deconvolved densities are more noticeable. However, differences between the three estimates of g are still small relative to the differences between $\hat f$ and any one estimate of g.

Assuming that the additive symmetric error model is reasonable, both figures suggest the interpretation that the long left tail of $\hat f$ is due to an underlying bimodal g smoothed by convolution. However, the data are 24-hour recall measurements of log saturated fat intake (STEFANSKI and CARROLL, 1989) and it seems prudent not to blindly accept the assumption of symmetric measurement error and the interpretations it renders. For example, it may be that a proportion of subjects systematically underreport foods high in saturated fats, resulting in a skewed or bimodal error density. This would also account for the long left tail in $\hat f$. In fact, if the density, g, of "true" log saturated fat intake is approximately normal and the error distribution, h, is bimodal, then the deconvolved density estimates, calculated under the assumption of symmetric errors, would be approximating h not g.

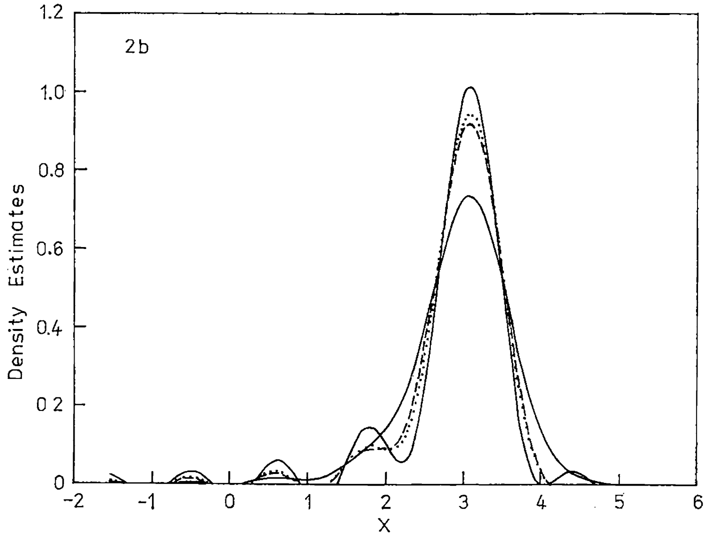


Fig. 2. Saturated-fat example: 2a. $\sigma_Z^2=(1/5)\,\sigma_X^2$; 2b. $\sigma_Z^2=(1/3)\,\sigma_X^2$; $\hat{f}$, solid line; $\hat{g}_{DE}$, dashed line; $\hat{g}_{HC}$, dotted line; $\hat{g}_N$, solid line; Ordering at the primary mode, $\hat{f}<\hat{g}_{DE}<\hat{g}_{HC}<\hat{g}_N$, both cases.

Acknowledgement

The authors gratefully acknowledge the financial support of the U.S. National Science Foundation, the U.S. Air Force Office of Scientific Research, the National Institutes of Health and the helpful comments of two referees.

References

Blum, T. and Susarla, V. (1977). Estimation of a mixing distribution function. *Ann. Probab.* **5**, 200–209.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83**, 1184–1186.

Choi, K. and Bulgren, W. (1968). An estimation procedure for mixtures of distributions. *J. Roy. Statist. Soc., Ser. B*, **30**, 444–460.

Davis, K. B. (1975). Mean square error properties of density estimates. *Ann. Statist.* **3**, 1025–1030.

Deely, J. J. and Kruse, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist.* **39**, 286–288.

DEVROYE, L. P. and WISE, G. L. (1979). On the recovery of discrete probability densities from imperfect measurements. *J. Franklin Inst.* **307**, 1–20.

EDDINGTON, A. S. (1913). On a formula for correcting statistics for the effects of a known probable error of observation. *Mon. Not. R. Astrom. Soc.* **73**, 359–360.

FAN, J. (1988). On the optimal rates of convergence for nonparametric deconvolution problem. Technical Report No. 157, Department of Statistics, University of California, Berkeley.

GAFFEY, W. R. (1959). A consistent estimator of a component of a convolution. *Ann. Math. Statist.* **30**, 198–205.

HALL, P. and SMITH, R. L. (1986). Unfolding a nonparametric density estimate. Preprint.

KAHN, F. D. (1955). The correction of observational data for instrumental bandwidth. *Proceedings of the Cambridge Philosophical Society.* **51**, 519–525.

LIU, M. C. and TAYLOR, R. L. (1989a). A consistent nonparametric density estimator for the deconvolution problem. Technical Report STA 73, University of Georgia.

LIU, M. C. and TAYLOR, R. L. (1989b). Simulations and computations of nonparametric density estimates for the deconvolution problem. Technical Report STA 74, University of Georgia.

MARITZ, J. S. (1970). *Empirical Bayes Methods.* Methuen, London.

MENDELSOHN, J. and RICE, J. (1982). Deconvolution of microfluorometric histograms with B splines. *J. Amer. Statist. Assoc.* **77**, 748–753.

PRESTON, P. F. (1971). Estimating the mixing distribution by piecewise polynomial arcs. *Austr. J. Statist.* **13**, 64–76.

SILVERMAN, B. W. (1981). Using kernel density estimators to investigate multi-modality. *J. Roy. Statist. Soc. Ser. B,* **43**, 97–99.

STEFANSKI, L. A. (1989). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics and Probability Letters,* to appear.

STEFANSKI, L. A. and CARROLL, R. J. (1989). Score tests in generalized linear measurement error models. Preprint. J. Roy. Statist. Soc. Ser. B, to appear.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Density Estimation.* John Hopkins University Press, Baltimore.

TAYLOR, C. C. (1982). A new method for unfolding sphere size distributions. *Journal of Microscopy* **132**, 57–66.

TRUMPLER, R. J. and WEAVER, H. F. (1953). *Statistical Astronomy.* University of California Press, Berkeley.

WISE, G. L., TRAGANITIS, A. P. and THOMAS, J. B. (1977). The estimation of a probability density function from measurements corrupted by Poisson noise. *IEEE Trans. Inform. Theory,* IT-**23**, 764–766.

Appendix

Using PARSEVAL's Identity, the fact that $\Phi_K(\bullet)$ is an indicator function, eveness of $|\Phi_f(\bullet)|^2$ and $|\Phi_h(\bullet)|^2$ and the relationship $\mathsf{E}\,\{|\hat{\Phi}(t)-\Phi_f(t)|^2\}=\{1-|\Phi_f(t)|^2\}/n$ it can be shown that the mean integrated squared error of $\hat{g}$, $MISE_g(\lambda)$, is given by

$$MISE_g(\lambda)=c+\pi^{-1}\int_0^{1/\lambda}\frac{1-(n+1)\,|\Phi_f(t)|^2}{n\,|\Phi_h(t)|^2}\,\mathrm{d}t \tag{A.1}$$

where c is a constant depending on f and h but not λ. The mean integrated squared error of $\hat{f}$, $MISE_f(\lambda)$, is obtained from (A.1) by setting $\Phi_h(t)\equiv 1$.

Suppose λ_g and λ_f minimize $MISE_g$ and $MISE_f$ respectively. By the Fundamental Theorem of Calculus, $I_n(\lambda_g)=0$ and $I_n(\lambda_f)=0$ where

$$I_n(\lambda)=1-(n+1)\ |\Phi_f(1/\lambda)|^2$$

Assume that $|\Phi_f(t)|^2$ is strictly decreasing on $[B, \infty)$ for some $B>0$. Since λ_g and λ_f necessarily converge to zero, λ_g^{-1} and λ_f^{-1} are contained in the interval $[B, \infty)$ for sufficiently large n. However, for $\lambda \in (0, 1/B]$, $I_n(\lambda)$ is one-to-one under the assumption on $|\Phi_f(t)|^2$ and thus the condition $I_n(\lambda)=0$ uniquely determines λ. It follows that λ_g and λ_f are equal for sufficiently large n.

Note that $\hat{I}_n(\lambda)$ defined in (4.2) can be regarded as an unbiased estimating equation for λ_g and λ_f in the sense that $\mathsf{E}\{\hat{I}_n(\lambda)\}=I_n(\lambda)$.

Received May 1988; revised June 1989.

Leonard A. Stefanski
Department of Statistics
North Carolina State University
Raleigh, NC 27695
U.S.A.

Raymond J. Carroll
Department of Statistics
Texas A & M University
College Station, TX 77843
U.S.A.

Approximate Quasi-likelihood Estimation in Models With Surrogate Predictors

RAYMOND J. CARROLL and LEONARD A. STEFANSKI*

We consider quasi-likelihood estimation with estimated parameters in the variance function when some of the predictors are measured with error. We review and extend four approaches to estimation in this problem, all of them based on small measurement error approximations. A taxonomy of the data sets likely to be available in measurement error studies is developed. An asymptotic theory based on this taxonomy is obtained and includes measurement error and Berkson error models as special cases.

KEY WORDS: Berkson error; Measurement error; Quasi-likelihood; Reliability data; Validation data.

1. INTRODUCTION AND PRELIMINARIES

1.1 Quasi-likelihood Models With Surrogate Predictors

The general quasi-likelihood/variance function model for a scalar response Y given a p-variate predictor $X = x$ is

$$E(Y \mid X = x) = f_m(x, \beta), \qquad \operatorname{var}(Y \mid X = x) = \sigma^2 f_v(x, \beta, \theta), \quad (1.1)$$

where σ^2 is a scalar parameter and β and θ are column-vector parameters designated collectively as $\Theta = (\beta', \theta', \sigma^2)'$. Model (1.1) includes the usual quasi-likelihood models and generalized linear models, as well as models in which the variance is an unknown function of the mean or predictors. For additional motivation and background on these models, see Carroll and Ruppert (1988) and McCullagh and Nelder (1989).

We consider fitting model (1.1) when a q-variate proxy W ($q \geq p$) is observed in place of the random predictor X in some subset of the available data. We assume that W is a *surrogate* for X, meaning that the conditional distribution of Y given (X, W) is identical to the conditional distribution of Y given X.

When the conditional distribution of X given W is specified parametrically, it is possible in principle to calculate the conditional mean and variance of Y given W and to estimate unknown parameters using standard quasi-likelihood/variance function techniques. This approach frequently requires numerical multiple integration and is computationally unattractive. In addition, experience indicates that this approach can be sensitive to specification of the conditional distribution of X given W; see, for example, Schafer (1987). The methods developed in this article depend only on the first two moments of the error given W, which is more in the spirit of quasi-likelihood/variance function techniques; see Carroll and Ruppert (1988, sec. 2.5). In Section 3 we describe a general approximate model for the first two conditional moments of Y given W when the relationship between X and W is specified either conditionally on X or conditionally on W. Thus we incorporate both measurement error and Berkson-error models.

Measurement error models are subsumed under the general model

$$W = c(X, \eta) + \delta U, \quad E(U \mid X = x) \equiv 0, \qquad \operatorname{cov}(U \mid X = x) = \Omega(x, \eta, \gamma), \quad (1.2)$$

where $c(\cdot, \cdot)$ and $\Omega(\cdot, \cdot, \cdot)$ are known and $\Lambda = (\eta', \gamma', \delta^2)'$ is a column vector of parameters. In certain models some components of Λ may be known. This includes the classical measurement model (Fuller 1987), $c(x, \eta) = x$, $\Omega(x, \eta, \gamma) = \Omega$, as well as linear models, $c(x, \eta) = \eta_0 + \eta_1^T x$, $\Omega(x, \eta, \gamma) = \Omega$, where η contains all of the unique elements of η_0 and η_1. Models in which some predictors are measured without error are handled by allowing $\Omega(\cdot, \cdot, \cdot)$ to have less than full rank. Of course we can benefit from the full generality of (1.2) only when sufficient validation/reliability data are available for estimating the unknown components of Λ; see Section 2.

In Section 3 we impose one significant restriction on (1.2). We require that the equation $t = c(s, \eta)$ can be solved for s as a smooth function of (t, η) in a neighborhood of the true parameter η, that is, $s = c^-(t, \eta)$, where c^- denotes an inverse to c. This is always possible when W and X are scalars and $c(\cdot, \cdot)$ is smooth and strictly monotonic in its first argument. The requirement is more stringent when the dimension of X exceeds 1. For example, in the multivariate regression version of (1.2), $W = \eta_0 + \eta_1 X + \delta U$, where η_1 is a $q \times p$ matrix, it requires that rank $(\eta_1) = p$ and, in this case, $X = \eta_1^-(W - \eta_0 - \delta U)$, where η_1^- is a generalized inverse of η_1 satisfying $\eta_1^- \eta_1 = I_{p \times p}$.

* Raymond J. Carroll is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Leonard A. Stefanski is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Carroll's research was supported by Grant GM-39015 from the National Institutes of Health. Stefanski's research was supported by Grant DMS-8613681 from the National Science Foundation and a cooperative agreement among Harvard School of Public Health, Societal Institute of Mathematical Sciences, and the Environmental Protection Agency. The authors thank Alice Whittemore, David Ruppert, and Donna Spiegelman for providing them with preprints of their papers. They also thank the referees and associate editor for their comments on a previous draft of the article.

Editor's Note: This article was reviewed under the direction of the Applications Editor, Clifford C. Clogg.

Journal of the American Statistical Association
September 1990, Vol. 85, No. 411, Theory and Methods

Berkson error models are subsumed under the general model

$$X = c^*(W, \eta) + \delta U, \quad \mathrm{E}(U \mid W = w) \equiv 0,$$
$$\mathrm{cov}(U \mid W = w) = \Omega^*(w, \eta, \gamma), \quad (1.3)$$

where $c^*(\cdot, \cdot)$ and $\Omega^*(\cdot, \cdot, \cdot)$ are known and $\Lambda = (\eta', \gamma', \delta^2)'$ is a column vector of parameters. In certain models some components of Λ may be known. The classical Berkson error model has $c^*(w, \eta) = w$ and $\Omega^*(w, \eta, \gamma) = \Omega$. When some components of U are equal to 0, $\Omega^*(\cdot, \cdot, \cdot)$ has less than full rank. As with (1.2), (1.3) includes most regression models used for data analysis.

New classes of estimates have been developed recently based on small measurement error approximations, that is, under the assumption that δ in (1.2) and (1.3) is small. We distinguish four approaches: (i) correct the naive estimators obtained by fitting (1.1) with X replaced by $g(W, \eta, 0)$, where g is given in (3.1) (Amemiya and Fuller 1988; Stefanski 1985; Stefanski and Carroll 1985); (ii) approximate the quasi-likelihood/variance function estimates (Whittemore and Keller 1988); (iii) approximate the quasi-likelihood/variance function model and then estimate the parameters in this model (Armstrong 1985; Carroll 1989; Fuller 1987, sec. 3.3; Rudemo, Ruppert, and Streibig 1989); and (iv) replace X by an estimate of $\mathrm{E}(X \mid W)$ and perform a standard analysis (Gleser 1989; Rosner, Willet, and Spiegelman 1989). These estimates are computable, widely applicable, and, in our experience, work well in applications.

In this article we develop a general model for the observed data that encompasses both (1.2) and (1.3) for small δ. We generalize previous models both with respect to the class of submodels relating Y to X and those linking X and W. The methods of estimation described previously are studied in the context of this general model. We show that Method (iv) can be viewed as a special case of Method (iii) and that Methods (i) and (ii) have a common underlying structure. These results allow us to present unified asymptotic distribution results for Methods (i) and (ii) and Methods (iii) and (iv), respectively. The asymptotic theory covers both the cases when the nuisance parameters Λ are known as well as when some or all of the components of Λ are unknown, provided that sufficient data are available for estimating these unknowns.

Section 1.2 explains some notations and definitions used in the article. In Section 2, we identify the common data structures that arise in the analysis of models with surrogate predictors. In Section 3, we define our general model, develop three approximate quasi-likelihood models for the observed data, and discuss several examples. Section 4 studies estimation based on the approximate quasi-likelihood/variance function models derived in Section 3. Section 5 discusses estimation via Methods (i) and (ii). Two examples are discussed in Section 6.

1.2 Notations and Definitions

In this article we will use the notation for derivatives of vector- and matrix-valued functions as described in Fuller (1987, app. A.4). To these results we add notations for $\partial A/\partial\theta$ and $\partial A/\partial\theta'$, where A is a $p \times q$ matrix and θ is an $r \times 1$ vector. The former is an $rp \times q$ matrix containing the $r \times 1$ vectors $\partial a_{ij}(\theta)/\partial\theta$ in a $p \times q$ block structure. The latter is a $p \times qr$ matrix containing the $1 \times r$ vectors $\partial a_{ij}(\theta)/\partial\theta'$ in a $p \times q$ block structure.

We extend the definition of the trace function to $ds \times s$ $(d, s = 1, 2, \ldots)$ matrices as follows: define $\mathrm{tr}(A_{ds\times s}) = \{\mathrm{trace}(A_1), \ldots, \mathrm{trace}(A_d)\}'$, where A_j is the square matrix containing rows $(j-1)s + 1$ through js of A. Although we use tr more extensively, it is particularly convenient in the handling of quadratic approximations to vector-valued functions of random vectors. For example, if $f(\cdot)$ is $p \times 1$, Z is $q \times 1$, $f_z(z) = (\partial/\partial z')f(z)$, and $f_{zz}(z) = (\partial/\partial z)f_z(z)$, then the quadratic Maclaurin series approximation to $f(Z)$, $f_Q(Z)$ can be represented as $f_Q(Z) = f(0) + f_z(0)Z + (1/2)\mathrm{tr}\{f_{zz}(0)ZZ'\}$ and its expectation as $\mathrm{E}\{f_Q(Z)\} = f(0) + f_z(0)\mathrm{E}(Z) + (1/2)\,\mathrm{tr}\{f_{zz}(0)\mathrm{E}(ZZ')\}$. The mnemonic simplicity of the latter two expressions is due to the fact that tr is a linear operator.

Finally, we use $\dim(v)$ to denote the dimension of a row- or column-vector v.

2. Data Structures

In this article we assume that the data available for estimating the parameters of (1.1) and (1.2) or (1.3) are composed of one or more of five types:

1. *primary* data containing n_1 observations (Y_i, W_i);
2. *internal validation* data containing n_2 observations (Y_i, X_i, W_i);
3. *internal reliability* data containing n_3 observations $(Y_i, W_{i1}, \ldots, W_{ik_i})$, where for fixed i, W_{ij} $(j = 1, \ldots, k_i)$ are iid;
4. *external validation* data containing n_4 observations (X_i, W_i);
5. *external reliability* data containing n_5 observations $(W_{i1}, \ldots, W_{ik_i})$, where for fixed i, W_{ij} $(j = 1, \ldots, k_i)$ are iid.

Observations within and between data types are stochastically independent.

When necessary for clarity we attach an additional preceding subscript to an observation to indicate its type. For example, the pairs (X_{2i}, W_{2i}) and (X_{4i}, W_{4i}) are from internal and external validation data, respectively. This notation permits identification of unobservables without ambiguity, for example, X_{1i} and U_{2i}.

All five types of data are relevant to model (1.2); when model (1.3) is assumed, however, only primary and validation data are relevant.

The primary data generally do not identify all of the components of Θ or the unknown components of Λ. External validation/reliability data generally identify the unknown components of Λ and allow for identification of Θ when combined with primary data. Internal validation/reliability data serve the same purpose as external data with regard to identification but are preferred over external data for obvious reasons. Note that internal validation data can be used to test the conditional independence assumption of Section 1.

Typically, the data required for a measurement error analysis have $n_1 > 0$ and one of $n_2, \ldots, n_5 > 0$, depending on model (1.2) or (1.3). For example, Tosteson, Stefanski, and Schafer (1989) have $n_2 = n_3 = n_5 = 0$, but $n_4 > 0$ (see Sec. 6). Rudemo et al. (1989), however, have $n_2 = n_3 = n_4 = n_5 = 0$, but they assume that $X = W + \delta U$ and they exploit the nonlinearity in their model to identify δ^2.

As noted previously, we assume that reliability data are collected only when the measurement error model (1.2) is assumed. In this case if $W_{ir} = c(X_i, \eta) + \delta U_{ir}$, $\mathrm{E}(U_{ir} \mid X_i = x_i) \equiv 0$, and $\mathrm{cov}\ (U_{ir} \mid X_i = x_i) = \Omega(x_i, \eta, \gamma)$ for $r = 1, \ldots, k_i$, then $\overline{W}_{i\cdot} = c(X_i, \eta) + k_i^{-1/2}\delta U_{i*}$, $\mathrm{E}(U_{i*} \mid X_i = x_i) \equiv 0$, and $\mathrm{cov}(U_{i*} \mid X_i = x_i) = \Omega(x_i, \eta, \gamma)$. Thus $(\overline{W}_{i\cdot}, X_i)$ satisfy (1.2) upon replacing δ with $k_i^{-1/2}\delta$. This fact is exploited later in the article.

When n_1 and $n_2 > 0$ all parameters are identifiable even when $n_3 = n_4 = n_5 = 0$. Typically, $n_1 \gg n_2$ because of additional costs associated with obtaining observations on X. Thus we have a small data set from which consistent estimates of Θ can be found and a larger data set from which approximately consistent estimates are available using the methods of Sections 4 and 5. In general, the effect of pooling the two types of data is to obtain a less variable and less biased estimate than would be obtained if only the primary data were available. The estimates are also typically much less variable than if the primary data were ignored, but the bias incurred means that a variance-bias trade-off determines whether the primary data should be used. Of course, the primary data are useful in testing for no association; see Tosteson and Tsiatis (1988) and Stefanski and Carroll (1990).

3. BASIC MODELS AND EXAMPLES

In this section, we determine three general approximate models for Y given W when the error in predicting X from W is small. The three approximations to the mean and variance functions of Y given W are given in Equations (3.2), (3.4), and (3.5), respectively. Several specific cases of the model are discussed for illustration and clarification of the general results.

Assume, in addition to (1.1), that as $\delta \to 0$, there are known functions (g, h_2, h_3), parameters $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$, and a mean zero random variable U, such that

$$X = g(W, \eta, \delta U),$$
$$\mathrm{E}(\delta U \mid W) = \delta^2 h_2(W, \eta, \gamma) + O_P(\delta^3),$$
$$\mathrm{cov}(\delta U \mid W) = \delta^2 h_3(W, \eta, \gamma) + O_P(\delta^3). \tag{3.1}$$

Define

$$f_{mx}(x, \beta) = (\partial/\partial x^t) f_m(x, \beta)$$

and

$$f_{mxx}(x, \beta) = (\partial/\partial x) f_{mx}(x, \beta),$$

letting subscripts denote derivatives. Define f_{vx} and f_{vxx} similarly. Also define $g_s(w, \eta, s) = (\partial/\partial s^t) g(w, \eta, s)$ and $g_{ss}(w, \eta, s) = (\partial/\partial s) g_s(w, \eta, s)$ and

$$\begin{aligned} H_m(w, \eta, \gamma, \beta) \\ &= f_{mx}\{g(w, \eta, 0), \beta\} g_s(w, \eta, 0) h_2(w, \eta, \gamma) \\ &\quad + (1/2)\mathrm{tr}[f_{mxx}\{g(w, \eta, 0), \beta\} \\ &\quad \times g_s(w, \eta, 0) h_3(w, \eta, \gamma) g_s(w, \eta, 0)^t] \\ &\quad + (1/2) f_{mx}\{g(w, \eta, 0), \beta\} \\ &\quad \times \mathrm{tr}\{g_{ss}(w, \eta, 0) h_3(w, \eta, \gamma)\}. \end{aligned}$$

Define $H_v(w, \eta, \gamma, \beta, \theta, \sigma^2)$ similarly except that f_v replaces f_m. In addition, define

$$\begin{aligned} S(w, \eta, \gamma, \beta) &= f_{mx}\{g(w, \eta, 0), \beta\} g_s(w, \eta, 0) \\ &\quad \times h_3(w, \eta, \gamma) g_s(w, \eta, 0)^t f_{mx}\{g(w, \eta, 0), \beta\}^t. \end{aligned}$$

Assumption (3.1), the relationships $\mathrm{E}(Y^p \mid W) = \mathrm{E}\{\mathrm{E}(Y^p \mid X) \mid W\}$ $(p = 1, 2)$, and Taylor series expansions of $f_m\{g(w, \eta, \delta u), \beta\}$ and $f_v\{g(w, \eta, \delta u), \beta, \theta\}$ around $\delta = 0$ are used to show that

$$\mathrm{E}(Y \mid W) \approx U_{mA,1}(W, \beta, \eta, \gamma, \delta^2), \tag{3.2a}$$

and

$$\mathrm{var}(Y \mid W) \approx U_{vA,1}(W, \beta, \theta, \sigma^2, \eta, \gamma, \delta^2), \tag{3.2b}$$

where

$$\begin{aligned} U_{mA,1}(w, \beta, \eta, \gamma, \delta^2) \\ &= f_m\{g(w, \eta, 0), \beta\} + \delta^2 H_m(w, \eta, \gamma, \beta) \end{aligned}$$

and

$$\begin{aligned} U_{vA,1}(w, \beta, \theta, \sigma^2, \eta, \gamma, \delta^2) \\ &= \sigma^2[f_v\{g(w, \eta, 0), \beta, \theta\} + \delta^2 H_v(w, \eta, \gamma, \beta, \theta, \sigma^2)] \\ &\quad + \delta^2 S(w, \eta, \gamma, \beta). \end{aligned}$$

The error in the approximation in (3.2) is of order $O_P(\delta^3)$. The subscript A in (3.2a) and (3.2b) indicates the approximate nature of model. The model specified in (3.2) is an extension of some approximate models studied by Armstrong (1985) and Fuller (1987, sec. 3.3) to allow for more general error structures in both the models relating Y to X and X with W.

A simpler and sometimes appropriate approximate model can be obtained by first modeling X as a function of W and then substituting the conditional expectation of X given W for X in a standard analysis of (1.1). This approach is most natural in the generalized Berkson error model, in which case $\mathrm{E}(X \mid W)$ is, in our notation, $g(W, \eta, 0)$, but it can also be employed when (1.2) is known to hold. In the latter case an approximation to $\mathrm{E}(X \mid W)$ would be employed:

$$\begin{aligned} \tilde{\mathrm{E}}(X \mid W) &= g(W, \eta, 0) + \delta^2 g_s(W, \eta, 0) h_2(W, \eta, \gamma) \\ &\quad + (\delta^2/2)\mathrm{tr}\{g_{ss}(W, \eta, 0) h_3(W, \eta, \gamma)\}. \end{aligned}$$

Note that the error in this approximation is of order $O_P(\delta^3)$ and that $\tilde{\mathrm{E}}(X \mid W)$ reduces to $\mathrm{E}(X \mid W)$ under the Berkson model (1.3). Thus we use it in the following, assuming that either (1.2) or (1.3) holds. Substituting $\tilde{\mathrm{E}}(X \mid W)$ for X in (1.1) induces a second approximate model:

$$U_{mA,2}(w, \beta, \eta, \gamma, \delta^2) = f_m\{\tilde{\mathrm{E}}(X \mid W = w), \beta\}, \quad (3.3a)$$

$$U_{vA,2}(w, \beta, \theta, \sigma^2, \eta, \gamma, \delta^2) = \sigma^2 f_v\{\tilde{\mathrm{E}}(X \mid W = w), \beta, \theta\}. \quad (3.3b)$$

Taylor series expansions of $U_{mA,2}$ and $U_{vA,2}$ and evaluation at $w = W$ show that, with an error of order $O_P(\delta^3)$,

$$U_{mA,1} - U_{mA,2} \approx (\delta^2/2)\mathrm{tr}[f_{mxx}\{g(W, \eta, 0), \beta\} g_s(W, \eta, 0) \times h_3(W, \eta, \gamma) g_s(W, \eta, 0)^t] \quad (3.4a)$$

and

$$U_{vA,1} - U_{vA,2} \approx \sigma^2(\delta^2/2)\mathrm{tr}[f_{vxx}\{g(W, \eta, 0), \beta\} g_s(W, \eta, 0) \times h_3(W, \eta, \gamma) g_s(W, \eta, 0)^t] + \delta^2 S(W, \eta, \gamma, \beta). \quad (3.4b)$$

It follows that the strategy of replacing X by $\tilde{\mathrm{E}}(X \mid W)$ can be justified whenever the right sides of (3.4) are negligible. Consider, for example, simple logistic regression, $\Pr(Y = 1 \mid X = x) = F(\beta_0 + \beta_1 x)$, where $F(t) = \{1 + \exp(t)\}^{-1}$, under the Berkson error model (1.3) with $c^*(w, \eta) = w$. In this case, noting that $\sigma^2 = 1$, (3.4a) and (3.4b) are

$$U_{mA,1} - U_{mA,2} \approx (\beta_1^2/2)\mathrm{var}(X \mid W) \times F^{(2)}\{\beta_0 + \beta_1 g(W, \eta, 0)\}$$

and

$$U_{vA,1} - U_{vA,2} \approx (\beta_1^2/2)\mathrm{var}(X \mid W) \times (F^{(2)}\{\beta_0 + \beta_1 g(W, \eta, 0)\} \times [1 - 2F\{\beta_0 + \beta_1 g(W, \eta, 0)\}])$$

where $F^{(k)}$ is the kth derivative of F. Thus the two approximations are essentially equal whenever $\beta_1^2 \mathrm{var}(X \mid W)$ is negligible. This occurs near the null model $\beta_1 = 0$, a situation not uncommon in epidemiologic research (Rosner et al. 1989). Similar justification for (3.3) can be found for other generalized linear models.

The strategy of replacing X by $\tilde{\mathrm{E}}(X \mid W)$ has the advantage of always producing a range-preserving model, whereas (3.2) does not. We observed earlier, however, that it is not always appropriate. We now describe a third approximate model that is range preserving and differs from (3.2) by $O_P(\delta^3)$. Let $a_m = a_m(w, \eta, \beta) = f_{mx}\{g(w, \eta, 0), \beta\}^t / \| f_{mx}\{g(w, \eta, 0), \beta\}\|^2$, and define $a_v = a_v(w, \eta, \beta, \theta)$ analogously. Define

$$U_{mA,3} = f_m\{g(w, \eta, 0) + \delta^2 a_m H_m(w, \eta, \gamma, \beta), \beta\} \quad (3.5a)$$

and

$$U_{vA,3} = \sigma^2 f_v[g(w, \eta, 0) + \delta^2 a_v\{H_v(w, \eta, \gamma, \beta, \theta, \sigma^2) + \sigma^{-2} S(w, \eta, \gamma, \beta)\}, \beta, \theta]. \quad (3.5b)$$

Taylor series expansions of $U_{mA,3}$ and $U_{vA,3}$ and evaluation at $w = W$ show that they differ from (3.2) by $O_P(\delta^3)$.

In Sections 4 and 5 we study estimation for the approximate model (U_{mA}, U_{vA}), where (U_{mA}, U_{vA}) can be any one of $(U_{mA,i}, U_{vA,i})$ $(i = 1, 2, 3)$ given by (3.2), (3.3), or (3.5), respectively. We now show by examples the flexibility and generality of our modeling framework.

Example 3.1 (general linear Berkson error model). Consider the model (1.3) with $c^*(w, \eta) = \eta_0 + \eta_1 w$. In the notation of (3.1), η is a vector containing the unique elements of η_0 and η_1, $g(w, \eta, \delta u) = \eta_0 + \eta_1 w + \delta u$, $h_2 = g_{ss} = 0$, $g_s = I_{\dim(u)}$, $h_3(w, \eta, \gamma) = \Omega^*(w, \eta, \gamma)$,

$$S(w, \eta, \gamma, \beta) = f_{mx}(\eta_0 + \eta_1 w, \beta) \times \Omega^*(w, \eta, \gamma) f_{mx}(\eta_0 + \eta_1 w, \beta)^t,$$

and

$$H_m(w, \eta, \gamma, \beta) = (1/2)\mathrm{tr}\{f_{mxx}(\eta_0 + \eta_1 w, \beta)\mathring{\Omega}^*(w, \eta, \gamma)\},$$

with a similar expression for H_v.

Example 3.2. Consider a homoscedastic linear regression Berkson error model $f_m(x, \beta) = x^t\beta$, $f_v(x, \beta, \theta) = 1$, $x = \eta_0 + \eta_1 w + \delta u$, and $\mathrm{cov}(U \mid W = w) = \Omega^*(w, \eta, \gamma)$. It follows that $\mathrm{E}(Y \mid W = w) = \beta^t(\eta_0 + \eta_1 w)$ and $\mathrm{var}(Y \mid W = w) = \sigma^2 + \delta^2 \beta^t \eta_1 \Omega^*(w, \eta, \gamma) \eta_1^t \beta$. If $\Omega^*(w, \eta, \gamma)$ is constant, then the observed data will have constant variance and we can estimate β only if η_0, η_1 are known, as in the classical Berkson case, or estimated from additional data. In the Berkson case, if only one variable, say the last, $X^{(p)}$, with proxy $W^{(p)}$, is measured with error, then the model is sufficiently identified to check for heteroscedastic measurement error. For example, if $e = (0, \ldots, 0, 1)^t$, then when $\Omega^*(w, \eta, \gamma) = ee^t \exp(\gamma w^{(p)})$, we have $\mathrm{var}(Y \mid W = w) = \sigma^2 + \delta_*^2 \exp(\gamma w^{(p)})$. Graphical and formal devices for checking whether $\gamma = 0$ here and estimating γ were given in Carroll and Ruppert (1988, sec. 2.7, chaps. 3 and 6). This example illustrates that our approach allows one to model various facets of the data while retaining the underlying measurement error structure.

Example 3.3. In assay models (Davidian, Carroll, and Smith 1988; Rudemo et al. 1989), $f_m(x, \beta)$ is usually nonlinear. A standard model assumes that the variance is proportional to a power of the mean, that is, $f_v(x, \beta, \theta) = f_m(x, \beta)^\theta$. Here X is the univariate log concentration, with zero concentration measured without error and handled separately. In the linear Berkson error version of model (1.3), $\eta_0 = 0$, $\eta_1 = 1$. In practice, both σ and δ are fairly small (Davidian et al. 1988). We have $h_2 = g_{ss} = 0$, $g_s =$

1, and to within order $O_p(\sigma^3 + \delta^3)$, as $\sigma \to 0$ and $\delta \to 0$, Equations (3.2) yield

$$\mathrm{E}(Y \mid W) \approx f_m(W, \beta) + (\delta^2/2) \times f_{mxx}(W, \beta) h_3(W, \gamma) \quad (3.6a)$$

and

$$\mathrm{var}(Y \mid W) \approx \sigma^2 f_v(W, \beta, \theta) + \delta^2 f_{mx}^2(W, \beta) h_3(W, \gamma). \quad (3.6b)$$

Model (3.6) is typically identifiable. Whether measurement error has constant variance can be assessed by positing forms for h_3, for example, $h_3(w, \gamma) = \exp(\gamma w)$, and then using standard techniques for variance analysis. Model (3.6a) may not be range preserving, and in such cases we suggest using the approximate mean model corresponding to (3.5a):

$$\mathrm{E}(Y \mid W) \approx f_m\left(W + \frac{\delta^2 f_{mxx}(W, \beta) h_3(W, \gamma)}{2 f_{mx}(W, \beta)}, \beta\right).$$

In this example, the variance function is range preserving.

Example 3.4 (general linear measurement error model). Consider the linear measurement error model version of (1.2). It follows that, for a generalized inverse η^- satisfying $\eta^- \eta = I_{p \times p}$, $g(w, \eta, \delta u) = \eta_1^-(w - \eta_0 - \delta u)$, $g_s = -\eta_1^-$, and $g_{ss} = 0$. It is easily shown that $\mathrm{E}(U \mid \eta_0 + \eta_1 X) = 0$ and $\mathrm{cov}(U \mid \eta_0 + \eta_1 X) = \mathrm{E}\{\Omega(X, \eta, \gamma) \mid \eta_0 + \eta_1 X\} = \Omega_*(X, \eta, \gamma)$, where $\Omega_*(x, \eta, \gamma) = \Omega\{\eta_1^-(x - \eta_0)\ \eta, \gamma\}$. Let κ_W be the marginal density of W with gradient $\kappa_W^{(1)}$. An appeal to Lemma A.1 in the Appendix shows that

$$h_2(w, \eta, \gamma) = -\left[\mathrm{tr}\left\{\frac{\partial \Omega_*(w, \eta, \gamma)}{\partial w}\right\} + \Omega_*(w, \eta, \gamma) \frac{\kappa_W^{(1)}(w)}{\kappa_W(w)}\right] \quad (3.7a)$$

and

$$h_3(w, \eta, \gamma) = \Omega_*(w, \eta, \gamma). \quad (3.7b)$$

This model includes the possibility that W is a biased measurement for X but can be calibrated with estimated parameters η_0, η_1. If W is unbiased so that $\eta_0 = g_{ss} = 0$, $\eta_1 = I_{\dim(X)} = -g_s$, we have the classical measurement error model. For identifiability, one of the diagonal elements of $\Omega(\cdot, \cdot, \cdot)$ corresponding to a predictor measured with error must have value 1.0.

In some instances, exact forms for h_2 and h_3 can be computed. For example, suppose that U and X are independent and normally distributed, the latter with mean μ_X and covariance Ω_X, and that $\mathrm{E}(W \mid X) = X$ and $\mathrm{cov}(W \mid X) = \Omega$. Here Ω has the first diagonal element equal to 1.0. In the notation of (3.1), $g(w, \eta, \delta u) = w - \delta u$, $\Lambda = (\gamma^t, \delta^2)^t$, γ contains the unique elements of μ_X, Ω_X, and Ω, and

$$h_2(w, \Lambda) = \Omega(\Omega_X + \delta^2 \Omega)^{-1}(w - \mu_X)$$

and

$$h_3(w, \Lambda) = \Omega\{I - (\Omega_X + \delta^2 \Omega)^{-1} \delta^2 \Omega\}.$$

Often, sample sizes are large enough that when the dimension of W is small, the location score $\kappa_W^{(1)}/\kappa_W$ in (3.7a) can be estimated nonparametrically. Related work on hypothesis testing (Stefanski and Carroll, 1990) shows that estimating $\kappa_W^{(1)}/\kappa_W$ is feasible and advantageous when $\dim(W) = 1$. Alternatively, a flexible parametric density could be fit to $\{W_i\}$, thereby providing an estimator of $\kappa_W^{(1)}/\kappa_W$. Note that $\kappa_W^{(1)}/\kappa_W$ is linear iff κ_W is normal.

Example 3.3 (continued). Consider the mean and variance functions for Example 3.3 but for a measurement error model instead of a Berkson error model, with X scalar so that $\Omega = 1$. Assuming normality and letting $\sigma \to 0$, $\delta \to 0$, to within $O_P(\sigma^3 + \delta^3)$ we obtain

$$\mathrm{var}(Y \mid W) \approx \sigma^2 f_v(W, \beta, \theta) + \delta^2 f_{mx}^2(W, \beta) \sigma_X^2 / (\delta^2 + \sigma_X^2);$$

$$\begin{aligned}\mathrm{E}(Y \mid W) &\approx f_m(W, \beta) + \{(1/2)\sigma_X^2 f_{mxx}(W, \beta) - f_{mx}(W, \beta)(W - \mu_X)\}\delta^2/(\delta^2 + \sigma_X^2) \\ &\approx f_m\left[W + \left(\frac{\delta^2}{\delta^2 + \sigma_X^2}\right) \times \left\{\frac{f_{mxx}(W, \beta)\sigma_X^2}{2 f_{mx}(W, \beta)} - W + \mu_X\right\}, \beta\right].\end{aligned}$$

The two approximations for the mean are from (3.2a) and (3.5a), respectively, the latter appropriate when the true mean is positive. Note that in this case a simpler model can be obtained by replacing $(\delta^2 + \sigma_X^2)$ with σ_X^2 in both the mean and variance function without affecting the order of the approximation.

Example 3.5. The error in using (3.7) is of order $O_P(\delta^3)$ when X and U are normally distributed. However, (3.7) suggests flexible models that can cope with nonnormal and/or heteroscedastic error. Consider, for example, a logistic regression study such as described by Jones et al. (1987). In this study, the outcome is incidence of breast cancer, and the predictor X is average daily dietary saturated fat intake. We have analyzed a subset of these data consisting of a cohort of 2,888 women under the age of 50. In this group there were 37 cases of breast cancer. Here W is derived from a 24-hour diet recall questionnaire and has a large variance. It is reasonable in this case to suppose that W is unbiased for X. Thus we set $\eta_0 = 0$ and $\eta_1 = 1$. If measurement error variance is log-linear in X, and if we replace the score $\kappa_W^{(1)}(w)/\kappa_W(w)$ by a linear function of w, then (3.7) and a little algebra yield

$$\begin{aligned}\mathrm{E}(W \mid X = x) &= x, \\ \mathrm{E}(\delta U \mid W = w) &= \delta^2 \exp(\alpha_3 w)(\alpha_1 + \alpha_2 w) \\ &= \delta^2 h_2(w, \Lambda); \quad (3.8a) \\ \mathrm{var}(\delta U \mid W = w) &= \delta^2 \exp(\alpha_3 w) = \delta^2 h_3(w, \Lambda). \quad (3.8b)\end{aligned}$$

Equations (3.8) identify h_2, h_3. For binary regression models, $\Pr(Y = 1 \mid X = x) = F(\beta_0 + \beta_1 x)$, (3.2) yields an approximate model that is not range preserving, and we suggest using (3.5). In this case (3.5a) gives the approximate mean to within $O_P(\delta^3)$,

$$\Pr(Y = 1 \mid W) \approx F[\beta_0 + \beta_1\{W - \delta^2 h_2(W, \Lambda)\} \\ + (\delta^2/2)\beta_1^2 F_{21}(\beta_0 + \beta_1 W) h_3(W, \Lambda)], \quad (3.9)$$

with $F_{21}(v) = F^{(2)}(v)/F^{(1)}(v)$, where $F^{(1)}$ and $F^{(2)}$ are the first two derivatives of F. Note that, for binary regression models, the approximations in (3.5) satisfy $U_{vA,3} = U_{mA,3}(1 - U_{mA,3})$.

4. ESTIMATORS BASED ON APPROXIMATE QUASI-LIKELIHOOD

Estimation of the parameters in the approximate quasi-likelihood/variance function models (3.2), (3.3), and (3.5) is complicated by the fact that data of different types are frequently available. We now describe a general method of obtaining estimators with asymptotically valid standard errors from data composed of combinations of observations of the types described in Section 2. Our approach to estimation is based on the principles set forth in Davidian and Carroll (1987) and Carroll and Ruppert (1988), but it is tailored to the specifics of the problem at hand.

Presentation of the general theory is facilitated by a judicious partitioning of the parameters in $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$, which we now describe. Recall that (U_{mA}, U_{vA}) is used to repesent the approximations in either (3.2), (3.3), or (3.5).

Depending on the particular model assumed in (1.2) or (1.3), some components of Λ may be known. Designate the unknown components by Λ_U. It is implicitly assumed that the components of Λ_U are identified by the available data. Let Λ_P^* designate those components of Λ_U that are identified by the primary data. Partition Λ_P^* into those components appearing in the mean function U_{mA}, Λ_{Pm}^*, and those not appearing in the mean function, Λ_{Pv}^*. Let Λ_{Pm} be a subset of Λ_{Pm}^*, and let Λ_{Pv} be a subset of the set of parameters in Λ_P^* that are not contained in Λ_{Pm}. The parameters in Λ_U that are capable of being estimated with the primary data are contained in Λ_P^*, whereas $\Lambda_P = (\Lambda_{Pm}^t, \Lambda_{Pv}^t)^t$ contains the parameters that we choose to estimate using the primary data. The two sets need not be equal and will sometimes differ for reasons of convenience and/or model robustness. Frequently, either Λ_{Pm} or Λ_{Pv} will be empty. Finally, let Λ_{VR} contain all of the parameters in Λ_U not contained in Λ_{Pm} or Λ_{Pv}. Thus, depending on context, we can write either $\Lambda_U = (\Lambda_{Pm}^t, \Lambda_{Pv}^t, \Lambda_{VR}^t)^t$ or $\Lambda = (\eta^t, \gamma^t, \delta^2)^t$ or $\Lambda = (\Lambda_U, \Lambda_U^c)$, where the last equality denotes set equivalence. It is also convenient to write $\hat{\Lambda} = (\hat{\eta}^t, \hat{\gamma}^t, \hat{\delta}^2)^t$, even when some components of Λ are known.

Most of the information for estimating the components of Λ_U is contained in the validation and/or reliability data, depending on the particular model under study. We assume that unbiased score equations, $\Upsilon_{R,\Lambda_{Pm}}$, $\Upsilon_{R,\Lambda_{Pv}}$, $\Upsilon_{R,\Lambda_{VR}}$, $\Upsilon_{V,\Lambda_{Pm}}$, $\Upsilon_{V,\Lambda_{Pv}}$, and $\Upsilon_{V,\Lambda_{VR}}$, are available for obtaining consistent M estimators of the components of Λ_U from the available reliability and validation data. For example, in Example 3.1, η_0 and η_1 would often be estimated by the usual normal equations for linear regression. The scores with first subscript R are functions of Λ and the replicates of W in the reliability data. The scores with first subscript V are functions of Λ and the pairs (X, W) from the validation data.

We propose to estimate Θ and Λ_U with M estimators $\hat{\Theta}$ and $\hat{\Lambda}_U$, solving equations of the form

$$0 = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} F_{ji}(\Theta, \Lambda_U), \quad (4.1)$$

where j indexes the five types of observations described in Section 2. We now describe the functional form of F_{ji} for each of the five types of data described in Section 2.

Write U_{mA} and U_{vA} for the right sides of (3.2), (3.3), and (3.5), and define $r_A = y - U_{mA}$. Let $\Psi_{P,\beta}$, $\Psi_{P,\theta}$, Ψ_{P,σ^2}, $\Psi_{P,\Lambda_{Pm}}$, $\Psi_{P,\Lambda_{Pv}}$, and $\Psi_{P,\Lambda_{VR}}$ be functions of (y, w, Θ, Λ) defined via componentwise matching in the following equations:

$$(\Psi_{P,\beta}^t, \Psi_{P,\Lambda_{Pm}}^t) = \left(\frac{y - U_{mA}}{U_{vA}}\right) \frac{\partial U_{mA}}{\partial(\beta^t, \Lambda_{pm}^t)};$$

$$(\Psi_{P,\theta}^t, \Psi_{P,\sigma^2}, \Psi_{P,\Lambda_{Pv}}^t) = \left(\frac{r_A^2 - U_{vA}}{U_{vA}}\right) \frac{\partial \log U_{vA}}{\partial(\theta^t, \sigma^2, \Lambda_{Pv}^t)};$$

$$\Psi_{P,\Lambda_{VR}} = 0_{\dim(\Lambda_{VR}) \times 1}.$$

With these definitions, $F_{1i}(\Theta, \Lambda_U) = (\Psi_{P,\beta}^t, \Psi_{P,\theta}^t, \Psi_{P,\sigma^2}, \Psi_{P,\Lambda_{Pm}}^t, \Psi_{P,\Lambda_{Pv}}^t, \Psi_{P,\Lambda_{VR}}^t)^t$ evaluated at $(Y_{1i}, W_{1i}, \Theta, \Lambda)$.

Write f_m and f_v for the mean and variance function in (1.1), and define $r = y - f_m$. Let $\Psi_{IV,\beta}$, $\Psi_{IV,\theta}$, Ψ_{IV,σ^2}, $\Psi_{IV,\Lambda_{Pm}}$, $\Psi_{IV,\Lambda_{Pv}}$, and $\Psi_{IV,\Lambda_{VR}}$ be functions of $(y, x, w, \Theta, \Lambda)$ defined via componentwise matching in the following equations:

$$\Psi_{IV,\beta} = \left(\frac{y - f_m}{f_v}\right) \frac{\partial f_m}{\partial \beta};$$

$$(\Psi_{P,\theta}^t, \Psi_{P,\sigma^2}) = \left(\frac{r^2 - f_v}{f_v}\right) \frac{\partial \log f_v}{\partial(\theta^t, \sigma^2)};$$

$$(\Psi_{IV,\Lambda_{Pm}}^t, \Psi_{IV,\Lambda_{Pv}}^t, \Psi_{IV,\Lambda_{VR}}^t) = (\Upsilon_{V,\Lambda_{Pm}}^t, \Upsilon_{V,\Lambda_{Pv}}^t, \Upsilon_{V,\Lambda_{VR}}^t).$$

With these definitions, $F_{2i}(\Theta, \Lambda_U) = (\Psi_{IV,\beta}^t, \Psi_{IV,\theta}^t, \Psi_{IV,\sigma^2}, \Psi_{IV,\Lambda_{Pm}}^t, \Psi_{IV,\Lambda_{Pv}}^t, \Psi_{IV,\Lambda_{VR}}^t)^t$ evaluated at $Y_{2i}, X_{2i}, W_{2i}, \Theta, \Lambda)$.

For internal reliability data we summarize the observation $(Y_i, W_{i1}, \ldots, W_{ik_i})$ as $(Y_i, \overline{\mathrm{W}}_i, k_i)$; see Section 2. Define U_{mA*} and U_{vA*} as functions of $(\overline{w}, k, \Theta, \Lambda)$ via

$$U_{mA*} = U_{mA}(\overline{w}, \beta, \eta, \gamma, \delta^2/k)$$

and

$$U_{vA*} = U_{vA}(\overline{w}, \beta, \theta, \sigma^2, \eta, \gamma, \delta^2/k),$$

and define $r_{A*} = y - U_{mA*}$. Using established conventions, define

$$\Psi_{IR,\beta} = \left(\frac{y - U_{mA*}}{U_{vA*}} \right) \frac{\partial U_{mA*}}{\partial \beta},$$

$$(\Psi^t_{IR,\theta}, \Psi_{IR,\sigma^2}) = \left(\frac{r^2_{A*} - U_{vA*}}{U_{vA*}} \right) \frac{\partial \log U_{vA*}}{\partial(\theta^t, \sigma^2)},$$

and

$$(\Psi^t_{IR,\Lambda_{Pm}}, \Psi^t_{IR,\Lambda_{Pv}}, \Psi^t_{IR,\Lambda_{VR}}) = (\Upsilon^t_{R,\Lambda_{Pm}}, \Upsilon^t_{R,\Lambda_{Pv}}, \Upsilon^t_{R,\Lambda_{VR}}).$$

With these definitions, $F_{3i}(\Theta, \Lambda_U) = (\Psi^t_{IR,\beta}, \Psi^t_{IR,\theta}, \Psi_{IR,\sigma^2}, \Psi^t_{IR,\Lambda_{Pm}}, \Psi^t_{IR,\Lambda_{Pv}}, \Psi^t_{IR,\Lambda_{VR}})^t$ evaluated at $(Y_{3i}, \overline{W}_{3i\cdot}, k_i, \Theta, \Lambda)$.

Finally, since external data generally provide no information on Θ,

$$F_{4i}(\Theta, \Lambda_U) = (0_{1 \times \dim(\Theta)}, \Upsilon^t_{V,\Lambda_{Pm}}, \Upsilon^t_{V,\Lambda_{Pv}}, \Upsilon^t_{V,\Lambda_{VR}})^t$$

and

$$F_{5i}(\Theta, \Lambda_U) = (0_{1 \times \dim(\Theta)}, \Upsilon^t_{R,\Lambda_{Pm}}, \Upsilon^t_{R,\Lambda_{Pv}}, \Upsilon^t_{R,\Lambda_{VR}})^t,$$

evaluated at (X_{4i}, W_{ri}) and $(W_{5i}, \ldots, W_{5ik_i})$, respectively.

The scores we defined for estimating θ, σ^2, and Λ_{Pv} use squared residuals. Other scores based on functions of absolute residuals can also be employed; see Davidian and Carroll (1987).

Let $n = n_1 + \cdots + n_5$ and $p_{j,n} = n_j/n$. The asymptotic distribution of $(\hat{\Theta}, \hat{\Lambda}_U)$ is given for the case in which $n \to \infty$ and $p_{j,n} \to p_j \geq 0$.

The M estimators $(\hat{\Theta}, \hat{\Lambda}_U)$ converge in probability to (Θ_*, Λ_{U*}), satisfying

$$0 = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} \mathrm{E}\{F_{ji}(\Theta_*, \Lambda_{U*})\}.$$

Let $G_{ji}(\Theta, \Lambda_U)$ be the matrix of partial derivatives of F_{ji} with respect to (Θ^t, Λ^t_U), and define

$$A = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} \mathrm{E}\{G_{ji}(\Theta_*, \Lambda_{U*})\}$$

and

$$B = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} \mathrm{E}\{F_{ji}(\Theta_*, \Lambda_{U*}) \times F_{ji}(\Theta_*, \Lambda_{U*})^t\}.$$

Moment estimators of these matrices are given by

$$\hat{A} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} G_{ji}(\hat{\Theta}, \hat{\Lambda}_U)$$

and

$$\hat{B} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} F_{ji}(\hat{\Theta}, \hat{\Lambda}_U) F_{ji}(\hat{\beta}, \hat{\Lambda}_U)^t.$$

Standard asymptotic results imply that, under sufficient regularity conditions,

$$n^{1/2}\{(\hat{\Theta} - \Theta_*)^t, (\hat{\Lambda}_U - \Lambda_{U*})^t\}^t \xrightarrow{D} \mathfrak{N}\{0, A^{-1}B(A^t)^{-1}\},$$

$$\hat{A}^{-1}\hat{B}(\hat{A}^t)^{-1} \xrightarrow{P} A^{-1}B(A^t)^{-1}. \quad (4.21)$$

Example 4.1 (externally validated studies). Suppose that $\Lambda_U = \Lambda$, Λ_{Pm} and Λ_{Pv} are both empty, and $n_2 = n_3 = n_5 = 0$. Thus estimation of Λ depends entirely on an external validation data set. In this case the only nonzero F_{ji} are

$$F_{1i}(\Theta, \Lambda_U) = (\Psi^t_{P,\beta}, \Psi^t_{P,\theta}, \Psi_{P,\sigma^2}, 0_{1 \times \dim(\Lambda_U)})^t$$

and

$$F_{4i}(\Theta, \Lambda_U) = (0_{1 \times \dim(\Theta)}, \Upsilon^t_{V,\Lambda_{Pm}}, \Upsilon^t_{V,\Lambda_{Pv}}, \Upsilon^t_{V,\Lambda_{VR}})^t.$$

Let Ψ and γ denote the nonidentically zero components of F_{11} and F_{41}, respectively. The matrices in the asymptotic covariance matrix of (4.2) for this case have the forms

$$A = \begin{pmatrix} p_1 A_{\Psi,\Theta} & p_1 A_{\Psi,\Lambda} \\ 0 & p_4 A_{\Upsilon,\Lambda} \end{pmatrix}, \qquad B = \begin{pmatrix} p_l B_{\Psi} & 0 \\ 0 & p_4 B_{\Upsilon} \end{pmatrix},$$

where, for example, $A_{\Psi,\Theta} = \mathrm{E}(\partial\Psi/\partial\Theta^t)$ and $B_\Psi = \mathrm{E}(\Psi\Psi^t)$. In this case the asymptotic covariance matrix of $n^{1/2}(\hat{\Theta} - \Theta_*)$ has the form

$$p_1^{-1} A^{-1}_{\Psi,\Theta} B_\Psi (A^t_{\Psi,\Theta})^{-1} + p_4^{-1}\Delta, \quad (4.3)$$

where Δ is a nonnegative matrix. The matrix Δ depends on the submatrices of A and B in a simple but not very informative way. The term $p_4^{-1}\Delta$ is the contribution to the asymptotic covariance matrix of $n^{1/2}(\hat{\Theta} - \Theta_*)$ due to estimating Λ.

If costs c_1, c_2 can be assigned to obtaining observations (Y, W), (X, W), respectively, then (4.3) could be used for design purposes.

5. ALTERNATIVE ESTIMATORS

Denote the right side of (4.1) by $L^*(\Theta, \eta, \gamma, \delta^2)$. Frequently, Λ_{Pm} and Λ_{Pv} are both empty, and then $L^* = (L^t, L^t_*)^t$, where L_* does not depend on Θ. It follows that $\hat{\Lambda}_U$ is found by solving $L_*(\Lambda_U) = 0$ and $\hat{\Theta}$ is found by solving $L(\Theta, \hat{\eta}, \hat{\gamma}, \hat{\delta}^2) = 0$. In this case there are two additional approaches to estimation that deserve attention. Below we adapt the methods proposed by Whittemore and Keller (1988) and Stefanski (1985) to our estimation problem.

In the following we drop the distinction between $\hat{\Lambda}_U$ and $\hat{\Lambda}$, keeping in mind that the latter is $(\hat{\Lambda}_U, \hat{\Lambda}^c_U)$. Note that

$$L(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} F^\dagger_{ji}(\Theta, \Lambda),$$

where $F^\dagger_{ji}(\Theta, \Lambda)$ denotes the first $\dim(\Theta)$ rows of $F_{ji}(\Theta, \Lambda)$ $(j = 1, 2, 3; i = 1, \ldots, n_j)$.

5.1 Approximating the Quasi-likelihood Estimators

Let $\hat{\Theta}(\tau)$ be a function of τ defined implicitly by the equation $L(\hat{\Theta}(\tau), \hat{\eta}, \hat{\gamma}, \tau\hat{\delta}^2) = 0$. Note that $\hat{\Theta}(1) = \hat{\Theta}$ and $\hat{\Theta}(0)$ is the so-called naive estimator obtained by fitting model (1.1) to the pairs $(Y_{ji}, \hat{X}_{Cji})_{i=1}^{n_j}$ $(j = 1, 2, 3)$, using standard estimation techniques, where $(Y_{ji}, \hat{X}_{Cji}) = (Y_{ji}, \tilde{X}_{ji}(\hat{\eta}, 0))$ and

$$\begin{aligned}(Y_{ji}, \tilde{X}_{ji}(\eta, \delta)) &= (Y_{1i}, g(W_{1i}, \eta, \delta U_{1i})), && \text{if } j = 1,\\ &= (Y_{2i}, X_{2i}), && \text{if } j = 2,\\ &= (Y_{3i}, g(\overline{W}_{3i\cdot}, \eta, \delta \overline{U}_{3i\cdot})), && \text{if } j = 3.\end{aligned} \tag{5.1}$$

In fact the naive estimator defined previously satisfies

$$0 = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} \Psi_T(Y_{ji}, \hat{X}_{Cji}, \hat{\Theta}(0)),$$

where Ψ_T is the appropriate score for model (1.1) in the absence of measurement error.

With this notation we now derive the estimator proposed by Whittemore and Keller (1988). Taylor series expansions of $L(\hat{\Theta}(\tau), \hat{\eta}, \hat{\gamma}, \tau\hat{\delta}^2)$ and $\hat{\Theta}(\tau)$ lead to the approximations to within $O_P(\delta^3)$,

$$L_\Theta(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0)\hat{\Theta}_\tau(0) + \hat{\delta}^2 L_{\delta^2}(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0) \approx 0$$

and

$$\hat{\Theta}_\tau(0) \approx \hat{\Theta}(1) - \hat{\Theta}(0),$$

where, for example, $L_\Theta = \partial L/\partial\Theta^t$ and $\hat{\Theta}_r = \partial\hat{\Theta}/\partial\tau$. Thus $\hat{\Theta} \approx \hat{\Theta}_{c,1}$, where

$$\hat{\Theta}_{c,1} = \hat{\Theta}(0) - \hat{\delta}^2\{L_\Theta(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0)\}^{-1}L_{\delta^2}(\hat{\Theta}(0), \hat{\eta}, \hat{\gamma}, 0).$$

The utility of $\hat{\Theta}_{c,1}$ lies in its computability. It is an explicit function of $\hat{\Theta}(0)$, which in turn often can be obtained using standard statistical software. The difference, $\hat{\Theta} - \hat{\Theta}_{c,1}$, is $O(\hat{\delta}^4)$ a.s. when L is a well-behaved function of Θ and δ^2.

Write $\hat{\Theta}_{c,1} = \hat{\Theta}_{c,1}(U_{mA}, U_{vA})$ to emphasize the dependence of $\hat{\Theta}_{c,1}$ on U_{mA} and U_{vA}, and let U_m and U_v denote the left sides of (3.2a) and (3.2b), respectively. Then since the approximations in (3.2) and (3.5) are of order $O_P(\delta^3)$, it follows that $\hat{\Theta}_{c,1}(U_{mA}, U_{vA}) = \hat{\Theta}_{c,1}(U_m, U_v)$ in these two cases. Thus $\hat{\Theta}_{c,1}$ is the estimator proposed by Whittemore and Keller (1988) whenever the approximations in (3.2) or (3.5) are employed.

Note that $\hat{\Theta}_{c,1}$ can be written in the form

$$\hat{\Theta}_{c,1} = \hat{\Theta}(0) + (\hat{\delta}^2/2)Q_n(\hat{\Theta}(0), \hat{\Lambda})^{-1}H_n(\hat{\Theta}(0), \hat{\Lambda}),$$

where

$$Q_n(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} q_{ij}(\Theta, \Lambda),$$

$$H_n(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} h_{ij}(\Theta, \Lambda),$$

$$h_{ij}(\Theta, \Lambda) = -2\left\{ \frac{\partial}{\partial \delta^2} F_{ji}^{\dagger}(\Theta, \Lambda)\right\}_{\delta^2=0},$$

$$q_{ij}(\Theta, \Lambda) = \Psi_{T\Theta}(Y_{ji}, \tilde{X}_{ji}(\eta, 0), \Theta),$$

and $\Psi_{T\Theta} = (\partial/\partial\Theta^t)\Psi_T$.

5.2 Correcting the Naive Estimators

Assume temporarily that for internal reliability data $k_i = k$ for $i = 1, \ldots, n_3$ and suppose that $\hat{\eta}$ and $\hat{\gamma}$ are consistent for η and γ. Then the naive estimator, $\hat{\Theta}(0)$, converges in probability to Θ_N satisfying

$$0 = \sum_{j=1}^{3} p_j \Psi_T(Y_{j1}, X_{Cj1}, \Theta_N),$$

where $(Y_{j1}, X_{Cj1}) = (Y_{j1}, \tilde{X}_{j1}(\eta, 0))$ $(j = 1, 2, 3)$. Let $(Y_{j1}, X_{j1}) = (Y_{j1}, \tilde{X}_{j1}(\eta, \delta))$ $(j = 1, 2, 3)$. In the Appendix we show that under both (1.2) and (1.3) there exist functions $d_{1j}(y, x, \eta, \gamma)$ and $d_{2j}(y, x, \eta, \gamma)$ such that

$$\mathrm{E}(X_{Cj1} - X_{j1} \mid Y_{j1}, X_{j1}) = \delta^2 d_{1j}(Y_{j1}, X_{j1}, \eta, \gamma) + O_P(\delta^3)$$

and

$$\begin{aligned}\mathrm{cov}(X_{Cj1} - X_{j1} \mid Y_{j1}, X_{j1}) = \delta^2 d_{2j}&(Y_{j1}, X_{j1}, \eta, \gamma)\\ &+ O_P(\delta^3).\end{aligned}$$

Note that d_{12} and d_{22} are identically zero under both (1.2) and (1.1). In addition, d_{13} and d_{23} are defined only under (1.2) and in this case $d_{13} = d_{11}/k$ and $d_{23} = d_{21}/k$.

An adaptation of the Taylor series argument in Stefanski (1985) shows that $\Theta_N = \Theta - (\delta^2/2)Q^{-1}H + O(\delta^3)$, where

$$Q = \mathrm{E}\left\{\sum_{j=1}^{3} p_j \Psi_{T\Theta}(Y_{j1}, X_{Cj1}, \Theta)\right\}$$

and

$$\begin{aligned}H = \mathrm{E}\Bigg(\sum_{j=1}^{3} p_j[2\Psi_{Tx}(Y_{j1}, X_{j1}, \Theta)\, d_{1j}\,(Y_{j1}, X_{j1}, \eta, \gamma)\\ + \mathrm{tr}(\Psi_{Txx}(Y_{j1}, X_{j1}, \Theta)\, d_{2j}\,(Y_{j1}, X_{j1}, \eta, \gamma)\}]\Bigg).\end{aligned}$$

Let $\kappa_{i1} = 1$, $\kappa_{i2} = 0$, and $\kappa_{i3} = k_i^{-1}$, and define

$$\begin{aligned}h_{ij}(\Theta, \Lambda) &= \kappa_{ij}(2\Psi_{Tx}\{Y_{ji}, \tilde{X}_{ji}(\eta, 0), \Theta\}\\ &\quad \times d_{11}\{Y_{ji}, \tilde{X}_{ji}(\eta, 0), \eta, \gamma\}\\ &\quad + \mathrm{tr}[\Psi_{Txx}\{Y_{ji}, \tilde{X}_{ji}(\eta, 0), \Theta\}\\ &\quad \times d_{21}\{Y_{ji}, \tilde{X}_{ji}(\eta, 0), \Theta\}]),\end{aligned}$$

$$q_{ij}(\Theta, \Lambda) = \Psi_{T\Theta}\{Y_{ji}, \tilde{X}_{ji}(\eta, 0), \Theta\},$$

$$Q_n(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} q_{ij}(\Theta, \Lambda),$$

and

$$H_n(\Theta, \Lambda) = n^{-1} \sum_{j=1}^{3} \sum_{i=1}^{n_j} h_{ij}(\Theta, \Lambda).$$

Then $\Theta_{c,2}$, defined as

$$\hat{\Theta}_{c,2} = \hat{\Theta}(0) + (\hat{\delta}^2/2)Q_n(\hat{\Theta}(0), \hat{\Lambda})^{-1}H_n(\hat{\Theta}(0), \hat{\Lambda}),$$

is the estimator proposed in Stefanski (1985), adapted to our model.

The estimators $\hat{\Theta}_{c,1}$ and $\hat{\Theta}_{c,2}$ can both be viewed as corrections to the naive estimator, but doing so obscures their fundamental difference. The former is an *approximation* to the quasi-likelihood estimator, whereas $\hat{\Theta}_{c,2}$ is a true correction for bias in the sense that it is obtained by subtracting an estimator of an approximation to the asymptotic bias in $\hat{\Theta}(0)$.

5.3 Asymptotic Distributions

We now derive the asymptotic joint distribution of $\hat{\Theta}(0)$, $\hat{\Lambda}$, and $\hat{\Theta}_c$, where $\hat{\Theta}_c = \hat{\Theta}_{c,1}$ or $\hat{\Theta}_{c,2}$. This is accomplished by representing $(\hat{\Theta}(0)^t, \hat{\Lambda}^t, \tilde{\Theta}_c^t)^t$ as an M estimator and appealing to standard asymptotic theory.

Let

$$\begin{aligned} \Psi_{Tji}(\Theta_1, \Lambda) &= \Psi_T(Y_{ji}, \tilde{X}_{ji}(\eta, 0), \Theta_1), && j = 1, 2, 3, \\ &= 0_{\dim(\Theta)\times 1}, && j = 4, 5; \\ \Upsilon_{ji}(\Lambda) &= 0_{\dim(\Lambda)\times 1}, && j = 1, \\ &= \Psi_{IV,\Lambda}(X_{ji}, W_{ji}, \Lambda), && j = 2, 4, \\ &= \Psi_{IR,\Lambda}(W_{ji1}, \ldots, W_{jik_i}, \Lambda), && j = 3, 5; \end{aligned}$$

and

$$\begin{aligned} \Delta_{ji}(\Theta_1, \Lambda, \Theta_2) &= q_{ij}(\Theta_1, \Lambda)(\Theta_2 - \Theta_1) \\ &\quad - (\delta^2/2)h_{ij}(\Theta_1, \Lambda), && j = 1, 2, 3, \\ &= 0_{\dim(\Theta)\times 1}, && j = 4, 5. \end{aligned}$$

Define $C_{ji}(\Theta_1, \Lambda, \Theta_2) = (\Psi_{Tji}^t, \Upsilon_{ji}^t, \Delta_{ji}^t)^t$. Then, $\hat{\Theta}(0)$, $\hat{\Lambda}$, and $\hat{\Theta}_c$ satisfy

$$0 = \sum_{j=1}^{5} \sum_{i=1}^{n_j} C_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c).$$

The M estimators $\hat{\Theta}(0)$, $\hat{\Lambda}$, and $\hat{\Theta}_c$ converge in probability to Θ_*, Λ_*, and Θ_{c*}, respectively, where

$$0 = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} \mathrm{E}\{C_{ji}(\Theta_*, \Lambda_*, \Theta_{c*})\}.$$

Let $D_{ji}(\Theta_1, \Lambda, \Theta_2)$ be the matrix of partial derivatives of C_{ji} with respect to $(\Theta_1^t, \Lambda^t, \Theta_2^t)$, and define

$$A = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\{D_{ji}(\Theta_*, \Lambda_*, \Theta_{c*})\}$$

and

$$B = \lim_{n\to\infty} n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} E\{C_{ji}(\Theta_*, \Lambda_*, \Theta_{c*}) \times C_{ji}(\Theta_*, \Lambda_*, \Theta_{c*})^t\}.$$

Moment estimators of these matrices are given by

$$\hat{A} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} D_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c)$$

and

$$\hat{B} = n^{-1} \sum_{j=1}^{5} \sum_{i=1}^{n_j} C_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c) C_{ji}(\hat{\Theta}(0), \hat{\Lambda}, \hat{\Theta}_c)^t.$$

Standard asymptotic results imply that, under sufficient regularity conditions,

$$n^{1/2}\{(\hat{\Theta}(0) - \Theta_*)^t, (\hat{\Lambda} - \Lambda_*)^t, (\hat{\Theta}_c - \Theta_{c*})^t\}^t \xrightarrow{\mathrm{D}} \mathfrak{N}\{0, A^{-1}B(A^t)^{-1}\}$$

and

$$\hat{A}^{-1}\hat{B}(\hat{A}^t)^{-1} \xrightarrow{\mathrm{P}} A^{-1}B(A^t)^{-1}.$$

6. EXAMPLES

We now present two examples, one with a Berkson error structure and the second with a measurement error structure.

Lung Function (Berkson model). For this example we use a subset of the data analyzed by Tosteson et al. (1989). As a means of studying the relationship between respiratory health and exposure to nitrogen dioxide in school-age children, these authors fit a probit regression model of Y = indicator of the presence of wheeze, on X = log(personal exposure to nitrogen dioxide), using a primary data set containing 231 observations on (Y, W), where W = (log(bedroom exposure), log(kitchen exposure))^t and two external validation data sets containing 81 and 564 observations, respectively, on (X, W).

Tosteson et al. elected to model personal exposure as a function of bedroom and kitchen exposure, thereby creating a Berkson error model. The log transformations were used to induce linearity and homoscedasticity in the regression of X on W. These authors fit the model $X = \eta_0 + \eta_1^t \mathrm{W} + \delta U$; $\mathrm{E}(U \mid W) = 0$; $\mathrm{var}(U \mid W) = 1$. In our notation, $g_{ss} = h_2 = 0$ and $g_s = h_3 = 1$. They assumed a probit binary model with U standard normal, thereby obtaining a probit model for Y on W. Standard errors reported by Tosteson et al. were computed without making allowance for the fact that parameters in the prediction equation for X had been estimated.

Fitting the model using the smaller of the two validation data sets (the Portage data), we find that $\hat{\delta} = .265$, $\hat{\eta}_0 = 1.28$, and $\hat{\eta}_1^t = (.28, .33)$. Let $W_\eta = \eta_0 + \eta_1^t W$. For probit regression, (3.9) becomes, with no assumptions on the distribution of U,

$$\begin{aligned} \Pr(Y = 1 \mid W_\eta) &\approx \Phi\{(\beta_0 + \beta_1 W)(1 - \delta^2\beta_1^2/2)\} \\ &\approx \Phi\{(\beta_0 + \beta_1 W_\eta)/(1 + \delta^2\beta_1^2)^{1/2}\}. \quad (6.1) \end{aligned}$$

The second approximation follows easily from the first and is exact when U is normally distributed. The observed regression slope estimate for the probit model of Y on W_η is .05. This estimate can be related to β_1 by (6.1), from which it is intuitively obvious that there is essentially no effect due to estimating Λ. This intuition is confirmed by an application of (4.2).

Diet and Breast Cancer (measurement error model). For illustration, we consider a sample of 2,888 women under the age of 50, the data being a subset of those used by Jones et al. (1987). The response Y is an indicator of breast cancer, and the predictor X is the logarithm of long-term average daily saturated fat intake. There were 37 cases of breast cancer in the study. Long-term average daily saturated fat intake is unobservable, and as the "observed" predictor we have W, the logarithm of a 24-hour diet-recall proxy for daily average saturated fat intake; see Jones et al. for elaboration on the study design and data collection. We model on the log scale under the assumption that on this scale the errors, that is, the differences $W - X$, are approximately normal and uncorrelated with X.

Dietary measures exhibit great within-person variability, and epidemiologists are concerned with the effects of such large measurement errors on standard statistical analyses. In this example we fit a logistic regression measurement error model for the purpose of illustrating the effect of measurement error on a logistic analysis of these data.

In the study, W had mean 2.98 and between-person standard deviation .635. A logistic regression fit to the (Y, W) data yielded an estimated slope of $-.40$, with estimated standard error .24 and p value .08. We fit a measurement error model, assuming the approximate model $W = X + \delta U$, where U given X has mean 0 and variance 1.

We do not have access to validation/reliability data for this study, but for illustrative purposes we use the validation results reported by Willett et al. (1985). They did not transform fat intake, and they used four seven-day diet record measurements, finding that the correlation between any two weekly measurements is approximately .55. If we assume normally distributed measurement error on the log scale, then after allowing for the log transformation their data suggest that, for their study, $\delta \approx .34$. Their seven-day diet records differ from our 24-hour recall measurements and should be more precise. It is not clear how to make the conversion from seven-day diet records to 24-hour recall measurements, but as a reasonable guess to illustrate our methods, we use $\delta \approx .53$; that is, we assume that seven-day diet records are about 2.5 times less variable than 24-hour recall.

Using this value for δ, the method of Section 5.2 yields an estimated slope of $-.62$. When we regard δ as known, the slope estimate has an estimated standard error of .24 and a p value of .01. To assess the effect of uncertainty in δ under the constraint that we do not have actual external reliability data, we make the assumption that U is normally distributed, so that with the sample size of 150 in the paper by Willett et al. (1985), the variance of $\hat{\delta}^2$ should be approximately $2\delta^2/150 \approx (.033)^2$. Using the theory of Section 5, we find that the adjusted standard error for the slope is .23, an insignificant change.

Because we expected the standard error of the corrected estimate to be larger than that of the naive estimate, we also performed a small bootstrap simulation. We resampled the (Y, W) pairs to form bootstrap samples and repeated the experiment 200 times. The usual analysis had bootstrap mean $-.39$ and standard error .16, and the corrected analysis had bootstrap mean $-.60$ and standard error .23. The results of the bootstrap study are displayed as kernel density estimates in Figure 1. A Gaussian kernel with bandwidth .20 was employed.

The asymptotic theory and the bootstrap analysis are remarkably similar for the corrected estimate, but for the usual analysis the estimated standard error seems to be a bit too high. A possible explanation for this finding is that the number of cases $(Y_i = 1)$ is small. Letting V_i be the vector containing a constant 1.0 and the observed diet measurement, the usual logistic regression analysis gives the estimated asymptotic covariance matrix $\{\Sigma_1^n V_i V_i' F^{(1)}(V_i'\hat{\beta})\}^{-1} = Q_n^{-1}$ [see (6.3)], whereas the theory of M estimation yields the estimated covariance $Q_n^{-1}[\Sigma_1^n V_i V_i' \{Y_i - F(V_i'\hat{\beta})\}^2] Q_n^{-1}$. Let $\hat{F}_i = F(V_i'\hat{\beta})$. Note that when $\hat{\beta} = 0$, $\hat{F}_i = \frac{1}{2}$, and the two estimated covariance matrices are equal. Thus they will be approximately equal whenever $\hat{\beta}$ is small. When $\hat{F}_i$ are not all near $\frac{1}{2}$, however, the two covariance matrices can differ substantially. In our example for most observations, $Y_i = 0$ and $\hat{F}_i \approx 0$. It follows that for most observations $(Y_i - \hat{F}_i)^2 = \hat{F}_i^2 \ll \hat{F}_i \approx \hat{F}_i(1 - \hat{F}_i)$. Thus one might expect that the ordinary logistic standard errors may be a bit too large.

Because of the imprecision in relating 24-hour recall measures with seven-day diet records, as well as the fact that we have not included other predictors, we wish to emphasize that the preceding analysis was illustrative only.

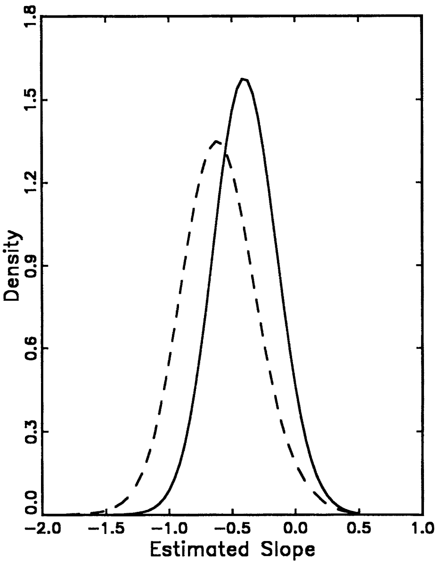


Figure 1. The results of the bootstrap study in Section 6 are displayed as kernel density estimates. The solid line represents the values obtained for the usual logistic regression estimate, and the dashed line represents the values for the corrected estimate.

7. CONCLUSION

The class of quasi-likelihood/variance function models (1.1) is broad and of recognized importance in statistical practice. We have examined some general methods of constructing parameter estimates when the predictors in (1.1) are measured with error. A major part of this paper, Sections 4 and 5, developed a comprehensive asymptotic theory for estimates derived using these methods. The theory provides usable standard error estimates and allows for the presence of validation and/or replication data.

In Sections 2 and 3, we developed range-preserving models for the observed data based upon (1.1). One such class of methods, (3.5), was shown to be correct to order $O(\delta^3)$. A second set of models, which replace X by an estimate of $E(X \mid W)$, is given by (3.3). We showed that these are correct only to order $O(\delta^2)$, although in (3.4) we note that in many applications the difference between (3.3) and (3.5) will be negligible. Taken together, the range-preserving model classes (3.3) and (3.5) include as special cases most of the suggestions made previously in the literature.

One can thus summarize the article as having (a) developed broad classes of models and estimators; (b) taken explicit account of the types of data sets, including validation and replication data, that are likely to be available in a measurement error model; and (c) provided the asymptotic distribution theory for parameter estimates, including formulas for obtaining consistent standard errors. We view these combined results as a necessary first step toward addressing the question of which method works best in practice. One advantage of the range-preserving models (3.3) and (3.5) is that, being models for the observed data, they can be checked for fit and compared with one another, as is done in a specific case by Rudemo et al. (1989). In our first example, we found that the two models were essentially the same for the likely values of the parameters. In other cases, the two models may differ. If forced to choose between the two in the absence of an applied context, we would recommend (3.5) over (3.3). The former is based on higher-order expansions in terms of δ, and it is more flexible with respect to modeling, as it can accommodate heterogeneous variability in the relationship between X and W. As seen in the discussion of logistic regression just following (3.4), such heterogeneity can be important in model fitting.

APPENDIX: TECHNICAL COMPLEMENTS

The following lemma is used in Example 3.4 and in the derivations of the functions d_{1j} and d_{2j} $(j = 1, 2, 3)$ defined in Section 5.2.

Lemma A.1. Suppose that V_1, V_2, and V_3 are random vectors such that $V_1 = V_2 + \delta V_3$, where $\delta > 0$ is a constant scalar. If V_1 and V_2 have joint density $f_{V_1V_2}(v_1, v_2)$, $E(V_3 \mid V_2) = 0$, $\operatorname{cov}(V_3 \mid V_2 = v_2) = \Omega(v_2)$, and $E(V_3 \mid V_1)$ and $\operatorname{cov}(V_3 \mid V_1)$ are three-times differentiable functions of δ a.s., then

$$\begin{aligned} E(V_3 \mid V_1) &= -\delta_1\left[\operatorname{tr}_1\left\{\frac{\partial\Omega(v_1)}{\partial v_1}\right\} + \Omega(v_1)\frac{\partial \log\{f_{V_2}(v_1)\}}{\partial v_1}\right]_{v_1=V_1} \\ &\quad + O_P(\delta^2) \\ &= -\delta_1\left[\operatorname{tr}_1\left\{\frac{\partial\Omega(v_1)}{\partial v_1}\right\} + \Omega(v_1)\frac{\partial \log\{f_{V_1}(v_1)\}}{\partial v_1}\right]_{v_1=V_1} \\ &\quad + \dot{O}_P(\delta^2) \end{aligned}$$

and $\operatorname{cov}(V_3 \mid V_1) = \Omega(V_1) + O_P(\delta)$.

Proof. Under the assumptions of the lemma, the assertion about $\operatorname{cov}(V_3 \mid V_1)$ is obvious and the second expression for $E(V_3 \mid V_1)$ follows easily from the first. Thus it only remains to identify the terms in the Maclaurin series expansion of $E(V_3 \mid V_1)$. Note that

$$E(V_3 \mid V_1) = \frac{\int v_3 f_{V_3|V_2}(v_3 \mid V_1 - \delta v_3) f_{V_2}(V_1 - \delta v_3)\, dv_3}{\int f_{V_2V_3}(V_1 - \delta v_3, v_3)\, dv_3}. \qquad \text{(A.1)}$$

The denominator in (A.1) equals $f_{V_2}(V_1) + O_P(\delta)$. After a Taylor series expansion, the numerator in (A.1) is shown to be

$$\begin{aligned} &-\delta\left[\int\!\!\int v_3v_3^t\left\{\frac{\partial f_{V_3|V_2}(v_3 \mid v_1)}{\partial v_1}\right\} dv_3 f_{V_2}(v_1) \right. \\ &\quad \left. + \int f_{V_3|V_2}(v_3 \mid v_1)v_3v_3^t\, dv_3\left\{\frac{\partial f_{V_2}(v_1)}{\partial v_1}\right\}\right]_{v_1=V_1} + O_P(\delta^2) = \\ &\quad -\delta\left[\operatorname{tr}\left\{\frac{\partial\Omega(v_1)}{\partial v_1}\right\} f_{V_2}(v_1) + \Omega(v_1)\frac{\partial f_{V_2}(v_1)}{\partial v_1}\right]_{v_1=V_1} + O_P(\delta^2), \end{aligned}$$

completing the proof.

We now derive the functions d_{1j} and d_{2j} $(j = 1, 2, 3)$ defined in Section 5.2. Note that X_{Cj1} is function of W_{j1} and X_{j1}, $(j = 1, 2, 3)$ and thus the conditional independence assumption of Section 1.1 implies that d_{1j} and d_{2j} are functions of X_{j1} a.s. for $j = 1, 2, 3$.

Consider first the generalized measurement error model (1.2). Note that $g\{c(X_{11}, \eta), \eta, 0\} = X_{11}$ a.s. A Taylor series expansion shows that

$$\begin{aligned} X_{C11} &= g(W_{11}, \eta, 0) \\ &= X_{11} + \delta g_w\{c(X_{11}, \eta), \eta, 0\}U \\ &\quad + (\delta^2/2)\operatorname{tr}[g_{ww}\{c(X_{11}, \eta), \eta, 0\}U_{11}U_{11}^t] + O_P(\delta^3). \end{aligned}$$

Thus under (1.2),

$$d_{11}(y_{j1}, x_{j1}, \eta, \gamma) = \operatorname{tr}[g_{ww}\{c(x_{j1}, \eta), \eta, 0\}\Omega(x_{j1}, \eta, \gamma)]$$

and

$$\begin{aligned} d_{21}(y_{j1}, x_{j1}, \eta, \gamma) &= g_w\{c(x_{j1}, \eta), \eta, 0\} \\ &\quad \times \Omega(x_{j1}, \eta, \gamma)g_w^t\{c(x_{j1}, \eta), \eta, 0\}. \end{aligned}$$

Both d_{12} and d_{22} are identically 0, and $d_{13} = d_{11}/k$ and $d_{23} = d_{21}/k$, where k is the number of replicates.

Now consider the generalized Berkson error model (1.3). Let $f_{c^*}(c^*)$ be the density of $C^*(W, \eta)$, assumed to exist. Conditioning first on $\{C^*(W, \eta), W\}$ and then on $\{C^*(W, \eta)\}$ shows that

$$E\{U \mid C^*(W, \eta)\} = 0$$

and

$$\operatorname{cov}\{U \mid C^*(W, \eta)\} = \Omega^{**}(W, \eta, \gamma),$$

where $\Omega^{**}(W, \eta, \gamma) = E\{\Omega^*(W, \eta, \gamma) \mid C^*(W, \eta)\}$. Frequently, $\Omega^{**}(W, \eta, \gamma)$ is a function of $C^*(W, \eta)$ and then $\Omega^{**}(W, \eta, \gamma) = \Omega^*(W, \eta, \gamma)$. However, this is not always the case.

It follows directly from Lemma A.1 that

$$d_{11}(y_{j1}, x_{j1}, \eta, \gamma) = \operatorname{tr}\left\{\frac{\partial\Omega^{**}(x_{j1}, \eta, \gamma)}{\partial x_{j1}}\right\} + \Omega^{**}(x_{j1}, \eta, \gamma)\,\frac{\partial \log\{f_{c^*}(x_{j1}, \eta)\}}{\partial x_{j1}}$$

and

$$d_{21}(y_{j1}, x_{j1}, \eta\ \gamma) = \Omega^{**}(x_{j1}, \eta, \gamma).$$

Both d_{12} and d_{22} are identically 0, and d_{13} and d_{23} are not defined for the generalized Berkson error model.

[*Received August 1988. Revised February 1990.*]

REFERENCES

Amemiya, Y., and Fuller, W. A. (1988), "Estimation for the Nonlinear Functional Relationship," *The Annals of Statistics,* 16, 147–160.

Armstrong, B. (1985), "Measurement Error in the Generalized Linear Model," *Communications in Statistics, Part B—Simulation and Computation,* 14, 529–544.

Carroll, R. J. (1989). "Covariance Analysis in Generalized Linear Measurement Error Models," *Statistics in Medicine,* 8, 1075–1093.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression,* London: Chapman & Hall.

Davidian, M., and Carroll, R. J. (1987), "Variance Function Estimation," *Journal of the American Statistical Association,* 82, 1079–1091.

Davidian, M., Carroll, R. J., and Smith, W. (1988), "Variance Functions and the Minimum Detectable Concentration in Assays," *Biometrika,* 75, 549–556.

Fuller, W. A. (1987), *Measurement Error Models,* New York: John Wiley.

Gleser, L. J. (1989), "Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-Variables Regression Models," Preprint.

Jones, D. Y., Schatzkin, A., Green, S. B., Block, G., Brinton, L. A., Ziegler, R. G., Hoover, R., and Taylor, P. R. (1987), "Dietary Fat and Breast Cancer in the National Health and Nutrition Survey I: Epidemiologic Follow-up Study," *Journal of the National Cancer Institute,* 79, 465–471.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models,* London: Chapman & Hall.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error," *Statistics in Medicine,* 8, 1051–1070.

Rudemo, M., Ruppert, D., and Streibig, J. C. (1989), "Random Effect Models in Nonlinear Regression With Applications to Bioassay," *Biometrics,* 45, 349–362.

Schafer, D. (1987), "Covariate Measurement Error in Generalized Linear Models," *Biometrika,* 74, 385–391.

Stefanski, L. A. (1985), "The Effects of Measurement Error on Parameter Estimation," *Biometrika,* 72, 583–592.

Stefanski, L. A., and Carroll, R. J. (1985), "Covariate Measurement Error in Logistic Regression," *The Annals of Statistics,* 13, 1335–1351.

——— (1990), "Score Tests in Generalized Linear Measurement Error Models," *Journal of the Royal Statistical Society,* Ser. B, 52, 345–359.

Tosteson, T., Stefanski, L. A., and Schafer, D. W. (1989), "A Measurement Error Model for Binary and Ordinal Regression," *Statistics in Medicine,* 8, 1139–1147.

Tosteson, T., and Tsiatis, A. (1988), "The Asymptotic Relative Efficiency of Score Tests in a Generalized Linear Model With Surrogate Covariates," *Biometrika,* 75, 507–514.

Whittemore, A. S., and Keller, J. B. (1988), "Approximations for Regression With Covariate Measurement Error," *Journal of the American Statistical Association,* 83, 1057–1066.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H., and Speizer, F. E. (1985), "Reproducibility and Validity of a Semiquantitative Food Frequency Questionnaire," *American Journal of Epidemiology,* 122, 51–65.

Case-Control Studies With Errors in Covariates

R. J. CARROLL, M. H. GAIL, and J. H. LUBIN*

We devise methods for estimating the parameters of a prospective logistic model with dichotomous response D and arbitrary covariates X from case-control data when these covariates are measured with error. We suppose that some fraction of the cases and controls provide only the error-prone covariate measurements, W (the "incomplete" or "reduced" data), whereas some of the cases and controls provide measurements on X and W (the "complete" data). We assume a measurement error density with a finite set of parameters α, namely $f_{W|XD}(w|x, d, \alpha)$, and nondifferential error is treated as a special case of this model, $f_{W|X}(w|x, \alpha)$. Our algorithm estimates both the logistic parameters and α from a pseudolikelihood. Because empirical distribution functions are used in place of needed distributions in the pseudolikelihoods, the required asymptotic theory is more elaborate than for pseudolikelihoods based on substitution for a finite number of nuisance parameters. We also examine computationally simpler methods under the assumptions that the disease is rare and that errors are nondifferential. Estimates of $m(W) = E(X|W)$ are substituted for X in the logistic model when X is not available. Such estimates of $m(W)$ can be obtained from the complete data described above or from an independent validation study. If measurements on X are not available, $m(W)$ can still be estimated from replicated W measurements in some circumstances. A final approach uses approximate logistic regression techniques and is appropriate when a more accurate approximation is required than obtained by simply substituting $m(W)$ for X. Asymptotic theory is presented for each of these procedures, and examples are used to illustrate the calculations.

KEY WORDS: Asymptotics; Case-control study; Differential misclassification; Errors in variables; Logistic regression; Pseudolikelihood.

1. INTRODUCTION

Since the work of Cornfield (1951), it has been appreciated that the prospective odds ratio of disease $\{P(D = 1|X)/P(D = 0|X)\}\{P(D = 1|X_0)/P(D = 0|X_0)\}$ could be estimated from case-control data as the odds of exposure, $\{P(X|D = 1)/P(X_0|D = 1)\}/\{P(X|D = 0)/P(X_0|D = 0)\}$, where X_0 is a baseline or reference level of exposure and other covariates and where $D = 1$ or 0 corresponds to the presence or absence of disease. Although one can use retrospective logistic models (Prentice 1976) to describe covariate outcomes, conditional on disease status, all but the simplest such models become unwieldy and only equivalent to a prospective logistic model if the retrospective model is saturated (Breslow and Powers 1978). For these reasons, and because it is more intuitive to think of covariates as causing disease than to think of disease as altering the distribution of exposures, it is common practice to fit a prospective logistic model to case-control data, namely

$$P(D = 1|X = x) \equiv \mathcal{H}_L(\beta_0^* + \beta^t x), \qquad (1)$$

where $\mathcal{H}_L(v) = \{1 + \exp(-v)\}^{-1}$. In model (1), $\beta_0^* = \log\{P(D = 1|X = x_0)/P(D = 0|X = x_0)\} - \beta^t x_0$. As discussed by Mantel (1973) and Farewell (1979), if model (1) describes the risk of disease in the source population, then, in the population constructed by case-control sampling, it will also describe the conditional probability of disease given covariates, except that β_0^* is replaced by $\beta_0 = \beta_0^* + \log(\pi_1/\pi_0)$ where π_1 is the probability that a case ($D = 1$) will be selected from the source population for the case-control sample and π_0 is the probability that a control will be selected. Remarkably, standard logistic regression, performed as if D were the dependent variable and the covariates X were fixed, leads to maximum likelihood estimates of β_0 and β^t for case-control sampling, whether X is discrete (Anderson 1972) or contains continuous components (Prentice and Pyke 1979). In particular, as the latter authors show, in the composed population obtained by randomly sampling n_1 cases ($D = 1$) and n_0 controls ($D = 0$) from the source population, the prospective maximum likelihood estimators (MLE's) $\hat{\beta}_0$ and $\hat{\beta}$, together with the empirical distribution function, $\bar{F}(x)$, maximize the retrospective likelihood

$$\Pi_{i=1}^{n_1} P(X_i|D_i = 1)\Pi_{i=n_1+1}^{n_1+n_0} P(X_i|D_i = 0),$$

subject to $\hat{P}(D = 1) = n_1/(n_1 + n_0)$.

In this article we extend the use of the prospective logistic model to case-control data in which the covariates X are measured with error. In most of the article, we assume that the true covariates X and error-prone measurements W are available in a validation study consisting of a random sample of n_{C1} cases and n_{C0} controls. These "complete" data are represented for $D = d$ by $\{(X_{Cid}, W_{Cid}), d = 0, 1, \text{and } i = 1, 2, \ldots, n_{Cd}\}$. In addition, we have incomplete or "reduced" case-control data, $\{W_{Rid}, d = 0, 1, \text{and } i = 1, 2, \ldots, n_{Rd}\}$, for $D = d$ obtained by sampling n_{R1} cases and n_{R0} controls from the source population. We define $n_C = n_{C0} + n_{C1}$ and $n_R = n_{R0} + n_{R1}$. It is assumed that selection as complete or reduced data is completely at random, so that the information on X is missing at random (Little and Rubin 1987). Disease status is assumed to be known without error.

The first approach we take is based on pseudolikelihoods. We express the retrospective likelihood for the case-control data in terms of the prospective logistic model (1), a parametric measurement error model $f_{W|X,D}(w|x, d, \alpha)$ with parameter α, and $F_1(X)$ and $F_0(X)$, the distributions of X

* R. J. Carroll is Professor of Statistics and Toxicology, Department of Statistics, Texas A & M University, College Station, TX 77843-3143. M. H. Gail is Head, Epidemiologic Methods Section and J. H. Lubin is Senior Health Statistician, National Cancer Institute, Bethesda, MD 20892. R. J. Carroll's research was supported by grants from the National Institute of General Medical Sciences and the National Cancer Institute and was partially completed during a visit to the National Cancer Institute. The authors thank A. Hildesheim and colleagues for supplying the data for Section 2; two referees, who provided detailed and very useful comments above and beyond the call of duty; and David Pee of Information Management Sciences, Inc. for his expert computational assistance.

among cases and controls. Pseudolikelihoods are obtained by inserting simple estimates of F_1 and F_0 or of F_1, F_0, and α, leading to pseudolikelihood estimates $\hat{\beta}$, together with estimates of its covariance. Nondifferential measurement error, $f_{W|X,D}(w|x, d, \alpha) = f_{W|X}(w|x, \alpha)$, is treated easily as a special case. We examine the special case of dichotomous X and W in Section 2 and compare pseudolikelihood estimates with maximum likelihood. In Section 3 we outline general theory for the pseudolikelihood approach. When X is discrete and takes on a finite number of values as in Section 2, the limit theory can be obtained by combining the Taylor series calculations of pseudolikelihood with case-control probability calculations (Appendix A), using techniques similar to those of Gong and Samaniego (1981). Otherwise, the limit theory is nonstandard.

Alternative approximate procedures are available if disease is rare and nondifferential error is assumed. Carroll, Spiegelman, Lan, Bailey, and Abbott (1984), Rosner, Willett, and Spiegelman (1989), Rosner, Spiegelman, and Willett (1990), Whittemore (1989), and Gleser (1990) have suggested or studied substituting $m(W) \equiv E(X|W)$, or an estimate of $m(W)$, for X in the analysis of cohort data if errors are nondifferential. Carroll and Stefanski (1990) provided a theoretical analysis. We investigate this idea for case-control studies with nondifferential error and rare diseases (Sec. 4.1). In Section 4.2 we consider this approximation for the case in which $m(W)$ is estimated from independent validation data. In Section 4.3 we consider the case that X cannot be obtained but instead replicate measures are available. Our results are related to those of Armstrong, Howe, and Whittemore (1989) and Buonaccorsi (1990) in the special case that, given $D = d$, $W = X + \varepsilon$, where ε is an independent normal error with covariance independent of d and X is normally distributed (Sec. 4.3). Buonaccorsi (1990) also considered differential measurement error and obtained an asymptotic theory for it. One drawback of the substitution approach in Section 4.1 is the assumption that the distribution of X given W can be adequately described by its conditional expectation. In Section 4.4 we consider briefly a more general approach.

In Section 5 we present simulations to compare general pseudolikelihood methods with the methods based on the rare disease and nondifferential error assumptions. In Section 6 we summarize the results.

The possibility of differential measurement error is greater in retrospective studies than in prospective studies; therefore, it is important to develop methods for retrospective studies that allow for differential measurement error. Differential error models have remained relatively unexplored in the literature pertaining to prospective studies. Our work includes a new pseudolikelihood algorithm that allows us to handle the differential error model in retrospective studies. These methods require complete (validation) data, however, and in some case-control studies of historical exposures it will not be possible to obtain such data.

2. DICHOTOMOUS X

To fix ideas and compare pseudolikelihood methods with maximum likelihood estimation, we consider the simple case of dichotomous X. Table 1 summarizes data on exposure to herpes simplex virus type 2 (HSV-2) measured by a refined western blot procedure (X) and by a less accurate western blot procedure (W), in women with invasive cervical cancer ($D = 1$) and in controls ($D = 0$). Most of the data are incomplete (or reduced), yet we are primarily interested in evaluating the relationship between D and X, which is only directly observable in the complete data. Note that the odds ratio 1.977 from the complete data exceeds the "crude" odds ratio 1.545 relating W with D in the incomplete data. There is a substantial amount of misclassification in this example, as indicated by low sensitivity and specificity (see Table 1). Moreover, the sensitivity seems higher for the cases than for the controls ($p = .049$ by Fisher's exact two-sided test). Thus there is evidence for differential measurement error. See Hildesheim et al. (1991) for a full description of these data.

Table 1. Case-Control Data

	D	X	W	Count
Complete data				
	1	0	0	13
	1	0	1	3
	1	1	0	5
	1	1	1	18
	0	0	0	33
	0	0	1	11
	0	1	0	16
	0	1	1	16
Incomplete data				
	1		0	318
	1		1	375
	0		0	701
	0		1	535

NOTE: Odds ratio (X:D) from complete data = 1.977. Odds ratio (W:D) from incomplete data = 1.545. Sensitivity = $\hat{P}(W = 1|X = 1, D = d) = .783$ and .500 for $d = 1, 0$. Specificity = $\hat{P}(W = 0|X = 0, D = d) = .812$ and .750 for $d = 1, 0$.

The simplest parameterization with which to analyze these data is the "retrospective model" $P(X, W|D) = P(X|D)P(W|X, D)$, which requires six parameters: $P(X = 1|D = 1)$, $P(X = 1|D = 0)$, $P(W = 1|X = 0, D = 0)$, $P(W = 1|X = 0, D = 1)$, $P(W = 1|X = 1, D = 0)$, and $P(W = 1|X = 1, D = 1)$. Dahm, Gail, Rosenberg, and Pee (1990) fitted this saturated retrospective model by directly maximizing the retrospective likelihood

$$\Pi_{dkl}\{P(X = l|D = d)P(W = k|X = l, D = d)\}^{V(d,k,l)}$$
$$\times \Pi_{dk}\left\{\sum_{l=0}^{1} P(X = l|D = d)P(W = k|X = l, D = d)\right\}^{Y(d,k)}, \qquad (2)$$

where $V(d, k, l)$ is the number of completely classified observations in cell $D = d$, $W = k$ and $X = l$, and $Y(d, k)$ is the number of incompletely classified observations with $D = d$ and $Y = k$. The assumption of nondifferential error reduces the number of required parameters to 4, because $P(W = 1|X = l, D = 0) = P(W = 1|X = l, D = 1)$ for $l = 0, 1$. As is clear from Table 1, one can regard the reduced data as a mixture, over X, of complete tables and use the EM algorithm and related methods in the literature on incomplete contingency tables and on misclassification in categorical data to maximize the retrospective likelihood (2), as reviewed by Espeland and Hui (1987), Ekholm and Palmgren (1987), and Chen (1989). The "matrix method" (Greenland 1988a; Greenland and Kleinbaum 1983) can also be used to estimate the parameters in (2). In this example, using the methods of Greenland (1988a), we found that the estimated variances of the estimated log odds ratio were about 22% and 69% larger than those of the MLE in the differential and nondifferential cases. Because model (2) is "saturated" with respect to the outcome space for X, maximum likelihood estimation under the retrospective parameterization will equal that under the prospective risk model (1).

Nonetheless, it is instructive to outline the maximization of (2) using the following prospective parameterization. As described in Section 1, $P_C(D = 1 \mid X = x) = \mathcal{H}_L(\beta_{0C} + \beta^t x)$ in the complete data, where $\beta_{0C} = \beta_0^* + \log(\pi_{1C}/\pi_{0C})$ and π_{1C} and π_{0C} are probabilities of selecting cases and controls for the complete case-control sample. In the incomplete data the same model applies for $P_R(D = 1 \mid X = x)$, except β_{0C} is replaced by $\beta_{0R} = \beta_0^* + \log(\pi_{1R}/\pi_{0R}) = \beta_{0C} + \log\{(\pi_{1R}/\pi_{0R})/(\pi_{1C}/\pi_{0C})\} = \beta_{0C} + \log(n_{R1}n_{C0}/n_{R0}n_{C1})$. The marginal distribution of X in the complete data is $Q_X^C(x) = \{n_{1C}F_1(x) + n_{0C}F_0(x)\}/n_C$, and in the reduced data the marginal distribution of X is $Q_X^R(x) = \{n_{1R}F_1(x) + n_{0R}F_0(x)\}/n_R$. The term $P(X = l \mid D = d)$ in (2) can be replaced by $q_X^R(l)P_R(D = d \mid X = l)/(n_{Rd}/n_R)$ for reduced data and by $q_X^C(l)P_C(D = d \mid X = l)/(n_{Cd}/n_C)$ for complete data. Here q_X^R and q_X^C are probability mass functions corresponding to Q_X^R and Q_X^C. Maximum likelihood estimates are obtained by maximizing (2) with these substitutions over β_{0C}, β^t, $q_X^C(X = 1)$, $q_X^R(X = 1)$, $P(W = 1 \mid X = l, D = d)$, and $d, l = 1, 2$ subject to $\hat{P}(D = 1) = n_{C1}/n_C$ in the complete data and $\hat{P}(D = 1) = n_{R1}/n_R$ in the reduced data. Note that these two constraints reduce the number of free parameters to 8-2 = 6 for the differential error model and to 6-2 = 4 for the nondifferential error model. Alternatively, one can express $q_X^R(X = 1)$ and $q_X^C(X = 1)$ in terms of the two parameters $P(X = 1 \mid D = 1)$ and $P(X = 1 \mid D = 0)$, the corresponding point masses from the distributions $F_1(x)$ and $F_0(x)$, before obtaining constrained maximum likelihood estimates. Rather than proceed in this way, however, we calculated MLE's using the simpler retrospective parameterization (Table 2).

The log odds ratio β from the complete data alone is estimated as $\log(\frac{23}{16}/\frac{32}{44}) = .681$ with standard deviation .400 (Table 2). Maximum likelihood estimation allowing for differential error yields $\hat{\beta} = .609$ (SD = .350); under the nondifferential error model, one obtains $\hat{\beta} = .958$ (SD = .237) (Table 2). The differences in these two estimates of β may be so large because the data appear to exhibit differential measurement error. As indicated previously, the sensitivity of the error-prone measurement, W, is higher among cases than among controls (see Table 1). A 2-degree-of-freedom likelihood ratio test of the hypothesis of nondifferential measurement error based on all the data yields a difference of deviances of 4.962 ($p = .073$). Compared to maximum likelihood using the complete data only, maximum likelihood based on both the complete and incomplete data improves the efficiency of estimates of β only by the factor $(.400/.350)^2 = 1.31$ in the presence of possible differential error. Under the assumption of nondifferential error, which is problematic for these case-control data, the efficiency improvement is much greater, namely $(.400/.237)^2 = 2.85$. These small improvements in efficiency obtained from using the entire data set instead of the complete data alone reflect the low sensitivity and specificity of the error-prone data (Table 1). In this example even large amounts of error-prone data add little information on the odds ratio of interest.

Table 2. Parameter Estimates (and Standard Errors) for the Data in Table 1.

Parameter	MLE	Complete	Pseudolikelihood
	Differential Error		
β	.609 (.350)	.681 (.400)	.622 (.355)
β_{0C}	−.980 (.187)		−.981 (.185)
Pr(W = 1 | X = 0, D = 0)	.311 (.055)	.250 (.065)	.317 (.057)
Pr(W = 1 | X = 0, D = 1)	.189 (.085)	.188 (.098)	.195 (.089)
Pr(W = 1 | X = 1, D = 0)	.578 (.067)	.500 (.088)	.577 (.067)
Pr(W = 1 | X = 1, D = 1)	.784 (.068)	.783 (.086)	.790 (.067)
Pr(X = 1 | D = 1)	.591 (.064)	.590 (.079)	.590 (.079)
Pr(X = 1 | D = 0)	.440 (.056)	.421 (.057)	.421 (.057)
	Nondifferential Error		
β	.958 (.237)	.681 (.400)	.959 (.226)
β_{0C}	−1.181 (.142)		−1.163 (.140)
Pr(W = 1 | X = 0)	.257 (.043)	.223 (.055)	.266 (.042)
Pr(W = 1 | X = 1)	.679 (.041)	.618 (.064)	.686 (.041)
Pr(X = 1 | D = 1)	.652 (.052)	.590 (.079)	.590 (.079)
Pr(X = 1 | D = 0)	.418 (.046)	.421 (.057)	.421 (.057)

NOTE: "Complete" refers to complete data only.

Note that the maximum likelihood estimates of $P(X = 1 \mid D = 1)$ and $P(X = 1 \mid D = 0)$ are not exactly equal to the empirical estimates of these quantities from the complete data only (Table 2), so that the complete data estimates of these quantities do not maximize the likelihood for the combined data. Thus the nice result in Prentice and Pyke (1979) for complete data that the empirical estimate of $Q_X^C(x)$ and the prospective logistic estimates of β_{0C} and β^t are maximum likelihood estimates does not generalize to the case of incomplete data, at least when $n_{1C}/n_C \neq n_{1R}/n_R$. This finding suggests that maximum likelihood estimates will be hard to compute exactly in the general case of continuous covariates X and leads to the following pseudolikelihood estimate.

The pseudolikelihood is obtained by reexpressing $P(X = l \mid D = d)$ in (2) in terms of the prospective risk models $P_R(D = d \mid X = l)$ and $P_C(D = d \mid X = l)$ as described previously and by substituting estimates $\hat{q}_X^R(l)$ and $\hat{q}_X^C(l)$ for q_X^R and q_X^C based only on the complete data. The estimates $\hat{q}_X^R(l)$ and $\hat{q}_X^C(l)$ are weighted averages of the empirical mass functions corresponding to the empirical distribution functions $\hat{F}_0(x)$ and $\hat{F}_1(x)$, which are obtained from noncases and cases in the complete data. The resulting pseudolikelihood is then maximized over β_{0C}, β, and the four misclassification parameters (two if nondifferential misclassification is assumed). Standard deviations were computed by combining standard Taylor series calculations with case-control theory (Appendix A). Pseudolikelihood yields results and precision very nearly equal to that of maximum likelihood (Table 2).

In Section 3 we outline the theory needed to apply pseudolikelihood methods to more complicated prospective risk models (1) with continuous and discrete covariates.

3. PSEUDOLIKELIHOOD WITH X OBSERVED IN A SUBSAMPLE

3.1 Notation

Let $n_C = n_{C0} + n_{C1}$ and $n_R = n_{R0} + n_{R1}$ be the sample sizes of the complete and incomplete case-control data sets. To apply the prospective model (1) to the complete data, we define $\theta = (\beta_{0C}, \beta^t)^t$ and the quantities

$$H_C(x, \theta) = \mathcal{H}_L(\beta_{0C} + x^t\beta) \text{ and } \dot{H}_C(x, \theta) = H_C(x, \theta)\{1 - H_C(x, \theta)\}.$$

For the incomplete or reduced data, we define H_R and $\dot{H}_R$ similarly, except that, as in Section 2, β_{0C} is replaced by $\beta_{0R} = \beta_{0C} + \log(n_{R1}n_{C0}/n_{R0}n_{C1})$. Let F_d be the marginal distribution function of X given $D = d$, and let $\hat{F}_d$ be the empirical distribution function of X in the complete data given $D = d$. Define $n_d = n_{Cd} + n_{Rd}$ and $n = n_0 + n_1$ and let $Q_X^C = \sum_{d=0}^1 (n_{Cd}/n_C)F_d$ and $Q_X^R = \sum_{d=0}^1 (n_{Rd}/n_R)F_d$ be the marginal distribution functions of X in the complete and incomplete case-control samples. Denote their density or mass functions by q_X^C and q_X^R and their empirical estimates by $\hat{Q}_X^C = \sum_{d=0}^1 (n_{Cd}/n_C)\hat{F}_d$ and $\hat{Q}_X^R = \sum_{d=0}^1 (n_{Rd}/n_R)\hat{F}_d$.

One can estimate the slope parameter β in (1) by regressing D on X in the complete case-control data, as in Prentice and Pyke (1979). We are interested in improving the precision of these estimates by also using the incomplete data. This will be accomplished by assuming that the conditional density/mass function of W given $(X, D = d)$, $f_{W|XD}(w|x, D = d, \alpha)$, depends on a finite set of parameters α.

By allowing $f_{W|XD}(w|x, D = d, \alpha)$ to depend on d, we explicitly allow for differential error; our results include nondifferential error as a special case. As an example, consider an analysis of covariance model for generating the differential error; that is, W given $X = x$, $D = d$ is normally distributed with mean $\eta_0 + \eta_1 d + \eta_2 x$ and variance σ^2. In this case $\alpha = (\eta_0, \eta_1, \eta_2, \sigma)^t$. In Section 2, for the case of differential error, α consists of the four parameters $P(W = 1 | X = l, D = d)$ for $l, d = 0, 1$, whereas for nondifferential error α consists of the two parameters $P(W = 1 | X = l)$, $l = 0, 1$.

3.2 Likelihoods

Cases ($d = 1$) and controls ($d = 0$) contribute factors

$$\begin{aligned} f_{XW|D}(x, w|D = d) &= f_{X|D}(x|D = d)f_{W|XD}(w|x, D = d, \alpha) \\ &= \frac{n_C}{n_{Cd}} q_X^C(x)H_C(x, \theta)^d\{1 - H_C(x, \theta)\}^{1-d} \\ &\quad \times f_{W|XD}(w|x, D = d, \alpha) \qquad (3) \end{aligned}$$

to the likelihood for the complete data. Corresponding factors for the incomplete data are

$$\begin{aligned} f_{W|D}(w|D = d) = \frac{n_R}{n_{Rd}} \int & H_R(x, \theta)^d\{1 - H_R(x, \theta)\}^{1-d} \\ & \times f_{W|XD}(w|x, D = d, \alpha)dQ_X^R(x). \qquad (4) \end{aligned}$$

If Q_X^R were known, then (4) would contribute information about θ and the incomplete data could be used to improve the estimation of θ. We will pursue this idea, but substitute $\hat{Q}_X^R$ for Q_X^R. This substitution leads to a pseudolikelihood factor for a single observation for the incomplete case-control data:

$$\begin{aligned} \hat{f}_{W|D}(w|D = d, \theta, \alpha) &= n_{Rd}^{-1} \sum_{k=0}^{1} \sum_{i=1}^{n_{Ck}} \frac{n_{Rk}}{n_{Ck}} H_R(X_{Cik}, \theta)^d\{1 - H_R(X_{Cik}, \theta)\}^{1-d} \\ &\quad \times f_{W|XD}(w|X_{Cik}, D = d, \alpha). \qquad (5) \end{aligned}$$

From (3) and (4), the combined likelihood and pseudolikelihood for the complete and incomplete data is proportional to

$$\begin{aligned} L(\theta, \alpha) = {} & \Pi_{d=0}^1 \Pi_{i=1}^{n_{Cd}} H_C^d(X_{Cid}, \theta)\{1 - H_C(X_{Cid}, \theta)\}^{1-d} \\ & \times \Pi_{d=0}^1 \Pi_{i=1}^{n_{Rd}} \hat{f}_{W|D}(W_{Rid} | D = d, \theta, \alpha) \\ & \times \Pi_{d=0}^1 \Pi_{i=1}^{n_{Cd}} f_{W|XD}(W_{Cid} | X_{Cid}, D = d, \alpha). \qquad (6) \end{aligned}$$

It is important to emphasize that (6) is not the likelihood of the data; instead, it is an estimated or pseudolikelihood.

Let $f_{W|XD}(w|x, d, \alpha)$ be the likelihood of W given (X, D) with unknown parameter α, and define $\Psi_d(x, w, \alpha) = (\partial/\partial\alpha)\log\{f_{W|XD}(w|x, D = d, \alpha)\}$.

3.3 Estimating Equations

The derivatives of the log of the terms in (6) that arise from complete data with respect to α and θ are

$$\mathcal{L}_{C1}(\alpha) = \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \Psi_d(X_{Cid}, W_{Cid}, \alpha)$$

and

$$\mathcal{L}_{C2}(\theta) = \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} (1, X_{Cid}^t)^t\{d - H_C(X_{Cid}, \theta)\}. \qquad (7)$$

Because $\mathcal{L}_{C1}(\alpha)$ is the likelihood estimating equation evaluated at α, $E\mathcal{L}_{C1}(\alpha) = 0$. Prentice and Pyke (1979), using the retrospective sampling distribution, showed that $E\mathcal{L}_{C2}(\theta) = 0$. The corresponding estimating equations of (θ, α) from the incomplete data are

$$\mathcal{L}_{R2}(\theta, \alpha, \hat{Q}_X^R) = \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} S_d(W_{Rid}, \theta, \alpha, \hat{Q}_X^R);$$

$$\mathcal{L}_{R1}(\theta, \alpha, \hat{Q}_X^R) = \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} S_{d*}(W_{Rid}, \theta, \alpha, \hat{Q}_X^R), \qquad (8)$$

where if

$$\begin{aligned} G_{d1}(w, x, \theta, \alpha) &= H_R^d(x, \theta)\{1 - H_R(x, \theta)\}^{1-d} f_{W|XD}(w|x, D = d, \alpha), \\ G_{d2}(w, x, \theta, \alpha) &= (2d - 1)\begin{pmatrix} 1 \\ x \end{pmatrix} \dot{H}_R(x, \theta)f_{W|XD}(w|x, D = d, \alpha), \\ G_{d3}(w, x, \theta, \alpha) &= H_R^d(x, \theta)\{1 - H_R(x, \theta)\}^{1-d} \\ &\quad \times \frac{\partial}{\partial\alpha} f_{W|XD}(w|x, D = d, \alpha), \end{aligned}$$

and

$$S_{dk}(w, \theta, \alpha, Q_X^R) = \int G_{dk}(w, x, \theta, \alpha)\, dQ_X^R(x),$$

then

$$S_d(w, \theta, \alpha, Q_X^R) = \frac{S_{d2}(w, \theta, \alpha, Q_X^R)}{S_{d1}(w, \theta, \alpha, Q_X^R)},$$

$$S_{d*}(w, \theta, \alpha, Q_X^R) = \frac{S_{d3}(w, \theta, \alpha, Q_X^R)}{s_{d1}(w, \theta, \alpha, Q_X^R)}.$$

By standard likelihood considerations, the estimating equation for α in the incomplete data is asymptotically unbiased; that is, $E\mathcal{L}_{R1}(\theta, \alpha, Q_X^R) = 0$. In Appendix A we show that the estimating equation for θ is also asymptotically unbiased; that is, $E\mathcal{L}_{R2}(\theta, \alpha, Q_X^R) = 0$.

In the notation of estimating equations, the pseudolikelihood algorithm solves $0 = \mathcal{L}_{C1}(\hat{\alpha}) + \mathcal{L}_{R1}(\hat{\theta}, \hat{\alpha}, \hat{Q}_X^R) = \mathcal{L}_{C2}(\hat{\theta}) + \mathcal{L}_{R2}(\hat{\theta}, \hat{\alpha}, \hat{Q}_X^R)$. We used the Newton–Raphson algorithm to solve these equations.

3.4 Limiting Distributions and Covariance Estimation

If X is a discrete random variable with a finite number of elements as in Section 2, then there are only a finite number of parameters to be estimated from the complete data. From (3) and (4), these are the parameters (θ, α) and the masses of X given $D = 0, 1$. One can maximize the pseudolikelihood (6), this being computationally convenient. Alternatively, one can combine (3) and (4) and maximize the likelihood, either slightly overparameterized or subject to the constraint $\int H_C(x, \theta)\, dQ_X^C(x) = n_{C1}/n_C$. The asymptotic normality and covariance matrix of $n^{1/2}(\hat{\theta} - \theta)$ is obtainable by combining Taylor series and case-control probability calculations (Appendix A). For continuous X the more complicated theory outlined in Appendix A is needed to take into account the fact than an estimated function, $\hat{Q}_X^C$, rather than an estimated finite set of parameters is used to obtain the pseudolikelihood (6). The resulting covariance estimates require extensive computations.

In some problems it may be possible that the effect of estimating Q_X^R can be ignored. This issue needs to be explored. As shown in Appendix A, this conjecture being true would result in far simpler covariance formulas. When X is categorical, Q_X^R consists of a finite number of parameters and standard calculations apply to this case (Appendix A).

Estimates of standard errors and confidence intervals also can be obtained from bootstrap sampling. In the complete data for $D = d$, obtain a bootstrap sample of size n_{Cd} by sampling with replacement from the set $\{X_{Cid}, W_{Cid}; d = 1, \ldots, n_{Cd}\}$. In the incomplete data for $D = d$, obtain a bootstrap sample of size n_{Rd} by sampling with replacement from the set $\{W_{Rid}; d = 1, \ldots, n_{Rd})\}$.

3.5 Nondifferential Error

In the case of nondifferential error, we have $f_{W|XD}(w|x, D = d, \alpha) = f_{W|X}(w|x, \alpha)$. Other than this minor change of notation, all of our previous results apply. Of course, in real data analysis there are important practical differences from the differential error case. The practical choice of when to use the nondifferential assumption will be critical in applications (Greenland 1988b).

4. RARE DISEASE APPROXIMATIONS WITH NONDIFFERENTIAL ERROR

In the previous sections the measurement error model concerned the distribution of W given X and D. In this section we will assume that disease is rare and that measurement error is nondifferential, and we work instead with distribution of X given W and D. In Section 4.1 we consider validation as described in Section 3, with a linear error model. Section 4.2, we assume an external validation gives information about W, but no complete data set is available. In Section 4.3 we consider the case that replicates of W are observable but X is not observable; in Section 4.4 we consider a method different from (9) that uses an approximation to the likelihood.

4.1 Internal Validation

Suppose that error is nondifferential and define $m(w) = E(X \mid W = w)$. Rosner et al. (1989, 1990), Whittemore (1989), Gleser (1989), Pierce, Stram, Vaeth, and Schafer (1992), and Carroll and Stefanski (1990) noted that for moderate effects, a reasonable approximation for cohort data is

$$\begin{aligned}\Pr(D = 1 \mid W = w) &= \int \Pr(D = 1 \mid X = x) f_{X|W}(x \mid w)\, dx \\ &\approx \mathcal{H}_L\{\beta_0 + \beta' m(w)\} \\ &= \Pr\{D = 1 \mid X = m(w)\}. \qquad (9)\end{aligned}$$

We will use (9) as if it were an equality. If $m(w)$ were known, one could perform a logistic regression with different intercepts and predictors $m(W_{Rid})$ and X_{Cid} in the incomplete and complete data. In essence we impute the value $m(W_{Rid})$ for the true but unobserved predictor.

In practice $m(w)$ is unknown. To estimate it for continuous X, we consider an analysis of covariance model for X given W, D; that is,

$$E(X \mid W, D) = \alpha_0 + \alpha_1 D + \alpha_2 W, \quad \text{so that}$$

$$X = \alpha_0 + \alpha_1 D + \alpha_2 W + V, \quad \text{where} \quad E(V \mid W, D) = 0. \qquad (10)$$

The linear model (10), while standard, is restrictive. We have chosen it to obtain easily computed standard errors. It is possible to use the techniques of Appendix B to obtain results for models more general than (10).

In general the term in D is needed in fitting (10), because nondifferential error does not imply that X and D are conditionally independent given W. However, if the disease is rare, then

$$\begin{aligned}m(W) &= \alpha_0 + \alpha_2 W + \alpha_1 \Pr(D = 1 \mid W) \\ &\approx \alpha_0 + \alpha_2 W. \qquad (11)\end{aligned}$$

Note here that X is a $(p \times 1)$ vector, as are α_0 and α_1, whereas W is $(q \times 1)$ and α_2 is $(p \times q)$. Substituting (11) into (9), we obtain $\Pr(D = 1 \mid W = w) \approx \mathcal{H}_L(\beta^{**} + \beta^t \alpha_2 w)$, where $\beta^{**} = \beta_0^* + \beta^t \alpha_0$. Note that β^{**} is a different intercept from β_0^* in the prospective model (1) applied to complete case-control data (D, X). Thus the intercepts for regressions of D on X and D on W in the retrospective data are different, even if $n_{C1}/n_{C0} = n_{R1}/n_{R0}$. Let $\hat{\alpha}_2$ be the ordinary least squares estimate of α_2 obtained by applying (10) to the complete data. Then the algorithm is to run a logistic regression with a common slope β, subject to the following:

- Complete data: intercept β_{0C}, predictors (X_{Cid})
- Incomplete data: intercept β_{0R}, predictors $(\hat{\alpha}_2 W_{Rid})$.

Computing the estimator requires nothing more than access to a standard logistic regression program, with a dummy variable for whether one is using complete or incomplete data. However, unlike in the Prentice and Pyke (1979) context, the estimated standard errors from such a logistic regression are *formally* inconsistent. The reason for this is the need to adjust the standard errors for estimation of α_2. In many instances, however, one can estimate α_2 much better than one can estimate β, in which case the estimated standard errors from the dummy variable logistic regression will be accurate enough for most purposes. A quick way to check this is to compare the logistic regression program's standard errors for β with the linear regression standard errors for α_2. When the effect of estimating α_2 cannot be ignored, asymptotic theory is given in Appendix B.

In some problems there are no complete cases, only complete controls. In these cases, the complete controls can be treated as an external validation data set (see Sec. 4.2).

4.2 External Validation and Unbiased Surrogates

In some instances, complete data will be unavailable. Approximation (11) still can be used if α_2 is estimated from an independent validation study; see Appendix B for details. This is the case discussed by Rosner et al. (1989), see also Spall (1989) for related theory.

In some instances, instead of observing X in the complete data, we observe only an unbiased surrogate $X_* = X + U$, where U is independent of W. In this case, one uses the complete data to estimate α_2, and then runs a logistic regression of all the data, complete and incomplete, using $\hat{\alpha}_2 W$ as the predictor. The theory required is only a minor modification of what appears in Appendix B.

4.3 Independent Replication

Even if gold standard measurements X are unavailable, approximate techniques can be developed as long as independent replicates are available, and W is unbiased for X. In this subsection we discuss one possible implementation of such techniques.

Our starting point is the approximate model (9), in which we need to ascertain $m(w)$. Suppose that, prospectively, $W = X + U$, where X and U are independent with means $(\mu_X, 0)$ and covariances (Σ_X, Σ_U). Define $\Lambda_1 = \Sigma_X(\Sigma_X + \Sigma_U)^{-1}$. As in Carroll and Stefanski (1990) and Gleser (1990), the best linear approximant to $m(W)$ is $(I - \Lambda_1)E(X) + \Lambda_1 W$, which suggests using $\Lambda_1 W$ as the predictor in a logistic regression.

With replication, estimating Λ_1 can be accomplished under two circumstances. First, one might entertain the normal discriminant model of Michalek and Tripathi (1980) and Armstrong et al. (1989) with normal errors, so that given D, the previous model holds for (W, X) with constant covariance matrices. Buonaccorsi (1990) also considered this model, but allowed X to be multivariate and did not mandate that surrogate W be unbiased for X; his work also included a comprehensive asymptotic theory, as well as allowing for differential measurement error. Alternatively, if the disease is sufficiently rare, the covariance matrices (Σ_X, Σ_U) are essentially the same as those computed from among the controls.

There are two possible sampling strategies appropriate for these cases: (1) that replication is done only among the controls or (2) that both cases and controls are replicated. In the latter case the covariances given $D = d$ do not depend on d. Our treatment will handle both cases. Represent the "incomplete" data by (D_{Rid}, W_{Rid}) for $d = 0$, 1 and $i = 1, \ldots, n_{Rd}$, with $n_R = n_{R0} + n_{R1}$, and represent the "complete" data by the replicated samples $(D_{Cid}, W_{Cid1}, \ldots, W_{CidM})$ for $d = 0$, 1 and $i = 1, \ldots, n_{Cd}$, $n_C = n_{C0} + n_{C1}$, where $M \geq 2$. In most instances the replicates are done in time, and we will let W_{Cid1} denote the first replicate. Define $\Lambda_M = \Sigma_X(\Sigma_X + M^{-1}\Sigma_U)^{-1}$ and let $\hat{\Lambda}_M$ be the estimate of Λ_M defined later. Except for differing intercepts, the best linear approximant to the regression of X_{id} on W_{id} is $\hat{\Lambda}_1 W_{id}$ for "incomplete" data and $\hat{\Lambda}_M \bar{W}_{id}$ for complete data, where $\bar{W}_{id}$ is the mean of the replicates for person i with disease status d. Our procedure is based on substituting $\hat{m}(W_{id})$ for X_{id} in the "incomplete" data and $\hat{m}(\bar{W}_{id})$ for X_{id} in the "complete" data:

- Estimate $\hat{\Lambda}_M$ and $\hat{\Lambda}_1$ by estimating Σ_U and Σ_X. Let

$$\hat{\Sigma}_U = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} (M-1)^{-1} \times \sum_{k=1}^{M} (W_{Cidk} - \bar{W}_{Cid\cdot})(W_{Cidk} - \bar{W}_{Cid\cdot})^t \qquad (12)$$

be the mean of the separate within-person estimates of Σ_U. Then let $\hat{\Sigma}_X$ be $\hat{\Sigma}_W - \hat{\Sigma}_U$, where

$$\begin{aligned}\hat{\Sigma}_W = {} & (n_{R0} + n_{C0} - 2)^{-1} \\ & \times \Sigma_{i=1}^{n_{R0}} (W_{Ri0} - \bar{W}_{R\cdot 0})(W_{Ri0} - \bar{W}_{R\cdot 0})^t \\ & + (n_{R0} + n_{C0} - 2)^{-1} \\ & \times \Sigma_{i=1}^{n_{C0}} M(\bar{W}_{Ci0\cdot} - \bar{W}_{C\cdot 0\cdot})(\bar{W}_{Ci0\cdot} - \bar{W}_{C\cdot 0\cdot})^t\end{aligned}$$

is estimated among the controls, because $\Sigma_W = \mathrm{cov}(W) \approx \mathrm{cov}(W \mid D = 0)$ if the disease is rare.

- If controls and cases are replicated, run a logistic regression with intercepts β_{0C} and β_{0R} and predictors $\hat{\Lambda}_M \bar{W}_{Cid\cdot}$ and $\hat{\Lambda}_1 W_{Rid}$ for the complete and incomplete data but with common slope β.
- If controls only are replicated, use the first replication and run a logistic regression with predictors $\hat{\Lambda}_1 W_{Cid1}$ and $\hat{\Lambda}_1 W_{Rid}$.

Operationally, the algorithm takes the following form. Define $\theta = (\beta_{0C}, \beta_{0R}, \beta^t)^t$ and

$$S_R(\theta, \Lambda_1) = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} \begin{pmatrix} 0 \\ 1 \\ \Lambda_1^t W_{Rid} \end{pmatrix} \times \{d - \mathcal{K}_L(\beta_{0R} + \beta^t \Lambda_1 W_{Rid})\},$$

$$S_{C,1}(\theta, \Lambda_1) = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \begin{pmatrix} 0 \\ 1 \\ \Lambda_1^t W_{Cid1} \end{pmatrix} \times \{d - \mathcal{K}_L(\beta_{0R} + \beta^t \Lambda_1 W_{Cid1})\},$$

and

$$S_{C,2}(\theta, \Lambda_M) = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} \begin{pmatrix} 1 \\ 0 \\ \Lambda_M^t \bar{W}_{Cid\cdot} \end{pmatrix} \times \{d - \mathcal{K}_L(\beta_{0C} + \beta^t \Lambda_M \bar{W}_{Cid\cdot})\}.$$

If cases and controls are replicated, we find estimates by solving $0 = S_R(\hat{\theta}, \hat{\Lambda}_1) + S_{C,2}(\hat{\theta}, \hat{\Lambda}_M)$; if only controls are replicated, $\theta = (\beta_{0R}, \beta^t)^t$, and estimates are found by solving $0 = S_R(\hat{\theta}, \hat{\Lambda}_1) + S_{C,1}(\hat{\theta}, \hat{\Lambda}_1)$. Theoretical details are given in Appendix C.

4.4 More General Approximations

In those cases where the first moment approximation (9) is not adequate, a second approximation may be used. For a rare disease in the source population, the logistic probability satisfies $\mathcal{K}_L(\beta_0^* + \beta^t x) \approx \exp(\beta_0^* + \beta^t x)$. For nondifferential error and a rare disease, our method requires an approximation to the distribution of X given W and $D = 0$ (that is, among noncases), which we denote by $f_{X|W,D}(x|w, D = 0, \alpha)$. It then follows that

$$\Pr(D = 1 | W = w) \approx \exp\{\beta_0^* + R(w, \beta, \alpha)\},$$

where

$$\begin{aligned} \exp\{R(w, \beta, \alpha)\} &= \int \exp(\beta^t x) f_{X|W}(x|w, \alpha)\, dx \\ &\approx \int \exp(\beta^t x) f_{X|W,D}(x|w, D = 0, \alpha)\, dx. \quad (13) \end{aligned}$$

Setting $\Pr(D = 0 | W = w) \approx 1$ in the odds ratio representation in the first line of this article, we find that for any w_0,

$$\frac{f_{W|D}(W/D\ = 1)/f_{W|D}(w_0 | D = 1)}{f_{W|D}(w | D = 0)/f_{W|D}(w_0 | D = 0)} \approx \exp\{R(w, \beta, \alpha) - R(w_0, \beta, \alpha)\}. \quad (14)$$

The calculations of Hsieh, Manski, and McFadden (1985) thus imply that for the complete and incomplete case-control data,

$$f_{XW|D}(x, w | D = d) = \frac{n_C}{n_{Cd}} q_X^C(x) \mathcal{K}_L^d(\beta_{0C} + \beta^t x) \times \{1 - \mathcal{K}_L(\beta_{0C} + \beta^t x)\}^{1-d} f_{W|XD}(w | x, d, \alpha) \quad (15)$$

and

$$f_{W|D}(w | D = d) \approx \frac{n_R}{n_{Rd}} q_W^R(w) \mathcal{K}_L^d\{\beta_{0R} + R(w, \beta, \alpha)\} \times [1 - \mathcal{K}_L\{\beta_{0R} + R(w, \beta, \alpha)\}]^{1-d}. \quad (16)$$

Here $q_W^R(w)$ is the density of W in the incomplete case-control data. Equations (13)–(16) are similar to the approach derived independently by Satten and Kupper (in press), who show that (16) is an exact expression for the likelihood. If X given W is normally distributed with linear mean structure and constant covariance matrix, $R(w, \beta, \alpha)$ is linear in w, essentially yielding the approximate likelihood used to derive the estimates in Section 4.1.

Estimation of the parameters in (15) and (16) is via maximizing the joint likelihood. Given α, $(\beta_{0C}, \beta_{0R}, \beta)$ can be estimated by the method of scoring, based on a logistic model with probabilities $\mathcal{K}_L(\beta_{0C} + \beta^t x)$ and $\mathcal{K}_L\{\beta_{0R} + R(w, \beta, \alpha)\}$ for the complete and incomplete data. The asymptotic theory for the estimates is effectively a special case of that of Section 3 (see App. A).

The ideas in this subsection are clearly connected to Section 3. The estimates are more appropriate than those of Section 4.1 for categorical X; see Satten and Kupper (in press) for details.

5. A SIMULATION EXPERIMENT

Our simulation concerns lognormally distributed predictors with nondifferential error, the parameters based on those discussed by Nero, Schwher, and Nazaroff (1986). They reviewed data on the distribution of radon in single-family homes in the United States and found that the lognormal distribution fit the observations reasonably well. In our simulations the prospective logistic regression model was given by $\Pr(Y = 1 | X = x) = \mathcal{K}_L(-3.09 + .50x)$. Data were generated prospectively as follows. First, X was generated as a lognormal random variable so that $\ln(X)$ had mean $(-1/2)\sigma_X^2$ and variance σ_X^2, where $\sigma_X = 1.08$. Disease outcome D was generated from the prospective model given earlier. The predictor W was also lognormal, with $\ln(W)$ given (X, D) having mean $\ln(X) + \mu_D$ and variance σ_D^2, the values given later. The results were accumulated into a case-control study. In this setup, with $\phi(\cdot)$ as the standard normal density function,

$$f_{W|X,D}(w|x, d) = (w\sigma_d)^{-1}\phi[\{\ln(w) - \ln(x) - \mu_d\}/\sigma_d].$$

We repeated each experiment 1,000 times.

We first describe results for the nondifferential case with $\mu_0 = \mu_1 = 0$, $\sigma_0 = \sigma_1 = .25, .50, 1.00$, $n_{C0} = n_{C1} = 40$, and $n_{R0} = n_{R1} = 80$. We computed the pseudolikelihood estimator both assuming nondifferential error and allowing for differential error with possibly different means and variances. In addition we computed a version of the approximate estimate of Section 4.1; but instead of approximating $E(X | W)$ by a linear function, we computed it exactly from the known structure, a computation that requires knowledge of the marginal distribution of X, which is usually not available. The results from computing $E(X | W)$ exactly shed some light on the need to use data analytic techniques to check the assumption that the regression of X on W is linear.

In the examples in which the assumption of nondifferential error holds (Cases 1, 2, and 3 in Table 3), each of the methods for using reduced data yields smaller mean squared error (MSE) and squared median absolute error (MAE^2) than do the methods using only the complete data. The gains in efficiency can be substantial. Surprisingly, the pseudolikelihood method that allows for differential error (PL − D) is nearly as efficient in these examples as the other methods, which are based on the assumption of nondifferential error. Thus in these examples there is little loss of efficiency from including the extra parameters needed to model differential error and there are large gains in robustness, as we will describe. Note that the method labeled APPROX, which assumes that the regression of X on W is linear, begins to lose efficiency as the measurement error increases (case 3), whereas the method that replaces X by an exact expression for $E(X \mid W)$ retains good efficiency. Thus in practice every effort should be made to model $E(X \mid W)$ correctly when using substitution methods (Pierce et al. 1992).

We also repeated the same experiment but allowed for differential measurement error (cases 4–6 in Table 3). We set $\mu_0 = -1.47$ and $\mu_1 = 1.47$, but still allowed $\sigma_0 = \sigma_1 = 1.0$. We then let the variances differ by setting $(\sigma_0^2, \sigma_1^2) = (.50, 1.00)$, $(1.00, 2.00)$, and $(1.00, 1.00)$. The methods that assume nondifferential error yield biased estimate of β and large MSE and MAE^2 values, giving a striking illustration of the need to carefully assess an assumption of nondifferential error (Greenland 1988b). This suggests that the PL − D should be computed even when the error is assumed to be nondifferential. This estimator improves on using the complete data only, and the gains are comparable to those seen using similar amounts of nondifferential error.

In summary, when the error is properly modeled, the pseudolikelihood estimates can lead to improvements over the complete data estimator. The version of this estimate that allows for differential error confers robustness without (in these simulations) much loss of efficiency; however, we do expect that in some situations a loss of efficiency will result from using the differential error procedure when the nondifferential method is applicable. The regression adjustments of Section 4.1 are reasonable to use when the regression of X on W is modeled carefully and the measurement error is nondifferential. The complete data estimate, although less efficient than the other methods, has the advantage that no error or regression modeling is required.

6. DISCUSSION

Although considerable research has been devoted to the analysis of error-prone covariates in cohort studies (Armstrong 1985; Carroll et al. 1984; Carroll and Stefanski 1990; Carroll and Wand 1991; Gleser 1990; Pepe and Fleming 1991; Pierce et al. 1992; Rosner et al. 1989, 1990; Schafer

Table 3. Lognormal Simulation

	μ_0	σ_{U0}	σ_{U1}	COMP	APPROX	LRMEAN	PL − ND	PL − D
Case #1	.00	.25	.25					
Mean				.57	.53	.52	.53	.53
MSE				1.00	.37	.28	.32	.31
Median				.52	.52	.51	.52	.51
MAE^2				1.00	.40	.33	.34	.33
Case #2	.00	.50	.50					
Mean				.56	.55	.52	.54	.53
MSE				1.00	.52	.31	.34	.44
Median				.52	.53	.50	.52	.51
MAE^2				1.00	.55	.34	.42	.42
Case #3	.00	1.00	1.00					
Mean				.57	.55	.51	.55	.55
MSE				1.00	.88	.39	.55	.65
Median				.53	.52	.49	.53	.53
MAE^2				1.00	.99	.54	.67	.76
Case #4	1.47	1.00	1.41					
Mean				.55	1.08	.91	1.28	.55
MSE				1.00	8.03	5.88	18.1	.76
Median				.52	1.00	.84	1.21	.52
MAE^2				1.00	12.9	5.74	25.5	.73
Case #5	1.47	.71	1.00					
Mean				.55	1.35	.73	1.35	.54
MSE				1.00	18.0	2.24	15.5	.59
Median				.52	1.29	.69	1.29	.51
MAE^2				1.00	31.8	2.00	32.3	.59
Case #6	1.47	1.00	1.00					
Mean				.56	1.35	.85	1.32	.55
MSE				1.00	15.8	3.68	13.3	.65
Median				.53	1.28	.79	1.27	.52
MAE^2				1.00	36.5	5.12	35.5	.70

NOTE: See Section 5 for details. The true value is $\beta = .5$. COMP refers to estimates of β using complete data only, APPROX refers to the estimate of Section 4.1, LRMEAN means replacing X by the exact expression for $E(X \mid W)$ and using this expression in the incomplete data, and PL − ND and PL − D mean the pseudolikelihood estimate computed assuming nondifferential and differential error structure. In all cases, $\mu_0 = -\mu_1$. MSE and MAE refer to the relative mean squared error and median absolute error from $\beta = .50$ relative to that of the complete data estimate.

1987; and Whittemore 1989), this article is the first to propose methodologies for the analysis of case-control data under the prospective logistic risk model (1). In the case that X is categorical these pseudolikelihood methods may be analyzed by previously developed methods (see Gong and Samaneigo 1981 and our App. A). But for the case of a continuous covariate, a more complex asymptotic theory (see App. A) is required. These pseudolikelihood methods are flexible in that they may be used with quite general error distributions, $f_{W|XD}(W|x, D = d, \alpha)$, and they allow for differential error.

Two other approaches are available. First, when X is discrete the retrospective logistic model (Prentice 1976) may be used, in which case methods for incomplete, mixed-up contingency table data are applicable (see for example, Chen 1989; Espeland and Hui 1987; Greenland and Kleinbaum 1983). As the support of X becomes more complex, however, these methods become unwieldy.

A second approach is based on the dual assumptions of nondifferential error and rare disease. Under these circumstances replacing X in (1) by $E(X \mid W)$ (see Secs 4.1, 4.2, and 4.3) or replacing $\exp(X'\beta)$ in (1) by $E\{\exp(X'\beta) \mid W\}$ (see Sec. 4.4) in the incomplete data leads to approximately consistent estimates of β. The former substitution has been suggested previously for cohort data and justified by Armstrong et al. (1989) under a normal discriminant model for case-control data. Satten and Kupper (1993) and Spiegelman and Robins (personal communication) have independently suggested a representation as in Section 4.4, although their use of the representation is different, see also Wang, Wang and Carroll (1992). As far as we know, however, this article is the first to develop the necessary distribution theory for case-control data, except for the special results in Armstrong et al. (1989) and Buonaccorsi (1990) and for the discrete case in Section 4.4 by Satten and Kupper (in press). These methods are easy to implement, even if some components of X (call them Z) are measured accurately in both the complete and incomplete data, whereas other components of X (call them X_*) are measured with error in the incomplete data. We assume that Z and X_* have no common elements. Then $X = (X_*, Z)$ is replaced by $E(X \mid Z, W) = \{E(X_* \mid Z, W), Z\}$ in the incomplete data. For the pseudolikelihood methods of Section 4.4, suppose that the prospective risk model (1) is rewritten as $\mathcal{H}_L(\beta_0^* + \beta_Z^t Z + \beta_X^t X_* + \delta^T Z X_*)$, where $Z X_*$ represents interactions that depend on Z and X_*. Then in the incomplete data, one replaces $\exp(\beta_Z^t Z + \beta_X^T X_* + \tau^{\mathrm{T}} Z X_*)$ by $\exp(\beta_Z^t Z)\ E\{\exp(\beta_X^t X_* + \delta^T Z X_*) \mid Z, W\}$ and proceeds as in Section 4.4. A major limitation of these methods is the requirement of nondifferential error structure.

All the methods mentioned require accurate specification of the error model to make use of the incomplete data. The contingency table literature describes formal methods for selecting error models (see, for example, Espeland and Odoroff 1985). More work is required to determine the robustness of the methods presented to misspecification of the error model. The work of Carroll et al. (1984) and Schafer (1987) indicates significant sensitivity to specification of the error model when the measurement error is moderately large.

Simulations suggest that the methods of Section 3 perform well for samples of practical size (Table 3). But if the complete data are sparse, so that $\hat{F}_0$ and $\hat{F}_1$ have few jump points, and if there is very little error in W, so that $f_{W|XD}(w|x, d, \alpha)$ is nearly degenerate at $w = x$, then (5) may be nearly 0, because then $f_{W|XD}(w|x, d, \alpha)$ may be near 0 where $\hat{Q}_X^R(x)$ has mass. Thus (5) can be a poor approximation to (4), and numerical instability may result. A possible improvement is to use smoothed versions of $\hat{F}_0$ and $\bar{F}_1$. Further numerical studies are also needed to determine how well the proposed estimates of covariance perform in samples of modest size.

If, as discussed, $X = (X_*, Z)$ where Z is measured without error but X_* is measured with error in the incomplete data, then the pseudolikelihood methods (see Sec. 3) become more complicated. Assume the prospective risk model $\mathcal{H}_L(\beta_0^* + \beta_Z^t Z + \beta_X^t X_* + \delta^T Z X_*)$. If Z is discrete, one can stratify on Z to obtain stratum-specific estimates of $\beta_X + \delta Z$ as in Section 3 and subsequently obtain an estimate of the common slope parameter β_X. Alternatively, one can replace (4) by

$$f_{Z\,W|D}(z, w|d)$$

$$= \frac{n_R}{n_{Rd}} \int H_R^d(z, x_*, \theta)\{1 - H_R(z, x_*, \theta)\}^{1-d}$$

$$\times f_{W|Z,X_*,D}(w|z, x_*, d, \alpha) q_Z^R(z) Q_{X_*|Z}^R(dx_*|z). \quad (17)$$

Note that in (17), $q_Z^R(\cdot)$ can be brought outside the integral sign and thus may be ignored. To apply the methods of Section 3, we construct estimates of $Q_{X_*|Z}^R(x_*|z)$ given by $\hat{Q}_{X_*|Z}^R(x_*|z) = \sum_{d=0}^1 (n_{RdZ}/n_{RZ})\hat{F}_d(x|d, z)$, where $\hat{F}_d(x|d, z)$ is the empirical distribution of X_* among members of the complete sample with $D = d$ and $Z = z$, n_{RdZ} is the number of persons with status $D = d$ and $Z = z$ in the incomplete data, and $n_{RZ} = \sum_{d=0}^1 n_{RdZ}$. Then Equation (17) is approximated by a sum analogous to (5), and the methods of Section 3 and Appendix A apply.

If Z is continuous, there are at least three possible approaches that we outline here, although additional research is necessary to develop the ideas. One approach is to ignore information on Z in the incomplete data and to compute the likelihood of W given D, with X replaced by (X, Z); this entails some loss of information. A second approach is to assume a flexible parametric model for the distribution of X given (Z, D), with parameters estimated from the complete data; this leads to a parametric form for $Q_{X_*|Z}^R(\cdot)$ in (17).

A third possibility is to reverse the conditioning in (17), so that

$$f_{ZW|D}(z, w|d) = \frac{n_R}{n_{Rd}} \int H_R^d(\cdot)\{1 - H_R(\cdot)\}^{1-d}$$

$$\times f_{W|Z,X_*,D}(\cdot) q_{Z|X_*}^R(z|x_*)\, dz\, dQ_{X_*}^R(x_*). \quad (18)$$

One might assume a flexible parametric model for the distribution of Z given (X_*, D) with parameters estimated in the complete data, use these distributions to compute $q_{Z|X_*}^R(\cdot)$ in (18), and then follow the pseudolikelihood algorithm.

In Section 3 we did not discuss the possibility that the measurement error model for W given X and D would be assessed by an independent data set, as in Section 4.2. In principle, however, earlier studies may be used to obtain better estimates of the error distribution $f_{W|XD}$. It is often reasonable to assume that this error distribution is invariant across studies. The alternative methods of Section 4 require the distribution of X given W and D; this distribution may vary from study to study. For example, in the nondifferential case, W given X might be normally distributed and mean X and variance σ_U^2. If X is normally distributed with variance $\sigma_{X,d}^2$ in the dth study population, then the slope of X on W is $\sigma_{X,d}^2/(\sigma_{X,d}^2 + \sigma_U^2)$ in the dth study. This lack of invariance of the distribution of X given W indicates a need for caution in incorporating external information on this error distribution into the analyses of Section 4.

Most of the techniques in this article apply to the more general relative risk model where $\beta' x$ on the right side of (1) is replaced by a general function $R(x, \beta)$, because the arguments of Prentice and Pyke (1979) still apply as in (13) and (16).

In summary, when X is discrete we recommend likelihood and pseudolikelihood methods over the matrix method. For continuous X and differential measurement error, pseudolikelihood provides the only consistent estimate among those studied. For rare diseases and nondifferential measurement error, when gold standard X measurements are unavailable, the methods of Sections 4.2 and 4.3 can be used. When X's are available and the measurement error is nondifferential, the possible choices are pseudolikelihood as in Section 3, the linear imputation methods of Sections 4.1 and 4.2, and the likelihood methods of Section 4.4. Pseudolikelihood is fundamentally different from the rest, because its basis is an error model for W given X, with the marginal of X estimated nonparametrically. The other methods are based on models for X given W and the assumptions of nondifferential measurement error and rare diseases. In applications one should assess the measurement error structure carefully when selecting and applying these methods. Ancillary analyses using variations of the error model may reveal how sensitive conclusions are to the assumed error model.

APPENDIX A: THEORY FOR SECTION 3

Two major results of this appendix are as follows. In Section A.2 we show that the estimating equations for pseudolikelihood are asymptotically unbiased, which leads to consistent and asymptotically normal estimates. In Section A.3 we sketch a technical result that is necessary to handle the contribution due to estimating Q_X^R. The main asymptotic theory, stated in Section A.4 and sketched in Section A.5, uses this result.

Unless otherwise stated, our results assume that n_C/n, n_{C0}/n_C, and n_{R0}/n_R converge to numbers strictly between 0 and 1. If this is not the case, then the results will fail and the rates of convergence will be slower than as stated.

A.1 A Covariance Result

Refer to (7) and (A.4). Let G be a matrix with all 0's except the first element, which is $(n_C/n_{C0}) + (n_C/n_{C1})$. Then, as shown by Prentice and Pyke (1979, p. 408), $\operatorname{cov}\{n_C^{-1/2} \mathcal{L}_{C2}(\theta)\} = -\Omega_{2\theta 2} - \Omega_{2\theta 2} G \Omega_{2\theta 2}$. This fact is used in (A.19), (B.3), (B.4), (C.3), and (C.4).

A.2 Asymptotic Unbiasedness of the Estimating Equations

Note that (8) converges to

$$\mathcal{L}_{R2}(\theta, \alpha, Q_X^R) = \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} (n_R/n_{Rd}) S_{d2}(W_{Rid}, \theta, \alpha, Q_X^R)/f_{W|D}(W_{Rid} \mid D = d),$$

with has exceptation

$$n_R \sum_{d=0}^{1} E\left\{ \frac{S_{d2}(W, \theta, \alpha, Q_X^R)}{f_{W|D}(W \mid D = d)} \middle| D = d \right\}$$
$$= n_R \sum_{d=0}^{1} \int S_{d2}(w, \theta, \alpha, Q_X^R)\, dw$$
$$= n_R \sum_{d=0}^{1} (2d - 1) \int\!\!\int \binom{1}{x} \dot{H}_R(x, \theta) f_{W|XD}(w|x, D = d, \alpha) \times dQ_X^R\, dw.$$

Because $\int f_{W|XD}(w|x, D = d, \alpha)\, dw = 1$, interchanging dw and dQ_X^R yields

$$E\mathcal{L}_{L2}(\theta, \alpha, Q_X^R) = n_R \sum_{d=0}^{1} (2d - 1) \int \binom{1}{x} \dot{H}_R(x, \theta)\, dQ_X^R(x) = 0,$$

because the integral does not depend on d and $\Sigma_d (2d - 1) = 0$.

A.3 A General Technical Lemma

Define $Z_n = \hat{Q}_X^R - Q_X^R$. Let Q_{dW} be the distribution function of W given $D = d$, and let $\hat{Q}_{dW}$ be the corresponding empirical distribution function. Define $T_{dE}(x, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R)$ and $T_d(x, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R)$ by

$$T_{dE}(\cdot) = \frac{n_{Rd}}{n_{Cd}} \sum_{k=0}^{1} \frac{n_{Rk}}{n_R} \times \int \frac{\{G_{k2}(w, x, \theta, \alpha) - G_{k1}(w, x, \theta, \alpha) S_k(w, \theta, \alpha, Q_X^R\}}{S_{k1}(w, \theta, \alpha, Q_X^R)}\, dQ_{kW}(w) \quad \text{(A1)}$$

and

$$T_d(\cdot) = T_{dE}(\cdot) - E\{T_{dE}(X, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R) \mid D = d\}.$$

Also T_{d*} is the same as T_d, except (G_{k2}, S_k) are replaced by (G_{k3}, S_{k*}). We wish to show that

$$n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} \{S_d(W_{Rid}, \theta, \alpha\, \hat{Q}_X^R) - S_d(W_{Rid}, \theta, \alpha, Q_X^R)\}$$
$$= n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} T_d(X_{Cid}, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R) + o_p(1), \quad \text{(A.2)}$$

and that a similar result holds if we replace S_d by S_{d*} and T_d by T_{d*}. For convenience will write $\hat{S}_d(w)$ and $S_d(w)$. Note that $\hat{S}_{dk}(w) - S_{dk}(w) = O_p(n^{-1/2})$, because when written out this is just an average of terms $\hat{S}_{dk}$ minus their expectation. Hence

$$\hat{S}_d(w) - S_d(w)$$
$$= \frac{\hat{S}_{d2}(w)}{\hat{S}_{d1}(w)} - \frac{S_{d2}(w)}{S_{d1}(w)}$$
$$= \frac{\hat{S}_{d2}(w) - S_{d2}(w) - S_d(w)\{\hat{S}_{d1}(w) - S_{d1}(w)\}}{\hat{S}_{d1}(w)}$$
$$= \frac{\hat{S}_{d2}(w) - S_{d2}(w) - S_d(w)\{\hat{S}_{d1}(w) - S_{d1}(w)\}}{S_{d1}(w)} + o_p(n^{-1/2})$$
$$= \int \frac{G_{d2}(w, x, \theta, \alpha) - G_{d1}(w, x, \theta, \alpha) S_d(w, \theta, \alpha, Q_X^R)}{S_{d1}(w, \theta, \alpha, Q_X^R)} \times dZ_n(x) + o_p(n^{-1/2})$$
$$= \int \mathcal{D}(w, x)\, dZ_n(x) + o_p(n^{-1/2}),$$

say. Hence the left side of (A.2) is, to order $O_p(n^{-1/2})$,

$$\begin{aligned}
&\approx n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} \int \mathcal{D}(W_{Rid}, x)\, dZ_n(x) \\
&= n_C^{-1/2} \sum_{d=0}^{1} n_{Rd} \int\int \mathcal{D}(w, x)\, dZ_n(x)\, d\hat{Q}_{dW}(w) \\
&\approx n_C^{-1/2} \sum_{d=0}^{1} n_{Rd} \int\int \mathcal{D}(w, x)\, dQ_{dW}(w)\, dZ_n(x) + o_p(1) \\
&= n_C^{-1/2} \sum_{d=0}^{1} n_{Rd} \int U_d(x)\, dZ_n(x) + o_p(1). \qquad \text{(A.3)}
\end{aligned}$$

We thus must show that the first terms on the right sides of (A.2) and (A.3) coincide. Note that the latter is

$$\begin{aligned}
&n_C^{-1/2} \sum_{d=0}^{1} n_{Rd} \sum_{k=0}^{1} \sum_{i=1}^{n_{Ck}} \frac{n_{Rk}}{n_R n_{Ck}} [U_d(X_{Cik}) - E\{U_d(x) | D = k\}] \\
&= n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \frac{n_{Rd}}{n_{Cd}} \sum_{k=0}^{1} \frac{n_{Rk}}{n_R} [U_k(X_{Cid}) - E\{U_k(x) | D = d\}] \\
&= n_C^{1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} T_d(X_{Cid}, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R),
\end{aligned}$$

completing the argument.

A.4 Asymptotic Theory

In this subsection we state the limiting distribution of the pseudolikelihood estimate of θ. A sketch of the proof is given in the next subsection. To this end, let subscripts θ and α denote derivatives and define

$$\Omega_{1\theta} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} E\{S_{d*\theta}(W_{Rid}, \theta, \alpha, Q_X^R) | D = d\},$$

$$\Omega_{1\alpha 1} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} E\{S_{d*\alpha}(W_{Rid}, \theta, \alpha, Q_X^R) | D = d\},$$

$$\Omega_{1\alpha 2} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\{\Psi_{d\alpha}(X_{Cid}, W_{Cid}, \alpha) | D = d\},$$

$$\Omega_{1\alpha} = \Omega_{1\alpha 1} + \Omega_{1\alpha 2};$$

$$\Omega_{2\alpha} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} E\{S_{d\alpha}(W_{Rid}, \theta, \alpha, Q_X^R) | D = d\},$$

$$\Omega_{2\theta 1} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} E\{S_{d\theta}(W_{Rid}, \theta, \alpha, Q_X^R) | D = d\},$$

$$\Omega_{2\theta 2} = -n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\left\{\begin{pmatrix} 1 \\ X_{Cid} \end{pmatrix} \begin{pmatrix} 1 \\ X_{Cid} \end{pmatrix}^t \dot{H}_C(X_{Cid}, \theta) \middle| D = d\right\}, \qquad \text{(A.4)}$$

$$\Omega_{2\theta} = \Omega_{2\theta 1} + \Omega_{2\theta 2},$$

and

$$\Omega_* = \begin{pmatrix} \Omega_{1\alpha} & \Omega_{1\theta} \\ \Omega_{2\alpha} & \Omega_{2\theta} \end{pmatrix}.$$

Remark A.1. Proofs of the results in this section are sketched informally in the next subsection. Technically, the major difficulty is in keeping the random variable $S_{d1}(w, \theta, \alpha\ \hat{Q}_X^R)$ away from 0, where $S_{d1}(w, \theta, \alpha, Q_X^R) = (n_{Rd}/n_R)\ _{W|D}(w | D = d, \theta, \alpha)$. This difficulty is evident from Section A.3. We have been able to prove the results only under fairly stringent regularity conditions, the key ones being as follows.

- For differential error, W given $D = d$ has compact support and $f_{W|D}(w | D = d, \theta_*, \alpha_*)$ is strictly positive on this support and in neighborhoods of (θ, α). This condition includes categorical W as a special case.
- For nondifferential error, either the previous condition holds or the sums in $(\mathcal{L}_{R1}, \mathcal{L}_{R2})$ are taken only over W_{Rid} on a fixed set independent of d and interior to the support of W given $D = d$.

Formal proofs even under these strict conditions are messy.

In what follows, rather than present detailed proofs under precise conditions, we sketch them only. We provide a level of detail that we hope will enable the reader to follow the main ideas.

When we use the notation "$A \approx B$", we mean that $A - B = o_p(1)$.

Result. With functions T_d and T_{d*} defined in the previous subsection, define for $s = 1, 2$

$$\mathcal{C}_1 = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} [S_d(W_{Rid}, \theta, \alpha, Q_X^R) - E\{S_d(W, \theta, \alpha, Q_X^R) | D = d\}],$$

$$\mathcal{C}_2 = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} [S_{d*}(W_{Rid}, \theta, \alpha, Q_X^R) - E\{S_{d*}(W, \theta, \alpha, Q_X^R) | D = d\}],$$

$$\mathcal{C}_{s+2} = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \begin{bmatrix} V_{ds}(X_{Cid}, W_{Cid}, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R) \\ -E\{V_{ds}(X, W, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R) | D = D\} \end{bmatrix},$$

$$V_{d1}(x, w, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R) = \begin{pmatrix} 1 \\ x \end{pmatrix} \{d - H(x, \theta)\} + T_d(x, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R),$$

and

$$V_{d2}(x, w, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R) = \Psi_d(x, w, \alpha) + T_{d*}(x, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R).$$

Then

$$n_C^{1/2} \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\theta} - \theta \end{pmatrix} = -\Omega_*^{-1} \begin{pmatrix} \mathcal{C}_3 + \mathcal{C}_2 \\ \mathcal{C}_4 + \mathcal{C}_1 \end{pmatrix} + o_p(1).$$

Hence $n_C^{1/2}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean 0 and covariance matrix the lower right block of

$$\Omega_*^{-1} \left\{ \operatorname{cov} \begin{pmatrix} \mathcal{C}_3 \\ \mathcal{C}_4 \end{pmatrix} + \operatorname{cov} \begin{pmatrix} \mathcal{C}_2 \\ \mathcal{C}_1 \end{pmatrix} \right\} \Omega_*^{-t}. \qquad \text{(A.5)}$$

A.5 Sketch of Proof of Result

We first note that because the estimating equations are asymptotically unbiased, under sufficient regularity conditions the estimates are consistent. Define

$$\mathcal{E}_1 = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} T_d(X_{Cid}, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R)$$

and

$$\mathcal{E}_2 = n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} T_{d*}(X_{Cid}, \theta, \alpha, Q_{0W}, Q_{1W}, Q_X^R).$$

Formally expanding in a Taylor series and using Section A.3, we find that

$$n_C^{-1/2} \mathcal{L}_{C1}(\hat{\alpha}) \approx n_C^{-1/2} \mathcal{L}_{C1}(\alpha) + \Omega_{1\alpha 2} n_C^{1/2}(\hat{\alpha} - \alpha),$$

$$n_C^{-1/2} \mathcal{L}_{C2}(\hat{\theta}) \approx n_C^{-1/2} \mathcal{L}_{C2}(\theta) + \Omega_{2\theta 2} n_C^{1/2}(\hat{\theta} - \theta),$$

$$n_C^{-1/2} \mathcal{L}_{R1}(\hat{\theta}, \hat{\alpha}, \hat{Q}_X^R) \approx n_C^{-1/2} \mathcal{L}_{R1}(\theta, \alpha, Q_X^R) + \Omega_{1\theta} n_C^{1/2}(\hat{\theta} - \theta) + \Omega_{1\alpha 1} n_C^{1/2}(\hat{\alpha} - \alpha) + \mathcal{E}_2,$$

and

$$n_C^{-1/2}\mathcal{L}_{R2}(\hat\theta, \hat\alpha, \hat Q_X^R) \approx n_C^{-1/2}\mathcal{L}_{R2}(\theta, \alpha, Q_X^R) + \Omega_{2\theta 1} n_C^{1/2}(\hat\theta - \theta) + \Omega_{2\alpha} n_C^{1/2}(\hat\alpha - \alpha) + \mathcal{E}_1.$$

By formal Taylor series expansions, we then obtain

$$\begin{aligned} 0 &= n_C^{-1/2}\{\mathcal{L}_{C2}(\hat\theta) + \mathcal{L}_{R2}(\hat\theta, \hat\alpha, \hat Q_X^R)\} \\ &\approx n_C^{-1/2}\{\mathcal{L}_{C2}(\theta) + \mathcal{L}_{R2}(\theta, \alpha, Q_X^R)\} \\ &\quad + \mathcal{E}_1 + \Omega_{2\alpha} n_C^{1/2}(\hat\alpha - \alpha) + \Omega_{2\theta} n_C^{1/2}(\hat\theta - \theta), \end{aligned} \qquad \text{(A.6)}$$

and

$$\begin{aligned} 0 &= n_C^{-1/2}\{\mathcal{L}_{C1}(\hat\alpha) = \mathcal{L}_{R1}(\hat\theta, \hat\alpha, \hat Q_X^R)\} \\ &\approx n_C^{-1/2}\{\mathcal{L}_{C1}(\alpha) + \mathcal{L}_{R1}(\theta, \alpha, Q_X^R)\} \\ &\quad + \mathcal{E}_2 + \Omega_{1\alpha} n_C^{1/2}(\hat\alpha - \alpha) + \Omega_{1\theta} n_C^{1/2}(\hat\theta - \theta). \end{aligned} \qquad \text{(A.7)}$$

Combining (A.6) and (A.7) yields

$$-\Omega_* n_C^{1/2}\begin{pmatrix} \hat\alpha - \alpha \\ \hat\theta - \theta \end{pmatrix} \approx n_C^{-1/2}\begin{pmatrix} \mathcal{L}_{C1}(\alpha) + \mathcal{L}_{R1}(\theta, \alpha, Q_X^R) + n_C^{1/2}\mathcal{E}_2 \\ \mathcal{L}_{C2}(\theta) + \mathcal{L}_{R2}(\theta, \alpha, Q_X^R) + n_C^{1/2}\mathcal{E}_1 \end{pmatrix}.$$

Each of the first three terms on the right side of (A.6) and (A.7) has mean 0, even though their individual summands do not. Thus, we may replace an individual summand, say U_{id}, by its centered version $U_{id} - E\{U_{id} \mid D = d\}$.

A.6 Covariance Estimates

We now provide consistent estimates of the covariance matrix (A.5). Estimating the various subscripted Ω-matrices is performed by removing the expectations in their definitions and replacing (θ, α, Q_X^R) by ($\hat\theta$, $\hat\alpha$, $\hat Q_X^R$). For example,

$$\hat\Omega_{1\alpha 2} = n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Cd}} \Psi_{d\alpha}(X_{Cid}, W_{Cid}, \hat\alpha);$$

$$\hat\Omega_{2\alpha} = n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Cd}} S_{d\alpha}(W_{Rid}, \hat\theta, \hat\alpha, \hat Q_X^R).$$

Estimation of $\operatorname{cov}(\mathcal{C}_1)$ is similar. Define

$$\hat S_d(W_{Rid}) = S_d(W_{Rid}, \hat\theta, \hat\alpha, \hat Q_X^R); \qquad \bar S_d = n_{Rd}^{-1}\sum_{i=1}^{n_{Rd}} \hat S_d(W_{Rid}).$$

We make similar definitions for S_{d*} and define

$$\widehat{\operatorname{cov}}\begin{pmatrix} \mathcal{C}_2 \\ \mathcal{C}_1 \end{pmatrix} = n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Rd}} \begin{pmatrix} \hat S_{d*}(W_{Rid}) - \bar S_{d*} \\ \hat S_d(W_{Rid}) - \bar S_d \end{pmatrix}\begin{pmatrix} \hat S_{d*}(W_{Rid}) - \bar S_{d*} \\ \hat S_d(W_{Rid}) - \bar S_d \end{pmatrix}.$$

We require estimates of T_d and T_{d*}. Refer to (A.1) and define $\hat T_{dE}(x) = T_{dE}(x, \hat\theta, \hat\alpha, \hat Q_{0W}, \hat Q_{1W}, \hat Q_X^R)$;

$$\bar T_{dE} = n_{Cd}^{-1}\sum_{i=1}^{n_{Cd}} \hat T_{dE}(X_{Cid}); \qquad \hat T_d(x) = \hat T_{dE}(x) - \bar T_{dE},$$

with $\hat T_{d*}$ defined similarly.

Define $\mathcal{B}_1 = -\Omega_{2\theta 2} - \Omega_{2\theta 2} G \Omega_{2\theta 2}$ and

$$\begin{pmatrix} V_{d1} \\ V_{d2} \end{pmatrix} = V_{d3} + V_{d4} = \begin{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} & \{d - H_C(x, \theta)\} \\ & 0 \end{pmatrix} + V_{d4},$$

and for $k = 3, 4$,

$$\mathcal{B}_k = n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Cd}} \{\hat V_{dk}(X_{Cid}, W_{Cid}) - \bar V_{dk}\} \times \{\hat V_{d4}(X_{Cid}, W_{Cid}) - \bar V_{d4}\}^t.$$

A consistent estimate of $\operatorname{cov}(\mathcal{C}_3^t, \mathcal{C}_4^t)^t$ is

$$\widehat{\operatorname{cov}}\begin{pmatrix} \mathcal{C}_3 \\ \mathcal{C}_4 \end{pmatrix} = \begin{pmatrix} \mathcal{B}_1 & 0 \\ 0 & 0 \end{pmatrix} + \mathcal{B}_4 + \mathcal{B}_3 + \mathcal{B}_3^t.$$

A.7 Generalizations

If Q_X^R is known or the effects of estimating it are ignored, the covariance estimates given previously simplify by setting $V_{d4} = \{0, \Psi_d^t(x, w, \alpha)\}^t$. The estimates are easy to compute in this case.

When there are a finite number of parameters, as when X is discrete, or for the likelihood estimates of Section 4, covariance estimates can be computed as follows. Let Θ include all parameters to be estimated and let $\mathcal{S}(\Theta)$ be the estimating equation, generally a sum of terms of the form

$$\mathcal{S}(\Theta) = \sum_{d=0}^{1}\left\{\sum_{i=1}^{n_{Cd}} \Psi_{idC}(\Theta) + \sum_{i=1}^{n_{Rd}} \Psi_{idR}(\Theta)\right\}.$$

Then $\hat\Theta$ is estimated by solving $0 = \mathcal{S}(\hat\Theta)$. Let $\mathcal{S}_\Theta(\Theta)$ be the derivative of the estimating equation. Let $\mathcal{T}(\Theta) = \operatorname{cov}\{\mathcal{S}(\Theta)\}$, which is calculated using the case-control probabilities. Then the estimated covariance matrix is $\mathcal{S}_\Theta^{-1}(\hat\Theta)\mathcal{T}(\hat\Theta)\mathcal{S}_\Theta^{-t}(\hat\Theta)$.

For example, in the 2 × 2 × 2 example of Section 2, $\mathcal{S}(\Theta)$ can be expressed as a function of the cell counts in the two quadrinomials defined by complete data (X, W) separately for $D = 0$, 1 and in the two binomials defined by the reduced data W separately for $D = 0$, 1. These four groups of cell counts constitute four independent multinomial random variables with cell probabilities determined by the parameters Θ; hence multinomial theory can be used to derive a formula for $\mathcal{T}(\Theta)$. These methods can be used either for maximum likelihood or pseudolikelihood estimates by defining appropriate estimating equations, $\mathcal{S}(\theta)$.

APPENDIX B: THEORY FOR SECTION 4.1

B.1 Statement of Results

Define $\alpha = (\alpha_0, \alpha_1, \alpha_2)^t$ and define e by $\alpha_2^t = e\alpha$. Define $\theta = (\beta_{0C}, \beta_{0R}, \beta^t)^t$, $H_C(x, \theta) = \mathcal{H}_L(\beta_{0C} + \beta^t x)$, $H_R(w, \theta, \alpha_2) = \mathcal{H}_L(\beta_{0R} + \beta^t \alpha_2 w)$,

$$\mathcal{L}_{C2}(\theta) = \sum_{d=0}^{1}\sum_{i=1}^{n_{Cd}} (1, 0, X_{Cid}^t)^t\{d - H_C(X_{Cid}, \theta)\},$$

and

$$\mathcal{L}_{R2}(\theta, \alpha_2) = \sum_{d=0}^{1}\sum_{i=1}^{n_{Rd}} \{0, 1, (\alpha_2 W_{Rid})^t\}^t\{d - H_R(W_{Rid}, \theta, \alpha_2)\}.$$

The estimates solve $\mathcal{L}_{C2}(\hat\theta) + \mathcal{L}_{R2}(\hat\theta, \hat\alpha_2) = 0$. In this section we show that in the limit,

$$n_C^{1/2}(\hat\theta - \theta) = \Omega_\theta^{-1} n_C^{-1/2}\{\mathcal{L}_{C2}(\theta) + \mathcal{L}_{C3}(\theta, \alpha_2) + \mathcal{L}_{R2}(\theta, \alpha_2)\} + o_p(1), \qquad \text{(B.1)}$$

where

$$\mathcal{L}_{C3}(\theta, \alpha) = \sum_{d=0}^{1}\sum_{i=1}^{n_{Cd}} \Omega_{R\alpha} e \Omega_\alpha^{-1}\begin{pmatrix} 1 \\ D_{Cid} \\ W_{Cid} \end{pmatrix} V_{Cid}^t \beta,$$

$$\begin{aligned} \Omega_\theta &= n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Cd}} E\left\{\begin{pmatrix} 1 \\ 0 \\ X_{Cid} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ X_{Cid} \end{pmatrix}^t \dot H_C(X_{Cid}, \theta) \,\middle|\, D = d\right\} \\ &\quad + n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Rd}} E\left\{\begin{pmatrix} 0 \\ 1 \\ \alpha_2 W_{Rid} \end{pmatrix}\begin{pmatrix} 0 \\ 1 \\ \alpha_2 W_{Rid} \end{pmatrix}^t \dot H_R(W_{Rid}, \theta, \alpha_2) \,\middle|\, D = d\right\} \\ &= \Omega_{C\theta} + \Omega_{R\theta}, \end{aligned}$$

$$\Omega_{R\alpha} = -n_C^{-1}\sum_{d=0}^{1}\sum_{i=1}^{n_{Rd}} E\left\{\begin{pmatrix} 0 \\ 1 \\ \alpha_2 W_{Rid} \end{pmatrix} W_{Rid}^t \dot H_R(W_{Rid}, \theta, \alpha_2) \,\middle|\, D = d\right\},$$

and

$$\Omega_\alpha = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\left\{ \begin{pmatrix} 1 \\ D_{Cid} \\ W_{Cid} \end{pmatrix} \begin{pmatrix} 1 \\ D_{Cid} \\ W_{Cid} \end{pmatrix}^t \middle| D = d \right\}.$$

Thus for a matrix $\mathcal{C}_{23}$ defined later, $n_C^{1/2}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean 0 and covariance matrix

$$\Omega_\theta^{-1} \begin{bmatrix} \operatorname{cov}\{n_C^{-1/2}\mathcal{L}_{C2}(\theta)\} + \operatorname{cov}\{n_C^{-1/2}\mathcal{L}_{C3}(\theta, \alpha_2)\} \\ + \operatorname{cov}\{n_C^{-1/2}\mathcal{L}_{R2}(\theta, \alpha_2)\} + \mathcal{C}_{23} + \mathcal{C}_{23}^t \end{bmatrix} \Omega_\theta^{-1}, \qquad \text{(B.2)}$$

where, by Section A.1,

$$\operatorname{cov}\{n_C^{-1/2}\mathcal{L}_{C2}(\theta)\} = \Omega_{C\theta} - \Omega_{C\theta} \begin{pmatrix} \frac{n_C}{n_{C0}} + \frac{n_C}{n_{C1}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \Omega_{C\theta} \qquad \text{(B.3)}$$

and

$$\operatorname{cov}\{n_C^{-1/2}\mathcal{L}_{R2}(\theta, \alpha_2)\}$$
$$= \Omega_{R\theta} - \frac{n_C}{n_R}\Omega_{R\theta} \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{n_R}{n_{R0}} + \frac{n_R}{n_{R1}} & 0 \\ 0 & 0 & 0 \end{pmatrix} \Omega_{R\theta}. \qquad \text{(B.4)}$$

Also, because $e = (0_{q\times 1}, 0_{q\times 1}, I_q)$, $V_{Cid} = X_{Cid} - \alpha_0 - \alpha_1 D_{Cid} - \alpha_2 W_{Cid}$, and if $V_{id*} = \Omega_{R\alpha} e \Omega_\alpha (1, D_{Cid}, W_{Cid}^t)^t V_{Cid}^t \beta$, then

$$\operatorname{cov}\{n_C^{-1/2}\mathcal{L}_{C3}(\theta, \alpha)\} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\{V_{id*} V_{id*}^t \mid D = d\}$$

and

$$\mathcal{C}_{23} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\left[\{d - H_C(X_{Cid}, \theta)\} \begin{pmatrix} 1 \\ 0 \\ X_{Cid} \end{pmatrix} V_{id*}^t \middle| D = d \right].$$

Remark B.1. Constructing an estimated covariance matrix is particularly easy: $\mathcal{C}_{23}$, $\Omega_{R\theta}$, $\Omega_{C\theta}$, $\Omega_{R\alpha}$, and Ω_α are estimating by replacing $(\theta, \alpha_0, \alpha_1, \alpha_2)$ by their estimated values, removing the expectations, and replacing V_{Cid} in their definitions by the residuals $\hat{V}_{Cid} = X_{Cid} - \hat{\alpha}_0 - \hat{\alpha}_1 D_{Cid} - \hat{\alpha}_2 W_{Cid}$.

We now sketch (B.1). By a Taylor series,

$$0 \approx n_C^{-1/2}\{\mathcal{L}_{C2}(\theta) + \mathcal{L}_{R2}(\theta, \alpha_2)\}$$
$$- \Omega_\theta n_C^{1/2}(\hat{\theta} - \theta) + \Omega_{R\alpha} n_C^{1/2}(\hat{\alpha}_2 - \alpha_2)^t \beta$$

If $\alpha = (\alpha_0, \alpha_1, \alpha_2)^t$, then the least squares estimates of α solves

$$0 = \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \begin{pmatrix} 1 \\ D_{Cid} \\ W_{Cid} \end{pmatrix} \left\{ X_{Cid}^t - \begin{pmatrix} 1 \\ D_{Cid} \\ W_{Cid} \end{pmatrix}^t \hat{\alpha} \right\},$$

so that

$$n_C^{1/2}(\hat{\alpha} - \alpha) \approx \Omega_\alpha^{-1} n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} (1, D_{Cid}, W_{Cid}^t)^t V_{Cid}^t.$$

Because $\alpha_2^t = e\alpha$, (B.1) follows.

B.2 External Validation

In the case of external validation, covariance calculations simplify because the estimate $\hat{\alpha}_2$ is independent of $\mathcal{L}_{R2}(\theta, \alpha_2)$, and only minor changes to the previous theory are necessary. Let $\hat{\alpha}_2$ be the estimate of α_2, and let Λ be the asymptotic covariance matrix of $n_R^{1/2}(\hat{\alpha}_2 - \alpha_2)^t\beta$. Referring to Appendix B, let $\Omega_{R\theta*}$ be obtained by deleting the first row and column of $\Omega_{R\theta}$, and let $\Omega_{R\alpha*}$ be obtained by deleting the first row of $\Omega_{R\alpha}$. Let $\theta = (\beta_{0R}, \beta^t)^t$. Solving $\mathcal{L}_{R2}(\theta, \hat{\alpha}_2) = 0$ yields $\hat{\theta}$. If G is the matrix with all 0s except a first element given by $(n_R/n_{R0}) + n_R/n_{R1})$, then $n_R^{1/2}(\hat{\theta} - \theta)$ has asymptotic covariance matrix

$$\Omega_{R\theta*}^{-1}(\Omega_{R\theta*} - \Omega_{R\theta*} G \Omega_{R\theta*} + \Omega_{R\alpha*} \Lambda \Omega_{R\alpha*}) \Omega_{R\theta*}^{-1}.$$

Remark B.1 then applies for covariance estimation.

APPENDIX C: RESULTS FOR SECTION 4.3

C.1 Results When Cases and Controls are Replicated

We first consider the situation in which cases and controls are replicated. Note that in either sampling strategy,

$$\hat{\Sigma}_W - \Sigma_W = (n_{C0} + n_{R0})^{-1} \left\{ \sum_{i=1}^{n_{R0}} (S_{Ri0} + \Sigma_W) + \sum_{i=1}^{n_{C0}} (S_{Ci0} - \Sigma_W) \right\}$$
$$+ o_p(n_C^{-1/2}) \qquad \text{(C.1a)}$$

and

$$\hat{\Sigma}_U - \Sigma_U = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} (S_{Cid}^* - \Sigma_U) + o_p(n_C^{-1/2}), \qquad \text{(C.1b)}$$

where $S_{Rid} = (W_{Rid} - \mu_d)(W_{Rid} - \mu_d)^t$, $S_{Cid} = (\bar{W}_{Cid\cdot} - \mu_d)(\bar{W}_{Cid\cdot} - \mu_d)^t$, $\mu_d = E(W \mid D = d)$, and S_{Cid}^* are the summands to (12), but with W replaced by U. Define $Q_{Rid} = I(d = 0)\Sigma_X^{-1}(S_{Rid} - \Sigma_W)\Sigma_X^{-1}(n_C + n_R)/(n_{C0} + n_{R0})$, $Q_{Cid} = I(d = 0)\Sigma_X^{-1}(S_{Cid} - \Sigma_W)\Sigma_X^{-1}(n_C + n_R)/(n_{C0} + n_{R0})$, and $Q_{Cid}^* = \Sigma_X^{-1}(S_{Cid}^* - \Sigma_U)\Sigma_X^{-1}$. Then, in the next subsection we show that

$$n_C^{1/2}(\hat{\theta} - \theta) = \Omega_\theta^{-1} \left\{ S_R(\theta, \Lambda_1) + n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} V_{Rid} \right\}$$
$$+ \Omega_\theta^{-1} \left\{ S_{C,2}(\theta, \Lambda_M) + n_C^{-1/2} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} V_{Cid} \right\} + o_p(1), \qquad \text{(C.2)}$$

where

$$V_{Cid} = \frac{n_C}{n_C + n_R}(\Omega_{R1}\Lambda_1^t Q_{Cid}\Lambda_1 + M^{-}\Omega_{CM}\Lambda_M^t Q_{Cid}\Lambda_M)\Sigma_U\beta$$
$$- (\Omega_{R1}\Lambda_1^t Q_{Cid}^*\Lambda_1 + M^{-1}\Omega_{CM}\Lambda_M^t Q_{Cid}^*\Lambda_M)\Sigma_W\beta,$$

$$V_{Rid} = \frac{n_C}{n_C + n_R}(\Omega_{R1}\Lambda_1^t Q_{Rid}\Lambda_1 + M^{-1}\Omega_{CM}\Lambda_M^t Q_{Rid}\Lambda_M)\Sigma_U\beta,$$

$$\Omega_{C\theta} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\left[\begin{pmatrix} 1 \\ 0 \\ \Lambda_M \bar{W}_{Cid\cdot} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \Lambda_M \bar{W}_{Cid\cdot} \end{pmatrix} \right.$$
$$\left. \times \dot{\mathcal{H}}(\beta_{0C} + \beta^t \Lambda_M \bar{W}_{Cid\cdot}) \mid D = d \right],$$

$$\Omega_{R\theta} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} E\left[\begin{pmatrix} 0 \\ 1 \\ \Lambda_1 W_{Rid} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ \Lambda_1 W_{Cid} \end{pmatrix}^t \right.$$
$$\left. \times \dot{\mathcal{H}}(\beta_{0R} + \beta^t \Lambda_1 W_{Rid}) \mid D = d \right],$$

$$\Omega_\theta = \Omega_{C\theta} + \Omega_{R\theta},$$

$$\Omega_{R1} = -n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} E\left[\begin{pmatrix} 0 \\ 1 \\ \Lambda_1 W_{Rid} \end{pmatrix} \right.$$
$$\left. \times \dot{\mathcal{H}}(\beta_{0R} + \beta^t \Lambda_1 W_{Rid}) W_{Rid}^t \mid D = d \right],$$

and

$$\Omega_{CM} = -n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} E\left[\begin{pmatrix} 1 \\ 0 \\ \Lambda_M \bar{W}_{Cid\cdot} \end{pmatrix} \right.$$

$$\left. \times \dot{\mathcal{K}}(\beta_{0C} + \beta^t \Lambda_M \bar{W}_{Cid\cdot}) \bar{W}^t_{Cid\cdot} \mid D = d \right].$$

From (C.2), $n_C^{1/2}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean 0. Its covariance matrix is estimated as follows. First note that

$$\operatorname{cov}\{S_R(\theta, \Lambda_1)\} = \Omega_{R\theta} - \frac{n_C}{n_R}\Omega_{R\theta} \begin{pmatrix} 0 & 0 & 0 \\ 0 & \dfrac{n_R}{n_{R0}} + \dfrac{n_R}{n_{R1}} & 0 \\ 0 & 0 & 0 \end{pmatrix} \Omega_{R\theta} \qquad \text{(C.3)}$$

and

$$\operatorname{cov}\{S_C(\theta, \Lambda_M)\} = \Omega_{C\theta} - \Omega_{C\theta} \begin{pmatrix} \dfrac{n_C}{n_{C0}} + \dfrac{n_C}{n_{C1}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \Omega_{C\theta}. \qquad \text{(C.4)}$$

Then the first term on the right side of (C.2) has a covariance matrix estimated by

$$\hat{\Omega}_\theta^{-1}\left\{(\widehat{C.3}) + n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} \hat{V}_{Rid}\hat{V}^t_{Rid} + \mathcal{C}_{41} + \mathcal{C}^t_{41}\right\}\hat{\Omega}_\theta^{-1},$$

where

$$\mathcal{C}_{41} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Rd}} \begin{pmatrix} 0 \\ 1 \\ \hat{\Lambda}_1 W_{Rid} \end{pmatrix} \hat{V}_{Rid}\{d - \mathcal{K}_L(\hat{\gamma}_1 + \hat{\beta}^t \hat{\Lambda}_1 W_{Rid}\},$$

where $\hat{V}_{Rid}$ is defined by replacing population quantities in V_{Rid} by sample quantities. Similarly, the second term in (C.2) has covariance matrix estimated by

$$\hat{\Omega}_\theta^{-1}\left\{(\widehat{C.4}) + n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \hat{V}_{Cid}\hat{V}^t_{Cid} + \mathcal{C}_{42} + \mathcal{C}^t_{42}\right\}\hat{\Omega}_\theta^{-1},$$

where

$$\mathcal{C}_{42} = n_C^{-1} \sum_{d=0}^{1} \sum_{i=1}^{n_{Cd}} \begin{pmatrix} 1 \\ 0 \\ \hat{\Lambda}_M \bar{W}_{Cid} \end{pmatrix} \hat{V}_{Cid}\{d - \mathcal{K}_L(\hat{\gamma}_0 + \hat{\beta}^t \hat{\Lambda}_M \bar{W}_{Cid}\}.$$

C.2 Sketch of Proof of (C.2)

Define $Q_W = \Sigma_X^{-1}(\hat{\Sigma}_W - \Sigma_W)\Sigma_X^{-1}$ and $Q_U = \Sigma_X^{-1}(\hat{\Sigma}_U - \Sigma_U)\Sigma_X^{-1}$. Then a simple expansion shows that

$$(\hat{\Lambda}_M - \Lambda_M)^t = M^{-1}\Lambda^t_M(Q_W\Lambda_M\Sigma_U - Q_U\Lambda_M\Sigma_W) + o_p(n_c^{-1/2}), \qquad \text{(C.5)}$$

and similarly for $(\hat{\Lambda}_M - \Lambda_M)^t$. By a standard Taylor series,

$$\begin{aligned} \Omega_\theta(\hat{\theta} - \theta) = {} & S_R(\theta, \Lambda_1) + S_{C,2}(\theta, \Lambda_M) \\ & + \Omega_{R1} n_C^{1/2}(\hat{\Lambda}_1 - \Lambda_1)^t\beta \\ & + \Omega_{CM} n_C^{1/2}(\hat{\Lambda}_M - \Lambda_M)^t\beta + o_p(1). \end{aligned} \qquad \text{(C.6)}$$

Substituting (C.5) into (C.6) and then using the expansion (C.1) yields (C.2).

C.3 Replication of Controls

In this case $n_{C1} = 0$ and $n_C = n_{C0}$. Here are the necessary changes. Change $\Omega_{C\theta}$ by replacing $(\Lambda_M, \bar{W}_{Cid\cdot})$ by (Λ_1, W_{Cid1}), and collapse the zero rows in $\Omega_{R\theta}$ and $\Omega_{C\theta}$; then $\Omega_{R\theta} = \Omega_{C\theta} = \Omega_\theta$, which can be estimated by pooling all the data. In the covariance formulas, replace (C.3) and (C.4) by

$$\Omega_\theta - \Omega_\theta \begin{pmatrix} n/n_0 + n/n_1 & 0 \\ 0 & 0 \end{pmatrix} \Omega_\theta,$$

where $n_0 = n_{C0} + n_{R0}$ and $n_1 = n_{C1} + n_{R1}$.

[*Received February 1991. Revised May 1992.*]

REFERENCES

Aitchinson, J., and Silvey, S. D. (1958), "Maximum Likelihood Estimation of Parameters Subject to Constraints," *Annals of Mathematical Statistics,* 29, 813–828.

Anderson, J. A. (1972), "Separate Sample Logistic Discrimination," *Biometrika,* 59, 19–35.

Armstrong, B. (1985), "Measurement Error in the Generalized Linear Model," *Communications in Statistics, Part B—Simulations and Computation,* 14, 529–544.

Armstrong, B., Howe, G., and Whittemore, A. S. (1989), "Correcting for Measurement Error in a Nutrition Study," *Statistics in Medicine,* 8, 1151–1163.

Breslow, N., and Powers, W. (1978), "Are There Two Logistic Regressions for Retrospective Studies?" *Biometrics,* 34, 100–105.

Buonaccorsi, J. P. (1990), "Double Sampling for Exact Values in the Normal Discriminant Model With Application to Binary Regression," *Communications in Statistics, Part A—Theory and Methods,* 19, 4569–4586.

Carroll, R. J., Spiegelman, C., Lan, K. K. G., Bailey, K. T., and Abbott, R. D. (1984), "On Errors in Variables for Binary Regression Models," *Biometrika,* 71, 19–26.

Carroll, R. J., and Stefanski, L. A. (1990), "Approximate Quasi-Likelihood Estimation in Models With Surrogate Predictors," *Journal of the American Statistical Association,* 85, 652–663.

Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society,* Ser. B, 53, 573–585.

Chen, T. T. (1980), "A Review of Methods for Misclassified Categorical Data in Epidemilogy," *Statistics in Medicine,* 8, 1095–1106.

Cornfield, J. (1951), "A Method of Estimating Comparable Rates From Clinical Data," *Journal of the National Cancer Institute,* 11, 1269–1275.

Dahm, P. F., Gail, M., Rosenberg, P., and Pee, D. (1990), "Determining How Much the Precision of an Estimated Odds Ratio can be Improved by Obtaining Additional Surrogate Exposure Data," preprint.

Ekholm, A., and Palmgren, J. (1987), "Correction for Misclassification Using Doubly Sampled Data," *Journal of Official Statistics,* 3, 419–429.

Espeland, M. A., and Hui, S. L. (1987), "A General Approach to Analyzing Epidemiologic Data That Contains Misclassification Errors," *Biometrics,* 43, 1001–1012.

Espeland, M. A., and Odoroff, C. L. (1985), "Log-Linear Models for Doubly Sampled Categorical Data Fitted by the EM Algorithm," *Journal of the American Statistical Association,* 80, 663–670.

Farewell, V. T. (1979), "Some Results on Estimation of Logistic Models Based on Retrospective Data," *Biometrika,* 66, 27–32.

Gleser, L. J. (1990), "Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-Variables Regression Models," in *Statistical Analysis of Measurement Error Models and Application,* eds. P. J. Brown and W. A. Fuller, Providence, RI: American Mathematics Society, pp. 99–114.

Gong, G., and Samaniego, F. (1981), "Pseudo-Maximum Likelihood Estimation: Theory and Applications," *The Annals of Statistics,* 9, 861–869.

Greenland, S. (1988a), "Variance Estimation for Epidemiologic Effect Estimates Under Misclassification," *Statistics in Medicine,* 7, 745–757.

——— (1988b), "Statistical Uncertainty Due to Misclassification: Implications for Validation Substudies," *Journal of Clinical Epidemiology,* 12, 1167–1174.

Greenland, S., and Kleinbaum, D. G. (1983), "Correcting for Misclassification in Two-Way Tables and Pair-Matching Studies," *International Journal of Epidemiology,* 12, 93–97.

Hildesheim, A., Mann, V., Brinton, L. A., Szklo, M., Reeves, W. C., and Rawls, W. E. (1991), "Herpes Simplex Virus Type 2: A Possible Interaction

With Human Papillomavirus Types 16/18 in the Development of Invasive Cervical Cancer," International Journal of Cancer, 49, 335–340.

Hsieh, D. A., Manski, C. F., and McFadden, D. (1985), "Estimation of Response Probabilities From Augmented Retrospective Observations," *Journal of the American Statistical Association,* 80, 651–662.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data,* New York: John Wiley.

Mantel, N. (1973), "Synthetic Retrospective Studies and Related Topics," *Biometrics,* 29, 470–486.

Michalek, J. E., and Tripathi, R. C. (1980), "The Effect of Errors in Diagnosis and Measurement on the Estimation of the Probability of an Event," *Journal of the American Statistical Association,* 75, 713–721.

Nero, A. V., Schwehr, M. B., and Nazaroff, W. M. (1986), "Distribution of Airborne Radon-222 Concentrations in U.S. Homes," *Science,* 234, 992–997.

Pepe, M. S., and Fleming, T. R. (1991), "A Nonparametric Method for Dealing With Mismeasured Covariate Data," *Journal of the American Statistical Association,* 86, 108–113.

Pierce, D. A. (1982), "The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics," *The Annals of Statistics,* 10, 475–487.

Pierce, D. A., Stram, D. O., Vaeth, M., and Schafer, D. W. (1992), "The Errors in Variables Problem: Considerations Provided by Radiation Dose-Response Analyses of the A-Bomb Survivor Data," *Journal of the American Statistical Association*, 87, 351–359.

Prentice, R. L. (1976), "Use of the Logistic Model in Retrospective Studies," *Biometrics,* 32, 599–606.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika,* 66, 403–411.

Randles, R. (1982), "On the Asymptotic Normality of Statistics With Estimated Parameters," *The Annals of Statistics,* 10, 462–474.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error," *Statistics in Medicine,* 8, 1051–1070.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Measurement Error: The Case of Multiple Covariates Measured With Error," *American Journal of Epidemiology,* 132, 734–745.

Satten, G. A., and Kupper, L. L. (in press), "Inferences About Exposure-Disease Association Using Probability of Exposure Information," *Journal of the American Statistical Association.*

Schafer, D. W. (1987), "Covariate Measurement Error in Generalized Linear Models," *Biometrika,* 74, 385-391.

Spall, J. C. (1989), "Effect of Imprecisely Known Nuisance Parameters on Estimates of Primary Parameters," *Communications in Statistics, Part A—Theory and Methods,* 18, 219-237.

Wang, C. Y., Wang, S. and Carroll, R. J. (1992). Likelihoods in Case-Control Studies With Nondifferential Measurement Error. Preprint.

Whittemore, A. S. (1989), "Errors in Variables Regression Using Stein Estimates," *American Statistician,* 43, 226–228.

Bias Analysis and SIMEX Approach in Generalized Linear Mixed Measurement Error Models

Naisyin WANG, Xihong LIN, Roberto G. GUTIERREZ, and Raymond J. CARROLL

We consider generalized linear mixed models (GLMMs) for clustered data when one of the predictors is measured with error. When the measurement error is additive and normally distributed and the error-prone predictor is itself normally distributed, we show that the observed data also follow a GLMM but with a different fixed effects structure from the original model, a different and more complex random effects structure, and restrictions on the parameters. This characterization enables us to compute the biases that result in common GLMMs when one ignores measurement error. For instance, in one common situation the biases in parameter estimates become larger as the number of observations within a cluster increases, both for regression coefficients and for variance components. Parameter estimation is described using the SIMEX method, a relatively new functional method that makes no assumptions about the structure of the unobservable predictors. Simulations and an example illustrate the results.

KEY WORDS: Asymptotic bias; Corrected penalized quasi-likelihood; Measurement error; Random effects; Variance components.

1. INTRODUCTION

Correlated data are frequently observed in various studies, such as longitudinal studies, clinical trials, and familial studies. Generalized linear mixed models (GLMMs) have become increasingly popular for analyzing such correlated and overdispersed data (see Breslow and Clayton 1993 for examples). A potential difficulty in making inference in GLMMs is that a full-likelihood analysis is burdened by often intractable numerical integration (although see McCulloch 1997 for Monte Carlo computation). Hence various approximate and Bayesian inference procedures have been proposed. The approximations include Laplace's approximations (Breslow and Lin 1995; Liu and Pierce 1993), penalized quasi-likelihood (PQL) (Breslow and Clayton 1993; Schall 1991), and corrected penalized quasi-likelihood (CPQL) (Lin and Breslow 1996). The Bayesian procedures include EM-type algorithms (Stiratelli, Laird, and Ware 1984) and the Gibbs sampler (Zeger and Karim 1991).

A common problem for analyzing correlated data is the presence of covariate measurement error. For example, it has been well documented in the literature that covariates such as blood pressure (Carroll, Ruppert, and Stefanski 1995), urinary sodium chloride (USC) level (Wang, Carroll, and Liang 1996), and exposure to pollutants (Tosteson, Stefanski, and Schafer 1989) are often subject to measurement error. In a longitudinal hypertension study, a patient's hypertension status may vary from one hospital visit to another due to different average USC levels prior to the hospital visit. A child's respiratory status may change from time to time depending on different amounts of pollutants (e.g., NO_2 or ozone) to which the child is exposed at different times. In the Framingham Heart Study data that we examine in Section 8, a binary outcome for the presence or absence of left ventricular hypertrophy (LVH) was observed every 2 years in an 8-year period for 75 coronary heart disease patients. The primary interest was to study the association between the risk of LVH and the time-varying covariate systolic blood pressure (SBP), after adjusting for other covariates including baseline age, smoking status, body mass index, and exam number (1–4). Because it is appropriate to assume biologically that the risk of LVH depends on the average SBP prior to the exam rather than on the SBP measured at the exam, one needs to model the measurement error in SBP while accounting for correlation among multiple observations measured repeatedly over time for each patient.

The problem of measurement error with independent observations has a vast literature in linear models (Fuller 1987) and a growing literature in generalized linear models and other nonlinear models (Carroll et al. 1995). Generally, the literature distinguishes between *functional* modeling, in which nothing is assumed about the unobserved predictors, and *structural* modeling, in which specific assumptions are made about the distributional structure of these unobserved predictors. The effects of measurement error on the analysis of clustered data and ways to correct for these effects are not well understood, however.

In this article, we propose a new class of models, generalized linear mixed measurement error models (GLMMeMs), which model the correlation and the measurement error simultaneously (Sec. 2). We explore GLMMeMs from two directions: bias analysis and functional inference using the SIMEX method (Cook and Stefanski 1994). To illustrate the fundamental impact of measurement error and our pri-

Naisyin Wang is Assistant Professor of Statistics and Raymond J. Carroll is University Distinguished Professor, Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843. Xihong Lin is Assistant Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109. Roberto G. Gutierrez is Assistant Professor, Department of Statistical Science, Southern Methodist University, Dallas, TX 75275. The research of Wang, Gutierrez, and Carroll was supported by National Cancer Institute grant CA-57030. Wang's research was also funded by National Science Foundation grant DMS-9423247 and by National Cancer Institute grant CA-74552. Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award. Lin's research was funded in part by National Cancer Institute grant R29 CA-76404. The authors gratefully acknowledge the editor, associate editor, and two referees for their helpful comments and suggestions that substantially improved the article.

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

mary findings, we concentrate on a simple but representative GLMMeM in our bias calculations (see Sec. 2.2); however, the proposed SIMEX method is applicable to the general GLMMeMs. In Sections 3 and 4 we study the bias in parameter estimation when the measurement error is not properly taken into account. The work here is facilitated by our showing that a GLMMeM can be viewed as a GLMM, with the same link function but with different fixed-effects and random-effects structures. This characterization enables us to compute the biases in parameter estimates resulting from ignoring measurement error. The bias analysis results are illustrated using several common GLMMs, including linear, logistic, and Poisson mixed models. The directions of the biases are complex and sometimes counterintuitive. For example, we show that in a particular but common setup, the biases in parameter estimates increase with the cluster size.

The description of the GLMMeM brings out the fact that likelihood estimation in this context requires the specification of a cluster-specific joint distribution for the unobserved covariates. Just as in the ordinary generalized linear model (GLM), concerns about robustness to distributional assumptions arise, but in GLMMeMs there are additional robustness concerns with respect to the covariance structure of the unobservables. In Section 5 in a special case we compute the biases resulting from a maximum likelihood analysis that accounts for measurement error but incorrectly specifies the within-cluster covariance structure of the unobservable.

In Section 6 we pursue functional estimation of regression coefficients and variance components and investigate the SIMEX procedure of Cook and Stefanski (1994). We also point out that the naive regression calibration approach often yields inconsistent estimates of certain parameters in GLMMeMs. We provide numerical results of a simulation and an example in Sections 7 and 8, and concluding remarks in Section 9.

2. THE GENERALIZED LINEAR MIXED MEASUREMENT ERROR MODEL

2.1 The General Model

Suppose that the data are obtained from m independent clusters with outcome variable Y_{ij}, unobserved true covariate $\mathbf{X}_{ij}(p_1 \times 1)$, observed $\mathbf{X}_{ij}$-related covariate $\mathbf{W}_{ij}$, and other observed covariates $\mathbf{Z}_{ij}(p_2 \times 1)$ and $\mathbf{A}_{ij}(q \times 1)$, where $i = 1, \ldots, m$ identifies the cluster; $j = 1, \ldots, n_i$ identifies subjects within clusters; and $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ and $\mathbf{A}_{ij}$ are associated with the fixed effects and random effects. That is, we consider the situations where the error-prone covariates $\mathbf{X}_{ij}$ are associated with fixed coefficients. Given the covariates $\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{A}_{ij}$ and an unobserved $q \times 1$ random-effects vector $\mathbf{b}_i$, the observations Y_{ij} in the ith cluster are assumed to be independent with means $\mu_{ij,x}^{\mathbf{b}_i}$ and variances $\phi\kappa_{ij}^{-1}v(\mu_{ij,x}^{\mathbf{b}_i})$, where ϕ is a scale parameter, κ_{ij} is a prior weight (e.g., binomial denominator), and $v(\cdot)$ is a variance function. The GLMM of $\mathbf{Y}$ given $\mathbf{X}$ and $\mathbf{Z}$ is constructed by assuming that the conditional mean $\mu_{ij,x}^{\mathbf{b}_i}$ is related to $\mathbf{X}_{ij}, \mathbf{Z}_{ij}$, and $\mathbf{A}_{ij}$ through a GLM,

$$g(\mu_{ij,x}^{\mathbf{b}_i}) = \beta_0 + \mathbf{X}_{ij}^T\boldsymbol{\beta}_x + \mathbf{Z}_{ij}^T\boldsymbol{\beta}_z + \mathbf{A}_{ij}^T\mathbf{b}_i, \tag{1}$$

where $g(\cdot)$ is a monotonic differentiable link function, the random effects $\mathbf{b}_i$ are independent of the covariates and are independent N(0, $\mathbf{D}(\boldsymbol{\theta})$), and $\boldsymbol{\theta}$ is an $l \times 1$ vector of variance components. Model (1) allows flexible correlation structures by assuming appropriate design matrix $\mathbf{A}_{ij}$ and covariance matrix $\mathbf{D}$ of the random effects $\mathbf{b}_i$.

Define $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T, \mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})^T$, and $\mathbf{Z}_i$ and $\mathbf{W}_i$ similarly. The integrated quasi-likelihood of $\mathbf{Y}_i$ given $(\mathbf{X}_i, \mathbf{Z}_i)$ in the ith cluster is

$$L_i(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\theta}) \propto |\mathbf{D}|^{-1/2} \times \int \exp\left\{\sum_{j=1}^{n_i} l_{ij}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i) - \frac{1}{2}\mathbf{b}_i^T\mathbf{D}^{-1}\mathbf{b}_i\right\} d\mathbf{b}_i, \tag{2}$$

where $l_{ij}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i) \propto \int_{Y_{ij}}^{\mu_{ij,x}^{\mathbf{b}_i}} \kappa_{ij}(Y_{ij} - u)/\{\phi v(u)\}\, du$ denotes the conditional log quasi-likelihood of Y_{ij} given $(\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)$ (see Breslow and Clayton 1993, eq. 2).

The model is completed by adding the measurement error structure. The most convenient structure is additive error, so that

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{U}_{ij}, \tag{3}$$

where the $\mathbf{U}_{ij}$ are independent of the $\mathbf{X}_{ij}$ and are independent N(0, $\boldsymbol{\Sigma}_{uu}$). When $\mathbf{W}$ and $\mathbf{X}$ are scalar, we write the measurement variance $\boldsymbol{\Sigma}_{uu}$ simply as σ_u^2. Model (3) is essential to our analytical closed-form bias calculations in Sections 3–5 but not for numerical bias calculations, estimation, and inference. Neither the independence of the $\mathbf{U}_{ij}$ nor the additivity is required (see Sec. A.5 in the Appendix). The joint integrated quasi-likelihood in the ith cluster is

$$L_i(\mathbf{Y}_i, \mathbf{W}_i|\mathbf{Z}_i) = \int L_i(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i)L_i(\mathbf{W}_i|\mathbf{X}_i, \mathbf{Z}_i)L_i(\mathbf{X}_i|\mathbf{Z}_i)\, d\mathbf{X}_i, \tag{4}$$

where $L_i(\mathbf{X}_i|\mathbf{Z}_i)$ is the likelihood function of $\mathbf{X}_i$ (so far unspecified) and $L_i(\mathbf{W}_i|\mathbf{X}_i, \mathbf{Z}_i)$ is the error distribution, which is often assumed to be independent of $\mathbf{Z}_i$ as in (3). The dependence of the quasi-likelihood on the within-cluster conditional distribution of the unobserved $\mathbf{X}$'s leads to issues of model robustness.

A special case is instructive. Suppose that X_{ij} is scalar and that $\mathbf{X}_i = \mathbf{1}_i\eta_0 + \mathbf{Z}_i\boldsymbol{\eta}_z + \mathbf{e}_{xi}$, where $\mathbf{1}_i$ is an $n_i \times 1$ vector of 1s and $\mathbf{e}_{xi}$ given $\mathbf{Z}_i$ is normally distributed with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{xxi}$. Denote an $n_i \times n_i$ identity matrix by $\mathbf{I}_i$ and the reliability matrix by $\boldsymbol{\Lambda}_i = \boldsymbol{\Sigma}_{xxi}\{\boldsymbol{\Sigma}_{xxi} + \text{cov}(\mathbf{U}_i)\}^{-1}$. Note that $\boldsymbol{\Sigma}_{xxi}$ and $\boldsymbol{\Lambda}_i$ depend on i through their dimensions n_i, but the set of unknown parameters in $\boldsymbol{\Sigma}_{xxi}$ and $\boldsymbol{\Lambda}_i$ does not depend on i. We can write $\mathbf{X}_i = (\mathbf{I}_i - \boldsymbol{\Lambda}_i)(\mathbf{1}_i\eta_0 + \mathbf{Z}_i\boldsymbol{\eta}_z) + \boldsymbol{\Lambda}_i\mathbf{W}_i + \mathbf{b}_i^*$, where the sum of the first two terms corresponds to $E(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ and $\mathbf{b}_i^* = \mathbf{X}_i - E(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i) = (\mathbf{I}_i - \boldsymbol{\Lambda}_i)\mathbf{e}_{xi} - \boldsymbol{\Lambda}_i\mathbf{U}_i$ is N$\{0, (\mathbf{I}_i - \boldsymbol{\Lambda}_i)\boldsymbol{\Sigma}_{xxi}\}$ and is independent of $\mathbf{b}_i$ and $\mathbf{W}_i$. The independence between $\mathbf{b}_i^*$ and $\mathbf{W}_i$ can be easily checked by showing $\mathrm{cov}(\mathbf{b}_i^*, \mathbf{W}_i) = 0$. The expression of $\mathbf{X}_i$ indicates that the jth component of $\mathbf{X}_i$ can be written as, say,

$$X_{ij} = \alpha_{0j} + \boldsymbol{\eta}_z^T \mathbf{Z}_i^T \boldsymbol{\alpha}_{zj} + \mathbf{W}_i^T \boldsymbol{\alpha}_{wj} + \mathbf{C}_{ij}^T \mathbf{b}_i^*. \qquad (5)$$

Because $\mathbf{b}_i^* = \mathbf{X}_i - E(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ and $\mathbf{Y}_i$ is independent of $\mathbf{W}_i$ given $\mathbf{X}_i, (\mathbf{Y}_i|\mathbf{W}_i, \mathbf{Z}_i, \mathbf{b}_i^*, \mathbf{b}_i)$ has the same distribution as $(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$ and follows the same conditional generalized linear model (1), except that $\mathbf{X}_i$ is replaced by (5). In other words, the observed data $(\mathbf{Y}_i|\mathbf{W}_i, \mathbf{Z}_i)$ follow a GLMM as follows:

$$\begin{aligned} g(\mu_{ij,w}^{\tilde{\mathbf{b}}_i}) &= (\beta_0 + \alpha_{0j}\beta_x) + \mathbf{W}_i^T \boldsymbol{\alpha}_{wj}\beta_x \\ &\quad + (\boldsymbol{\eta}_z^T \mathbf{Z}_i^T \boldsymbol{\alpha}_{zj}\beta_x + \mathbf{Z}_{ij}^T \boldsymbol{\beta}_z) + (\mathbf{A}_{ij}^T \mathbf{b}_i + \mathbf{C}_{ij}^T \beta_x \mathbf{b}_i^*), \qquad (6) \end{aligned}$$

where $\tilde{\mathbf{b}}_i^T = \{\mathbf{b}_i^T, \mathbf{b}_i^{*^T}\}$ denotes a vector of the new random effects.

Equation (6) shows that ignoring the measurement error may result in misspecifying both the fixed-effects and random-effects structures. The cluster size n_i, which is implicit in the reliability matrix $\boldsymbol{\Lambda}_i$, also plays an important role in the asymptotic bias in maximum likelihood estimator (MLE) under a misspecified model as $m \to \infty$. A general approach for bias analysis is to maximize the probability limit of the log quasi-likelihood of the misspecified model, which is equal to its expectation, when model (6) is true. Calculations of this expectation often involve numerical integration. Appendix Section A.5 gives a brief discussion on numerical integration techniques using Gaussian–Hermite quadrature. When the misspecified model and the observed data $(\mathbf{Y}_i|\mathbf{W}_i, \mathbf{Z}_i)$ models follow the same type of GLMMs, bias calculations can be greatly simplified by using the correspondence of their mean models (see Sec. 3). The foregoing general bias calculation strategy applies to an arbitrary GLMMeM. A more complicated GLMMeM structure requires no extra procedures than the ones used for a simple structure, except that the results will be more complicated. We hence consider simple GLMMeMs in our bias analysis to show the fundamental impact of measurement error and our primary findings, and to demonstrate the basic techniques used in bias calculations.

2.2 Specific Models Considered in Bias Analysis

In the bias analyses in Sections 3–5, for simplicity we assume that $n_i = n$, the $\mathbf{X}_{ij}$ are scalar, and simple random intercept GLMM $g(\mu_{ij,x}^{b_i}) = \beta_0 + \beta_x X_{ij} + b_i$, where the b_i are independent N(0, θ). We distinguish between two cases depending on the likelihood structure $L_i(\mathbf{X}_i)$ of $\mathbf{X}_i$. The *homogeneous* case occurs when the X_{ij} are marginally independent and have the same distribution irrespective of the cluster, so that

$$X_{ij} = \mu_x + e_{ij}, \qquad (7)$$

where the e_{ij} are independent N(0, σ_x^2). In the *heterogeneous* case, conditional on a cluster random effect a_i, the distributions of the X_{ij} differ from cluster to cluster. In the version that we study here theoretically, we assume that the conditional cluster means differ, so

$$X_{ij} = \mu_x + a_i + e_{ij}, \qquad (8)$$

where the a_i, independent of the model random effect b_i, are independent N(0, $\sigma_{x\mu}^2$).

Although the two models that we consider here have simple structures, the bias calculation techniques used in Sections 3–5 are applicable to more complicated cases. Specifically, as indicated in Section 2.1, we can accommodate the covariates measured without error $\mathbf{Z}_i$ by treating $\mathbf{Z}_i$ as fixed and further allow for multivariate $\mathbf{X}_{ij}$ and a more complicated structure of the random effects $\mathbf{b}_i$. For example, to accommodate multivariate $\mathbf{X}_{ij}$, we can define $\mathbf{X}_i^T = (\mathbf{X}_{i1}^T, \ldots, \mathbf{X}_{in}^T)$, define $\mathbf{W}_i$ analogously, and modify (6).

To appreciate the practical differences between the homogeneous and heterogeneous models, we consider an ozone exposure example. When the subjects are from the same site, one distribution can be used to describe the behavior of the short-term average ozone exposures for all subjects, and the homogeneous model is appropriate. In this case the variations in ozone exposure may be mainly seasonal. On the other hand, if these subjects are from different neighborhoods, then the heterogeneous model, which accommodates cross-cluster variation, should be used. Clearly, a homogeneous model is a special case of the heterogeneous model ($\sigma_{x\mu}^2 = 0$), and this seems to indicate that one should consider only the heterogeneous model. However, assuming a heterogeneous model while the homogeneous model holds results in estimators with unnecessarily large variance.

The next three sections are devoted to studying the asymptotic biases in regression coefficients and variance component estimators under three misspecified models. To facilitate the bias analysis, it is helpful to rewrite (6) in the special cases under consideration. The calculations outlined in Appendix Section A.1 show that the observed data under the heterogeneous GLMMeM satisfy

$$g(\mu_{ij,w}^{\tilde{\mathbf{b}}_i}) = \beta_0' + \beta_x' W_{ij} + \beta_2' \bar{W}_{i\cdot} + b_i' + b_{ij}'', \qquad (9)$$

where

$$\beta_0' = \beta_0 + (1 - \lambda)\tilde{\lambda}\mu_x\beta_x,$$

$$\beta_x' = \lambda\beta_x,$$

$$\beta_2' = (1 - \lambda)(1 - \tilde{\lambda})\beta_x,$$

$$\tilde{\lambda} = (\sigma_x^2 + \sigma_u^2)/(\sigma_x^2 + \sigma_u^2 + n\sigma_{x\mu}^2),$$

$$\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2),$$

$$\bar{W}_{i\cdot} = n^{-1}\sum_{j=1}^{n} W_{ij},$$

and

$$\tilde{\mathbf{b}}_i = (b_i', b_{i1}'', \ldots, b_{in}'')^T.$$

The random effects b_i' and b_{ij}'' are independent of W_{ij}, and are mutually independent and distributed as N(0, θ') and N(0, γ), where $\theta' = \theta + (1 - \lambda)(1 - \tilde{\lambda})\beta_x^2\sigma_u^2/n$ and $\gamma = \lambda\sigma_u^2\beta_x^2$. The exact expressions of b_i' and b_{ij}'' are given in Appendix Section A.1.

3. BIAS IN THE NAIVE ESTIMATOR UNDER THE HOMOGENEOUS MODEL

In this section we study the asymptotic biases in naive estimators of regression coefficients and variance component when the homogeneous model (7) holds. The naive estimator is defined as the estimator under the model that ignores the measurement error,

$$g(\mu_{ij,w}^{b_i}) = \beta_0 + \beta_w W_{ij} + b_i. \qquad (10)$$

From (9), the homogeneous model has $\tilde{\lambda} = 1, \beta_2' = 0$, and $\theta = \theta'$ and thus can be written as

$$g(\mu_{ij,w}^{\tilde{\mathbf{b}}_i}) = \beta_0' + \beta_x' W_{ij} + b_i + b_{ij}'', \qquad (11)$$

where the b_i are independent N(0, θ) and the b_{ij}'' are independent N(0, γ).

Because conditional on the b_i and the W_{ij}, the model for the observed data (11) corresponds to an overdispersed GLM, the bias analysis of the naive estimators can proceed by comparing the conditional mean $\mathbf{E}(Y_{ij}|W_{ij}, b_i)$ and the conditional variance $\text{var}(Y_{ij}|W_{ij}, b_i)$ under the naive model (10) with those under the homogeneous model (11). Note that this conditional mean and variance under the homogeneous model can be easily obtained by integrating out b_{ij}'' from (11).

Although the naive model correctly assumes that the Y_{ij} are independent conditional on W_{ij} and b_i, it may misspecify both the mean and variance conditional on W_{ij} and b_i. When there is a correspondence between $E(Y_{ij}|W_{ij}, b_i)$ and $\text{var}(Y_{ij}|W_{ij}, b_i)$ under these two models, they follow the same type of GLMM, and hence the asymptotic biases in regression coefficients and the variance component can be easily derived. Otherwise, the calculations are often difficult, and closed-form expressions for the biases are not always available.

3.1 The Linear Mixed Model for Gaussian Data

It can be easily shown that in the linear mixed model, the observed data also follow a linear random intercept model, with the terms b_{ij}'' in (11) absorbed into the within-cluster variance in the responses. The naive estimators hence asymptotically converge to $\beta_{0,\text{naive}} = \beta_0 + (1 - \lambda)\mu_x\beta_x, \beta_{x,\text{naive}} = \lambda\beta_x$, and $\theta_{\text{naive}} = \theta$. Thus the naive estimators of the regression coefficients are asymptotically biased in a usual way, but the naive estimator of the variance component is asymptotically unbiased.

3.2 The Probit, Logistic, and Log-linear Mixed Models for Binary Data

When $\mathbf{Y}_i$ given $\mathbf{X}_i$ follows a probit random intercept model, so too do the observed data $\mathbf{Y}_i$ given $\mathbf{W}_i$. Let $\tau = (1 + \gamma)^{1/2} = (1 + \lambda\sigma_u^2\beta_x^2)^{1/2}$; the derivations outlined in Appendix Section A.2 show that $\beta_{0,\text{naive}} = \{\beta_0 + (1 - \lambda)\mu_x\beta_x\}/\tau, \beta_{x,\text{naive}} = \lambda\beta_x/\tau$, and $\theta_{\text{naive}} = \theta/\tau^2$. These results indicate that the naive estimators of both β_x and θ are asymptotically biased toward 0. For the logistic model, exact closed-form results are not available. But by approximating the standard logistic distribution function by the distribution function of a mean 0 normal random variable that has standard deviation $c = 15\pi/(16\sqrt{3}) \approx 1.7$, similar calculations show that the same bias expressions obtain, but with τ replaced by $\tau^* = (1 + \lambda\sigma_u^2\beta_x^2/c^2)^{1/2}$.

The log-linear model assumes a log link function and is useful in ecological studies, where the disease rates may be low. Using the identity $\int \exp(a + t)\, d\Phi(t/\sigma) = \exp(a + \sigma^2/2)$ for any constants a and σ, where $\Phi(\cdot)$ denotes the cumulative probability function of a standard normal random variable, similar calculations to those given in Appendix Section A.2 show that $(\mathbf{Y}_i|\mathbf{W}_i)$ also follows a log-linear random intercept model and that the naive estimators converge to $\beta_{0,\text{naive}} = \beta_0 + (1 - \lambda)\mu_x\beta_x + \gamma/2, \beta_{x,\text{naive}} = \lambda\beta_x$, and $\theta_{\text{naive}} = \theta$. These results, except for the intercept, agree with those in the linear mixed model.

3.3 The Poisson Mixed Model for Count Data

Define $\theta^* = \theta + \gamma, d_{ij} = \exp(\theta/2 + \beta_0 + \beta_x X_{ij})$, and $c_{ij} = \exp(\theta^*/2 + \beta_0' + \beta_x' W_{ij})$, where β_0' and β_x' are defined in Section 2.2 with $\tilde{\lambda} = 1$. Then the conditional mean and covariance of $\mathbf{Y}_i$ given $\mathbf{W}_i$ under the Poisson mixed model are $E(Y_{ij}|W_{ij}) = c_{ij}$, $\text{var}(Y_{ij}|W_{ij}) = c_{ij} + c_{ij}^2\{\exp(\theta^*) - 1\}$, and $\text{cov}(Y_{ij}, Y_{ik}|W_{ij}, W_{ik}) = c_{ij}c_{ik}\{\exp(\theta^* - \gamma) - 1\}$, and those of $\mathbf{Y}_i$ given $\mathbf{X}_i$ follow the same structure except that c_{ij} is replaced by d_{ij} and that $\text{cov}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = d_{ij}d_{ik}\{\exp(\theta) - 1\}$. The lack of correspondence between the two covariances reveals that $(\mathbf{Y}|\mathbf{W})$ and $(\mathbf{Y}|\mathbf{X})$ do not follow the same GLMM structure. Thus the approach based on the conditional mean correspondence is not applicable.

Using the techniques of reparameterization and applying the properties of sufficient statistics and maximum likelihood estimators, the calculations outlined in Appendix Section A.3 show that $\beta_{0,\text{naive}} = \beta_0' + (\theta + \gamma - \theta_{\text{naive}})/2, \beta_{x,\text{naive}} = \lambda\beta_x$, and

$$\theta_{\text{naive}} = \theta + \log\left\{\frac{(n-1) + \exp(\beta_x^2\sigma_x^2)}{(n-1) + \exp(\lambda\beta_x^2\sigma_x^2)}\right\}. \qquad (12)$$

Equation (12) suggests that the naive estimator θ_{naive} overestimates θ and that its bias depends on the cluster size n and decreases monotonically to 0 as the cluster size $n \uparrow \infty$.

4. BIAS IN THE NAIVE ESTIMATOR UNDER THE HETEROGENEOUS MODEL

In this section we study the asymptotic bias of the naive estimator when the heterogeneous model (8) holds. We see from (9) that the heterogeneous conditional quasi-likelihood $(\mathbf{Y}|\mathbf{W})$ also corresponds to a GLMM, with the same random-effects structure as in the homogeneous case but with an additional fixed effect—namely, the within-cluster mean of the W's. When the heterogeneous model is true, a comparison of (9) and (11) suggests that the naive model (10) misspecifies both the fixed-effects and the random-effects structures. This double misspecification makes the bias analysis much more complicated, and closed-form solutions are often not available.

Nonetheless, we can calculate the asymptotic bias in the naive estimator when the cluster size n goes to infinity. Specifically, our calculations show that the bias in the naive estimator becomes the same as in the homogeneous case when $n \to \infty$, except that θ is replaced by $\theta + (1 - \lambda)^2\sigma^2_{x\mu}\beta^2_x$. This can be easily shown by noting that for the heterogeneous model (9), as $n \to \infty$, we have $\tilde{\lambda} \to 0, \beta'_2 \to (1 - \lambda)\beta_x, \theta' \to \theta$, and $\bar{W}_{i\cdot} = \mu_x + a_i + \bar{U}_{i\cdot} + \bar{e}_{i\cdot} \to \mu_x + a_i$. Consequently, the heterogeneous GLMM in (9) becomes

$$g(\mu^{\tilde{\mathbf{b}}_i}_{ij,w}) = \beta''_0 + \beta'_x W_{ij} + b'_{*i} + b''_{ij}, \tag{13}$$

where $\beta''_0 = \beta_0 + (1 - \lambda)\mu_x\beta_x, b'_{*i} = b_i + (1 - \lambda)\beta_x a_i$ now follows N(0, $\theta + (1-\lambda)^2\sigma^2_{x\mu}\beta^2_x$), and b''_{ij} stays the same. Because within a cluster (13) has the same structure as (11), and because the cluster size is infinite, the maximum likelihood estimated within-cluster intercept (and slope) must have the same form, except that b_i is replaced by b'_{*i} and hence θ is replaced by $\theta + (1 - \lambda)^2\sigma^2_{x\mu}\beta^2_x$. In the next two sections we use analytic and numerical approaches to study the asymptotic bias in the naive estimator in the linear and logistic mixed models when the cluster size n is finite.

4.1 The Linear Mixed Model

Denote the residual variance by σ^2, which corresponds to the scale parameter ϕ in Section 2. Let the probability limits of the naive estimators as $m \to \infty$ be $\boldsymbol{\beta}_{\text{naive}} = (\beta_{0,\text{naive}}, \beta_{x,\text{naive}})^T$ and $\boldsymbol{\vartheta} = (\theta_{\text{naive}}, \sigma^2_{\text{naive}})$. Then $\boldsymbol{\beta}_{\text{naive}}$ and $\boldsymbol{\vartheta}$ are solutions of the following equations, which are the probability limits of the linear mixed model score equations as $m \to \infty$ (Harville 1977):

$$E\{\mathcal{W}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathcal{W}\boldsymbol{\beta}_{\text{naive}})\} = 0 \tag{14}$$

and

$$\frac{1}{2}\left\{E(\mathbf{Y} - \mathcal{W}\boldsymbol{\beta}_{\text{naive}})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\vartheta}_j}\right.$$
$$\left.\times\ \mathbf{V}^{-1}(\mathbf{Y} - \mathcal{W}\boldsymbol{\beta}_{\text{naive}}) - \text{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\vartheta}_j}\right)\right\} = 0, \tag{15}$$

where $\mathcal{W} = (\mathbf{1}, \mathbf{W}), \mathbf{V} = \sigma^2_{\text{naive}}\mathbf{I} + \theta_{\text{naive}}\mathbf{J}, \mathbf{J}$ is an $n \times n$ matrix of 1s, and the expectations are taken with respect to both $\mathbf{W}$ and $\mathbf{Y}$. For simplicity, here we have removed the subscripts i of $\mathbf{Y}_i$ and $\mathbf{W}_i$, because they are identically distributed. Repeatedly using the equality

$$E(\mathbf{X}^T\mathbf{B}\mathbf{X}) = \text{tr}(\mathbf{B}\mathbf{V}_x) + \boldsymbol{\mu}^T_x\mathbf{B}\boldsymbol{\mu}_x, \tag{16}$$

which holds for any positive definite matrix $\mathbf{B}$ and any random vector $\mathbf{X}$ following N($\boldsymbol{\mu}_x, \mathbf{V}_x$), the calculations outlined in Appendix Section A.4 yield $\beta_{0,\text{naive}} = \beta_0 + (1 - \lambda_*)\mu_x\beta_x, \beta_{x,\text{naive}} = \lambda_*\beta_x$,

$$\theta_{\text{naive}} = \theta + (1 - \lambda_*)^2\sigma^2_{x\mu}\beta^2_x,$$

and

$$\sigma^2_{\text{naive}} = \sigma^2 + \{(1 - \lambda_*)^2\sigma^2_x + \lambda^2_*\sigma^2_u\}\beta^2_x, \tag{17}$$

where

$$\lambda_* = \frac{\sigma^2_{x\mu} + \sigma^2_x\{1 + (n - 1)\theta_{\text{naive}}/\sigma^2_{\text{naive}}\}}{\sigma^2_{x\mu} + (\sigma^2_x + \sigma^2_u)\{1 + (n - 1)\theta_{\text{naive}}/\sigma^2_{\text{naive}}\}}. \tag{18}$$

A closed-form solution to (17) and (18) does not seem to be available. However, defining $\rho = \theta_{\text{naive}}/\sigma^2_{\text{naive}}$ using (17) and writing (17) and (18) as joint equations of (λ_*, ρ), one can easily solve them numerically.

Because $\lambda \leq \lambda_* \leq 1$, the naive estimator $\beta_{x,\text{naive}}$ is still attenuated, but to a lesser extent compared to that in the homogeneous case. In contrast, θ_{naive} and σ^2_{naive} generally overestimate their true counterparts. Our theoretical calculations show that the values of $\beta_{0,\text{naive}}, \beta_{x,\text{naive}}, \theta_{\text{naive}}$, and σ^2_{naive} all depend on the cluster size n. Some tedious calculations show that $\partial\lambda_*/\partial n < 0$. Hence λ_* is a decreasing function of n, and the biases in all naive estimators except σ^2_{naive} become more serious when n increases. As $n \uparrow \infty, \lambda_* \downarrow \lambda$, and we obtain the results in the last paragraph before Section 4.1.

In **Figures 1a** and **1b,** we numerically evaluate the biases in $\beta_{x,\text{naive}}$ and θ_{naive} for σ^2_u varying between 0 and 1. For each plot, we obtained four curves corresponding to $n = 2, 5, 10$, and ∞. The parameter configurations were $\beta_0 = 0, \beta_x = 2, \theta = .5, \sigma^2 = 1, \sigma^2_{x\mu} = 1.5$, and $\mu_x = 0, \sigma^2_x = 1$. The relative bias is defined as the bias of a parameter divided by its true value. Note the major features of the plot—the naive estimator of β_x is attenuated, the naive estimator θ_{naive} overestimates θ, and the biases in $\beta_{x,\text{naive}}$ and θ_{naive} increase as n increases.

4.2 The Logistic Mixed Model

Our interest here is in calculating the bias in naive estimator in the logistic mixed model when n is finite and the heterogeneous model is true. Because there is no closed-form expression for these biases, even in the probit case, we calculated the asymptotic bias by numerically maximizing the probability limit of the log-likelihood of the naive model, which is calculated by its expectation, when the heterogeneous model is true. We briefly describe our numerical methods in Appendix Section A.5. The parameter configurations used in our numerical calculations are identical to those used in Section 4.1. As shown in **Figures 1c** and **1d,** the naive estimate $\beta_{x,\text{naive}}$ underestimates β_x as usual, and its bias becomes larger as σ^2_u and n increase, whereas the bias in θ_{naive} is no longer monotonic in σ^2_u, and its direction depends on σ^2_u and $\sigma^2_{x\mu}$. For example, θ_{naive} underestimates θ when σ^2_u is close to 0 and overestimates θ when σ^2_u increases. For $n = 2$, the measurement error effect on θ_{naive} is less pronounced compared to its effect on $\beta_{x,\text{naive}}$; however, there is substantial bias when $n = \infty$. This simply points out once again that cluster size is important in the bias of estimates computed by ignoring measurement error.

5. BIAS IN THE HOMOGENEOUS MAXIMUM LIKELIHOOD ESTIMATOR UNDER THE HETEROGENEOUS MODEL

In this section we study the asymptotic bias in the MLE assuming the homogeneous model (7) when the heterogeneous model (8) in fact is true. The major issue is: What happens when one accounts for measurement error in a likelihood analysis, but incorrectly models the covariance of

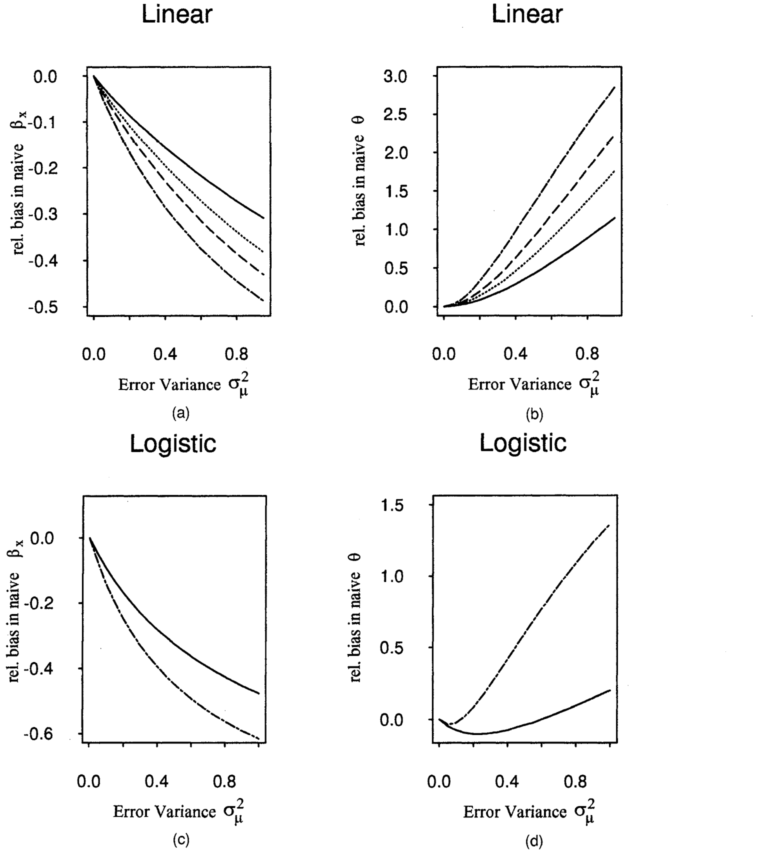


Figure 1. Asymptotic Relative Biases in Naive Estimates of β_x and θ in the Linear and Logistic Mixed Models When the Heterogeneous **X** *Model is True. The true parameter values are $\beta_0 = 0$, $\beta_x = 2$, $\theta = .5$, $\sigma^2 = 1$, $\mu_x = 0$, $\sigma_x^2 = 1$, and $\sigma_{x\mu}^2 = 1.5$. The four plots correspond to (a) relative bias in $\beta_{x,naive}$ against σ_u^2 for the linear model; (b) relative bias in θ_{naive} against σ_u^2 for the linear model; (c) relative bias in $\beta_{x,naive}$ against σ_u^2 for the logistic model; and (d) relative bias in θ_{naive} against σ_u^2 for the logistic model. The four curves in (a) and (b) are ——, $n = 2$; - - -, $n = 5$; — — —, $n = 10$; and — - —, $n = \infty$. The two curves in (c) and (d) are ——, $n = 2$ and — - —, $n = \infty$.*

the unobserved predictors? The calculations in this special case give some idea of the biases that can occur in structural modeling with an incorrectly specified structural model.

In our calculations we assume for simplicity that σ_u^2 is known. The unknown parameters under the homogeneous model are $(\beta_0, \beta_x, \theta)$ and (μ_x, σ_x^2). By writing (4) as $L_i(\mathbf{Y}_i, \mathbf{W}_i) = L_i(\mathbf{Y}_i|\mathbf{W}_i)L_i(\mathbf{W}_i)$, a comparison of the homogeneous model (11) with the heterogeneous model (9) reveals that the bias in the homogeneous MLE comes from two sources: one from misspecification of the marginal likelihood of $\mathbf{W}_i$ and the other from misspecification of the likelihood structure of $\mathbf{Y}_i$ given $\mathbf{W}_i$. It is easily seen that the former ignores the cluster-level random effect a_i. In contrast to the naive model, the homogeneous GLMM (11) assumed by the latter misspecifies only the fixed-effects structure.

Denote the asymptotic limits of the homogeneous MLEs of $(\beta_0, \beta_x, \theta)$ and $(\mu_x, \sigma_x^2, \lambda)$ when the heterogeneous model is true by $(\beta_{0,\text{hom}}, \beta_{x,\text{hom}}, \theta_{\text{hom}})$ and $(\mu_{x,\text{hom}}, \sigma_{x,\text{hom}}^2, \lambda_{\text{hom}})$. Because the $\mathbf{W}_i$ are sufficient statistics for (μ_x, σ_x^2), some calculations give $\mu_{x,\text{hom}} = \mu_x, \sigma_{x,\text{hom}}^2 = \sigma_x^2 + \sigma_{x\mu}^2$, and $\lambda_{\text{hom}} = \sigma_{x,\text{hom}}^2/(\sigma_{x,\text{hom}}^2 + \sigma_u^2) = (\sigma_x^2 + \sigma_{x\mu}^2)/(\sigma_x^2 +$

$\sigma^2_{x\mu} + \sigma^2_u$). The bias analysis of $(\beta_{0,\text{hom}}, \beta_{x,\text{hom}}, \theta_{\text{hom}})$ may then proceed by comparing the heterogeneous GLMM (9) and the homogeneous GLMM (11) with λ replaced by λ_{hom} in (11). Because the homogeneous GLMM (11) misspecifies the fixed-effects structure, such bias analyses are often difficult, and closed-form solutions may not be obtained.

But when $n \to \infty$, closed-form results are available. A comparison of the homogeneous GLMM (11) and the heterogeneous $n \to \infty$ GLMM (13) shows that the two models have the same structure. Consequently, we have $\beta_{0,\text{hom}} + (1 - \lambda_{\text{hom}})\mu_{x,\text{hom}}\beta_{x,\text{hom}} = \beta_0 + (1-\lambda)\mu_x\beta_x$, $\lambda_{\text{hom}}\beta_{x,\text{hom}} = \lambda\beta_x$, and $\theta_{\text{hom}} = \theta + (1-\lambda)^2\sigma^2_{x\mu}\beta^2_x$. Simple calculations give $\beta_{0,\text{hom}} = \beta_0 + (1-\lambda')\mu_x\beta_x$ and $\beta_{x,\text{hom}} = \lambda'\beta_x$, where $\lambda' = \lambda/\lambda_{\text{hom}}$. When the homogeneous model is true, we have $\sigma^2_{x\mu} = 0$ and $\lambda' = 1$, and the homogeneous MLEs of $(\beta_0, \beta_x, \theta)$ are hence unbiased. Because $\lambda' \le 1$, the homogeneous MLE of β_x always underestimates β_x and the homogeneous MLE of θ always overestimates θ as $n \to \infty$. Because $\lambda' \ge \lambda$, the bias in homogeneous MLE of β_x is often less than its naive counterpart as $n \to \infty$. The homogeneous MLE and the naive estimator of θ are the same as $n \to \infty$ in the linear and Poisson mixed models, whereas θ_{hom} is larger than θ_{naive} in the logistic and probit mixed models. In the next two sections we study the bias in the homogeneous MLE in the linear and logistic mixed models when the cluster size n is finite.

5.1 The Linear Mixed Model

Using (11), it can be easily shown that under the homogeneous linear mixed model, the observed data $\mathbf{Y}_i$ given $\mathbf{W}_i$ also assume a linear random intercept model. Thus we can simply use the results in Section 4.1 to calculate the asymptotic bias in homogeneous MLE. Specifically, we replace $(\beta_{0,\text{naive}}, \beta_{x,\text{naive}})$ by $(\beta_{0,\text{hom}} + (1 - \lambda_{\text{hom}})\mu_{x,\text{hom}}\beta_{x,\text{hom}}, \lambda_{\text{hom}}\beta_{x,\text{hom}})$ and replace $(\theta_{\text{naive}}, \sigma^2_{\text{naive}})$ by $(\theta_{\text{hom}}, \sigma^2_{\text{hom}} + \lambda_{\text{hom}}\sigma^2_u\beta^2_{x,\text{hom}})$ in all equations in Section 4.1. This gives

$$\beta_{0,\text{hom}} + (1 - \lambda_{\text{hom}})\mu_x\beta_{x,\text{hom}} = \beta_0 + (1 - \lambda_{**})\mu_x\beta_x,$$

$$\lambda_{\text{hom}}\beta_{x,\text{hom}} = \lambda_{**}\beta_x,$$

$$\sigma^2_{\text{hom}} + \lambda_{\text{hom}}\sigma^2_u\beta^2_{x,\text{hom}} = \sigma^2 + \{(1-\lambda_{**})^2\sigma^2_x + \lambda^2_{**}\sigma^2_u\}\beta^2_x,$$

and

$$\theta_{\text{hom}} = \theta + (1-\lambda_{**})^2\sigma^2_{x\mu}\beta^2_x,$$

where λ_{**} is equal to λ_* in (18) except that θ_{naive} and σ^2_{naive} are replaced by θ_{hom} and $\sigma^2_{\text{hom}} + \lambda_{\text{hom}}\sigma^2_u\beta^2_{x,\text{hom}}$. Using the foregoing equations for $(\theta_{\text{hom}}, \sigma^2_{\text{hom}})$, one can easily show that the value of λ_{**} is the same as that of λ_* in (18) for a given set of parameter values.

Simple calculations yield $\beta_{0,\text{hom}} = \beta_0 + (1 - \lambda'_*)\mu_x\beta_x, \beta_{x,\text{hom}} = \lambda'_*\beta_x$, and $\sigma^2_{\text{hom}} = \sigma^2 + \{(1-\lambda_*)^2\sigma^2_x + (\lambda^2_* - \lambda_*\lambda'_*)\sigma^2_u\}\beta^2_x$, where $\lambda'_* = \lambda_*/\lambda_{\text{hom}}$. Because $\lambda_* \le \lambda'_* \le 1$, the homogeneous MLE $\beta_{x,\text{hom}}$ is still attenuated, but its bias is less than that of its naive counterpart. The homogeneous MLE θ_{hom} is identical to the naive estimator θ_{naive}, and both overestimate θ. As $n \uparrow \infty$, we have $\lambda_* \downarrow \lambda$ and $\lambda'_* \downarrow \lambda'$, and hence $\beta_{x,\text{hom}} \downarrow \lambda'\beta_x$ and $\theta_{\text{hom}} \downarrow \theta + (1-\lambda)\sigma^2_{x\mu}\beta^2_x$, which agree with the general $n \to \infty$ results in the last paragraph before Section 5.1. Similar to the naive case, the biases in $\beta_{x,\text{hom}}$ and θ_{hom} increase with n.

In Figure 2 we numerically study the asymptotic biases in $\beta_{x,\text{hom}}$ and θ_{hom}. The parameter configurations and setup in Figure 2 are identical to those used in Figure 1. These figures reflect our theoretical results. Specifically, the bias in $\beta_{x,\text{hom}}$ is less than the bias in $\beta_{x,\text{naive}}$ and increases with n (see Fig. 2a). As expected, Figures 1b and 2b are identical.

5.2 The Logistic Mixed Model

We now study the bias in the homogeneous MLE in the logistic mixed model when the heterogeneous model is true. Because no closed-form solution is available, we evaluated the bias by numerically maximizing the expectation of the log-likelihood of the homogeneous model when the heterogeneous model is true. We used numerical integration techniques similar to those in Section 4.2 to calculate this expectation; see Appendix Section A.5 for details.

The same parameter configurations as those in Figure 1 were used in our numerical bias calculations. The results are given in Figures 2c and 2d. Our calculations show that for $n = 2, \beta_{x,\text{hom}}$ slightly overestimates β_x. Its bias slightly increases with σ^2_u in the range that we consider. Contrary to the naive case, here the regular attenuation in β_x estimator is not present for $n = 2$. The homogeneous MLE θ_{hom} overestimates θ for both $n = 2$ and $n = \infty$, and the bias tends to be larger as n increases.

It is interesting to compare the biases in naive estimates with the biases in homogeneous MLEs. The naive estimate of β_x is much more biased than its homogeneous MLE counterpart. However, in figures not provided here, comparisons of θ_{hom} and θ_{naive} for various values of $\sigma^2_{x\mu}$ indicate that when $\sigma^2_{x\mu}$ is small, the biases for θ_{hom} and θ_{naive} have different directions.

6. SIMULATION EXTRAPOLATION AND FUNCTIONAL METHODS

Carroll et al. (1995) have drawn a distinction between functional modeling, in which nothing is assumed about the distribution of the X's, and structural modeling, in which a parametric model (e.g., homogeneous or heterogeneous normal) is assumed and the MLE is computed. Functional methods have the advantage that when they apply, they are model robust. Two common functional methods are *regression calibration* and *simulation extrapolation* (SIMEX). We discuss their application in GLMMeMs in this section.

6.1 Inconsistency of the Regression Calibration Approach

The *regression calibration* method simply replaces $\mathbf{X}$ by an estimate of $E(\mathbf{X}|\mathbf{W}, \mathbf{Z})$, and applies the naive method to these imputed values. Using (6) and (9) and noticing the sum of the first three terms is $E(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$, Wang, Lin, and Gutierrez (1997) showed that regular regression calibration in GLMMeMs often correctly specifies the fixed-effects structure but may misspecify the random-effects structures.

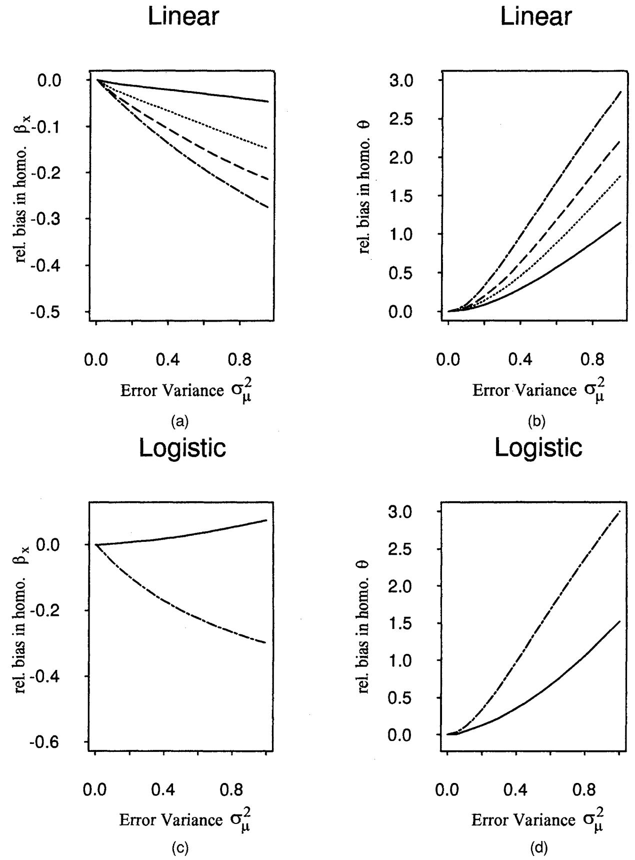


*Figure 2. Asymptotic Relative Biases in Homogeneous MLEs of β_x and θ in the Linear and Logistic Mixed Models When the Heterogeneous **X** Model is True. The true parameter values are $\beta_0 = 0$, $\beta_x = 2$, $\theta = 0.5$, $\sigma^2 = 1$, $\mu_x = 0$, $\sigma_x^2 = 1$, and $\sigma_{x\mu}^2 = 1.5$. The four plots correspond to (a) relative bias in $\beta_{x,hom}$ against σ_u^2 for the linear model; (b) relative bias in θ_{hom} against σ_u^2 for the linear model; (c) relative bias in $\beta_{x,hom}$ against σ_u^2 for the logistic model; and (d) relative bias in θ_{hom} against σ_u^2 for the logistic model. The four curves in (a) and (b) are ——, n = 2; - - -, n = 5; – – –, n = 10; and – - –, n = ∞. The two curves in (c) and (d) are ——, n = 2 and – - –, n = ∞.*

Because the regression coefficients and variance components are often not orthogonal in GLMMs, regression calibration can yield biased estimates of both (β_x, β_z) and θ, especially the latter.

Specifically, under the homogeneous X model, the regression calibration estimators of β_0, β_x, and θ are unbiased in the linear mixed model and biased in the logistic mixed model, with the asymptotic limits equal to $\beta_0/\tau^*, \beta_x/\tau^*$, and $(\tau^*)^{-2}\theta$, where τ^* is defined in Section 3.2. Under the heterogeneous X model, the regression calibration estimators of β_0, β_x, and θ converge to β_0, β_x, and θ' in the linear case, where θ' is defined in (9), and converge to $\beta_0/\tau^*, \beta_x/\tau^*$, and $(\tau^*)^{-2}\theta'$ in the logistic case. (For other bias analysis results and how to correct for the bias in naive regression calibration estimator, see Wang, Lin, and Gutierrez 1997.)

6.2 Simulation Extrapolation Estimation

SIMEX is a simulation-based measurement error method, a full description of which has been is given by Carroll et al. (1995) and Cook and Stefanski (1994). Rather than repeating the content of these references, we explain the

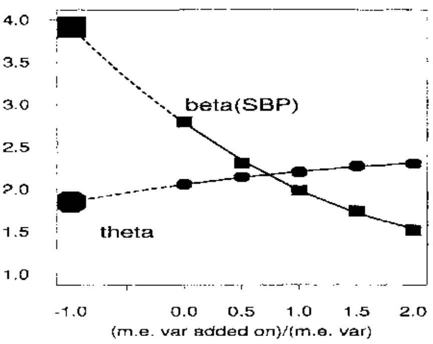


Figure 3. SIMEX/CPQL Extrapolations in the Framingham Data When All Intra-Individual Variability in Systolic Blood Pressures is Due To Measurement Error ($\sigma_x^2 = 0$). The x-axis is the ratio of the measurement error added on to observed SBP in the SIMEX simulation steps divided by the estimate of measurement error variance $\sigma_u^2 = .014$. ●, θ; ■, β (SBP).

SIMEX procedure using Figure 3, which shows application of SIMEX to the LVH example in Section 8. The two parameters of interest are the regression coefficient of the log-transformed SBP β_x and the variance component θ. For more details about this example, see Section 8.

Let the estimated value of σ_u^2 be $\hat{\sigma}_u^2$. The SIMEX method consists of two steps. The first step, the simulation step, is to establish the naive estimates if the measurement error were $(1+\xi)\sigma_u^2$. A simple empirical method is to add to the terms W_{ij} a normally distributed random variable with mean 0 and variance $\xi\hat{\sigma}_u^2$, then recompute the naive estimates. Doing this only once may be misleading, because it introduces simulation variability, so instead one repeats the procedure a large number B times and computes the median of the resulting parameter estimates. For example, to estimate β_x in Figure 3, one does so for each $\xi = (0, .5, 1.0, 1.5, 2.0)$ and plots the resulting naive estimates of β_x versus ξ. These are shown in small solid squares in Figure 3. Comparing the solid quadratic line connecting them to a plot such as Figure 1c, which is the bias curve of the naive estimate of β_x resulting from ignoring measurement error, the solid curve in Figure 3 corresponds to part of the curve in Figure 1c where $\sigma_u^2 \geq \hat{\sigma}_u^2$. The rest of the curve where $\sigma_u^2 < \hat{\sigma}_u^2$ is "hidden." Therefore, the solid curve in Figure 3 is referred as *partial bias plot.*

In the second step, the extrapolation step, a model is fit to the partial bias plot. A typical default is the quadratic, which we used. This is because quadratic curves often approximate the bias curves in Figure 1 well, and quadratic extrapolation works well in our simulation. We also experimented with fitting a quadratic to the log-transformed naive variance component estimates, to little positive effect for most of the cases considered. After a model is fit, the "hidden" parts of the figure are filled in by extrapolating the model to the values less than $\hat{\sigma}_u^2$, which are the dashed curves in Figure 3. The extrapolated value at $\xi = -1$ (zero error variance) is the SIMEX estimator.

Our calculations in Sections 3–5 show that the biases in the naive estimators are continuous functions of σ_u^2. Straightforward derivations using the M estimator arguments given by Wang, Lin, Gutierrez, and Carroll (1997, appendix) indicate that the asymptotic results given by Carroll et al. (1995) and Stefanski and Cook (1995) are directly applicable. Note that both of the latter works accommodate nonadditive or dependent measurement errors provided that the exact extrapolants are used and that the error distributions are normal. In our work we chose the corrected penalized quasi-likelihood method (CPQL) as our naive estimator, because compared to the naive MLE, it is more stable, easier to implement, and converges much faster and its performance is comparable to its MLE counterpart when the variance components are small or moderate. Because the CPQL estimator is an M estimator, the results of Stefanski and Cook (1995) can be applied. The CPQL procedure is briefly described in Appendix Section A.6. (For more details, see Breslow and Lin 1995 and Lin and Breslow 1996.)

7. SIMULATIONS

We conducted a simulation study to evaluate the finite-sample performance of various estimators. Binary observations Y_{ij} were generated within each cluster with conditional success probabilities satisfying logit $\{\Pr(Y_{ij}|X_{ij}, Z_{ij}, b_i)\} = \beta_0 + \beta_x X_{ij} + \beta_z Z_{ij} + b_i, i = 1, 2, \ldots, m, j = 1, 2, \ldots, n$. The following combinations of experiments were considered: (a) $m = 50, 100$, $n = 3, 8$, which are common sample sizes in longitudinal studies; and (b) homogeneous model (7) with $\mu_x = 0$, within-cluster variance $\sigma_x^2 = 1$, and between-cluster variance $\sigma_{x\mu}^2 = 0$ and heterogeneous model (8) with $\mu_x = 0, \sigma_x^2 = 1$, and $\sigma_{x\mu}^2 = 1.5$. A moderate measurement error variance $\sigma_u^2 = .5$ was considered for both the homogeneous and heterogeneous models. The exactly measured covariate Z was generated independently from a standard normal distribution. Other parameters used to specify the $\mathbf{Y}|\mathbf{X}$ and $\mathbf{W}|\mathbf{X}$ models were $\theta = .5, \beta_0 = 0, \beta_z = 1$, and $\beta_x = 2$. There were 1,000 simulations for each parameter setting. A single run for one dataset using our C program with $m = 100$ and $n = 3$ took about 1.5 minutes on a SUN UltraSparc station, and about 4 minutes when $n = 8$. The estimators considered in the simulation study included the (artificial) estimates based on the true X's, naive CPQL, and SIMEX/CPQL. For the SIMEX estimates, we set $B = 100$ and used quadratic extrapolations for all parameters (SIMEX-Q). The results are displayed in Tables 1 and 2.

We first comment on the homogeneous case ($\sigma_{x\mu}^2 = 0$). These results are largely consistent with our theory. The estimates of β_x and β_z reflect attenuation and are reasonably well corrected by SIMEX/CPQL. The theoretical value of $\beta_{\text{naive},z}$ is $\beta_z/\tau^* \approx .9$ (τ^* defined in Sec. 3.2), which is less biased than the estimate of β_x. As expected, there is a bias–variance trade-off, so that in estimating β_x, SIMEX/CPQL is less biased but more variable than the naive estimate, which ignores measurement error. A similar phenomenon occurs for estimating the variance component θ. The only

Table 1. Simulation of Logistic Regression in the Homogeneous Case; that is, the Between-Cluster Variance is $\sigma^2_{x\mu} = 0$

Cluster size	*Parameter*	*Method*	*Mean*	*SE*	*MSE*	*Mean*	*SE*	*MSE*
				m = 50			*m = 100*	
$n = 3$	$\beta_x = 2$	TRUE X	2.168	.531	.310	2.093	.313	.106
		NAIVE	1.516	.348	.356	1.470	.217	.328
		SIMEX-Q	2.100	.633	.411	2.010	.335	.125
	$\beta_z = 1$	TRUE X	1.088	.335	.120	1.051	.219	.050
		NAIVE	.944	.283	.083	.918	.189	.042
		SIMEX-Q	1.074	.364	.138	1.036	.232	.055
	$\theta = 0.5$	TRUE X	.528	.560	.314	.500	.390	.152
		NAIVE	.438	.474	.229	.407	.335	.121
		SIMEX-Q	.570	.627	.397	.532	.428	.170
$n = 8$	$\beta_x = 2$	TRUE X	2.090	.256	.074	2.066	.185	.038
		NAIVE	1.461	.175	.322	1.447	.131	.322
		SIMEX-Q	1.987	.290	.084	1.959	.210	.046
	$\beta_z = 1$	TRUE X	1.039	.185	.036	1.033	.130	.018
		NAIVE	.907	.162	.035	.902	.113	.022
		SIMEX-Q	1.018	.196	.039	1.010	.136	.019
	$\theta = 0.5$	TRUE X	.465	.286	.083	.456	.198	.041
		NAIVE	.369	.237	.073	.359	.168	.048
		SIMEX-Q	.466	.293	.087	.451	.207	.045

NOTE: Here n refers to the number of observations per cluster, $\beta_0 = 0$, $\beta_x = 2$, $\beta_z = 1$, and $\theta = .5$. The measurement error variance is $\sigma^2_u = .5$. The Z's are generated as standard normal and random variables.

feature that is somewhat at odds with the theory is that the naive estimates for $n = 8$ and $n = 3$ seem to have biases of different magnitudes. This discrepancy may be due to a much larger sampling variation of the variance component estimates when $n = 3$ compared to $n = 8$. As pointed out by one of the referees, when $n = 3$, the estimates of θ had large variation even when the true X were used. For a study with a small cluster size and a small-to-moderate number of clusters, one needs to be aware that variance component estimates may not be reliable; however, the fixed-effects coefficients could be obtained with good precision.

In the heterogeneous case ($\sigma^2_{x\mu} = 1.5$), a similar pattern repeats itself, although the results are less definitive. For $\sigma^2_u = .5$, as expected, there is relatively little bias in estimating θ. The magnitude of bias in estimating β is also smaller than that in the homogeneous case. SIMEX/CPQL does a good job of correcting bias in both estimates. Note that the quadratic extrapolation function works well for all scenarios considered in our simulations.

8. FRAMINGHAM HEART STUDY

We illustrate the SIMEX/CPQL method using the left ventricular hypertrophy (LVH) data discussed in Section 1. The study includes 75 patients who have coronary heart disease (CHD) developed before or during the study period and have not received diuretics treatment. Binary indicators Y for the presence or absence of LVH diagnosed by electrocardiogram (ECG) were observed every 2 years in an 8-year period. Two systolic blood pressure readings (SBP) were taken during each exam and were transformed to log(SBP-50) as suggested by Carroll et al. (1995) to achieve approximate normality. The covariates considered are X, average log-transformed SBP, and $\mathbf{Z}$, age, smoking status, body mass index, and the exam number (values 1–4). A logistic mixed model with random intercept was fit. Analysis of this dataset using our C program took about 4 minutes. The objective is to study the association between the risk of LVH and SBP after adjusting for the other covariates.

Initial analysis of the observed blood pressures themselves, or their residuals after regressing on Z, shows strong evidence in favor of the heterogeneous model, with approximately $\frac{2}{3}$ of the observed variability due to cluster-to-cluster variation. Thus even in the absence of measurement error, we would conclude that $\sigma^2_x \approx (1/2)\sigma^2_{x\mu}$.

Because blood pressures are obtained only every 2 years, and it is entirely possible that a person's SBP changes over time, there is no direct estimate of σ^2_u; this would require that SBP be obtained over a number of days within a relatively shorter period. To see this, consider the following argument. Suppose that within a cluster, $\mathbf{X}_i$ given $\mathbf{Z}_i$ follows a normal linear model with mean $\eta_0\mathbf{1} + \mathbf{Z}_i\eta_x$ and covariance matrix $\sigma^2_x\mathbf{I} + \sigma^2_{x\mu}\mathbf{J}$. Then $\mathbf{W}_i$ given $\mathbf{Z}_i$ follows a normal linear model with the same mean but with covariance matrix $(\sigma^2_x + \sigma^2_u)\mathbf{I} + \sigma^2_{x\mu}\mathbf{J}$. Thus the observed Framingham (W, Z) data alone can identify only the sum of σ^2_x and σ^2_u, but not either component separately. One means of identification is to fix a value of σ^2_x, (e.g., $\sigma^2_x = 0$), which assumes that latent blood pressure does not vary over the course of the study. Alternatively, identification is possible only from the assumed model for Y given (X, Z) and (outside the linear GLMM) from distributional assumptions concerning X.

For illustrative purposes, we vary the measurement error variance between the two extremes $\sigma^2_u = 0$ and $\sigma^2_x = 0$, using the method-of-moments estimator of their sum (.016) to identify σ^2_u exactly. In this illustration σ^2_u is treated as fixed and known, and we thus used the standard error estimation methods of Stefanski and Cook (1995).

Table 2. Simulation of Logistic Regression in the Heterogeneous Case

Cluster size	*Parameter*	*Method*	*Mean*	*SE*	*MSE*	*Mean*	*SE*	*MSE*
			$m = 50$			$m = 100$		
$n = 3$	$\beta_x = 2$	TRUE X	2.180	.479	.261	2.094	.337	.123
		NAIVE	1.729	.378	.216	1.661	.251	.178
		SIMEX-Q	2.173	.623	.418	2.067	.388	.155
	$\beta_z = 1$	TRUE X	1.082	.382	.153	1.037	.253	.066
		NAIVE	.951	.355	.128	.911	.229	.060
		SIMEX-Q	1.086	.454	.213	1.032	.280	.080
	$\theta = 0.5$	TRUE X	.541	.603	.365	.500	.451	.203
		NAIVE	.510	.559	.312	.457	.401	.163
		SIMEX-Q	.603	.697	.496	.545	.489	.241
$n = 8$	$\beta_x = 2$	TRUE X	2.106	.282	.091	2.056	.185	.037
		NAIVE	1.641	.200	.169	1.618	.139	.165
		SIMEX-Q	2.041	.311	.098	2.006	.211	.044
	$\beta_z = 1$	TRUE X	1.048	.215	.048	1.028	.142	.021
		NAIVE	.919	.194	.044	.906	.126	.025
		SIMEX-Q	1.030	.239	.058	1.012	.151	.023
	$\theta = 0.5$	TRUE X	.467	.330	.110	.453	.239	.059
		NAIVE	.416	.300	.097	.407	.211	.053
		SIMEX-Q	.443	.359	.132	.434	.248	.066

NOTE: Here $\beta_0 = 0$, $\beta_x = 2$, $\beta_z = 1$, and $\theta = .5$. The measurement error variance is $\sigma_u^2 = .5$. The within-cluster variance of the *X*'s is $\sigma_x^2 = 1$. The between-cluster variance is $\sigma_{x\mu}^2 = 1.5$. The *Z*'s are generated as standard normal random variables.

We expect from our theory that as σ_u^2 increases, the measurement error–corrected estimate of β_x will increase and the estimate of θ will decrease. The results confirm this. The estimate of β_x increased from 2.79 with a standard error of 1.44 (p value = .052) when $\sigma_u^2 = 0$ to 3.92 with a standard error of 1.73 (p value = .023) when $\sigma_u^2 = .016$. This suggests statistically significant effect of SBP on LVH. Higher SBP is associated with a higher risk of LVH. The evidence becomes stronger when the measurement error is taken into account.

The estimate of θ decreased from 2.05 with a standard error of 1.57 when $\sigma_u^2 = 0$ to 1.85 with a standard error of 1.52 when $\sigma_u^2 = .016$. We note that the nominal standard error of $\hat{\theta}$ cannot be used directly for testing $\theta = 0$, because the null hypothesis is on the boundary of the parameter space and the Wald statistic is not asymptotically distributed in a chi-square (Lin 1997). A SIMEX/score test for the variance component developed in an earlier version of this article (Wang, Lin, Gutierrez and Carroll 1997) indicates strong evidence for a nonzero variance component. Specifically, the p value of the score test increased from $<.0001$ when $\sigma_u^2 = 0$ to .006 when $\sigma_u^2 = .016$. Figure 3, discussed in Section 6.2, illustrates the SIMEX/CPQL extrapolations of β_x and θ when $\sigma_x^2 = 0$.

The conclusions for the rest of the regression coefficients stay the same with or without taking the measurement errors into account; namely, except for the intercept, none is significant at .05 level. The SIMEX/CPQL and the naive CPQL estimates (standard errors) of the regression coefficients for intercept, age, smoking status, body mass index, and the exam number are −23.92(8.45), .03(.07), −.75(.73), .02(.11), and .43(.25) and −17.7(6.88), .03(.06), −.60(.67), .05(.11), and .43(.25).

9. CONCLUDING REMARKS

In this article, we have shown that the effect of ignoring measurement error in GLMMeMs can be to bias regression coefficient and variance component estimators. The reason is that even under normality assumptions, the observed data follow a GLMM but with the *structure* of the fixed effects and random effects being misspecified. For example, typically the observed data are overdispersed relative to the assumed model ignoring measurement error (Sec. 2).

We have been able to compute the biases in a number of cases (Secs. 3–5). Typically, the broad direction of the biases in regression coefficient estimators is similar to that in ordinary generalized linear models (GLIM), whereas the direction of bias when estimating the variance component varies from case to case. Our results show that there is an important effect of the within-cluster sample size on the biases, and indeed some of the worst biases may occur when such sample sizes are the largest. We note that the techniques that we provided in bias calculations apply to more complicated cases.

The measurement error literature makes a distinction between functional and structural modeling. The former makes no assumptions about the distribution of the unobservables, and the latter typically makes distributional assumptions. The appeal of functional modeling is one of model robustness. We showed in Section 5 that in GLMMeMs there is an additional component of concern in structural modeling with respect to model robustness—namely, the assumed covariance structure of the unobservable predictors within a cluster.

The functional estimator that we considered (Sec. 6) is the SIMEX/CPQL method, largely because it has the potential to estimate parameters with minimal bias. Because CPQL is often fast even when the random-effects structures become more complicated, with respect to computational concern we expect that using SIMEX/CPQL to fit the general GLMMeM in (1) would still be practical. Using the general theory of Carroll et al. (1995) and Stefanski and Cook (1995), one can show that the SIMEX approach can be applied to the cases with dependent measurement errors and multivariate $\mathbf{X}_{ij}$.

In our example (Sec. 8), we noted a difficulty with the functional approach in GLMMs—namely, that without careful attention to experiment design, the structure of the measurement error (additivity, covariance) may not be identifiable from the observed covariates themselves. Thus in our example of logistic regression, estimating the measurement error covariance with any precision would require a structural approach; in the linear GLMM, only the covariance structure of the unobserved covariate must be specified. Clearly, if one is to consider the issue of measurement error in GLMMs, one must also pay attention to the design, and allow for estimation of the error covariance structure.

The appeal of functional methods in general, and the SIMEX procedure in this article in particular, is robustness against modeling the structure of the latent unobservable covariates $\mathbf{X}$. But the formidable appeal of functionality and the ease of implementation of the SIMEX procedure must be balanced by the potential loss of information and efficiency incurred relative to a correctly modeled structural analysis. Based on some preliminary calculations, we conjecture that the loss of efficiency in functional estimation may not be very severe in many problems for estimating the regression parameters $(\beta_0, \beta_x, \beta_z)$, but can in some cases be considerable for estimating variance components. Even in the ordinary GLIM, the tradeoff between model robustness and efficiency is a subject under vigorous development, with many attempts to model the latent covariates flexibly. In the GLMMeM context, with its more complex structure involving both distributions and covariances of the latent variables, such flexible modeling is a challenging problem that clearly deserves attention.

APPENDIX: DETAILED BIAS CALCULATIONS

A.1 Derivation of Equation (9)

Under the heterogeneous X model, we have $\mathbf{\Lambda}_i = (\sigma_x^2\mathbf{I} + \sigma_{x\mu}^2\mathbf{J})\{(\sigma_x^2 + \sigma_u^2)\mathbf{I} + \sigma_{x\mu}^2\mathbf{J}\}^{-1}$ and $\mathbf{e}_{xi} = a_i\mathbf{1} + \mathbf{e}_i$. Using the equality $(a\mathbf{I} + b\mathbf{J})^{-1} = a^{-1}\{\mathbf{I} - b(a+bn)^{-1}\mathbf{J}\}$ for any constants a and b, we have $\mathbf{\Lambda}_i = \sigma_u^2\{\lambda\mathbf{I} + n^{-1}(1-\lambda)(1-\tilde{\lambda})\mathbf{J}\}$ and $\mathbf{X}_i = \lambda\mathbf{W}_i + (1-\lambda)(1-\tilde{\lambda})\overline{W}_{i\cdot}\mathbf{1} + (1-\lambda)\tilde{\lambda}\mu_x\mathbf{1} + \mathbf{b}_i^*$, where $\mathbf{b}_i^* = \mathbf{X}_i - E(\mathbf{X}_i|\mathbf{W}_i) = b_{1i}^*\mathbf{1} + \mathbf{b}_{2i}^*, b_{1i}^* = \tilde{\lambda}(1-\lambda)a_i - (1-\lambda)(1-\tilde{\lambda})\overline{e}_{i\cdot} - (1-\lambda)1-\tilde{\lambda})\overline{U}_{i\cdot}$, and the jth component of $\mathbf{b}_{2i}^*$ is $\mathbf{b}_{2ij}^* = (1-\lambda)e_{ij} - \lambda U_{ij}$. Note that b_{1i}^* and $\mathbf{b}_{2i}^*$ are independent and are independent of $\mathbf{W}_i$ and b_i. Define $b_i' = b_i + \beta_x b_{1i}^* b_{ij}'' = \beta_x b_{2ij}^*$. We then have equation (9).

A.2 Bias Derivations in the Homogeneous Probit GLMMeM

The naive probit model is $\Phi^{-1}(\mu_{ij,w}^{b_i}) = \beta_0 + \beta_w W_{ij} + b_i$, and the Y_{ij} are binary with conditional means $\mu_{ij,w}^{b_i}$ and conditional variances $\mu_{ij,w}^{b_i}(1 - \mu_{ij,w}^{b_i})$. The homogeneous probit model is $\Phi^{-1}(\mu_{ij,w}^{\tilde{\mathbf{b}}_i}) = \beta_0' + \beta_x' W_{ij} + b_i + b_{ij}''$. To integrate out b_{ij}'', one can use the identity $\int \Phi(a+t)\, d\Phi(t/\gamma^{1/2}) = \Phi\{a(1+\gamma)^{-1/2}\}$ for any constants a and γ. It follows that $E(Y_{ij}|W_{ij}, b_i)$ satisfies $\Phi^{-1}(\mu_{ij,w}^{b_i}) = (1+\gamma)^{-1/2}(\beta_0' + \beta_x' W_{ij} + b_i)$. We hence have the bias results in Section 3.2.

A.3 Bias Derivations in the Homogeneous Poisson GLMMeM

Let $\zeta = \theta/2 + \beta_0$ and $\zeta^* = \theta^*/2 + \beta_0'$. The marginal means of $(Y_{ij}|X_{ij})$ and $(Y_{ij}|W_{ij})$ are $d_{ij} = \exp(\zeta + \beta_x X_{ij})$ and $c_{ij} = \exp(\zeta^* + \beta_x' W_{ij})$. This slight reparameterization allows θ to occur only in the conditional variances and covariances of $(Y_{ij}|X_{ij})$ and $(Y_{ij}|W_{ij})$ and not in the conditional means. Simple calculations using the correspondence between the conditional means of the $(\mathbf{Y}|\mathbf{X})$ and the $(\mathbf{Y}|\mathbf{W})$ models show that the naive estimators of (ζ, β_x) converge to (ζ^*, β_x').

Given (ζ, β_x), if the X's are observable, then the sufficient statistics for θ are $(\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{m\cdot})$. Denoting $\xi = \exp(\theta) - 1$, the asymptotic limit of the MLE of ξ statistics $\xi = \{E(\bar{Y}_{i\cdot} - \bar{d}_{i\cdot})^2 - n^{-1}E(\bar{d}_{i\cdot})\}/E(\bar{d}_{i\cdot}^2)$. The asymptotic limit of the naive estimator θ_{naive} must retain the same characterization and satisfies $\xi_{\text{naive}} = \{E(\bar{Y}_{i\cdot} - \bar{c}_{i\cdot})^2 - n^{-1}E(\bar{c}_{i\cdot})\}/E(\bar{c}_{i\cdot}^2)$, where $\xi_{\text{naive}} = \exp(\theta_{\text{naive}}) - 1$, and the expectation is taken with respect to (Y_{ij}, W_{ij}, X_{ij}) under the homogeneous Poisson model. Some calculations show that $\xi_{\text{naive}} = \exp(\theta_{\text{naive}}) - 1 = \exp(\theta^*) - 1 + (n-1)\{(n-1) + \exp(\lambda\beta_x^2\sigma_x^2)\}^{-1}\{\exp(-\gamma) - 1\}\exp(\theta^*)$. Equation (12) follows immediately.

A.4 Bias Derivations in the Heterogeneous Linear GLMMeM

Let $\boldsymbol{\beta} = (\beta_0, \beta_x)^T$ and $\mathcal{X} = (\mathbf{1}, \mathbf{X})$. Equation (14) can be written as $E(\mathcal{W}^T\mathbf{V}^{-1}\mathcal{W})\boldsymbol{\beta}_{\text{naive}} = E(\mathcal{X}^T\mathbf{V}^{-1}\mathcal{X})\boldsymbol{\beta}$. Applying (16) to both sides of this equation, some algebra yields $\beta_{0,\text{naive}}, \beta_{x,\text{naive}}$, and λ_* given in (18). To solve (15) for θ_{naive} and σ_{naive}^2, we first calculate the mean and covariance of $\mathbf{T} = \mathbf{Y} - \mathcal{W}\boldsymbol{\beta}_{\text{naive}}$ as $\boldsymbol{\mu}_T = E(\mathbf{T}) = 0$ and

$$\begin{aligned}\mathbf{V}_T &= \operatorname{cov}(\mathbf{T})\\ &= \{(\sigma^2 + \beta_{x,\text{naive}}^2\sigma_u^2) + \sigma_x^2(\beta_x - \beta_{x,\text{naive}})^2\}\mathbf{I}\\ &\quad + \{\theta + \sigma_{x\mu}^2(\beta_x - \beta_{x,\text{naive}})^2\}\mathbf{J}\\ &= \{\sigma^2 + (\lambda_*^2\sigma_u^2 + (1-\lambda_*)^2\sigma_x^2)\beta_x^2\}\mathbf{I}\\ &\quad + \{\theta + (1-\lambda_*)^2\sigma_{x\mu}^2\beta_x^2\}\mathbf{J}.\end{aligned}$$

Using the equalities $\partial\mathbf{V}/\partial\theta_{\text{naive}} = \mathbf{J}, \partial\mathbf{V}/\partial\sigma_{\text{naive}}^2 = \mathbf{I}$ and (16), one can show that (15) is equivalent to $\operatorname{tr}(\mathbf{V}^{-1}\mathbf{J}\mathbf{V}^{-1}\mathbf{V}_T) = \operatorname{tr}(\mathbf{V}^{-1}\mathbf{J})$ and $\operatorname{tr}(\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{V}_T) = \operatorname{tr}(\mathbf{V}^{-1})$. Because $\mathbf{V}$ and $\mathbf{V}_T$ have the same matrix structure, we have $\mathbf{V} = \mathbf{V}_T$. Equivalently, θ_{naive} and σ_{naive}^2 satisfy (17).

A.5 Numerical Bias Calculations of the Naive Estimators

Denote the log-likelihood by $l = \log L$. The naive estimators $\hat{\boldsymbol{\beta}}_{\text{naive}} = (\hat{\beta}_{0,\text{naive}}, \hat{\beta}_{x,\text{naive}})$ and $\hat{\theta}_{\text{naive}}$ maximize $m^{-1}\sum_{i=1}^m l^{\text{naive}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i; \boldsymbol{\beta}_*, \boldsymbol{\theta}_*)$, where $L^{\text{naive}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i; \boldsymbol{\beta}_*, \boldsymbol{\theta}_*)$ takes the same form as (2) except $\mathbf{X}_i$ is replaced by $\mathbf{W}_i$. Thus, the probability limit of the naive estimators $\boldsymbol{\beta}_{\text{naive}}$ and θ_{naive} maximizes $E\{l^{\text{naive}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i; \boldsymbol{\beta}_*, \boldsymbol{\theta}_*)|\mathbf{Z}_i\}$ as $m \to \infty$, where the expectation is taken with respect to $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i)$ conditional on $\mathbf{Z}_i$. For simplicity, we remove the subscript i in the ensuing discussion. Using the identity $E(\cdot|\mathbf{Z}) = E\{E(\cdot|\mathbf{X}, \mathbf{Z})\}$ and the independence of $\mathbf{Y}$ and $\mathbf{W}$ given $\mathbf{X}, \mathbf{Z}$, we first calculate

$$\begin{aligned}&E\{l^{\text{naive}}(\mathbf{Y}, \mathbf{W}, \mathbf{Z}; \boldsymbol{\beta}_*, \boldsymbol{\theta}_*)|\mathbf{X}, \mathbf{Z}\}\\ &\quad = \int l^{*\text{naive}}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_*, \boldsymbol{\theta}_*)L(\mathbf{Y}|\mathbf{X}, \mathbf{Z}; \boldsymbol{\beta}_*, \boldsymbol{\theta}_*)\, d\nu(\mathbf{Y}), \qquad \text{(A.1)}\end{aligned}$$

where $\nu(\mathbf{Y})$ denotes an appropriate probability measure of $\mathbf{Y}$ and

$$l^{*\text{naive}}(\mathbf{Y},\mathbf{X},\mathbf{Z};\boldsymbol{\beta}_*,\boldsymbol{\theta}_*) = \int l^{\text{naive}}(\mathbf{Y},\mathbf{X}+\mathbf{U},\mathbf{Z};\boldsymbol{\beta}_*,\boldsymbol{\theta}_*)L(\mathbf{U})\,d\mathbf{U}. \quad \text{(A.2)}$$

In the heterogeneous logistic GLMMeM considered in Section 4.2, we have

$$l^{*\text{naive}}(\mathbf{Y},\mathbf{X},\mathbf{Z};\boldsymbol{\beta}_*,\boldsymbol{\theta}_*) = \int\cdots\int \log\left\{\int\prod_{j=1}^{n}\mu_{y_j,w}(b)\,d\Phi(b)\right\} d\Phi(u_1)\ldots\Phi(u_n),$$

where $\mu_{y_j,w}(b) = \{H_{w_j}(b)\}^{y_j}\{1 - H_{w_j}(b)\}^{1-y_j}, H_{w_j}(b) = H(\beta_{0*} + \beta_{x*}X_j + \beta_{x*}\sigma_u u_j + \boldsymbol{\theta}_*^{1/2}b)$, and $H(v) = \{1+\exp(-v)\}^{-1}$.

We next need to further take expectation of (A.1) with respect to $\mathbf{X}$ conditional on Z. For the heterogeneous X model considered in Section 4.2, $\mathbf{X}$ follows $\mathrm{N}(\mu_x\mathbf{1},\mathbf{V}_x)$, with $\mathbf{V}_x = \sigma_x^2\mathbf{I} + \sigma_{x\mu}^2\mathbf{J}$. Let $\mathbf{T} = \mathbf{A}^{-1}(\mathbf{X} - \mu_x\mathbf{1})$, where $\mathbf{A}$ is a lower triangular matrix satisfying $\mathbf{A}\mathbf{A}^T = \mathbf{V}_x$ obtained using the Cholesky decomposition of $\mathbf{V}_x$. Denoting (A.1) by $f(\mathbf{X}, \mathbf{Z})$, we have $E\{l^{\text{naive}}(\mathbf{Y},\mathbf{W},\mathbf{Z};\boldsymbol{\beta}_*,\boldsymbol{\theta}_*)|\mathbf{Z}\} = E\{f(\mathbf{X},\mathbf{Z})|\mathbf{Z}\} = E\{f(\mu_x\mathbf{1}+\mathbf{A}\mathbf{T}, \mathbf{Z})|\mathbf{Z}\} = \int f(\mu_x\mathbf{1}+\mathbf{A}\mathbf{T},\mathbf{Z})\,d\Phi(T_1)\ldots d\Phi(T_n)$. Evaluation of required integration can be carried out by repeatedly using 20-point Gauss–Hermite quadrature for small n. Monte Carlo simulations can be used for large n. Using the change-of-variable technique for numerical integration has been discussed in detail by Davidian and Giltinan (1995, chap. 7). An optimization routine was used for maximization. This numerical technique can be applied to accommodate the cases with nonadditive and/or dependent measurement errors provided that the conditional distribution of $(\mathbf{W}|\mathbf{X})$ is normal.

The foregoing procedures can be used to calculate the biases of the homogeneous estimators discussed in Section 5. Specifically, we can replace $L^{\text{naive}}(\mathbf{Y},\mathbf{W},\mathbf{Z};\boldsymbol{\beta}_*,\boldsymbol{\theta}_*)$ in (A.1) and (A.2) by $L^{\text{hom}}(\mathbf{Y},\mathbf{W},\mathbf{Z};\boldsymbol{\beta}_*,\boldsymbol{\theta}_*,\mu_{x*},\sigma_{x*}^2)$, which is defined as $L(\mathbf{Y},\mathbf{W}|\mathbf{Z})$ in (4) with $L(\mathbf{X}|\mathbf{Z})$ being normal with mean μ_{x*} and covariance $\sigma_{x*}^2\mathbf{I}$. We then maximize the resulting expectation with respect to $\boldsymbol{\beta}_*,\boldsymbol{\theta}_*,\mu_{x*}$, and σ_{x*}^2. The maximizers are $\boldsymbol{\beta}_{\text{hom}},\boldsymbol{\theta}_{\text{hom}},\mu_{x,\text{hom}}$, and $\sigma_{x,\text{hom}}^2$.

A.6 The CPQL Method

A popular approximate inference procedure in the GLMM without measurement error (1) is the penalized quasi-likelihood (PQL) method of Schall (1991) and Breslow and Clayton (1993). Denote the right side of equation (1) by $\eta_{ij,x}^{\mathbf{b}_i}$. A key feature of PQL is that it can be easily implemented by iteratively fitting a linear mixed model to a modified dependent variable $y_{ij} = \eta_{ij,x}^{\mathbf{b}_i} + g'(\mu_{ij,x}^{\mathbf{b}_i})(Y_{ij} - \mu_{ij,x}^{\mathbf{b}_i})$ as

$$y_{ij} = \beta_0 + \mathbf{X}_{ij}^T\boldsymbol{\beta}_x + \mathbf{Z}_{ij}^T\boldsymbol{\beta}_z + \mathbf{A}_{ij}^T\mathbf{b}_i + \varepsilon_{ij},$$

where the random effects $\mathbf{b}_i$ follow N(0, $\mathbf{D}(\boldsymbol{\theta})$), ε_{ij} follows N(0, K_{ij}^{-1}), and K_{ij} is the working weight and equals $\{\phi\kappa_{ij}^{-1}v(\mu_{ij,x}^{\mathbf{b}_i})(g'(\mu_{ij,x}^{\mathbf{b}_i}))^2\}^{-1}$. The recently developed SAS macro GLIMMIX using the MIXED procedure has made this method readily accessible to practitioners. Simulation studies of Breslow and Clayton (1993) show that the PQL estimates may be subject to serious bias when the data are sparse (e.g., binary). Breslow and Lin (1995) and Lin and Breslow (1996) have proposed corrected PQL estimates to improve the performance of PQL. Specifically, the corrected PQL (CPQL) estimators take the form $\hat{\theta}_{\text{CP}} = C_\theta\hat{\theta}_{\text{P}}$ and $\hat{\alpha}_{\text{CP}} = \hat{\alpha}_{\text{P}} + C_\alpha\hat{\theta}_{\text{CP}}$, where $(\hat{\beta}_{\text{P}},\hat{\theta}_{\text{P}})$ and $(\hat{\beta}_{\text{CP}},\hat{\theta}_{\text{CP}})$ denote the PQL and CPQL estimators, and the correction matrices C_α and C_θ are easy to calculate and are given in equations (16) and (20) of Lin and Breslow (1996). In this article we used CPQL as our naive estimators.

[Received November 1996. Revised September 1997.]

REFERENCES

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N. E., and Lin, X. (1995), "Bias Correction in Generalized Linear Mixed Models With a Single Component of Dispersion," *Biometrika*, 82, 81–91.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996), "Asymptotics for the SIMEX Estimator in Structural Measurement Error Models," *Journal of the American Statistical Association*, 91, 242–250.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Cook, J., and Stefanski, L. A. (1995), "A Simulation Extrapolation Method for Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Davidian, M., and Giltinan, D. M. (1995), *Nonlinear Models for Repeated Measurement Data*, London: Chapman and Hall.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–340.

Lin, X. (1997), "Variance Components Testing in Generalized Linear Models With Random Effects," *Biometrika*, 84, 309–326.

Lin, X., and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

Liu, Q., and Pierce, D. A. (1993), "Heterogeneity in Mantel–Haenzel–Type Models," *Biometrika*, 80, 543–556.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Schall, R. (1991), "Estimation in Generalized Linear Models With Random Effects," *Biometrika*, 40, 917–927.

Stefanski, L. A., and Cook, J. R. (1995), "Simulation-Extrapolation: The Measurement Error Jackknife," *Journal of the American Statistical Association*, 90, 1247–1256.

Stiratelli, R., Laird, N., and Ware, J. (1984), "Random Effect Models for Serial Observations With Binary Response," *Biometrics*, 40, 961–971.

Tosteson, T., Stefanski, L. A., and Schafer, D. W. (1989), "A Measurement Error Model for Binary and Ordinal Regression," *Statistics in Medicine*, 8, 1139–1147.

Wang, N., Carroll, R. J., and Liang, K. Y. (1996), "Quasi-Likelihood and Variance Functions in Measurement Error Models With Replicates," *Biometrics*, 52, 401–411.

Wang, N., Lin, X., and Gutierrez, R. G. (1997), "A Corrected Regression Calibration Approach in Generalized Linear Mixed Measurement Error Models," Technical Report 285, Texas A&M University, Dept. of Statistics.

Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1997), "Bias Analysis and SIMEX Inference in Generalized Linear Mixed Measurement Error Models," Technical Report 275, Texas A&M University, Dept. of Statistics.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Biometrika (1999), **86**, 3, *pp.* 541–554
© 1999 Biometrika Trust
Printed in Great Britain

Nonparametric regression in the presence of measurement error

By RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

carroll@stat.tamu.edu

JEFFREY D. MACA

Department of Statistics, Novartis Pharmaceuticals Corporation, 59 Route 10, East Hanover, New Jersey 07936-1080, U.S.A.

jeff.maca@pharma.novartis.com

and DAVID RUPPERT

School of Operational Research and Industrial Engineering, Cornell University, Ithaca, New York 14853, U.S.A.

davidr@orie.cornell.edu

Summary

In many regression applications the independent variable is measured with error. When this happens, conventional parametric and nonparametric regression techniques are no longer valid. We consider two different approaches to nonparametric regression. The first uses the SIMEX, simulation-extrapolation, method and makes no assumption about the distribution of the unobserved error-prone predictor. For this approach we derive an asymptotic theory for kernel regression which has some surprising implications. Penalised regression splines are also considered for fixed number of known knots. The second approach assumes that the error-prone predictor has a distribution of a mixture of normals with an unknown number of components, and uses regression splines. Simulations illustrate the results.

Some key words: Estimating equation; Local polynomial regression; Measurement error; Regression spline; Sandwich estimation; SIMEX.

1. Introduction

We consider the problem of nonparametric regression function estimation in the presence of measurement error in the predictor. Suppose that the regression of a response Y on a predictor X is given by $E(Y \mid X) = m(X)$. Instead of observing X, we can only observe W, an error-prone measurement related to X by an additive error model, $W = X + U$, where U is a mean-zero normal random variable with variance σ_u^2. The question is how to estimate $m(.)$ when observations on Y and W are all that are available.

This problem has been addressed previously, most notably by Fan & Truong (1993), who found the following discouraging result. Suppose that we allow $m(.)$ to have up to

k derivatives. They showed that, if the measurement error is normally distributed, even with known error variance, then, based on a sample of size n, no consistent nonparametric estimator of $m(.)$ converges faster than the rate $\{\log(n)\}^{-k}$. Since, for example, $\log(10\,000\,000) \simeq 16$, effectively this result suggests that consistent nonparametric regression function estimation in the presence of measurement error is impractical.

The Fan & Truong result can be interpreted in another way. As reviewed by Carroll, Ruppert & Stefanski (1995), much of the enormous practical progress made in the field of measurement error for nonlinear models has been through the use of approximately consistent estimators, i.e. estimators which correct for most of the measurement error induced bias, but not all. Furthermore, when the measurement error variance is zero then the associated convergence rate is of order $n^{-\frac{1}{2}}$ rather than $\{\log(n)\}^{-k}$. We might expect, then, that estimation will be of practical use if the measurement error variance is not too large. Theoretically, for small errors, that is $\sigma_u^2 \to 0$, the bias of naive estimators is of the order $O(\sigma_u^2)$, while the approximate error correctors have a bias of order $O(\sigma_u^6)$ or less.

A second positive interpretation is to remember that the Fan & Truong result pertains to globally consistent estimation, i.e. estimators of $E(Y|X)$ which are consistent without anything but smoothness assumptions. Such results say nothing about estimators which are consistent for a flexible yet parametric subclass of the nonparametric family. For example, regression splines are a well-known parametric family with the capability of estimating wide classes of regression functions. If one is willing to estimate $E(Y|X)$ by a regression spline, then effective semiparametric estimation of $E(Y|X)$ should be possible even in the presence of measurement error.

This paper develops the two ideas of approximately consistent and regression spline estimation in the presence of measurement error. In § 2 we show how to implement the SIMEX, simulation-extrapolation, method (Cook & Stefanski, 1994; Stefanski & Cook, 1995) in ordinary nonparametric kernel regression, cubic smoothing splines and penalised regression splines. The SIMEX method is a functional method, i.e. one that can be applied without estimation of the distribution of the unobservable X. In § 3, we take up the structural approach in the context of regression splines, showing that the observed data follow a type of regression spline depending on the conditional distribution of X given W. If W given X is normally distributed, X given W depends on the marginal distribution of X, which we model flexibly by a mixture of normal distributions with an unknown number of components. This flexible distribution is estimated by modifying the Gibbs sampling algorithm of Wasserman & Roeder (1997). Section 5 gives a number of simulations. Section 6 has concluding remarks.

While the discussion to follow is easiest in the case that the measurement error variance σ_u^2 is known, in practice that is usually not the case. In some instances, σ_u^2 is estimated by an external dataset. Otherwise, internal replicates are used, so that we observe $W_{ij} = X_i + U_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, \kappa_i \geqslant 1$, where the measurement errors U_{ij} are independent, mean zero, normally distributed random variables with variance σ_u^2; a components of variance estimate is given as equation (3·2) in Carroll et al. (1995). In theory, for either external or internal data, σ_u^2 is estimated at ordinary parametric rates $O_p(n^{-\frac{1}{2}})$, and so the asymptotic effect of such estimation on nonparametric regression functions is often nil.

2. The SIMEX estimator

The SIMEX estimator was developed by Cook & Stefanski (1994); see Carroll et al. (1996) and Stefanski & Cook (1995) for related theory. The idea behind the method is most clearly seen in simple linear regression when the independent variable is subject to measurement error. Suppose the regression model is $E(Y|X)=\alpha+\beta X$ and that $W=X+U$, rather than X, is observed where U has mean zero and variance σ_u^2, and σ_u^2 is known. It is well known that the ordinary least squares estimate of the slope from regressing Y on W converges to $\beta\sigma_x^2(\sigma_x^2+\sigma_u^2)^{-1}$, where σ_x^2 denotes the variance of X.

For any fixed $\lambda > 0$, suppose one repeatedly 'adds on', via simulation, additional error with mean zero and variance $\sigma_u^2\lambda$ to W, computes the ordinary least squares slope each time and then takes the average. This simulation estimator consistently estimates $g(\lambda)=\beta\sigma_x^2/\{\sigma_x^2+\sigma_u^2(1+\lambda)\}$. Since, formally at least, $g(-1)=\beta$, the idea is to plot $g(\lambda)$ against $\lambda \geqslant 0$, fit a model to this plot and then extrapolate back to $\lambda=-1$. Cook & Stefanski (1994) show that this procedure will yield a consistent estimate of λ if one uses the model $g(\lambda)=\gamma_0+\gamma_1(\gamma_2+\lambda)^{-1}$.

Here is the precise definition of the SIMEX estimator for nonparametric regression. First consider the case that number of replicates $\kappa_i=1$ and that σ_u is known. Fix $B>0$ to be a large but finite integer, 50–200 in practice, and consider estimation of $E(Y|X)$ at x_0. For $b=1,\ldots,B$ and any $\lambda>0$, let ε_{ib} $(i=1,\ldots,n)$ be a set of independent standard normal random variables which are then transformed to have sample mean zero, variance one and to be uncorrelated with the Y's and the W's. Define $W_{ib}(\lambda)=W_i+\sigma_u\lambda^{\frac{1}{2}}\varepsilon_{ib}$. The resulting estimate from these simulated data is $\hat{m}_{b,\lambda}(x_0)$. The average of these estimates over $b=1,\ldots,B$ is $\hat{m}_\lambda(x_0)$.

With any nonparametric regression estimator, the SIMEX estimator is then defined by a three-step process: (a) select a finite set of λ's, such as $\lambda=0,\frac{1}{2},1,\frac{3}{2},2$, and compute $\hat{m}_\lambda(x_0)$; (b) fit a convenient function of λ, such as a quadratic, to the terms $\hat{m}_\lambda(x_0)$; (c) extrapolate this function back to $\lambda=-1$, resulting in $m(x_0)$.

When σ_u is unknown, it is replaced by an estimate. If the number of replicates is constant and equal to κ, then $W_{ib}(\lambda)=\bar{W}_{i.}+(\sigma_u\kappa^{-\frac{1}{2}}\lambda^{\frac{1}{2}}\varepsilon_{ib})$. The only remaining step is how to handle the case that the number of replicates is not constant. Carroll et al. (1995) use the definition of $W_{ib}(\lambda)$ given immediately above, but there is a theoretical difficulty, namely that $E(Y|\bar{W}_{i.},\kappa_i=1)\neq E(Y|\bar{W}_{i.},\kappa_i=2)$. This causes some problems of theory and even more of notation because, if we define $m_\lambda(x_0,\kappa_i)=E(Y|\bar{W}_{i.},\kappa_i)$, then the naive kernel regression estimator which ignores measurement error converges to $n^{-1}\sum_{i=1}^n m_\lambda(x_0,\kappa_i)$, which is a mixture of regression functions depending on the design. Despite this technical complication, the results derived in the Appendix extend immediately.

The estimators as implemented in this paper are as follows.

Kernel estimators. For a symmetric density function $K(.)$ and bandwidth h, define $K_h(u)=h^{-1}K(u/h)$. Local linear kernel estimates solve the weighted least squares equation in $\beta=(\beta_0,\beta_1)^{\mathrm{T}}$,

$$0=\sum_{i=1}^{n}[Y_i-G_2^{\mathrm{T}}\{W_{ib}(\lambda)-x_0\}\lambda]G_2\{W_{ib}(\lambda)-x_0\}K_h\{W_{ib}(\lambda)-x_0\}, \tag{1}$$

where $G_2(v)=(1,v)^{\mathrm{T}}$. The kernel estimate is $\hat{m}_{b,\lambda}(x_0,h)=\hat{\beta}_0$. In general, one must estimate h as well, and we do this using empirical bias bandwidth selection; see Ruppert (1997). The resulting implemented estimate is $\hat{m}_{b,\lambda}(x_0)$. The average of these estimates over $b=1,\ldots,B$ is $\hat{m}_\lambda(x_0)$. The MATLAB programs for computing empirical bias bandwidths selection are freely available at http://www.orie.cornell.edu/~davidr/matlab.

Smoothing splines. Cubic smoothing splines (Green & Silverman, 1994, Ch. 2) form another method of computing the estimates $\hat{m}_\lambda(x_0)$. These are available in S-Plus in the command 'smooth.spline'; the tuning constant is estimated by generalised crossvalidation.

Regression splines. We write the regression spline of order p and with l knots $(\xi_1, \ldots, \xi_l)$ as

$$m_{pl}(x; \beta) = \sum_{j=0}^{p} \beta_j x^j + \sum_{j=1}^{l} \beta_{p+j}(x - \xi_j)_+^p, \tag{2}$$

where $v_+ = vI(v > 0)$ and $I(.)$ is the indicator function. Eilers & Marx (1996) propose fixing the knots and estimating the regression parameters by penalised least squares, with the penalty term estimated by generalised crossvalidation, see their formula (29). With this parameterisation, and for a fixed penalty term α, the idea is to minimise

$$\sum_{i=1}^{n} \left\{ Y_i - \beta_0 - \beta_1 x - \ldots - \beta_p x^p - \sum_{j=1}^{l} \beta_{p+j}(x - k_j)_+^p \right\}^2 + \alpha \sum_{j=1}^{l} \beta_{p+j}^2. \tag{3}$$

The resulting functions serve as $\hat{m}_\lambda(x_0)$. We implemented this method in MATLAB.

Eilers & Marx (1996) use the B-spline basis rather than the truncated power function in (3). However, their approach is identical to ours for equally spaced knots and similar in other cases. Let λ, $\tilde{Z}$ and Y be the parameter vector, design matrix and response vector, respectively, of the linear model in (3). Then $\hat{\beta}$ is $(\tilde{Z}^{\mathrm{T}}\tilde{Z} + \alpha D)^{-1}\tilde{Z}^{\mathrm{T}}Y$, where D is the diagonal matrix with zeros in the first $p+1$ diagonal places and ones elsewhere along the diagonal.

3. Structural approach to regression splines

The SIMEX estimators described in § 2 have in common the fact that they make no assumption about the distribution of the unobserved X's. In contrast, in structural estimation in measurement error models one hypothesises a distribution for X depending on a parameter Θ. Since W given X is normal with mean X and variance σ_u^2, (σ_u, Θ) together produce the conditional distribution of X given W. Furthermore, if Y given X has mean determined by the spline (2), Y given W has mean

$$E(Y|W) = \sum_{j=0}^{p} \beta_j E(X^j|W) + \sum_{j=1}^{l} \beta_{p+j} E\{(X - \xi_j)_+^p|W\}. \tag{4}$$

We write β for the vector β_j's. Under a parametric model for X given W, all the conditional expectations in (4) are easily calculated numerically. The β_j's can be estimated by penalised least squares, along the lines of (3) with the obvious substitutions of $E(X^j|W)$ for X^j and $E\{(X - \xi_j)_+^p|W\}$ for $(X - \xi_j)_+^p$.

Our proposal then is as follows. Use the values of W to estimate a distribution for X and hence for X given W; one flexible method for doing this is described below. Then for given α estimate the β_j's by minimising

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} \beta_j \hat{E}(X^j|W_i) - \sum_{j=1}^{l} \beta_{p+j} \hat{E}\{(X - \xi_j)_+^p|W_i\} \right]^2 + \alpha \sum_{j=1}^{l} \beta_{p+j}^2. \tag{5}$$

Here $\hat{E}(.|W_i)$ means conditional expectation given W_i calculated using $\hat{\Theta}$ in place of the unknown Θ. The estimated function at x_0 is then

$$\hat{m}(x_0; \alpha) = \sum_{j=0}^{p} \hat{\beta}_j x_0^j + \sum_{j=1}^{l} \hat{\beta}_{p+j}(x_0 - \xi_j)_+^p.$$

Estimation of the smoothing parameter α is complicated by the fact that the structural regression spline has an enormous variance when $\alpha = 0$ because of the near singularity of the resulting design matrix, induced by the shrinkage inherent in computing $\hat{E}(X^j|W)$ and $\hat{E}\{(X-\xi_j)_+^p|W\}$. The generalised crossvalidation method, which has a tendency to undersmooth, is thus unacceptable unless one places a lower bound on the smoothing parameter. Although not reported here, this procedure works remarkably well in our simulations when the smallest possible value of α equals 10^{-6} times the sample size times the sample variance of the observed W's.

Instead, we replace generalised crossvalidation by a mean squared error estimation procedure, based on doubling smoothing and defined as follows. Write the design matrix implied by (4) as $\tilde{Z}$, and when predicting at an individual value call the design vector $\tilde{z}$. The bias of the fitted line at a design vector $\tilde{z}$ is $B(\tilde{z}, \alpha, \beta) = \tilde{z}^{\mathrm{T}}M(\alpha)\beta$, where $M(\alpha) = (\tilde{Z}^{\mathrm{T}}\tilde{Z} + \alpha D)^{-1}\tilde{Z}^{\mathrm{T}}\tilde{Z} - I$. If we fix α_0 and use $\hat{\beta}(\alpha_0)$ to estimate bias, then the variance of this estimated bias is

$$V(\tilde{z}, \alpha, \alpha_0) = \tilde{z}^{\mathrm{T}}M(\alpha)C(\alpha_0)M^{\mathrm{T}}(\alpha)\tilde{z},$$

where $C(\alpha_0) = \hat{\sigma}^2(\alpha_0)(\tilde{Z}^{\mathrm{T}}\tilde{Z} + \alpha_0 D)^{-1}\tilde{Z}^{\mathrm{T}}\tilde{Z}(\tilde{Z}^{\mathrm{T}}\tilde{Z} + \alpha_0 D)^{-1}$. If we start from α_0 and $\hat{\beta}(\alpha_0)$, a biased-corrected estimate of squared bias is $B^2\{\tilde{z}, \alpha, \hat{\beta}(\alpha_0)\} - V(\tilde{z}, \alpha, \alpha_0)$. Now the variance for the fitted line at a point $\tilde{z}$ for a given α is $R(\tilde{z}, \alpha) = \tilde{z}^{\mathrm{T}}C(\alpha)\tilde{z}$. This leads to the following algorithm.

ALGORITHM. *Fix α_0 at a value that implies considerable smoothing. Define α_1 to minimise the average, over a grid of values $\tilde{z}$, of the function $B^2\{\tilde{z}, \alpha, \hat{\beta}(\alpha_0)\} - V(\tilde{z}, \alpha, \alpha_0) + R(\tilde{z}, \alpha)$. Set $\alpha_0 = \alpha_1$ and then repeat until convergence.*

Since the regression of Y on W is generally heteroscedastic, we used a weighted version of this procedure, with weights calculated as follows. First a naive spline was estimated using generalised crossvalidation. Absolute residuals were formed and regressed on W with a spline using generalised crossvalidation. The weights are the inverse of the square of the fitted values of this last smooth. However, it is well known that weighting can be disastrously variable if the weights are allowed to vary too much, so we computed the median weight and then truncated the fitted weights to be within a factor of 3 of this median.

The remaining issue is to specify a distribution of X. The obvious one is the normal distribution, in which case $W = X + U$ would be marginally normally distributed, so that the assumption of normal X can be checked empirically from the observed data. To build some model robustness, one could use instead a flexible parametric family which includes the normal distribution, e.g. the seminonparametric family of Davidian & Gallant (1993) or the mixture of normals family.

A mixture of k normals has the means $\tilde{\mu}_k = (\mu_{1k}, \ldots, \mu_{kk})$, standard deviations $\tilde{\sigma}_k = (\sigma_{1k}, \ldots, \sigma_{kk})$ and proportions $\tilde{p}_k = (p_{1k}, \ldots, p_{kk})$, where $\sum_{j=1}^{k} p_{jk} = 1$. When X is observable, Wasserman & Roeder (1997) propose a Bayesian method for estimating $(k, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k)$ when k is constrained to lie in the set $1 \leqslant k \leqslant L$ for some fixed L. Here we modify their method to account for the measurement error. Suppose that we observe $W_{ij} = X_i + U_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Let σ_u have the inverse-chi prior density (A_u, r_u), where r_u is known, i.e.

$$[\sigma_u] \sim A_u^{r_u/2}\sigma_u^{-r_u-1}\exp\left(-\frac{1}{2}\frac{A_u}{\sigma_u^2}\right) \bigg/ \{2^{(r_u/2)-1}\Gamma(r_u/2)\}.$$

Fix k. Let $\tilde{W}$ consist of all the observed W's, $\tilde{X}$ and latent X's, and $\tilde{G}_k$ the latent group assignment indicators $(G_{k1}, \ldots, G_{kn})$ that identify from which of the k normal subpopulations $\tilde{X}$ is drawn; let $[A_k]$ be proportional to a scaling constant, and $[\tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k]$ be the prior defined by Wasserman & Roeder (1997).

The joint density for given k is

$$[\tilde{W}, \tilde{X}, \tilde{G}_k, \sigma_u, A_k, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k] \sim [\tilde{W} \mid \tilde{X}, \sigma_u][\sigma_u][\tilde{X} \mid \tilde{G}_k, A_k, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k] \times [\tilde{G}_k \mid A_k, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k][A_k, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k]. \qquad (6)$$

Inspection of (6) reveals that the Gibbs sampler has an especially convenient form. Once one has generated the latent variables $\tilde{X}$ and σ_u in a Gibbs step, the generation of $(\tilde{G}_k, A_k, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k)$ is exactly the same as if the $\tilde{X}$ were known and there was no measurement error. Thus we can adapt without change the Gibbs steps derived by Wasserman & Roeder (1997). Implementing the Gibbs steps for generating σ_u and $\tilde{X}$ is also easy. One sees that σ_u given all the rest is inverse-chi with parameters

$$A_u + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (W_{ij} - X_i)^2, \quad r_u + \sum_{i=1}^{n} m_i,$$

while any X_i, given all the rest and given that $G_{ki} = j$, is normal with mean and variance

$$\mu = (W_i \sigma_{jk}^2 + \mu_{jk} \sigma_u^2)/(m\sigma_{jk}^2 + \sigma_u^2), \quad \sigma^2 = \sigma_u^2 \sigma_{jk}^2/(m\sigma_{jk}^2 + \sigma_u^2),$$

respectively, where $W_i = \sum_{j=1}^{m} W_{ij}$.

Following Wasserman & Roeder (1997), having generated estimates of $\Theta_k = (\sigma_u, \tilde{\mu}_k, \tilde{\sigma}_k, \tilde{p}_k)$ for given k, namely the median of the value $(\sigma_k, \tilde{\mu}_k, \tilde{\sigma}_k)$ and the mean of the values $\tilde{p}_k$ in the Gibbs steps, we estimate the posterior probability that there are k mixture components as $n^{-3k/2} l(\hat{\Theta}_k)$, where $l(\Theta_k)$ is the likelihood of $\tilde{W}_k$ evaluated at the parameters Θ_k. This likelihood is

$$l(\Theta_k) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \sum_{l=1}^{k} p_{lk} (\sigma_{lk}^2 + \sigma_u^2)^{-\frac{1}{2}} (2\pi)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(W_{ij} - \mu_{lk})^2/(\sigma_{lk}^2 + \sigma_u^2)\}.$$

We now return to (4). To implement this, we need the conditional distribution of X_i given $(W_{i1}, \ldots, W_{im_i})$ for $i = 1, \ldots, n$. When X is a mixture of k normals, this conditional distribution is easily seen to be a mixture of k normals with the jth mean

$$(\mu_{jk} \sigma_u^2 + \bar{W} m \sigma_{jk}^2)(\sigma_u^2 + m\sigma_{jk}^2)^{-1},$$

the jth variance equal to $\sigma_{jk}^2 \sigma_u^2 (\sigma_u^2 + m\sigma_{jk}^2)^{-1}$, and the jth proportion given by

$$p_{jk} \left(\tilde{\sigma}_{jk} \sum_{i=1}^{k} [p_{jk} \tilde{\sigma}_{ik}^{-1} \exp\{-\tfrac{1}{2}(\bar{W} - \mu_{ik})^2/\tilde{\sigma}_{ik}^2\}] \right)^{-1} \exp\{-\tfrac{1}{2}(\bar{W} - \mu_{jk})^2/\tilde{\sigma}_{jk}^2\},$$

where $\tilde{\sigma}_{jk} = (\sigma_{jk}^2 + m^{-1}\sigma_u^2)^{\frac{1}{2}}$. If $(\hat{\eta}_1, \ldots, \hat{\eta}_L)$ are the estimated posterior probabilities formed from Gibbs sampling, we take X_i given $(W_{i1}, \ldots, W_{im_i})$ to be a mixture of the previously defined mixture normals, with mixing proportions $(\hat{\eta}_1, \ldots, \hat{\eta}_L)$.

4. Theoretical development

4·1. *Introduction*

As is essentially always the case, theoretical results are most readily obtained for kernel methods. The results then apply at least heuristically to cubic smoothing splines through the use of equivalent kernels (Silverman, 1984). Results for regression splines are more difficult. In unpublished work, S. Zhou has derived bias and variance formulae for unpenalised regression splines, but the formulae are not straightforward and the conditions require that the number of knots be of order $n^{1/5}$. For the sample sizes considered in our simulations, namely 200–500, this means a very small number of knots, which is exactly against the Eilers & Marx (1996) approach of using a fairly large number of knots.

4·2. *Kernels and regression splines*

We phrase our main theoretical result in a general way, and prove it explicitly in the case of local linear kernel regression; the details are in the Appendix. In any given problem, suppose that the measurement error is σ_u^2. Write the density of W with this measurement error as $f(w, \sigma_u^2)$ and the regression of Y on W with this measurement error as $m(w, \sigma_u^2)$. Thus, for instance, in the SIMEX steps we work with the derived variables $W(\lambda)$ which have measurement error $(1+\lambda)\sigma_u^2$, and the corresponding density and regression functions are $f\{w, (1+\lambda)\sigma_u^2\}$ and $m\{w, (1+\lambda)\sigma_u^2\}$.

Let the kernel function be $K(.)$. Suppose further that, when the bandwidth is h and the measurement error is σ_u^2, for fixed constants q_1 and q_2, the bias and variance of the kernel regression are given by

$$h^{q_1}\mathcal{G}_b\{x, m(x, \sigma_u^2), f(x, \sigma_u^2), \sigma_u^2, K\}, \quad (nh^{q_2})^{-1}\mathcal{G}_v\{x, m(x, \sigma_u^2), f(x, \sigma_u^2), \sigma_u^2, K\},$$

respectively. We run the SIMEX algorithm with B simulation replications at each value λ in a finite set Λ. We extrapolate using a polynomial of order q_s. Define e_s to be the (q_s+1)-vector with jth element $(-1)^{j+1}$, $s(\lambda)$ to be the (q_s+1)-vector with jth element λ^{j-1} and E_s to be the $(q_s+1)\times(q_s+1)$ matrix all of whose elements are zero except the first, which equals one. Finally, define $c^{\mathrm{T}}(x, \Lambda) = e_s^{\mathrm{T}}\{\sum_{\lambda\in\Lambda} s(\lambda)s^{\mathrm{T}}(\lambda)\}^{-1}$.

THEOREM 1. *Assume that the polynomial extrapolant is exact. For any λ, denote the bandwidth by h_λ. Then, as $n\rightarrow\infty$ and then $B\rightarrow\infty$, the* SIMEX *kernel estimator is asymptotically equivalent to an estimator with the bias and variance given respectively by*

$$c^{\mathrm{T}}(x, \Lambda)\sum_{\lambda\in\Lambda} h_\lambda^{q_1}\mathcal{G}_b[x, m\{x, (1+\lambda)\sigma_u^2\}, f\{x, (1+\lambda)\sigma_u^2\}, (1+\lambda)\sigma_u^2, K]s(\lambda), \tag{7}$$

$$(nh_0^{q_2})^{-1}\mathcal{G}_v\{x, m(x, \sigma_u^2), f(x, \sigma_u^2), \sigma_u^2, K\}c^{\mathrm{T}}(x, \Lambda)E_s c(x, \Lambda)O(1+B^{-1/4}). \tag{8}$$

The variance (8) is the more surprising, implying that the variance of the SIMEX estimate is asymptotically the same as if measurement error were ignored, but multiplied by $c^{\mathrm{T}}(x, \Lambda)E_s c(x, \Lambda)$, a factor which is independent of the regression function. Thus, we can easily compare the various extrapolants on the basis of variance. For instance, suppose that the set of possible values of λ is $\Lambda = (0{\cdot}0, 0{\cdot}5, 1{\cdot}0, 1{\cdot}5, 2{\cdot}0)$. Then direct calculation shows that use of the quadratic extrapolant leads to an estimator which is asymptotically 9 times more variable than that based on the linear extrapolant, while the cubic extrapolant is asymptotically 52 times more variable than the linear extrapolant.

The results (7)–(8) also apply at least roughly to linear and cubic smoothing splines, through the 'equivalent kernel' approach (Silverman, 1984). These results say that such smoothing splines behave away from the boundary like a Nadaraya–Watson kernel regression estimator with a locally chosen bandwidth and a higher-order kernel. If we make the identification, the major consequence is that the ratio of the variances of a kernel regression estimate and a linear or cubic smoothing spline when using SIMEX should be roughly the same as if measurement error were ignored. If the two methods are calibrated so that they have roughly the same variance when measurement error is ignored, then they should have roughly the same variance after use of the SIMEX procedure.

4·3. *Comparisons of kernels, regression splines and smoothing splines in* SIMEX

For regression splines, with or without penalties, there is no known equivalent kernel, although clearly the same results ought to apply approximately since regression splines with many knots and smoothing splines behave similarly.

One can gain some insight in the case that one uses a finite, fixed number of knots as $n \to \infty$, and one fixes the penalising factor α. Let $\mathscr{S}(X)$ be the design vector associated with (2), and as before let β be the collection of regression coefficients. Then, if X is observable, the penalised regression spline estimator is the solution to $\sum_{i=1}^{n} \psi(Y_i, X_i, \beta, \alpha) = 0$, where

$$\psi(Y_i, X_i, \beta, \alpha) = \mathscr{S}(X)\{Y - \mathscr{S}^{\mathrm{T}}(X)\beta\} - 2(\alpha/n) \sum_{j=1}^{l} \beta_{p+j}. \tag{9}$$

Equation (9) is an estimating equation, and, if we ignore the dependence on n and formally treat it as an unbiased estimating equation, i.e. as if it had mean zero, then the asymptotic theory for SIMEX given by Carroll et al. (1996) applies. Asymptotically of course the term α/n in (9) disappears and the estimator has the same behaviour as if there were no penalty. However, we have found that keeping this term in gives a somewhat better approximation to what actually happens in the simulations.

In the five models described in the simulation, § 5, we have computed the resulting asymptotic variances using numerical integration for 10 knots at the $k/11$ quantiles of the distribution of W ($k = 1, \ldots, 10$). When $\alpha = 0{\cdot}0$, the cubic extrapolant is approximately 4 times more variable than the quadratic extrapolant, and approximately 20 times more variable than the linear extrapolant. There is thus less variance inflation for using cubic extrapolation than in the kernel and smoothing spline case, although the variance inflation is still substantial. This suggests that, while the cubic extrapolant function should perform poorly for kernel and smoothing spline SIMEX, it will have better behaviour for regression spline SIMEX. This is at least qualitatively what happens in our simulations.

4·4. *Structural regression splines*

Asymptotic analysis for structural splines is also possible, and is most convenient in the case that X is known to have a normal distribution, and the knots and α are fixed. The estimating equation for β is the derivative of a typical term in (5). The estimating equation for the mean of X, μ_x, is $W - \mu_x$, the estimating equation for the variance of X, σ_x^2, is $(W - \mu_x)^2 - \sigma_x^2$, and a standard parametric analysis using estimating equation theory is easily obtained.

Unfortunately, because of the near collinearity of the terms in (5), numerical computation of the asymptotic variance of the structural spline estimator is difficult. By a mixture of exact calculations and 10 001-point Gaussian quadrature, we have been able to compute this variance in what we call Case 1 in the simulations, namely that $m(x) = 1000x_+^3(1 - x)_+^3$ for $n = 200$, $\mathrm{var}(Y \mid X) = 0{\cdot}0015^2$, Y normally distributed given X, with $X \sim N(0{\cdot}5, 0{\cdot}25^2)$ and $\sigma_u^2 = 3\sigma_x^2/7$. This calculation shows that the variance of the structural spline essentially blows up for 10 or more knots as $\alpha \to 0$. As long as there is substantial smoothing, the variance of the fitted function is small, while the bias is negligible. Thus, we would expect the structural spline to have good bias and variance behaviour if we ensure that the smoothing parameter never becomes too small, an expectation fulfilled in the simulations.

5. Simulations

We performed simulations on five test cases, for only one of which was the distribution of X from a normal distribution. The cases are as follows.

Case 1. We have $m(x) = 1000x_+^3(1-x)_+^3$ for $n = 200$, $\text{var}(Y|X) = 0{\cdot}0015^2$, Y normally distributed given X, with $X \sim N(0{\cdot}5, 0{\cdot}25^2)$ and $\sigma_u^2 = 3\sigma_x^2/7$.

Case 2. This is the same as Case 1, except that X followed a skew-normal distribution. The skew-normal base density is $2\phi(x)\Phi(10x)$, where ϕ and Φ are the density and distribution function of the standard normal distribution, respectively. We used the translation and scaling of this density which had the same mean and variance for X as in Case 1. The skew-normal as investigated here is quite skew, and cannot be exactly modelled as a mixture of normals.

Case 3. This is the same as Case 2, but with $n = 400$.

Case 4. We have $m(x) = 10\sin(4\pi x)$, $n = 500$, X uniform on $[0, 1]$, $\text{var}(Y|X) = 0{\cdot}05^2$ and $\sigma_u = 0{\cdot}141$.

Case 5. This is the same as Case 4, except that $\text{var}(Y|X) = 0{\cdot}02^2$ and $m(x) = 100/\{2 - \sin(2\pi x)\}$.

We used S-Plus for the estimation of smoothing splines and mixtures of normals, and MATLAB for kernel regression and regression splines. We generated and saved 200 datasets from each of the five cases, so that all methods are computed using the same simulated datasets.

In our simulations, it is obviously impractical to monitor convergence of the Gibbs sampler for every case. Instead, we examined a few test cases, examined their behaviour graphically, and noted that convergence of the sampler with good mixing appeared to have been reached at 2000 runs in each case. We then ran the sampler in the simulation 12 000 times.

There is no known bandwidth estimator for Fan & Truong's (1993) deconvoluting kernel estimator. We followed their approach, and compared our methods to a method which cannot be calculated without knowing the true regression function. Specifically, we used their kernels (5·1) and (5·2), and in each case found the global bandwidth which had minimum mean squared error. The mean squared error at this minimum was then reported. We experimented with using local linear deconvolution, but found that this led to higher mean squared errors. It is important to note that the mean squared errors listed for deconvolution are sensitive to the bandwidth, and of course the results we present are more favourable, perhaps far more favourable, to the deconvolution estimate than one would expect in practice.

In the first case, our mixture of normals method always selected that there was one population, and gave essentially no probability to two or more components. For the other cases, many simulations gave significant probability to two populations, but none gave probability to three populations. In results not reported here, for Cases 2 and 3, the estimated mixture density reproduces some of the features of the skew normal, but does not nearly reproduce it exactly. For Cases 4 and 5, the results are far worse, because the uniform density appears to be poorly modelled by a mixture of normals. Qualitatively, we expect the most problems for the structural regression spline in these latter two cases, because the mixture density provides a poor representation of the actual density of X. The results from the simulations are given in Table 1.

Table 1. *Squared bias and mean squared error results from the simulations*

				Method				
Case	Naive kernel	Naive regression spline	Naive smoothing spline	SIMEX kernel	SIMEX smoothing spline	SIMEX regression spline	Structural spline	Decon. kernel
				Squared bias				
1	11·8	12·6	13·0	2·9	3·8	2·8	0·4	6·9
2	13·8	14·2	14·7	4·9	5·3	4·5	1·2	8·8
3	13·9	14·1	14·2	5·0	5·0	4·5	1·0	8·6
4	32·3	30·7	31·3	15·7	13·5	12·2	1·4	17·6
5	77·5	73·5	76·8	23·9	25·4	23·2	16·9	58·9
				Mean squared error				
1	12·4	13·2	13·5	6·4	6·26	5·6	2·1	8·9
2	14·4	14·6	15·1	8·1	7·3	6·6	5·5	10·9
3	14·4	14·6	14·4	7·7	5·8	6·9	5·6	13·7
4	32·8	31·2	31·8	18·8	15·5	14·5	7·6	20·3
5	80·5	76·0	79·1	32·6	35·2	34·4	37·9	62·1

We see that the 'naive' kernel, smoothing spline and regression spline all have similar behaviour, as do the SIMEX kernel, smoothing spline and regression spline with quadratic extrapolant. The SIMEX methods have much smaller bias and mean squared errors than the methods which ignore the measurement error. The deconvolving kernel methods have behaviour somewhat intermediate between the naive and SIMEX approaches, although these results are of limited relevance because the former has bandwidth optimised to have smallest mean squared error.

Where possible, in results not reported here, we have compared the simulation results with the asymptotic theory, and found them roughly in accord with one another.

The structural regression spline has good performance overall, although it is biased in Case 5. As described above, the reason for this is a mixture of the regression function chosen and the fact that mixtures of normals do not give a good approximation to the uniform density, especially with measurement error. We re-ran the structural spline in Case 5 but with X normally distributed with the same mean and variance as the uniform, and hence in the mixture family, and as expected the bias essentially disappeared.

Our results, both theoretical and numerical, indicate that within the SIMEX context kernels, smoothing splines and penalised regression splines with a large number of knots behave fairly similarly. The quadratic extrapolant seems the one of choice among the polynomials, because of better bias behaviour than the linear extrapolant but far smaller variability than the cubic extrapolant. The structural regression spline approach has overall the best numerical behaviour in these simulations, and sometimes is far more efficient than the other methods. It clearly has considerable potential, although the simulation results on which this potential is based do not exhaust the possible regression functions one might observe in practice.

6. Discussion

We have assumed without comment that $W = X + U$, with U normally distributed and having mean zero. In fact, for purposes of nearly nonparametric estimation, it suffices merely that some monotone transformation of originally observed W's follow this additive error model, that is $g(W) = g(X) + U$, because if $g(.)$ is any strictly monotone function, $E(Y \mid X = x_0) = E\{Y \mid g(X) = g(x_0)\}$.

Acknowledgement

R. J. Carroll's research was supported by a grant from the National Cancer Institute and through the Texas A&M Center for Environmental and Rural Health by a grant from the National Institute of Environmental Health Sciences. D. Ruppert's research was supported by grants from the National Science Foundation and the National Security Agency. We wish to acknowledge with thanks conversations with Simon Sheather. Two referees made many helpful comments, for which we are most appreciative.

Appendix

Theory for simex *estimate in kernel regression*

The goal of this Appendix is to sketch a proof of the main result (7)–(8) for local linear kernel regression. In the simex algorithm, we add normally distributed measurement error to the observed W's $b = 1, \ldots, B$ times, for each value of λ. Let $f_\lambda(.)$ be the density function of $W + \lambda^{\frac{1}{2}} \sigma_u \varepsilon$, where ε has a standard normal distribution. Hence, if $f_W(.)$ is the density of W,

$$f_\lambda(x_0) = \int (\lambda^{1/2} \sigma_u^{1/2})^{-1} f_W(z) \phi\{(x_0 - z)/(\lambda^{1/2} \sigma_u)\}\, dz.$$

Let $m_\lambda(x_0) = E(Y \mid W + \lambda^{1/2} \sigma_u = x_0)$. Carroll, Ruppert & Welsh (1998) show that, for any fixed b, as $h \to 0$ and $nh \to \infty$, with $\int z^2 K(z)\, dz = 1$,

$$\hat{m}_{b,\lambda}(x_0, h) - m_\lambda(x_0) - (h^2/2) m_\lambda^{(2)}(x_0) \doteq \{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} [Y_i - m_\lambda\{W_{ib}(\lambda)\}] K_h\{W_{ib}(\lambda) - x_0\}, \quad \text{(A1)}$$

where the error is of order $o_p\{h^2 + (nh)^{-\frac{1}{2}}\}$.

In what follows, it is convenient notationally to use the same bandwidth for every $b = 1, \ldots, B$, but to allow this bandwidth to depend on λ; hence we write h_λ. In practice and as in our simulations, one might estimate h for each λ and b, but as $n \to \infty$ the error in estimating this bandwidth becomes negligible, and hence asymptotically the same bandwidths are being used for all b.

Using (A1) and the decomposition of Carroll et al. (1996), since B is fixed and since $\hat{m}_\lambda(x_0, h) = B^{-1} \sum_{b=1}^{B} \hat{m}_{b,\lambda}(x_0, h_\lambda)$, we have

$$\begin{aligned} &\hat{m}_\lambda(x_0, h_\lambda) - m_\lambda(x_0) - (h_\lambda^2/2) m_\lambda^{(2)}(x_0) \\ &\qquad \doteq \{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} \left(B^{-1} \sum_{b=1}^{B} [Y_i - m_\lambda\{W_{ib}(\lambda)\}] K_{h_\lambda}\{W_{ib}(\lambda) - x_0\} \right). \qquad \text{(A2)} \end{aligned}$$

In what follows, we will use the following slight abuse of notation. We will write expressions for moments of $\hat{m}_\lambda(x_0, h_\lambda)$, but these will actually apply to the asymptotically equivalent version on the right-hand side of (A2). The terms inside the parentheses on the right-hand side of (A2) are independent mean zero random variables. Letting $\tilde{Y} = (Y_1, \ldots, Y_n)$ and $\tilde{W} = (W_1, \ldots, W_n)$, and using the right-hand side of (A2) as equivalent to the left-hand side, consider

$$\begin{aligned} \operatorname{var}\{\hat{m}_\lambda(x_0, h_\lambda)\} &\doteq E[\operatorname{var}\{\hat{m}_\lambda(x_0, h_\lambda) \mid \tilde{Y}, \tilde{W}\}] \\ &\quad + \operatorname{var}[E\{\hat{m}_\lambda(x_0, h_\lambda) - m_\lambda(x_0) - (h_\lambda^2/2) m_\lambda^{(2)}(x_0) \mid \tilde{Y}, \tilde{W}\}]. \qquad \text{(A3)} \end{aligned}$$

If $\lambda = 0$ or if $\sigma_u^2 = 0$, then $m_0(x_0) = E(Y \mid W = x_0)$, $f_0(x_0) = f_W(x_0)$ and (A2) becomes

$$\hat{m}_0(x_0, h_0) - m_0(x_0) - (h_0^2/2) m_0^{(2)}(x_0) \simeq \{n f_0(x_0)\}^{-1} \sum_{1=i}^{n} \{Y_i - m_0(W_i)\} K_{h_0}(W_i - x_0),$$

which has mean zero and asymptotic variance

$$\{n h_0 f_0(x_0)\}^{-1} \operatorname{var}(Y \mid W = x_0) \int K^2(v)\, dv. \tag{A4}$$

If $\lambda > 0$ and $\sigma_u^2 > 0$, we study the terms of (A3) in turn. For the first, note that, given $\tilde{Y}$ and $\tilde{W}$, the only remaining random variables are the (ε_{ib}), which are all mutually independent. Hence

$$\begin{aligned} &\operatorname{var}\{\hat{m}_\lambda(x_0, h_\lambda) \mid \tilde{Y}, \tilde{W}\} \\ &\quad \simeq \{nB f_\lambda^2(x_0)\}^{-1} n^{-1} \sum_{i=1}^{n} \operatorname{var}[\{Y_i - m_\lambda(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon)\} K_{h_\lambda}(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon - x_0) \mid Y_i, W_i] \\ &\quad = \{nB f_\lambda^2(x_0)\}^{-1} n^{-1} \sum_{i=1}^{n} \int \{Y_i - m_\lambda(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon)\}^2 K_{h_\lambda}^2(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon - x_0) \phi(\varepsilon)\, d\varepsilon \\ &\quad\quad - \{nB f_\lambda^2(x_0)\}^{-1} n^{-1} \sum_{i=1}^{n} \left[\int \{Y_i - m_\lambda(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon)\} K_{h_\lambda}(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon - x_0) \phi(\varepsilon)\, d\varepsilon \right]^2. \end{aligned}$$

Setting $z = (W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon - x_0)/h_\lambda$ so that $W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon = x_0 + z h_\lambda$ and $\varepsilon = (x_0 + z h_\lambda - W_i)/\sigma_u \lambda^{\frac{1}{2}}$, we compute the two terms as

$$\begin{aligned} &\{n h_\lambda B f_\lambda^2(x_0) \sigma_u \lambda^{\frac{1}{2}}\}^{-1} n^{-1} \sum_{i=1}^{n} \int \{Y_i - m_\lambda(x_0 + z h_\lambda)\}^2 K^2(z) \phi\left(\frac{z h_\lambda + x_0 - W_i}{\sigma_u \lambda^{\frac{1}{2}}}\right) dz \\ &\quad - \{nB f_\lambda^2(x_0) \sigma_u^2 \lambda\}^{-1} n^{-1} \sum_{i=1}^{n} \left[\int \{Y_i - m_\lambda(x_0 + z h_\lambda)\} K(z) \phi\left(\frac{z h_\lambda + x_0 - W_i}{\sigma_u \lambda^{\frac{1}{2}}}\right) dz \right]^2. \end{aligned}$$

The second term is $O(n^{-1})$, so we are left with

$$\operatorname{var}\{\hat{m}_\lambda(x_0, h_\lambda) \mid \tilde{Y}, \tilde{W}\} \simeq \{n h_\lambda B f_\lambda^2(x_0) \sigma_u \lambda^{\frac{1}{2}}\}^{-1} \left\{ \int K^2(z)\, dz \right\} n^{-1} \sum_{i=1}^{n} \{Y_i - m_\lambda(x_0)\}^2 \phi\left(\frac{W_i - x_0}{\sigma_u \lambda^{\frac{1}{2}}}\right). \tag{A5}$$

Note the curious fact that there is a B in the denominator. This means that, if B is large, (A5) is small in comparison to what happens when $\lambda = 0$; see (A4). In fact, as $B \to \infty$, (A2) converges to

$$\{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} E[\{Y_i - m_\lambda(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon)\} K_{h_\lambda}(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon) \mid Y_i, W_i], \tag{A6}$$

and this random variable has zero variance given $(\tilde{Y}, \tilde{W})$, just as predicted by (A5).

We next turn to the second term in (A3). If we continue to assume that $\lambda > 0$ and $\sigma_u^2 > 0$, the expectation in question is just

$$\begin{aligned} &\{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} \int \{Y_i - m_\lambda(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon)\} K_{h_\lambda}(W_i + \sigma_u \lambda^{\frac{1}{2}} \varepsilon - x_0) \phi(\varepsilon)\, d\varepsilon \\ &\quad = \{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} \int \{Y_i - m_\lambda(x_0 + z h_\lambda)\} K(z) \phi\left(\frac{z h_\lambda + x_0 - W_i}{\sigma_u \lambda^{\frac{1}{2}}}\right) (\sigma_u \lambda^{\frac{1}{2}})^{-1}\, dz \\ &\quad \simeq \{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} \{Y_i - m_\lambda(x_0)\} \phi\left(\frac{W_i - x_0}{\sigma_u \lambda^{\frac{1}{2}}}\right) (\sigma_u \lambda^{\frac{1}{2}})^{-1}, \end{aligned} \tag{A7}$$

which has variance of order $O(n^{-1})$. We have thus shown that, for $\lambda > 0$, $\sigma_u^2 > 0$,

$$\operatorname{var}\{\hat{m}_\lambda(x_0, h)\} = O\{(nhB)^{-1}\} + O(n^{-1}). \tag{A8}$$

It is important to note that the second term in (A3) is $O(n^{-1})$ only when $\lambda > 0$ and $\sigma_u^2 > 0$. If either equals 0, the expectation calculated above is

$$\{nf_0(x_0)\}^{-1} \sum_{i=1}^{n} \{Y_i - m_0(W_i)\} K_{h_0}(W_i - x_0),$$

which has mean zero and variance $O\{(nh)^{-1}\}$. The difference is that, when $\lambda > 0$ and $\sigma_u^2 > 0$, (A8) represents a 'double-smooth', i.e. summation and integration, and it is well known that double smoothing increases rates of convergence.

If we compare (A4) with (A8), we note that, for n and B sufficiently large, the latter will be negligible with respect to the former, at least in practice. Hence, in what follows, we will ignore this variability by treating B as if it were equal to infinity. This makes the analysis of the SIMEX extrapolants easy. In our notation we are minimising in $\mathscr{A}$, say, the sum of squares $\sum_\lambda \{\hat{m}_\lambda(x_0, h_\lambda) - s^{\mathrm{T}}(\lambda)\mathscr{A}\}^2$. Thus, we are solving $0 = \sum_\lambda \{\hat{m}_\lambda(x_0, h_\lambda) - s^{\mathrm{T}}(\lambda)\mathscr{A}\} s(\lambda)$. Using standard least-squares results, we obtain

$$\hat{\mathscr{A}} - \mathscr{A} = \left\{\sum_\lambda s(\lambda)s^{\mathrm{T}}(\lambda)\right\}^{-1} \sum_\lambda \{\hat{m}_\lambda(x_0, h_\lambda) - s^{\mathrm{T}}(\lambda)\mathscr{A}\} s(\lambda). \tag{A9}$$

If we assume the terms $m_\lambda(x_0)$ actually follow the extrapolant function, this means that the left-hand side of (A9) has approximate mean

$$\left\{\sum_\lambda s(\lambda)s^{\mathrm{T}}(\lambda)\right\}^{-1} \sum_\lambda (h_\lambda^2/2) m_\lambda^{(2)}(x_0) s(\lambda),$$

and, because B is large, its approximate variance is

$$\{nh_\lambda f_0(x_0)\}^{-1} \left\{\int K^2(z)\,dz\right\} \operatorname{var}(Y \mid W = x_0) \left\{\sum_\lambda s(\lambda)s^{\mathrm{T}}(\lambda)\right\}^{-1} E_s \left\{\sum_\lambda s(\lambda)s^{\mathrm{T}}(\lambda)\right\}^{-1}.$$

The SIMEX estimate is just $e_s^{\mathrm{T}}\hat{\mathscr{A}}$, so that its asymptotic bias is $c^{\mathrm{T}}(x_0, \Lambda)\sum_\lambda h_\lambda^2 m_\lambda^{(2)}(x_0)s(\lambda)/2$, and its asymptotic variance is

$$\{nh_0 f_0(x_0)\}^{-1} \operatorname{var}(Y \mid W = x_0) c^{\mathrm{T}}(x_0, \Lambda) E_s c(x_0, \Lambda) \int K^2(z)\,dz,$$

as claimed.

References

Carroll, R. J., Küchenhoff, H., Lombard, F. & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *J. Am. Statist. Assoc.* **91**, 242–50.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D. & Welsh, A. H. (1998). Local estimating equations. *J. Am. Statist. Assoc.* **93**, 214–27.

Cook, J. R. & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Am. Statist. Assoc.* **89**, 1314–28.

Davidian, M. & Gallant, R. A. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 475–88.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

Fan, J. & Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–25.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. New York: Chapman and Hall.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Assoc.* **92**, 1049–62.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.

Stefanski, L. A. & Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *J. Am. Statist. Assoc.* **90**, 1247–56.

Wasserman, L. & Roeder, K. (1997). Bayesian density estimation using mixtures of normals. *J. Am. Statist. Assoc.* **92**, 894–902.

[*Received February* 1997. *Revised December* 1998]

The Annals of Statistics
1999, Vol. 27, No. 5, 1519–1535

ESTIMATION IN A SEMIPARAMETRIC PARTIALLY LINEAR ERRORS-IN-VARIABLES MODEL

By Hua Liang,[1] Wolfgang Härdle[2] and Raymond J. Carroll[3]

Chinese Academy of Sciences, Humboldt-Universität zu Berlin and Texas A & M University

We consider the partially linear model relating a response Y to predictors (X, T) with mean function $X^{\mathrm{T}}\beta + g(T)$ when the X's are measured with additive error. The semiparametric likelihood estimate of Severini and Staniswalis leads to biased estimates of both the parameter β and the function $g(\cdot)$ when measurement error is ignored. We derive a simple modification of their estimator which is a semiparametric version of the usual parametric correction for attenuation. The resulting estimator of β is shown to be consistent and its asymptotic distribution theory is derived. Consistent standard error estimates using sandwich-type ideas are also developed.

1. Introduction and background. Consider the semiparametric partially linear model based on a sample of size n,

$$Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + \varepsilon_i, \tag{1}$$

where X_i is a possibly vector-valued covariate, T_i is a scalar covariate, the function $g(\cdot)$ is unknown and the model errors ε_i are independent with conditional mean zero given the covariates. The partially linear model was introduced by Engle, Granger, Rice and Weiss (1986) to study the effect of weather on electricity demand and further studied by Heckman (1986), Chen (1988), Speckman (1988), Cuzick (1992a, b), Liang and Härdle (1997) and Severini and Staniswalis (1994).

We are interested in the estimation of the unknown parameter β and the unknown function $g(\cdot)$ in model (1) when the covariates X_i are measured

Received July 1997; revised March 1999.

[1] Supported by an award from the Alexander von Humboldt Foundation and by Sonderforschungsbereich 373 “Quantifikation und Simulation Ökonomischer Prozesse.”

[2] Supported by Sonderforschungsbereich 373 “Quantifikation und Simulation Ökonomischer Prozesse.”

[3] Supported by National Cancer Institute Grant CA-57030, by the Texas A & M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences Grant P30-ESO9106 and by Sonderforschungsbereich 373 “Quantifikation und Simulation Ökonomischer Prozesse.”

AMS 1991 *subject classifications.* Primary 62J99, 62H12, 62E25, 62F10; secondary 62H25, 62F10, 62F12, 60F05.

Key words and phrases. Errors-in-variables, measurement error, nonparametric likelihood, orthogonal regression, partially linear model, semiparametric models, structural relations.

with error. Instead of observing X_i, we observe

$$W_i = X_i + U_i, \tag{2}$$

where the measurement errors U_i are independent and identically distributed, independent of (Y_i, X_i, T_i), with mean zero and covariance matrix Σ_{uu}. We will assume that Σ_{uu} is known, taking up the case that it is estimated in Section 5. The measurement error literature has been surveyed by Fuller (1987) and Carroll, Ruppert and Stefanski (1995).

If the X's are observable, estimation of β at ordinary rates of convergence can be obtained by a local-likelihood algorithm, as follows. For every fixed β, let $\hat{g}(T, \beta)$ be an estimator of $g(T)$. For example, in the Severini and Staniswalis implementation, $\hat{g}(T, \beta)$ maximizes a weighted likelihood assuming that the model errors ε_i are homoscedastic and normally distributed, with the weights being kernel weights with symmetric kernel density function $K(\cdot)$ and bandwidth h. Having obtained $\hat{g}(T, \beta)$, β is estimated by a least squares operation,

$$\text{minimize} \sum_{i=1}^{n} \{Y_i - X_i^{\mathrm{T}}\beta - \hat{g}(T_i, \beta)\}^2.$$

In this particular case, the estimate for β can be determined explicitly. Let $\hat{g}_{y,h}(\cdot)$ and $\hat{g}_{x,h}(\cdot)$ be the kernel regressions with bandwidth h of Y and X on T, respectively. Then

$$\begin{aligned} \hat{\beta}_n = {} & \left[\sum_{i=1}^{n} \{X_i - \hat{g}_{x,h}(T_i)\}\{X_i - \hat{g}_{x,h}(T_i)\}^{\mathrm{T}}\right]^{-1} \\ & \times \sum_{i=1}^{n} \{X_i - \hat{g}_{x,h}(T_i)\}\{Y_i - \hat{g}_{y,h}(T_i)\}. \end{aligned} \tag{3}$$

One of the important features of the estimator (3) is that it does not require undersmoothing, so that bandwidths of the usual order $h \sim n^{-1/5}$ lead to the result

$$n^{1/2}\left(\hat{\beta}_n - \beta\right) \Rightarrow \text{Normal}(0, B^{-1}CB^{-1}), \tag{4}$$

where B is the covariance matrix of $X - E(X|T)$ and C is the covariance matrix of $\varepsilon\{X - E(X|T)\}$.

The least squares form of (3) can be used to show that if one ignores the measurement error and replaces X by W, the resulting estimate is inconsistent for β. The form, though, suggests even more. It is well known that in linear regression, inconsistency caused by the measurement error can be overcome by applying the so-called "correction for attenuation." In the context of semiparametric models, this suggests that we use the estimator

$$\begin{aligned} \hat{\beta}_n = {} & \left[\sum_{i=1}^{n} \{W_i - \hat{g}_{w,h}(T_i)\}\{W_i - \hat{g}_{w,h}(T_i)\}^{\mathrm{T}} - n\Sigma_{uu}\right]^{-1} \\ & \times \sum_{i=1}^{n} \{W_i - \hat{g}_{w,h}(T_i)\}\{Y_i - \hat{g}_{y,h}(T_i)\}. \end{aligned} \tag{5}$$

The estimator (5) can be derived in much the same way as the Severini–Staniswalis estimator. For every β, let $\hat{g}(T, \beta)$ maximize the weighted likelihood, ignoring the measurement error. Then form the estimators of β via a negatively penalized operation

$$(6) \qquad \text{minimize} \sum_{i=1}^{n} \left\{Y_i - W_i^{\mathrm{T}}\beta - \hat{g}(T_i, \beta)\right\}^2 - \beta^{\mathrm{T}}\Sigma_{uu}\beta.$$

The negative sign in the second term in (6) looks odd until one remembers that the effect of the measurement error is attenuation, that is, to underestimate β in absolute value when it is scalar, and thus one must correct for attenuation by making β larger, not by shrinking it further towards zero.

In this paper, we analyze the estimate (5), and show that it is consistent, asymptotically normally distributed with a variance different from (4). Just as in the Severini–Staniswalis algorithm, the kernel weight with ordinary bandwidths of order $h \sim n^{-1/5}$ may be used.

The outline of the paper is as follows. In Section 2, we define the weighting scheme to be used and hence the estimators of β and $g(\cdot)$. Section 3 is the statement of the main results for β, while the results for $g(\cdot)$ are stated in Section 4. Section 5 states the corresponding results when the measurement error variance Σ_{uu} is estimated. Section 6 gives a numerical illustration. Final remarks are given in Section 7. All proofs are delayed until the Appendix.

2. Definition of the estimators. For technical convenience we will assume that the T_i are confined to the interval $[0, 1]$. Throughout, we shall employ $C(0 < C < \infty)$ to denote some constant not depending on n, but which may assume different values at each appearance. In our proofs and statement of results, we will let the X's be independent random variables.

Let $\omega_{ni}(t) = \omega_{ni}(t; T_1, \dots, T_n)$ be weight functions depending only on the design points $T_1, \dots, T_n$. For example,

$$(7) \qquad \omega_{ni}(t) = \frac{1}{h_n}\int_{s_{i-1}}^{s_i} K\left(\frac{t-s}{h_n}\right) ds, \qquad 1 \le i \le n,$$

where $s_0 = 0$, $s_n = 1$ and $s_i = (1/2)(T_{(i)} + T_{(i+1)})$, $1 \le i \le n-1$, $T_{(i)}$ are the order statistics of T_i, h_n is a sequence of bandwidth parameters which tends to zero as $n \to \infty$ and $K(\cdot)$ is a nonnegative kernel function, which is supposed to have compact support and to satisfy

$$\operatorname{supp}(K) = [-1, 1], \sup|K(x)| \le C < \infty,$$

$$\int K(u)\, du = 1 \quad \text{and} \quad K(u) = K(-u).$$

In this paper, for any sequence of variables or functions $(S_1, \dots, S_n)$, we always denote $\mathbf{S}^{\mathrm{T}} = (S_1, \dots, S_n)$, $\tilde{S}_i = S_i - \sum_{j=1}^{n}\omega_{nj}(T_i)S_j$, $\tilde{S}^{\mathrm{T}} = (\tilde{S}_1, \dots, \tilde{S}_n)$. For example, $\tilde{\mathbf{W}}^{\mathrm{T}} = (\tilde{W}_1, \dots, \tilde{W}_n)$, $\tilde{W}_i = W_i - \sum_{j=1}^{n}\omega_{nj}(T_i)W_j$; $\tilde{g}_i = g(T_i) - \sum_{k=1}^{n}\omega_{nk}(T_i)g(T_k)$, $\tilde{\mathbf{G}} = (\tilde{g}_1, \dots, \tilde{g}_n)^{\mathrm{T}}$.

The fact that $g(t) = E(Y_i - X_i^{\mathrm{T}}\beta | T = t) = E(Y_i - W_i^{\mathrm{T}}\beta | T = t)$ suggests

$$\hat{g}_n(t) = \sum_{j=1}^{n} \omega_{nj}(t)\big(Y_j - W_j^{\mathrm{T}}\hat{\beta}_n\big) \tag{8}$$

as the estimator of $g(t)$.

In some cases, it may be reasonable to assume that the model errors ε_i are homoscedastic with common variance σ^2. In this event, since $E\{Y_i - X_i^{\mathrm{T}}\beta - g(T_i)\}^2 = \sigma^2$ and $E\{Y_i - W_i^{\mathrm{T}}\beta - g(T_i)\}^2 = E\{Y_i + X_i^{\mathrm{T}}\beta - g(T_i)\}^2 + \beta^{\mathrm{T}}\Sigma_{uu}\,\beta$, we define

$$\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}\left(\tilde{Y}_i - \tilde{W}_i^{\mathrm{T}}\hat{\beta}_n\right)^2 - \hat{\beta}_n^{\mathrm{T}}\Sigma_{uu}\,\hat{\beta}_n \tag{9}$$

as the estimator of σ^2.

3. Main results.

Let the components of X_i be $X_i = (X_{ij})$ be denoted by X_{ij}. Denote $h_j(T_i) = E(X_{ij}|T_i)$, $V_i = X_i - E(X_i|T_i)$, $1 \le i \le n$, $1 \le j \le p$. We make the following assumptions.

ASSUMPTION 1.1. $\sup_{0 \le t \le 1} E(\|X_1\|^4 | T = t) < \infty$ and $B = E(V_1 V_1^{\mathrm{T}})$ is a positive definite matrix.

ASSUMPTION 1.2. $g(\cdot)$ and $h_j(\cdot)$ are Lipschitz continuous of order 1.

ASSUMPTION 1.3. The weight functions $\omega_{ni}(\cdot)$ satisfy:

$$\text{(i)}\qquad \max_{1\le i\le n}\sum_{j=1}^{n}\omega_{nj}(T_i) = O_P(1),$$

$$\text{(ii)}\qquad \max_{1\le i,j\le n}\omega_{ni}(T_j) = O_P(b_n),$$

$$\text{(iii)}\qquad \max_{1\le i\le n}\sum_{j=1}^{n}\omega_{nj}(T_i)I\big(|T_j - T_i| > c_n\big) = O_P(c_n),$$

where $b_n = n^{-4/5}$, $c_n = n^{-1/5}\log n$.

ASSUMPTION 1.4. $E(\varepsilon_i) = E(U_i) = 0$ and $\sup_i E(\varepsilon_i^4 + \|U_i\|^4) < \infty$.

Our two main results concern the limit distributions of the estimates of β and σ^2.

THEOREM 3.1. *Suppose that Assumptions* 1.1–1.4 *hold. Then* $\hat{\beta}_n$ *is an asymptotically normal estimator*; *that is,*

$$n^{1/2}\big(\hat{\beta}_n - \beta\big) \to_d N(0, B^{-1}\Gamma B^{-1}),$$

with $\Gamma = E[(\varepsilon - U^{\mathrm{T}}\beta)\{X - E(X|T)\}]^{\otimes 2} + E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2} + E(UU^{\mathrm{T}}\varepsilon^2)$, *where* $A^{\otimes 2} = AA^{\mathrm{T}}$. *Note that* $\Gamma = E(\varepsilon - U^{\mathrm{T}}\beta)^2 B = E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2} + \Sigma_{uu}\sigma^2$ *if* ε *is homoscedastic and independent of* (X, T).

THEOREM 3.2. *Suppose that the conditions of Theorem* 3.1 *hold, and that the* ε*'s are homoscedastic with variance* σ^2 *and independent of* (X_i, T_i). *Then*

$$n^{1/2}\left(\hat{\sigma}_n^2 - \sigma^2\right) \to_d N(0, \sigma_*^2),$$

where $\sigma_*^2 = E\{(\varepsilon - U^{\mathrm{T}}\beta)^2 - (\beta^{\mathrm{T}}\Sigma_{uu}\beta + \sigma^2)\}^2$.

REMARKS. (i) It is relatively easy to estimate the covariance matrix of $\hat{\beta}_n$. Let $\dim(X)$ be the number of the components of X. A consistent estimate of B is just

$$\{n - \dim(X)\}^{-1}\sum_{i=1}^{n}\{W_i - \hat{g}_{w,h}(T_i)\}^{\otimes 2} - \Sigma_{uu} =_{\mathrm{def}} B_n.$$

In the general case, one can use (25) below to construct a consistent sandwich-type estimate of Γ, namely,

$$n^{-1}\sum_{i=1}^{n}\left\{\tilde{W}_i\left(\tilde{Y}_i - \tilde{W}_i^{\mathrm{T}}\hat{\beta}_n\right) + \Sigma_{uu}\hat{\beta}_n\right\}^{\otimes 2}.$$

In the homoscedastic case, namely that ε_i is independent of (X_i, T_i, U_i) with variance σ^2 and with U being normally distributed, a different formula can be used. Let $\mathscr{C}(\beta) = E\{(UU^{\mathrm{T}} - \Sigma_{uu})\beta\}^{\otimes 2}$. Then a consistent estimate of Γ is

$$\left(\hat{\sigma}_n^2 + \hat{\beta}_n^{\mathrm{T}}\Sigma_{uu}\hat{\beta}_n\right)\hat{B}_n + \hat{\sigma}_n^2\Sigma_{uu} + \mathscr{C}\left(\hat{\beta}_n\right).$$

(ii) In the classical functional model [Kendall and Stuart (1992)], instead of obtaining an estimate of Σ_{uu} through replication, it is instead assumed that the ratio of Σ_{uu} to σ^2 is known. Without loss of generality, we set this ratio equal to the identity matrix. The resulting analogue of the parametric estimators to the partially linear model is to solve the following minimization problem:

$$\sum_{i=1}^{n}\left|\frac{\tilde{Y}_i - \tilde{W}_i^{\mathrm{T}}\beta}{\sqrt{1 + \|\beta\|^2}}\right|^2 = \min!,$$

here and in the sequel $\|\cdot\|$ denotes the Euclidean norm. One can use the techniques of this paper to show that this estimator is consistent and asymptotically normally distributed. The asymptotic variance of the estimate of β for the case where ε_i is independent of (X_i, T_i) can be shown to be

$$B^{-1}\left[\left(1 + \|\beta\|^2\right)^2\sigma^2 B + \frac{E\left\{\left(\varepsilon - U^{\mathrm{T}}\beta\right)^2\Gamma_1\Gamma_1^{\mathrm{T}}\right\}}{1 + \|\beta\|^2}\right]B^{-1},$$

where $\Gamma_1 = (1 + \|\beta\|^2)U + (\varepsilon - U^{\mathrm{T}}\beta)\beta$.

4. Asymptotic results for the nonparametric part.

THEOREM 4.1. *Suppose that Assumptions* 1.1–1.4 *hold and that* $\omega_{ni}(t)$ *are Lipschitz continuous of order* 1 *for all* $i = 1,\dots,n$. *Then for fixed* T_i, *the asymptotic bias and asymptotic variance of* $\hat{g}_n(t)$ *are, respectively,* $\sum_{i=1}^n \omega_{ni}(t) g(T_i) - g(t)$ *and* $\sum_{i=1}^n \omega_{ni}^2(t)(\beta^{\mathsf T}\Sigma_{uu}\beta + \sigma^2)$. *These are all of order* $O(n^{-2/5})$ *for the kernel estimators.*

5. Estimated error variance. Although in some cases the measurement error covariance matrix Σ_{uu} has been established by independent experiments, in others it is unknown and must be estimated. The usual method of doing so [Carroll, Ruppert and Stefanski (1995), Chapter 3] is by partial replication, so that we observe $W_{ij} = X_i + U_{ij}$, $j = 1,\dots,m_i$.

For notational convenience, we consider here only the case that $m_i \le 2$ and assume that a fraction δ of the data has such replicates. Let $\overline{W}_i$ be the sample mean of the replicates. Then a consistent, unbiased method of moments estimate for Σ_{uu} is

$$\hat{\Sigma}_{uu} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} \left(W_{ij} - \overline{W}_i\right)\left(W_{ij} - \overline{W}_i\right)^{\mathsf T}}{\sum_{i=1}^n (m_i - 1)}.$$

The estimator changes only slightly to accommodate the replicates, becoming

$$\hat{\beta}_n = \left[\sum_{i=1}^n \left\{\overline{W}_i - \hat{g}_{w,h}(T_i)\right\}^{\otimes 2} - n(1 - \delta/2)\hat{\Sigma}_{uu}\right]^{-1} \times \sum_{i=1}^n \left\{\overline{W}_i - \hat{g}_{w,h}(T_i)\right\}\left\{Y_i - \hat{g}_{y,h}(T_i)\right\}, \tag{10}$$

where $\hat{g}_{w,h}(\cdot)$ is the kernel regression of the $\overline{W}_i$'s on T_i.

Using the techniques in the Appendix, one can show that the limit distribution of (10) is Normal$(0, B^{-1}\Gamma_2 B^{-1})$, with

$$\begin{aligned}\Gamma_2 &= (1 - \delta)E\left[(\varepsilon - U^{\mathsf T}\beta)\{X - E(X|T)\}\right]^{\otimes 2} \\ &\quad + \delta E\left[(\varepsilon - \overline{U}^{\mathsf T}\beta)\{X - E(X|T)\}\right]^{\otimes 2} \\ &\quad + (1 - \delta)E\left(\left[\{UU^{\mathsf T} - (1 - \delta/2)\Sigma_{uu}\}\beta\right]^{\otimes 2} + UU^{\mathsf T}\varepsilon^2\right) \\ &\quad + \delta E\left(\left[\left\{\overline{UU}^{\mathsf T} - (1 - \delta/2)\Sigma_{uu}\right\}\beta\right]^{\otimes 2} + \overline{UU}^{\mathsf T}\varepsilon^2\right).\end{aligned} \tag{11}$$

In (11), $\overline{U}$ refers to the mean of two U's. In the case that ε is independent of (X, T), the sum of the first two terms simplifies to $\{\sigma^2 + \beta^{\mathsf T}(1 - \delta/2)\Sigma_{uu}\beta\}B$.

Standard error estimates can also be derived. A consistent estimate of B is

$$\hat{B}_n = \{n - \dim(X)\}^{-1} \sum_{i=1}^n \left\{\overline{W}_i - \hat{g}_{w,h}(T_i)\right\}^{\otimes 2} - (1 - \delta/2)\hat{\Sigma}_{uu}.$$

Estimates of Γ_2 can also be easily developed. In the homoscedastic case with normal errors, the sum of the first two terms can be estimated by $(\hat{\sigma}_n^2 + (1 - \delta/2)\hat{\beta}_n^{\mathsf{T}} \hat{\Sigma}_{uu} \hat{\beta}_n)\hat{B}_n$. The sum of the last two terms is a deterministic function of $(\beta, \sigma^2, \Sigma_{uu})$, and these estimates are simply substituted into the formula.

A general sandwich-type estimator is developed as follows. Define $\kappa = n^{-1}\sum_{i=1}^{n} m_i^{-1}$, and define

$$R_i = \bar{\tilde{W}}_i\left(\tilde{Y}_i - \bar{\tilde{W}}_i^{\mathsf{T}}\hat{\beta}_n\right) + \frac{\hat{\Sigma}_{uu}\,\hat{\beta}_n}{m_i} + \frac{\kappa}{\delta}(m_i - 1)\left\{\frac{1}{2}(W_{i1} - W_{i2})(W_{i1} - W_{i2})^{\mathsf{T}} - \hat{\Sigma}_{uu}\right\}.$$

Then a consistent estimate of Γ_2 is the sample covariance matrix of the R_i's.

6. Numerical example.

To illustrate our method, we consider data from the Framingham Heart Study. We consider $n = 1615$ males with Y being their average blood pressure in a fixed two-year period, T being their age and W being the logarithm of the observed cholesterol level, for which there are two replicates.

We do two analyses. In the first, we use both cholesterol measurements, so that in the notation of Section 5, $\delta = 1$. In this analysis, there is not a great deal of measurement error. Thus, in our second analysis, which is given for illustrative purposes, we use only the first cholesterol measurement, but fix the measurement error variance at the value obtained in the first analysis, in which case $\delta = 0$. For nonparametric fitting, we chose the bandwidth using cross-validation to predict the response. In precise terms, we compute the squared error using a geometric sequence of 191 bandwidths ranging in [1, 20]. The optimal bandwidth is selected to minimize the squared error among these 191 candidates. An analysis ignoring the measurement error found some curvature in T; see Figure 1 for the estimate of $g(T)$. All calculations were performed in XploRe [Härdle, Klinke and Turlach (1995)].

Our results are as follows. First, consider the case that the measurement error is estimated and both cholesterol values are used to estimate Σ_{uu}. The estimator of β ignoring the measurement error is 9.438, with estimated standard error 0.187. When we account for the measurement error, the estimate increases to $\hat{\beta} = 12.540$ and the standard error increases to 0.195.

In the second analysis, we fix the measurement error variance and use only the first cholesterol value. The estimator of β ignoring the measurement error was 10.744, with estimated standard error 0.492. When we account for the measurement error, the estimate increases to $\hat{\beta} = 13.690$ and the standard error increases to 0.495.

7. Discussion.

The nonparametric regression estimator (8) is based on locally weighted averages. Clearly, results such as Theorem 3.1 should apply if (8) is replaced by a locally linear kernel regression estimator or by a spline estimator, although our proofs do not apply to these estimators.

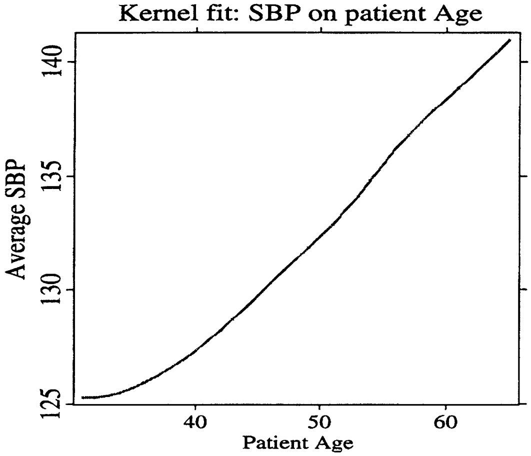


FIG. 1. *Estimate of the function $g(T)$ in the Framingham data ignoring measurement error.*

We have treated the case that the parametric part X of the model has measurement error and the nonparametric part T is measured exactly. An interesting problem is to interchange the roles of X and T, so that the parametric part is measured exactly and the nonparametric part is measured with error, that is, $E(Y|X,T) = \theta T + g(X)$. Fan and Truong (1993) have shown in this case that with normally distributed measurement error, the nonparametric function $g(\cdot)$ can be estimated only at logarithmic rates and not with rate $n^{-2/5}$. We conjecture even so that θ can be estimated at parametric rates, but this remains an open problem.

APPENDIX

In this Appendix, we prove several required lemmas. Lemma A.1 provides bounds for $h_j(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)h_j(T_k)$ and $g(T_i) - \sum_{k=1}^n \omega_{nk}(T_i)g(T_k)$. The proof is immediate.

LEMMA A.1. *Suppose that Assumptions* 1.1–1.4 *hold. Then*

$$\max_{1\le i\le n}\left|G_j(T_i) - \sum_{k=1}^{n} \omega_{nk}(T_i)G_j(T_k)\right| = O_p(c_n) \quad \mathit{for}\, j = 0,\dots,p,$$

where $G_0(\cdot) = g(\cdot)$ and $G_l(\cdot) = h_l(\cdot)$ for $l = 1,\dots,p$.

LEMMA A.2. *If Assumptions* 1.1–1.4 *hold, then $n^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}} = B + o_P(1)$.*

PROOF. Denote $\overline{h}_{ns}(T_i) = h_s(T_i) - \sum_{k=1}^{n} \omega_{nk}(T_i) X_{ks}$. It follows from $X_{js} = h_s(T_j) + V_{js}$ that the (s, m)th element of $\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}}$ $(s, m = 1, \ldots, p)$ is

$$\begin{aligned}\sum_{j=1}^{n} \tilde{X}_{js}\tilde{X}_{jm} &= \sum_{j=1}^{m} V_{js}V_{jm} + \sum_{j=1}^{n} \overline{h}_{ns}(T_j)V_{jm} \\ &\quad + \sum_{j=1}^{n} \overline{h}_{nm}(T_j)V_{js} + \sum_{j=1}^{n} \overline{h}_{ns}(T_j)\overline{h}_{nm}(T_j) \\ &=_{\text{def}} \sum_{j=1}^{n} V_{js}V_{jm} + \sum_{q=1}^{3} R_{nsm}^{(q)}.\end{aligned}$$

The strong law of large numbers implies that $n^{-1}\sum_{i=1}^{n} V_i V_i^{\mathsf{T}} = B + o_P(1)$, and Lemma A.1 means $R_{nsm}^{(3)} = o_P(n)$, which together with the Cauchy–Schwarz inequality shows that $R_{nsm}^{(1)} = o_P(n)$ and $R_{nsm}^{(2)} = o_P(n)$. This completes the proof of the lemma. □

LEMMA A.3 (Bernstein's inequality). *Let $\Gamma_1, \ldots, \Gamma_n$ be independent random variables with zero means and bounded ranges, $|\Gamma_i| \le M$. Then for each $\eta > 0$,*

$$P\left\{\left|\sum_{i=1}^{n} \Gamma_i\right| > \eta\right\} \le 2\exp\left\{-\eta^2 \Big/ \left[2\left\{\sum_{i=1}^{n} \operatorname{var}(\Gamma_i) + M\eta\right\}\right]\right\}.$$

Denote $\varepsilon_j' = \varepsilon_j I(|\varepsilon_j| \le n^{1/4})$ and $\varepsilon_j'' = \varepsilon_j - \varepsilon_j' = \varepsilon_j I(|\varepsilon_j| > n^{1/4})$, $j = 1, \ldots, n$. We next establish several results for nonparametric regression.

LEMMA A.4. *Assume that Assumptions* 1.3 *and* 1.4 *hold. Then*

$$\max_{1 \le i \le n} \left|\sum_{k=1}^{n} \omega_{nk}(T_i)\,\varepsilon_k\right| = o_P\{n^{-2/5}\log(n)\}.$$

PROOF. Fix $L > 0$ but arbitrarily large. Let

$$B_{nL} = \left\{\max_{1 \le i \le n} \sum_{j=1}^{n} w_{nj}(T_i) \le L, \max_{1 \le i, j \le n} w_{nj}(T_i) \le L b_n\right\}.$$

Then

$$\begin{aligned} & P\left\{\max_{1 \le i \le n}\left|\sum_{j=1}^{n} w_{nj}(T_i)\,\varepsilon_j\right| > n^{-2/5}\log(n)\right\} \\ & \quad \le P\{I(B_{nL}) = 0\} \qquad (12) \\ & \qquad + P\left\{\max_{1 \le i \le n}\left|\sum_{j=1}^{n} w_{nj}(T_i)\,\varepsilon_j\right| > n^{-2/5}\log(n), I(B_{nL}) = 1\right\}.\end{aligned}$$

Since by Assumption 1.3, $P\{I(B_{nL}) = 1\}$ can be made arbitrarily small by choosing L sufficiently large, it suffices to show that the second term in (12) converges to zero for any L.

Application of Bernstein's inequality to (12) is complicated by the fact that the terms $w_{nj}(T_i)$ and $I(B_{nL}) = 1$ are random. We first condition on these terms and will later uncondition. For sufficiently large C, first note that

$$P\left\{\max_{1\le i\le n}\left|\sum_{j=1}^n w_{nj}(T_i)\{\varepsilon_j' - E(\varepsilon_j')\}\right| > Cn^{-2/5}\log(n)|\{w_{nj}(T_i)\},\quad I(B_{nL}) = 1\right\}$$
$$\le \sum_{i=1}^n P\left\{\left|\sum_{j=1}^n w_{nj}(T_i)\{\varepsilon_j' - E(\varepsilon_j')\}\right| > Cn^{-2/5}\log(n)|\{w_{nj}(T_i)\},\quad I(B_{nL}) = 1\right\}.$$

Now apply Bernstein's inequality with $\eta = Cn^{-2/5}\log(n)$ and $M = 2Lb_n n^{1/4}$. Then the right-hand side of the last expression is bounded by

$$(13)\qquad 2I(B_{nL})\sum_{i=1}^n \exp\left\{-\frac{C^2 n^{-4/5}\log^2(n)}{4LCb_n n^{1/4-2/5}\log(n) + 2\Sigma_{j=1}^n w_{nj}^2(T_i)\operatorname{var}(\varepsilon_j')}\right\}.$$

First note that $b_n = n^{-4/5}$ and $\operatorname{var}(\varepsilon_j') < \infty$. On the set that $I(B_{nL}) = 1$, we have thus that

$$\sum_{j=1}^n w_{nj}^2(T_i) \le \sum_{j=1}^n w_{nj}(T_i)\max_{1\le i,j\le n} w_{nj}(T_i) \le L^2 b_n.$$

This means that (13) is bounded by $2nI(B_{nL})\exp\{-(C/L)\log(n)\} \le n^{-3/2}$ for sufficiently large C. Since this last expression is independent of the $\{w_{nj}(T_i)\}$ except through $I(B_{nL})$, we have that

$$P\left\{\max_{1\le i\le n}\left|\sum_{j=1}^n w_{nj}(T_i)\{\varepsilon_j' - E(\varepsilon_j')\}\right| > Cn^{-2/5}\log(n)|I(B_{nL}) = 1\right\} \le n^{-3/2}.$$

This shows that

$$(14)\qquad \max_{1\le i\le n}\left|\sum_{j=1}^n w_{nj}(T_i)\{\varepsilon_j' - E(\varepsilon_j')\}\right| = o_p\{n^{-2/5}\log(n)\}.$$

Now consider $V_n = \max_{1\le i\le n}\Sigma_{j=1}^n w_{nj}(T_i)\{\varepsilon_j'' - E(\varepsilon_j'')\}$. Let p and q be such that $1 \le p < 2$, $1/p + 1/q = 1$ and $1/q < 2/5 - 1/4$. By Hölder's inequality,

$$|V_n| \le \max_{1\le i\le n}\left\{\sum_{j=1}^n w_{nj}^q(T_i)\right\}^{1/q}\left\{\sum_{j=1}^n\left|\varepsilon_j'' - E(\varepsilon_j'')\right|^p\right\}^{1/p}.$$

By Assumption 1.3(ii), $w_{nj}^q(T_i) = O_P(b_n^q)$ so that $\sum_j w_{nj}^q(T_i) = O_P(nb_n^q) = O_P(n^{1-4q/5})$, and thus

$$|V_n| \le O_P\{n^{(1-4q/5)/q}\}\left\{\sum_{j=1}^{n} \left|\varepsilon_j'' - E(\varepsilon_j'')\right|^p\right\}^{1/p}.$$

Clearly,

$$n^{-1}\sum_{j=1}^{n}\left[\left|\varepsilon_j'' - E(\varepsilon_j'')\right|^p - E\left\{\left|\varepsilon_j'' - E(\varepsilon_j'')\right|^p\right\}\right] = o_P(1). \tag{15}$$

Also, again using Hölder's inequality,

$$E|\varepsilon_j''|^p = E\left\{|\varepsilon_j|^p I\left(\varepsilon_j > n^{1/4}\right)\right\} \le \left(E|\varepsilon_j|^4\right)^{p/4}\left\{P\left(|\varepsilon_j| > n^{1/4}\right)\right\}^{1-p/4},$$

which by Chebyshev's inequality is bounded by $\le n^{-1+p/4}(E|\varepsilon_j|^4)^{p/4}$. It thus follows that

$$\sum_{j=1}^{n} E\left|\varepsilon_j'' - E(\varepsilon_j'')\right|^p = O_P(n^{p/4}). \tag{16}$$

Replacing (16) into (15), we get

$$\sum_{j=1}^{n}\left|\varepsilon_j'' - E(\varepsilon_j'')\right|^p = O_P(n^{p/4}),$$

where, along with the fact that $1/q < 2/5 - 1/4$, we find that

$$\max_{1\le i\le n}\sum_{j=1}^{n} w_{nj}(T_i)\left\{\varepsilon_j'' - E(\varepsilon_j'')\right\} = O_P(n^{(1-4q/5)/q+1/4}) = o_P(n^{-2/5}).$$

This completes the proof of Lemma A.4. □

LEMMA A.5. *Suppose that Assumptions* 1.1–1.4 *hold. Then*

$$\sum_{i=1}^{n} U_i\tilde{g}_i = o_p(n^{1/2}),$$
$$\sum_{i=1}^{n} \varepsilon_i\tilde{g}_i = o_p(n^{1/2}).$$

The same holds if $g(T_i)$ *is replaced by* $h_j(T_i)$.

PROOF. We prove only the first step, as the other steps follow in a similar fashion. Let $\xi_n = n^{1/2}/\log(n)$:

$$P\left(\left|\sum_{i=1}^{n} U_i\tilde{g}_i\right| > \xi_n\right) \le P\left(\left|\sum_{i=1}^{n} U_i\tilde{g}_i\right| > \xi_n, \max_i|\tilde{g}_i| \le c_n\log n\right) + P\left(\max_i|\tilde{g}_i| > c_n\log n\right).$$

The second term is $o_P(1)$ by Lemma A.1. For the first term, let r_i be the event that $|\tilde{g}_i| \le c_n \log(n)$. Then,

$$\begin{aligned} &P\left[\left|\sum_{i=1}^{n} U_i\tilde{g}_i\right| > \xi_n, \{I(r_i) = 1\,\forall\, i)\}\right] \\ &\quad \le \xi_n^{-2} \sum_{i=1}^{n} E\big[U_i\tilde{g}_i\{I(r_i) = 1\}\big]^2 \qquad (17) \\ &\qquad + \xi_n^{-2} \sum_{i \ne k}^{n} E\big[U_iU_k\tilde{g}_i\tilde{g}_k I(r_k) = 1\,\forall\, k\}\big]. \end{aligned}$$

Since $\tilde{g}_i\{I(r_i) = 1\} \le c_n \log(n)$ is independent of U_i, the first term in (17) is $O\{n\xi_n^{-2}c_n^2 \log^2(n)\} = o(1)$. The second term is easily seen to equal zero. □

LEMMA A.6. *Suppose that Assumptions* 1.1–1.4 *hold. Then*

$$n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j U_i = o_P(1),$$

$$n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j \varepsilon_i = o_P(1),$$

$$n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)U_j U_i = o_P(1).$$

PROOF. We prove only the first step, as the other steps follow in a similar fashion. Let r_{ij} be the event that $|w_{nj}(T_i)| \le Cb_n \log n$:

$$\begin{aligned} &P\left\{n^{-1/2}\left|\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j U_i\right| > \xi\right\} \\ &\quad \le P\left\{n^{-1/2}\left|\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j U_i\right| > \xi,\ I(r_{ij} = 1\,\forall\, i, j)\right\} \\ &\qquad + P\left\{\max_{i,j}\left|w_{nj}(T_i)\right| > Cb_n \log n\right\}. \end{aligned}$$

The second term tends to zero by Assumption 1.3(ii). For the first term, note that

$$\begin{aligned} &P\left\{n^{-1/2}\left|\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j U_i\right| > \xi,\ I(r_{ij} = 1\,\forall\, i, j)\right\} \\ &\quad \le n^{-1}\xi^{-2} E\left\{\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j U_i I(r_{ij} = 1\,\forall\, i, j)\right\}^2 \\ &\quad = n^{-1}\xi^{-2} \sum_{i=1}^{n} E\left\{\sum_{j=1}^{n} \omega_{nj}(T_i)\varepsilon_j I(r_{ij} = 1\,\forall\, i, j)\right\}^2 EU_i^2. \end{aligned}$$

The last equation holds because U_i and $\sum_{j=1}^{n}\omega_{nj}(T_i)\varepsilon_j I(r_{ij} = 1\ \forall\, i, j)$ are independent for each i, and U_i are iid with mean zero. It suffices to prove

$$\max_i E\left\{\sum_{j=1}^{n} \omega_{nj}(T_i)\,\varepsilon_j I\big(r_{ij} = 1\,\forall\, i, j\big)\right\}^2 \to 0.$$

In fact,

$$\begin{aligned} &E\left\{\sum_{j=1}^{n} \omega_{nj}(T_i)\,\varepsilon_j I\big(r_{ij} = 1\,\forall i, j\big)\right\}^2 \\ &\quad = \sum_{j=1}^{n} E\big\{\omega_{nj}(T_i)\,\varepsilon_j I\big(r_{ij} = 1\,\forall\, i, j\big)\big\}^2 \\ &\qquad + \sum_{j\neq k}^{n} E\big\{\omega_{nj}(T_i)\,\varepsilon_j\,\omega_{nk}(T_i)\,\varepsilon_k I\big(r_{ij} = 1\,\forall\, i, j\big)\big\}. \end{aligned}$$

The second term equals zero. The first term equals

$$\sum_{j=1}^{n} E\Big[\big\{\omega_{nj}(T_i)\,\varepsilon_j\big\}^2\big\{I(r_{ij}) = 1\,\forall\, i, j\big\}\Big],$$

and this is $O\{nb_n^2 \log^2(n)\} = o(1)$, as required. □

LEMMA A.7. *Assume that Assumptions* 1.1–1.4 *hold. Then*

$$p\lim_{n\to\infty} n^{-1}\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}} = B + \Sigma_{uu}, \tag{18}$$

$$p\lim_{n\to\infty} n^{-1}\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{Y}} = B\beta, \tag{19}$$

$$p\lim_{n\to\infty} n^{-1}\tilde{\mathbf{Y}}^{\mathsf{T}}\tilde{\mathbf{Y}} = \beta^{\mathsf{T}}B\beta + \sigma^2. \tag{20}$$

PROOF. Since $W_i = X_i + U_i$ and $\tilde{W}_i = \tilde{X}_i + \tilde{U}_i$, for the (s, m) matrix element we obtain

$$\begin{aligned} n^{-1}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}})_{sm} &= n^{-1}(\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}})_{sm} + n^{-1}(\tilde{\mathbf{U}}^{\mathsf{T}}\tilde{\mathbf{X}})_{sm} \\ &\quad + n^{-1}(\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{U}})_{sm} + n^{-1}(\tilde{\mathbf{U}}^{\mathsf{T}}\tilde{\mathbf{U}})_{sm}. \end{aligned} \tag{21}$$

First, we prove that the second and third terms converge to zero. It follows from the strong law of large numbers and Lemma A.2 that

$$n^{-1}\sum_{j=1}^{n} X_{js}U_{jm} \to 0 \quad \text{a.s.} \tag{22}$$

Observe that

$$n^{-1}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{U}})_{sm} = n^{-1}\Bigg[\sum_{j=1}^{n} X_{js}U_{jm} - \sum_{j=1}^{n}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)X_{ks}\Bigg\}U_{jm} - \sum_{j=1}^{n}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\Bigg\}X_{js} + \sum_{j=1}^{n}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)X_{ks}\Bigg\}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\Bigg\}\Bigg].$$

Similarly to the proof of Lemma A.4, we can prove that

$$\sup_{1\le j\le n}\left|\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\right| = o_P(1),$$

which, together with (22) and Assumption 1.3(ii), imply that each term above tends to zero. The same reason implies that $n^{-1}(\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{X}})_{sm}$ also tends to zero.

Second, we prove

$$n^{-1}(\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{U}})_{sm} \to \sigma_{sm}^2, \tag{23}$$

where σ_{sm}^2 is the (s,m)th element of Σ_{uu},

$$n^{-1}(\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{U}})_{sm} = n^{-1}\Bigg[\sum_{j=1}^{n} U_{js}U_{jm} - \sum_{j=1}^{n}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{ks}\Bigg\}U_{jm} - \sum_{j=1}^{n}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\Bigg\}U_{js} + \sum_{j=1}^{n}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{ks}\Bigg\}\Bigg\{\sum_{k=1}^{n}\omega_{nk}(T_j)U_{km}\Bigg\}\Bigg].$$

Obviously, $n^{-1}\sum_{j=1}^{n}U_{js}U_{jm} \to \sigma_{sm}^2$. It follows from Lemmas A.4 and A.6 that (23) holds. Using (21), (23) and the arguments for $n^{-1}(\tilde{\mathbf{U}}^{\mathrm{T}}\tilde{\mathbf{X}})_{sm} \to 0$ and $n^{-1}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{U}})_{sm} \to 0$, we complete the proof of (18).

We now prove (19). Note that $\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{Y}} = \tilde{\mathbf{W}}^{\mathrm{T}}(\tilde{\mathbf{X}}\beta + \tilde{\mathbf{G}} + \tilde{\varepsilon})$. From Lemma 1, $\sum_{j=1}^{n}\tilde{g}_j^2 = O_P(c_n^2 n)$, so that

$$\left|\sum_{j=1}^{n}X_{js}\tilde{g}_j\right| \le \left(\sum_{j=1}^{n}X_{js}^2\sum_{j=1}^{n}\tilde{g}_j^2\right)^{1/2} \le O_P(c_n n^{1/2})\left(\sum_{j=1}^{n}X_{js}^2\right)^{1/2} = O_P(Cnc_n)$$

and

$$(\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{G}})_s = \sum_{j=1}^{n}\tilde{X}_{js}\tilde{g}_j + \sum_{j=1}^{n}\tilde{U}_{js}\tilde{g}_j = \sum_{j=1}^{n}\Bigg\{X_{js} - \sum_{k=1}^{n}\omega_{nk}(T_j)X_{ks}\Bigg\}\tilde{g}_j + \sum_{j=1}^{n}\tilde{U}_{js}\tilde{g}_j.$$

Obviously, $n^{-1}\sum_{j=1}^{n}\tilde{U}_{js}\tilde{g}_j$ tends to zero. Therefore $n^{-1}(\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{G}})_s$ tends to zero.

The proof that $n^{-1}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\varepsilon})_s$ tends to zero is similar to that of $n^{-1}(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{U}})_s \to 0$. Combining the above arguments and (18), we complete the proof of (19). The proof of (20) can be completed by similar arguments. The details are omitted. □

LEMMA A.8. *Assume that Assumptions* 1.1–1.4 *hold. Then*

$$n^{-1/2}\sum_{i=1}^{n}\tilde{\varepsilon}_i\tilde{X}_i = n^{-1/2}\sum_{i=1}^{n}\varepsilon_i V_i + o_P(1),$$

$$n^{-1/2}\sum_{i=1}^{n}\tilde{X}_i\tilde{U}_i^{\mathsf{T}} = n^{-1/2}\sum_{i=1}^{n}V_i U_i^{\mathsf{T}} + o_P(1).$$

PROOF. We show only the first step, as the second step follows in a similar fashion. Let $h(T) = E(X|T)$ and $h_i = h(T_i)$. By a direct calculation,

$$n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\big(V_i - \tilde{X}_i\big) = n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\tilde{h}_i - n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\sum_{j=1}^{n}w_{nj}(T_i)\big\{X_j - h(T_j)\big\}.$$

The first term is $o_P(1)$ by Lemma A.4. The second term follows, using Assumption 1.1 by using the same method of proof as in Lemma A.6, upon remembering that for $j \neq k$,

$$E\big[\big\{X_j - h(T_j)\big\}\{X_k - h(T_k)\}|T_1,\dots,T_n)\big] = 0. \qquad \square$$

PROOF OF THEOREM 3.1. Denote $\Delta_n = (\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}} - n\Sigma_{uu})/n$. By Lemma A.7 and a direct calculation,

$$\begin{aligned} n^{1/2}\big(\hat{\beta}_n - \beta\big) &= n^{1/2}\Delta_n^{-1}\Big(\tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{Y}} + \tilde{\mathbf{W}}^{\mathsf{T}}\tilde{\mathbf{W}}\beta + n\Sigma_{uu}\,\beta\Big) \\ &= n^{-1/2}\Delta_n^{-1}\Big(\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{G}} + \tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\varepsilon} + \tilde{\mathbf{U}}^{\mathsf{T}}\tilde{\mathbf{G}} + \tilde{\mathbf{U}}^{\mathsf{T}}\tilde{\varepsilon} \\ &\qquad - \tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{U}}\beta - \tilde{\mathbf{U}}^{\mathsf{T}}\tilde{\mathbf{U}}\beta + n\Sigma_{uu}\,\beta\Big). \end{aligned}$$

By Lemmas A.1–A.2, A.4–A.6 and A.8, it is an easy calculation to show that

$$\begin{aligned} n^{1/2}\big(\hat{\beta}_n - \beta\big) &= n^{-1/2}\Delta_n^{-1} \\ &\quad \times \sum_{i=1}^{n}\big(V_i\varepsilon_i - V_iU_i^{\mathsf{T}}\beta + U_i\varepsilon_i - U_iU_i^{\mathsf{T}}\beta + \Sigma_{uu}\,\beta\big) \tag{24} \\ &\quad + o_P(1) \\ &=_{\text{def}} n^{-1/2}\sum_{i=1}^{n}\zeta_{in} + o_P(1). \tag{25} \end{aligned}$$

Since $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}V_i = 0$ and $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}V_iV_i^{\mathsf{T}} = B$ and $\sup_i E(\varepsilon_i^4 + \|U\|^4) < \infty$, it follows that the sequence of kth elements $\{\zeta_{in}^{(k)}\}$ of $\{\zeta_{in}\}$ $(k = 1,\dots,p)$ satisfy, for any given $\zeta > 0$, $n^{-1}\sum_{i=1}^{n}E\{\zeta_{in}^{(k)2}I(|\zeta_{in}^{(k)}| > \zeta n^{1/2})\} \to 0$ as $n \to \infty$. This means that the Lindeberg condition for the central limit theorem holds. Moreover, note that

$$\operatorname{cov}(\zeta_{ni}) = E\{V_i(\varepsilon_i - U_i^{\mathsf T}\beta)\}^{\otimes 2} + E\{(U_iU_i^{\mathsf T} - \Sigma_{uu})\beta\}^{\otimes 2} + E(U_iU_i^{\mathsf T}\varepsilon_i^2) + E(V_iU_i^{\mathsf T}\beta\beta^{\mathsf T}U_iU_i^{\mathsf T}) + E(U_iU_i^{\mathsf T}\beta\beta^{\mathsf T}U_i)V_i.$$

These arguments ensure that

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^n \operatorname{cov}(\zeta_{ni}) = E[(\varepsilon - U^{\mathsf T}\beta)\{X - E(X|T)\}]^{\otimes 2} + E\{(UU^{\mathsf T} - \Sigma_{uu})\beta\}^{\otimes 2} + E(UU^{\mathsf T}\varepsilon^2).$$

Theorem 3.1 now follows. □

PROOF OF THEOREM 3.2. Denote

$$A_n = n^{-1}\begin{bmatrix} \tilde{\mathbf{Y}}^{\mathsf T}\tilde{\mathbf{Y}} & \tilde{\mathbf{Y}}^{\mathsf T}\tilde{\mathbf{W}} \\ \tilde{\mathbf{W}}^{\mathsf T}\tilde{\mathbf{Y}} & \tilde{\mathbf{W}}^{\mathsf T}\tilde{\mathbf{W}} \end{bmatrix}; \qquad A = \begin{bmatrix} \beta^{\mathsf T}B\beta + \sigma^2 & \beta^{\mathsf T}B \\ B\beta & B + \Sigma_{uu} \end{bmatrix};$$

$$\tilde{A}_n = n^{-1}\begin{bmatrix} (\varepsilon + \mathbf{V}\beta)^{\mathsf T}(\varepsilon - \mathbf{V}\beta) & (\varepsilon + \mathbf{V}\beta)^{\mathsf T}(\mathbf{U} + \mathbf{V}) \\ (\mathbf{U} + \mathbf{V})^{\mathsf T}(\varepsilon + \mathbf{V}\beta) & (\mathbf{U} + \mathbf{V})^{\mathsf T}(\mathbf{U} + \mathbf{V}) \end{bmatrix}.$$

Note that $\hat{\sigma}_n^2 = (1, -\hat{\beta}_n^{\mathsf T})A_n(1, -\tilde{\beta}_n^{\mathsf T})^{\mathsf T} - \tilde{\beta}_n^{\mathsf T}\Sigma_{uu}\tilde{\beta}_n^{\mathsf T}$. A direct calculation using Lemma A.6 yields that $n^{1/2}(\hat{\sigma}_n^2 - \sigma^2) = n^{1/2}\sum_{j=1}^4 S_{jn} + n^{-1/2}(\varepsilon - \mathbf{U}\beta)^{\mathsf T}(\varepsilon - \mathbf{U}\beta) - n^{1/2}(\beta^{\mathsf T}\Sigma_{uu}\beta + \sigma^2) + o_P(1)$, where $S_{1n} = (1, -\hat{\beta}_n^{\mathsf T})(A_n - \tilde{A}_n)(1, -\hat{\beta}_n^{\mathsf T})^{\mathsf T}$, $S_{2n} = (1, -\hat{\beta}_n^{\mathsf T})(\tilde{A}_n - A)(0, \beta^{\mathsf T} - \hat{\beta}_n^{\mathsf T})^{\mathsf T}$, $S_{3n} = (0, \beta^{\mathsf T} - \hat{\beta}_n^{\mathsf T})(\tilde{A}_n - A)(1, -\beta^{\mathsf T})^{\mathsf T}$, $S_{4n} = -(\beta - \hat{\beta}_n)^{\mathsf T}B(\beta - \hat{\beta}_n)$. It follows from Theorem 3.1 and Lemma A.7 that $n^{1/2}S_{jn}$ converges to zero in probability for $j = 2, 3, 4$. To show that $n^{1/2}S_{1n} = o_P(1)$ is more detailed, but follows from Lemmas A.1, A.4–A.6. This means that

$$n^{1/2}(\hat{\sigma}_n^2 - \sigma^2) = n^{-1/2}\sum_{i=1}^n \{(\varepsilon_i - U_i^{\mathsf T}\beta)^2 - (\beta^{\mathsf T}\Sigma_{uu}\beta + \sigma^2)\} + o_P(1).$$

Theorem 3.2 now follows immediately. □

PROOF OF THEOREM 4.1. Since $\hat{\beta}_n$ is a consistent estimator of β, its asymptotic bias and variance equal the relative ones of $\sum_{j=1}^n \omega_{nj}(t)(Y_j - W_j^{\mathsf T}\beta)$, which is denoted by $\hat{g}_n^*(t)$. By a simple calculation,

$$E\hat{g}_n^*(t) - g(t) = \sum_{i=1}^n \omega_{ni}(t)g(T_i) - g(t),$$

$$\hat{g}_n^*(t) - E\hat{g}_n^*(t) = \sum_{i=1}^n \omega_{ni}^2(t)(\beta^{\mathsf T}\Sigma_{uu}\beta + \sigma^2).$$

Both terms are $O(n^{-2/5})$ by Lemma A.1 and Assumption 1.3(iii). Theorem 4.1 then follows. □

Acknowledgments. The authors are extremely grateful to the Editor, Associate Editor and an anonymous referee for their many valuable suggestions and comments, which greatly improved the presentation of the paper.

REFERENCES

CARROLL, R. J., RUPPERT, D. and STEFANSKI, L. A. (1995). *Nonlinear Measurement Error Models*. Chapman and Hall, New York.

CHEN, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–146.

CUZICK, J. (1992a). Semiparametric additive regression. *J. Roy. Statist. Soc. Ser. B* **54** 831–843.

CUZICK, J. (1992b). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Ann. Statist.* **20** 1129–1136.

ENGLE, R. F., GRANGER, C. W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

FAN, J. and TRUONG, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925.

FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York.

HÄRDLE, W., KLINKE, S. and TURLACH, B. A. (1995). *XploRe: An Interactive Statistical Computing Environment*. Springer, New York.

HECKMAN, N. E. (1986). Spline smoothing in partly linear models. *J. Roy. Statist. Soc. Ser. B* **48** 244–248.

KENDALL, M. and STUART, A. (1992). *The Advanced Theory of Statistics* **2**, 4th ed. Griffin, London.

LIANG, H. and HÄRDLE, W. (1997). Asymptotic normality of parametric part in partially linear heteroscedastic regression models. DP 33, SFB 373, Humboldt Univ. Berlin.

SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

H. LIANG
INSTITUTE OF SYSTEMS SCIENCE
CHINESE ACADEMY OF SCIENCES
BEIJING 100080
CHINA
E-MAIL: hliang@iss01.iss.ac.cn

W. HÄRDLE
INSTITUT FÜR STATISTIK UND ÖKONOMETRIE
HUMBOLDT-UNIVERSITÄT ZU BERLIN
D-10178 BERLIN
GERMANY
E-MAIL: haerdle@wiwi.hu-berlin.de

R. J. CARROLL
DEPARTMENT OF STATISTICS
TEXAS A & M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143
E-MAIL: carroll@stat.tamu.edu

Bayesian Smoothing and Regression Splines for Measurement Error Problems

Scott M. BERRY, Raymond J. CARROLL, and David RUPPERT

In the presence of covariate measurement error, estimating a regression function nonparametrically is extremely difficult, the problem being related to deconvolution. Various frequentist approaches exist for this problem, but to date there has been no Bayesian treatment. In this article we describe Bayesian approaches to modeling a flexible regression function when the predictor variable is measured with error. The regression function is modeled with smoothing splines and regression P-splines. Two methods are described for exploration of the posterior. The first, called the iterative conditional modes (ICM), is only partially Bayesian. ICM uses a componentwise maximization routine to find the mode of the posterior. It also serves to create starting values for the second method, which is fully Bayesian and uses Markov chain Monte Carlo (MCMC) techniques to generate observations from the joint posterior distribution. Use of the MCMC approach has the advantage that interval estimates that directly model and adjust for the measurement error are easily calculated. We provide simulations with several nonlinear regression functions and provide an illustrative example. Our simulations indicate that the *frequentist* mean squared error properties of the fully Bayesian method are better than those of ICM and also of previously proposed frequentist methods, at least in the examples that we have studied.

KEY WORDS: Bayesian methods; Efficiency; Errors in variables; Functional method; Generalized linear models; Kernel regression; Measurement error; Nonparametric regression; P-splines; Regression splines; SIMEX method; Smoothing splines; Structural modeling.

1. INTRODUCTION

In this article we present a fully Bayesian approach to the problem of nonparametric regression when the independent variables are measured with error. This is known to be an extremely difficult problem in terms of global rates of convergence. Fan and Truong (1993) showed that for additive normally distributed measurement error, the optimal rate of convergence is $\{\log(n)\}^2$ for a globally consistent estimator when making no assumptions other than the existence of two continuous derivatives. They constructed an estimator using kernel methods that achieve this very slow rate of convergence.

Carroll, Maca, and Ruppert (1999) relaxed the assumption of global consistency. They suggested two estimators: (a) a semiparametric estimator based on the SIMEX method of Cook and Stefanski (1994), which makes no assumptions about the unknown and unobserved covariates, and (b) a more parametric estimator that assumes that the unobserved covariates follow a mixture of normals distribution. Their methods are based on fixed-knot regression splines (or polynomial splines; see Eilers and Marx 1996; Ruppert and Carroll 2000; Sec. 2.2). In simulations, they showed that their methods are generally far superior to those of Fan and Truong (1993), with only moderate bias and smaller variability. The more parametric model tended to have by far the best performance in their simulation study.

There is at least one major difficulty with the efficient but more parametric approach of Carroll et al (1999). Let X be the true covariate and let W be the measured covariate. Basically, they propose fitting a modified polynomial spline. They start with the power function basis in X described in Section 2.2. Then the modified polynomial spline has as its basis functions the regression of the true basis functions in X on the observed covariate W. The resulting modified polynomial basis functions of W are very highly correlated. Their method, like ours and any other nonparametric method, requires the choice of smoothing parameters. As described by them, standard approaches to smoothing parameter estimation, such as generalized cross-validation (GCV) cannot be used, because GCV occasionally does no smoothing. This matters because in such cases their formulation leads to unusually great instability of function estimation. They developed two ad hoc methods for handling this problem: put a positive lower bound on the smoothing parameter, and use an entirely different method based on estimating the mean squared error. However, they gave evidence that shows nonetheless that their method remains numerically unstable if there are more than 15 knots.

One way to deal with this numerical instability is to use a different set of basis functions in X that are nearly orthogonal, for example, the B-spline basis. One would then conjecture that when the B-spline basis functions in X are regressed on the observed covariate W, they will not be highly correlated, and the method of Carroll et al. (1999), will be more stable. In practice, this conjecture remains to be proven.

Rather than trying to tweak the method of Carroll et al. in this way, we set out to do something radically different. Specifically, we conjectured that a fully Bayesian approach had the potential to achieve large gains in efficiency of estimation compared to previously proposed methods. One additional assumption is necessary—namely, the error distribution of the response Y about its mean was specified up to parameters.

In this article we propose a new method for nonparametric function estimation when the covariate is measured with error. Our procedure can be looked at as the natural fully Bayesian extension of the techniques of Carroll et al. (1999). It can also

Scott M. Berry is Statistical Scientist, Berry Consultants, Sycamore, IL 60178 (E-mail:*scott@berryconsultants.com*). Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station TX 77843 (E-mail: *carroll@stat.tamu.edu*). David Ruppert is Professor, School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853 (E-mail: *davidr@orie.cornell.edu*). Carroll's research was supported by National Cancer Institute grant CA-57030, and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30-ES09106. Ruppert's research was supported by National Science Foundation grant DMS-9804058. The authors thank the associate editor and two referees for many helpful comments.

be viewed as the extension to measurement error models of the Markov chain Monte Carlo (MCMC) technique of Hastie and Tibshirani (1998) or, viewed more broadly, the entire Bayesian formulation of smoothing splines (e.g., Wahba 1978, 1983; Nychka 1988, 1990).

The methodology that we present is new in two respects. First, the adjustment for bias due to measurement error comes automatically from the Bayesian machinery. In contrast, other methods explicitly analyze the bias and devise a correction in a more ad hoc fashion. Second, and perhaps more importantly, the smoothing parameter selector, which also comes automatically from the Bayesian approach, is designed for the measurement error problem. Earlier work either did not propose a smoothing parameter selector (Fan and Truong 1993) or applied a smoothing parameter selector that ignores the effects of measurement error. However, measurement error has large effects on both bias and variance, and a smoothing parameter that is optimal for correctly measured covariates may be far from optimal in the presence of measurement error.

In Section 2 we describe some background information on smoothing and regression P-splines that is necessary for our development. In Section 3 we present our methodology. Two approaches are used. One is straightforward from a calculation standpoint, but estimates only the conditional mode. The second method uses the fully Bayes approach and finds the entire posterior distribution. In Section 4 we presents simulations of these two algorithms. The results indicate that even as a frequentist estimator, our fully Bayesian method is at least competitive with that of Carroll et al. (1999), and sometimes is much better. In Section 5 we provide an illustrative example and in Section 6 we present a discussion of the results.

2. SMOOTHING AND REGRESSION P-SPLINES

Here we present a brief introduction to smoothing and P-splines. For additional information, see the work of Wahba (1978, 1990), Green and Silverman (1994), Hastie and Tibshirani (1998), and Eubank (1999) on smoothing splines and Eilers and Marx (1996) and Ruppert and Carroll (2000) on P-splines.

2.1 Smoothing Splines

Assume that $Y_i = m(X_i) + \epsilon_i$, where ϵ_i has mean 0 and variance σ_ϵ^2. Let $[a, b]$ be the interval for which an estimate of m is sought. Let g be the best natural cubic spline (NCS) approximator of m, that is, the NCS that minimizes $\sum_{i=1}^n \{m(X_i) - g(X_i)\}^2$. If m is smooth, then the error in approximating m by g typically is negligible compared to the estimation error, so we assume that $m = g$.

A *smoothing spline* is defined as the minimizer over g of the penalized sum of squares,

$$S(g) = \sum_{i=1}^n \{Y_i - g(X_i)\}^2 + \alpha \int_a^b \{g''(x)\}^2\, dx, \qquad (1)$$

for $\alpha > 0$. This minimizer is a NCS with knots at the distinct X_i values. The integral term of (1) is a roughness penalty, and α is the smoothing parameter. Let $\mathbf{g} = \{g(X_1), g(X_2), \ldots, g(X_n)\}^\mathsf{T}$. The penalty term can be written as $\alpha \int_a^b \{g''(x)\}^2\, dx = \alpha \mathbf{g}^\mathsf{T}\mathbf{K}\mathbf{g}$, where $\mathbf{K}$ is an $n \times n$-dimensional matrix of rank $n-2$, defined by Eubank (1999). The smoothing spline minimizing $S(g)$ is $\hat{\mathbf{g}} = \mathbf{A}(\alpha)\mathbf{Y}$, where $\mathbf{A}(\alpha) = (\mathbf{I} + \alpha\mathbf{K})^{-1}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$. The vector $\hat{\mathbf{g}}$ uniquely defines the smoothing spline.

The Bayesian approach to smoothing splines gives the vector $\mathbf{g}$ a prior density proportional to the "partially improper" Gaussian process,

$$(\alpha/2\sigma_\epsilon^2)^{M/2} \exp\{-(\alpha/\sigma_\epsilon^2)\mathbf{g}^\mathsf{T}\mathbf{K}\mathbf{g}\}, \qquad (2)$$

where $M = n - 2$ and $\mathbf{K}$ is defined as before. Although both $\mathbf{K}$ and $\mathbf{g}$ depend on the knot locations, because $\mathbf{g}'\mathbf{K}\mathbf{g} = \int_a^b \{g''(x)\}^2\, dx$, this prior is independent of the knot locations. If the observations Y_i are independent and normally distributed with mean $g(X_i)$ and variance σ_ϵ^2, then the posterior distribution for $\mathbf{g}$ is multivariate normal with mean $\hat{\mathbf{g}} = \mathbf{A}(\alpha)\mathbf{y}$ and covariance matrix $\sigma_\epsilon^2 \mathbf{A}(\alpha)$.

2.2 Regression P-Splines

Smoothing splines become less practical when n is large, because they use n knots. A more general approach to spline fitting is *penalized splines*, or simply *P-splines*, a term borrowed from Eilers and Marx (1996). Let $\mathbf{B}(x) = \{B_1(x), \ldots, B_N(x)\}^\mathsf{T}$, $N \le n$ be a spline basis. The P-spline model specifies that for some N-dimensional $\boldsymbol{\beta}$, $g(x) := \mathbf{B}(x)^\mathsf{T}\boldsymbol{\beta}$. Let $\mathbf{D}$ be a fixed, symmetric, positive semidefinite $N \times N$ matrix and let α be a smoothing parameter. The penalized least squares estimator $\hat{\boldsymbol{\beta}}(\alpha)$ minimizes

$$\sum_{i=1}^n \left\{ Y_i - \mathbf{B}(X_i)^\mathsf{T}\boldsymbol{\beta} \right\}^2 + \alpha \boldsymbol{\beta}^\mathsf{T}\mathbf{D}\boldsymbol{\beta}.$$

Let $\mathcal{B}$ be the $n \times N$ matrix with ith row equal to $\mathbf{B}(X_i)^\mathsf{T}$. Then the penalized least squares estimator is

$$\hat{\boldsymbol{\beta}}(\alpha) = (\mathcal{B}^\mathsf{T}\mathcal{B} + \alpha\mathbf{D})^{-1}\mathcal{B}^\mathsf{T}\mathbf{Y}.$$

The choice of k has been discussed by Ruppert (2000) who found that the exact value of k is not important, provided that k is at least a certain minimum value. Generally, $k = 20$ suffices for the types of regression functions found in practice, and $k = 40$ provides a margin of safety. Of course, there will be exceptions where more knots are required, for example, a long periodic time series. Also, functions whose higher derivatives are large may not be well approximated by, say, quadratic splines; see Figure 3.

Here we use the term "P-splines" to refer to both P-splines and smoothing splines as a special case. Convenient classes of P-splines are the penalized B-splines of Eilers and Marx (1996) and the closely related splines of Ruppert and Carroll (2000). The latter are the pth-degree polynomial splines with k fixed knots, $t_1, \ldots, t_k$. The knots could be equally spaced on the range of the X_i's, although we prefer to select them at the quantiles of the X's. A convenient basis is $\mathbf{B}(x) = (1, x, x^2, \ldots, x^p, (x-t_1)_+^p, \ldots, (x-t_k)_+^p)^\mathsf{T}$. Then $\beta_{2+p}, \ldots, \beta_N$ are the sizes of the jumps in the pth derivative of $g(x) = \mathbf{B}(x)^\mathsf{T}\boldsymbol{\beta}$ at the knots. Ruppert and Carroll (2000) penalize these jumps by letting $\mathbf{D}$ be the $N \times N$ diagonal matrix with $p+1$ 0's followed by k 1's along the diagonal. Then $\boldsymbol{\beta}^\mathsf{T}\mathbf{D}\boldsymbol{\beta} = \sum_{j=1}^k \beta_{1+p+j}^2$ is the sum of the squared jumps.

2.2.1 Bayesian P-Splines. We partition $\boldsymbol{\beta}$ into the coefficients of the monomial basis functions of the truncated power basis functions by letting $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$ where $\boldsymbol{\beta}_1$ is of length $p+1$ and $\boldsymbol{\beta}_2$ is of length k.

The penalized least squares estimator is the mean of the posterior distribution of $\boldsymbol{\beta}$ when $\boldsymbol{\beta}_1$ has an improper uniform (on $\mathbf{R}^{p+1}$) prior density and $\boldsymbol{\beta}_2$ has a proper prior proportional to $\gamma^{k/2}\exp\{-(\gamma/2)\boldsymbol{\beta}_2^{\mathsf{T}}\boldsymbol{\beta}_2\}$ where $\gamma = \alpha/\sigma_\epsilon^2$ and, as before, k is the number of knots. This prior on $\boldsymbol{\beta}$ induces a prior on $g(\cdot)$ and $\mathbf{g}$ because $g(x) = \mathbf{B}(x)^{\mathsf{T}}\boldsymbol{\beta}$.

The posterior of $\boldsymbol{\beta}$, conditional on σ_ϵ^2 and α, is $N\{(\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D})^{-1}\mathcal{B}^{\mathsf{T}}\mathbf{Y}, \sigma_\epsilon^2(\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D})^{-1}\}$. Let $\mathbf{A}(\alpha) = \mathcal{B}(\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D})^{-1}\mathcal{B}^{\mathsf{T}}$. Then the posterior distribution of $\mathbf{g} = \mathcal{B}\boldsymbol{\beta}$, conditional on $(\alpha, \sigma_\epsilon^2)$, is $N\{\mathbf{A}(\alpha)\mathbf{Y}, \sigma_\epsilon^2\mathbf{A}(\alpha)\}$, the same result obtained for smoothing splines. Often $\mathbf{D}$ is singular but $\mathcal{B}^{\mathsf{T}}\mathcal{B} + \alpha\mathbf{D}$ is nonsingular, so that the prior is improper but the posterior is proper.

3. GENERAL MODEL

We consider the measurement error model

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where the ϵ_i are the independent normal random variables with mean 0 and variance σ_ϵ^2. The X's are not observable (i.e., they are latent variables), but W that are surrogates for the Xs are observed,

$$W_{ij} = X_i + U_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \tag{4}$$

where the U_{ij} are independent normal errors with mean 0 and variance σ_u^2.

Model (4) is more general that it first appears. It can be interpreted as stating that a *known* function of the observed covariates (W) is the same function of the latent variables, plus independent, homoscedastic, normally distributed measurement errors. We have written (4) as if the function were the identity function, but it could be anything (e.g., the logarithm). The reason for this generality is that in (3), the function $m(\cdot)$ is unknown, so that, for example, if $m_*(v) = m\{\exp(v)\}$, then $m_*\{\log(x)\} = m(x)$.

The mean function m is modeled as a P-spline with smoothing parameter α. We use the notation $[A]$ and $[A|B]$ to represent prior densities and conditional densities.

Denote $\boldsymbol{\theta} = (\mathbf{g}, \mathbf{X}, \sigma_\epsilon^2, \sigma_u^2, \alpha)$. The posterior density is

$$[\boldsymbol{\theta}|\mathbf{Y}, \mathbf{W}] \propto [\mathbf{Y}|\mathbf{g}, \mathbf{X}, \sigma_\epsilon^2][\mathbf{W}|\mathbf{X}, \sigma_u^2][\mathbf{g}|\alpha][\sigma_u^2][\sigma_\epsilon^2][\mathbf{X}][\alpha]. \tag{5}$$

The form of (5) has a structure that is common in latent variable models. We exploit an important feature of such models, namely that in the Gibbs sampler or other Monte Carlo computational approaches, once $X_1, \ldots, X_n$ are generated from the posterior, estimation of g becomes a standard problem for which much software exists. A basic modeling issue and computational problem is how to generate $X_1, \ldots, X_n$.

Two approaches to the estimation of $\mathbf{g}$ are taken. The first method, which is "quick and dirty," estimates the posterior mode of $\mathbf{g}$ by the iterative conditional modes (ICM) algorithm (Besag 1986). The calculation is fast and easy to blend with a program that calculates a P-spline. The ICM method also serves to create starting values for the second method, which is fully Bayesian but involves more complex and time-consuming calculation. The former is described in the next section; the latter, in Section 3.2.

3.1 The Iterative Conditional Modes Algorithm

In this section we describe an iterative method of estimating g in the presence of measurement error by finding the mode of the posterior in (5). This methodology sacrifices the philosophical advantages of a fully Bayesian analysis for computational ease. In the first step, we estimate the three variance components $\sigma_\epsilon^2, \sigma_u^2$, and α; these estimates are held fixed for the rest of the procedure. We first describe the estimation of these variance components.

If $m_i = 1$ for all $i = 1, \ldots, n$, then the user must supply an estimate for σ_u^2; otherwise, the usual pooled sample variance of the W's from analysis of variance calculations is used. If s_i^2 is the sample variance of $W_{i1}, \ldots, W_{im_i}$, then $\hat{\sigma}_u^2 = \sum_{i=1}^n (m_i - 1)s_i^2 / \sum_{i=1}^n (m_i - 1)$.

The initial estimate for the $\mathbf{X}$ parameter is $\mathbf{X}^{(0)} = (\overline{W}_1, \ldots, \overline{W}_n)$, where $\overline{W}_i = \sum_{j=1}^{m_i} W_{ij}/m_i$. A *naive* smoothing spline, $\hat{\mathbf{g}}^{(0)}$, is estimated by assuming $\mathbf{X} = \mathbf{X}^{(0)}$ and fitting the standard nonmeasurement error smoothing spline. The smoothing parameter for the naive estimator, $\hat{\alpha}$, is fit using cross-validation (CV) or GCV; we use CV in our numerical work in this article. We use the standard estimate of σ_ϵ^2 (see Green and Silverman 1994),

$$\hat{\sigma}_\epsilon^2 = \sum_{i=1}^n \{Y_i - \hat{g}^{(0)}(\widehat{X}_i^{(0)})\}^2 / \mathrm{trace}\{\mathbf{I} - \mathbf{A}(\hat{\alpha})\}.$$

A normal prior distribution is used for each X_i, with mean μ_x and variance σ_x^2, where μ_x and σ_x are constants. In the algorithm, we replace these by the mean and standard deviation of the Ws.

Conditional on $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_u^2$, and $\hat{\alpha}$, the posterior distribution is proportional to

$$\exp\left[-\frac{1}{2\hat{\sigma}_\epsilon^2}\sum_{i=1}^n\{Y_i - g(X_i)\}^2 - \frac{1}{2\hat{\sigma}_u^2}\sum_{i=1}^n\sum_{j=1}^{m_i}(W_{ij} - X_i)^2 - \frac{1}{2\sigma_x^2}\sum_{i=1}^n(X_i - \mu_x)^2 - \frac{\hat{\alpha}}{2\hat{\sigma}_\epsilon^2}\mathbf{g}^{\mathsf{T}}\mathbf{K}\mathbf{g}\right]. \tag{6}$$

We use the ICM approach described by Besag (1986) to find the posterior mode of (6). This approach can be also phrased as generalized EM (Meng and Rubin 1993). The approach is to find the posterior mode by sequentially finding the mode of the complete conditional distributions. This is in contrast to Gibbs sampling, where observations are iteratively *drawn* from the complete conditionals. In summary, ICM iteratively updates the parameters with their modal values.

ICM Algorithm

1. Find the estimates $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_u^2$, $\hat{\alpha}$, and the naive spline $\mathbf{g}^{(0)}$. Set $\mathbf{i} = 1$.

2. Fixing **K**, find the vector **X** that maximizes (6) conditional on $\mathbf{g}^{(i-1)}$. These are labeled $\mathbf{X}^{(i)}$. There is no analytic solution available. We use a grid search to find the maximum for each component of **X**. The complete conditional for each component of **X** is independent of the other components of **X**, because we have fixed **K**. Therefore, this maximization can be done individually. We maximize each component using a uniform grid evaluation, followed by a finer grid used in the region of the maximum value from the first grid.

3. Find the vector $\mathbf{g}^{(i)}$ that maximizes (6) conditional on $\mathbf{X}^{(i)}$. This is the usual P-spline for a nonmeasurement error problem, which is described in Section 2.

4. Set $i = i+1$ and repeat steps 2 and 3 until the estimate $\mathbf{g}^{(i)}$ converges.

Figure 1 demonstrates the ICM method with simulated data. The data consist of 100 X's generated from a standard normal distribution. The responses, Y, were generated from a normal distribution with a standard deviation of .3 and a mean function of

$$m(x) = \frac{\sin(\pi x/2)}{1+2x^2\{\text{sign}(x)+1\}}. \qquad (7)$$

Each $m_i = 2$ and the W_{ij} for $j = 1, 2$, are normally distributed with a mean of X_i and a standard deviation of .8. Figure 1 shows the (x, y) pairs with open diamonds and the $(\overline{W}, y)$ pairs with solid diamonds. The regression function $m(x)$ and the naive regression spline, $g^{(0)}$, are presented, and the estimated $(X^{(1)}, Y)$ pairs from the first ICM iteration are shown by the "1" symbols. The naive spline, the estimated curve from one iteration of the ICM procedure, $g^{(1)}$, and the curve judged to have converged, $g^{(\infty)}$, are presented.

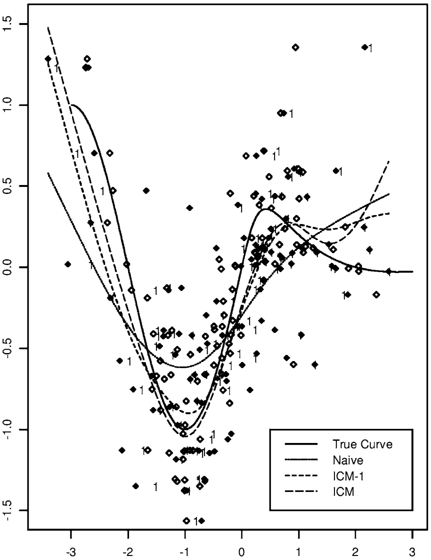


Figure 1. The (x,y) Pairs Are Shown With Open Diamonds Whereas the ($\overline{W}$,Y) Observations Are Shown With Solid Diamonds. The true regression function is shown with the solid line. The naive spline estimate, the ICM spline from one iteration (ICM-1), and the converged ICM spline (ICM) are also shown. The 1's represent the ($X^{(1)}$,Y) pairs.

The motivation behind the ICM approach is that it uses some of the strengths of the Bayesian approach but is very easy to program and fast to compute. An S-PLUS function for the ICM approach is available from the first author.

The major difficulty with the ICM method as a general method for measurement error models can be seen most clearly by considering parametric models $g(X) = g(X, \beta)$ of known form but with an unknown parameter β. If in (6) we set $\alpha = 0$, delete the term $\sum_{i=1}^{n}(X_i - \mu_x)^2/(2\sigma_x^2)$, replace $g(X_i)$ by $g(X_i, \beta)$, and maximize in the unknown parameters $(X_1, \ldots, X_n, \beta, \sigma_u^2, \sigma_\epsilon^2)$, then we are computing what is known in the literature as the *functional maximum likelihood estimate* (Fuller 1987). Functional models assume that $X_1, \ldots, X_n$ are fixed parameters to be estimated. In contrast, in a *structural* model, the X_i's are latent variables from a distribution depending on structural parameters; one integrates the X_i's out of the likelihood and maximizes simultaneously over the structural and other parameters. In linear regression, the functional approach is known to yield consistent estimates of regression parameters. In nonlinear regression, this need not be the case. Fuller (1987) and Amemiya and Fuller (1988) studied parametric nonlinear regression problems as $n \to \infty$ and $\sigma_u^2 \to 0$ in such a way that $\sigma_u^2 \propto n^{-1/2}$. They found that in the terms of rate of convergence, the functional approach is no better than the naive approach, because both have bias of order $O(\sigma_u^2) = O(n^{-1/2})$; see eq. (2.11) of Amemiya and Fuller (1988).

This similarity with functional modeling suggests that although the ICM approach is computationally simple, it need not yield consistent estimates of the regression function, not even in parametric problems. The next section describes the fully Bayesian approach, which is computationally more difficult but has the benefits of the Bayesian machinery. The fully Bayesian approach is structural, and with diffuse priors, one gets essentially the structural maximum likelihood estimate (MLE), a consistent estimator. In this context, the ICM method largely serves to produce starting values for the full Bayesian approach.

3.2 Fully Bayesian Approach

In this section we develop the fully Bayesian approach to this problem. The ICM approach estimates the variance components and keeps them fixed, allowing the smoothing spline estimate and the X's to fluctuate. In this section, prior distributions are placed on all parameters, including the structural parameters (μ_x, σ_x^2) and the variance components $(\sigma_\epsilon^2, \sigma_u^2)$, and the joint posterior distribution is calculated. One of the benefits of this approach is that observations of the smoothing spline are generated from the posterior, and thus we estimate the entire posterior distribution of g, not just its mode. Thus calculation of the various forms of "error bars" is straightforward. These credible sets take into account the measurement error of the independent variables and the use of a data-based smoothing parameter.

The method is as follows. Without loss of generality, we replace α/σ_ϵ^2 by γ. The prior distributions for σ_ϵ^2 and σ_u^2 are inverse-gamma distributions, and the prior distribution for γ is a gamma distribution: $\sigma_\epsilon^2 \sim \mathrm{IG}(A_\epsilon, B_\epsilon), \sigma_u^2 \sim \mathrm{IG}(A_U, B_U)$, and $\gamma \sim \mathrm{G}(A_\gamma, B_\gamma)$. We use the definitions of the inverse-gamma and gamma distributions (respectively) from Berger (1985):

$$f(x|A,B) = \frac{1}{\Gamma(A)B^A x^{A+1}} \exp\left(-\frac{1}{Bx}\right) I_{(0,\infty)}(x)$$

and

$$f(x|A,B) = \frac{1}{\Gamma(A)B^A} x^{A-1} \exp\left(-\frac{x}{B}\right) I_{(0,\infty)}(x).$$

There is no reasonable prior distribution for the X's, which eases the computational burden. This prior distribution also can easily change from application to application. In some examples, a flat reference prior may be reasonable, whereas in others, a normal hierarchical distribution may be appropriate. A mixture of normals is a flexible approach that has some intuitive appeal (see Carroll, Roeder, and Wasserman 1999). A difficulty is that the X_is continually change throughout the MCMC algorithm, and updating this mixture at every iteration is chronically slow. We leave the choice of prior distribution for X an open choice for the particular application. For the simulations and examples in this article, we use a hierarchical Bayes approach. A normal distribution with mean μ_x and variance σ_x^2 is used, where $\mu_x \sim \mathrm{normal}(d_x, t_x^2)$ and $\sigma_x^2 \sim \mathrm{IG}(A_x, B_x)$.

The hyperparameters that are fixed a priori and thus are "tuning constants" are denoted by Roman fonts. These are $A_Y, B_Y, A_U, B_U, A_\gamma, B_\gamma, d_x, t_x^2, A_x$, and B_x.

Assuming the hierarchical normal structure for $[X]$, the joint posterior is proportional to

$$\begin{aligned} &\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\{Y_i - g(X_i)\}^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2 \right.\\ &\qquad\left. - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(X_i-\mu_x)^2 - \frac{1}{2t_x^2}(\mu_x - d_x)^2\right\} \\ &\times \exp\left\{-(\gamma/2)\mathbf{g}^{\mathsf{T}}\mathbf{K}\mathbf{g} - \frac{1}{B_\epsilon\sigma_\epsilon^2} - \frac{1}{B_U\sigma_u^2}\right.\\ &\qquad\left. - \frac{\gamma}{B_\gamma} - \frac{1}{B_x\sigma_x^2}\right\} \\ &\times \sigma_\epsilon^{-2(n/2+A_\epsilon+1)}\sigma_u^{-2(1/2\sum_{i=1}^{n} m_i + A_U+1)} \\ &\times \sigma_x^{-2(n/2+A_X+1)}\gamma^{(A_\gamma+M/2-1)}, \end{aligned} \tag{8}$$

where $M = n-2$ as in (2). The sampling is done using a successive substitution algorithm (Gelfand and Smith 1990). The complete conditional distributions for the parameters are

$$\mathbf{g}|\mathbf{X}, \gamma, \sigma_\epsilon^2, \mathbf{Y}, \mathbf{W} \sim \mathrm{normal}\{\mathbf{A}(\sigma_\epsilon^2\gamma)\mathbf{y}, \sigma_\epsilon^2\mathbf{A}(\sigma_\epsilon^2\gamma)\},$$

$$\begin{aligned} &[X_i|\mathbf{W}_i, \mathbf{g}, \sigma^2, \sigma_u^2, \mathbf{Y}, \mathbf{W}] \\ &\quad\propto \exp\left(-\frac{1}{2\sigma_u^2}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2 \right.\\ &\qquad\left. -\frac{1}{2\sigma_\epsilon^2}\{Y_i - g(X_i)\}^2 - \frac{1}{2\sigma_x^2}(X_i-\mu_x)^2\right), \end{aligned} \tag{9}$$

$$\sigma_\epsilon^2|\mathbf{g}, \mathbf{X}, \mathbf{Y}, \mathbf{W}$$
$$\sim \mathrm{IG}\left(A_\epsilon + n/2, [1/B_\epsilon + (1/2)\sum_{i=1}^{n}\{Y_i - g(X_i)\}^2]^{-1}\right),$$

$$\sigma_u^2|\mathbf{X} \sim \mathrm{IG}\left(A_U + (1/2) \sum_{i=1}^{n} m_i, \left[1/B_U + (1/2)\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2\right]^{-1}\right),$$

$$\gamma|\mathbf{g}, \mathbf{X} \sim \mathrm{G}\left(A_\gamma + \frac{M}{2}, \left[1/B_\gamma + \frac{1}{2}\mathbf{g}^{\mathsf{T}}\mathbf{K}\mathbf{g}\right]^{-1}\right),$$

$$\mu_x|\mathbf{X} \sim \mathrm{normal}\{(n\overline{X}t_x + d_x\sigma_x^2)/(nt_x^2+\sigma_x^2), \sigma_x^2 t_x^2/(nt_x^2+\sigma_x^2)\},$$

$$\sigma_x^2|\mathbf{X} \sim \mathrm{IG}\left[A_x + n/2, \left\{B_x^{-1} + (1/2)\sum_{i=1}^{n}(X_i-\mu_x)^2\right\}^{-1}\right].$$

The estimates $\hat\sigma_\epsilon^2$, $\hat\sigma_u^2$, $\hat\gamma$, $\mathbf{x}^{(\infty)}$, and $\mathbf{g}^{(1)}$ from the ICM approach are used as starting values for the MCMC algorithm. Observations from each of the complete conditionals are drawn iteratively in the order just presented. The generation of an observation of $\mathbf{g}$ is computationally difficult for smoothing splines, because they have n knots. Because the values of $\mathbf{X}$, $\mathbf{K}$ (for smoothing splines), and γ are continually changing in the algorithm, the matrix $\mathbf{A}(\sigma_\epsilon^2\gamma)$ (which is $n \times n$) and its inverse must be recomputed for each iteration of the MCMC algorithm. Hastie and Tibshirani (1998) discussed an algorithm for generating observations of $\mathbf{g}$ in $O(n)$ operations. Computations can be reduced by using P-splines with fewer than n knots, with no real loss of precision (Ruppert 2000).

The complete conditionals for the X_i's require a Metropolis–Hastings step. This is done by generating a candidate observation of X_i from a normal distribution with a mean of the current value of X_i and a standard deviation of $2\sigma_u^{(i)}/\sqrt{m_i}$, where $\sigma_u^{(i)}$ is the current value of σ_u in the MCMC algorithm. Using $\overline{W}_i$ as an estimate of X_i, without the information in the regression function, has a standard error of $\sigma_u^{(i)}/\sqrt{m_i}$. Using the rule of thumb of a candidate value with a standard deviation twice the standard deviation of the marginal posterior provides a conservative candidate distribution for X_i. In terms of efficiency of the Metropolis–Hastings step, in our experience it is better to overestimate this standard deviation than to underestimate it. The evaluation of the complete conditional for X_i is computationally straightforward.

Generating observations from each of the other complete conditionals is straightforward and fast. Because the position of $\mathbf{X}$ changes throughout the algorithm, when using smoothing splines, we keep track of the value of g at a uniformly distributed grid of points. For each realization of g in the sampler, the value of g for each grid point is recorded. This enables us to keep track of pointwise moments and percentiles. For fixed-knot P-splines, g is defined by $\boldsymbol{\beta}$, the coefficients of the basis functions. Because the basis functions stay fixed as $\mathbf{X}$ varies, there is no need to record the values of g on a grid. Rather, one keeps track of the realizations of $\boldsymbol{\beta}$. For any realization of $\boldsymbol{\beta}$ there is a corresponding realization of $\mathbf{g}$ given by $\mathbf{g} = \mathcal{B}\boldsymbol{\beta}$. Details of implementation are given in the Appendix.

Having observations from the joint posterior distribution provides a powerful tool for inference. The pointwise mean curve is a natural estimate of the regression mean function m. Pointwise credible intervals can also be calculated very easily from the observations of $\mathbf{g}$. Functions (linear or nonlinear) of the regression function can also be estimated, along with standard errors. This is the approach used by Wahba (1983) in nonmeasurement error cases and by Hastie and Tibshirani (1998) in nonmeasurement error semiparametric models. Wahba's work predated the revolution in Bayesian computations, and she treated the smoothing parameter as fixed. Hastie and Tibshirani used the Gibbs sampler to adjust the credible sets for uncertainty in the variance components that define the smoothing parameter. In this article, use of the Gibbs sampler also adjusts the credible sets for measurement error.

Although the regression function and its functionals are the main focus of this article, inferences about the mismeasured X_i's can also be made. Posterior means and credible intervals can easily be constructed for each of the individual X_i's. The variance components may also be of interest, and likewise constructing estimates and credible intervals for them is straightforward.

For an example of this method, we use the same data from the ICM example and use smoothing splines. A pointwise posterior mean curve is used for the estimate of m. Credible curves are calculated by interpolating the pointwise $100(1-\alpha)\%$ credible intervals. A burn-in time of 500 observations is used with 1,000 observations from the posterior. Figure 2 shows the estimate of g and the 90% pointwise credible curves.

There are at least two possible methods for choosing the smoothing parameter for a smoothing spline. We place a prior distribution on γ; Hastie and Tibshirani (1998) used an identical procedure within semiparametric models. It is worth noting that by placing a continuous density prior on γ, we have automatically given zero prior probability to the possibility of doing no smoothing at all. This is an automatic way of avoiding the possibility of gross undersmoothing that caused so much trouble for the methods of Carroll et al. (1999).

An alternative to this method that we recommend is to choose the smoothing parameter using a criterion such as CV or GCV during each iteration of the successive substitution sampling. This method loses some of the Bayesian interpretation, is computationally expensive, and does not account for the effects of measurement error. However, a referee asked that we evaluate this procedure; see Section 4.

4. SIMULATIONS

4.1 Basic Simulations

We performed a series of simulations to compare our methods with those of Carroll et al. (1999). The results presented here are based on smoothing splines, but checks show that P-splines with 30 knots give much the same results. In each case, 200 simulated datasets were generated, with X_i generated as independent normal random variables with mean μ_x and variance σ_x^2, with two replicates ($m_i = 2$), and with ϵ_i also normally distributed. In each simulation, for the fully Bayesian method, the following prior distributions are used: $\sigma_\epsilon^2 \sim \mathrm{IG}(1, 1)$, $\sigma_u^2 \sim \mathrm{IG}(1, 1)$, $\gamma \sim \mathrm{G}(3, 1000)$, $\mu_x \sim N(0, 10^2)$, and $\sigma_x^2 \sim \mathrm{IG}(1, 1)$. These priors were selected because of their relative flexibility. They are all proper, yet they are not strong, in the sense of bringing a lot of information to the problem. We found the results insensitive to moderate modifications of these priors. The flexibility of these priors is demonstrated by their success in the different regression functions used in the simulations.

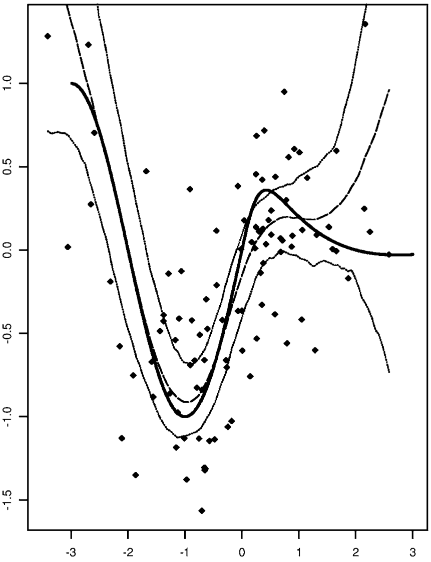


Figure 2. An Example of the Fully Bayesian Spline. The solid curve is the true regression function. The dashed middle curve is the mean of the posterior of the regression function. The dotted error bars represent the piecewise 90% credible intervals.

For purposes of bias and mean squared error calculations, the smoothing spline estimates of g were computed on a grid of 101 points in the interval $[a, b]$, the interval chosen to contain most of the distribution for X. The mean squared biases and mean squared errors were computed over this grid.

It is impossible to assess convergence of the MCMC chain for all simulated data sets. Instead, for a few selected datasets, we used tests of convergence (Gelman and Rubin 1992) separately for each parameter and also for the estimated function on a few selected grid points.

The first five cases considered were as follows:

Case 1: The regression function, m, is given in (7), with $n = 100$, $a = -2.0$, $b = 2.0$, $\sigma_\epsilon^2 = .3^2$, $\sigma_u^2 = .8^2$, $\mu_x = 0$, and $\sigma_x^2 = 1$.

Case 2: Same as case 1, except $n = 200$.

Case 3: A modification of case 1 except that $n = 500$.

Table 1. The Mean Squared Bias and MSE for the Simulation

Method	*Case 1*	*Case 2*	*Case 3*	*Case 4*	*Case 5*	*Case 6*	*Case 7*	*Case 8*
Mean squared bias $\times 10^2$								
Naive	5.59	4.92	5.21	1108	3733	4.83	4.80	15.27
ICM	2.98	2.22	2.04	629	1541	2.20	1.94	8.51
Bayes	.78	.38	1.04	17.4	468	1.74	1.68	6.20
Structural(5)	1.38	.62	.46	3.7	838	1.47	1.48	12.82
Structural(15)	1.44	.60	.66	3.3	226	1.75	1.70	12.36
MSE $\times 10^2$								
Naive	6.91	5.57	5.38	1155	3793	5.77	5.84	16.48
ICM	5.93	3.87	3.29	751	1948	4.36	3.93	12.38
Bayes	**2.84**	**1.56**	**1.47**	**195**	1031	**2.69**	**2.49**	**7.41**
Structural(5)	8.17	3.82	1.73	217	2032	7.27	7.91	16.84
Structural(15)	9.90	5.40	1.85	237	**799**	6.94	9.91	20.22

NOTE: The regression functions are $m(x) = \sin(\pi x/2)/[1 + 2x^2\{\text{sign}(x) + 1\}]$ (cases 1, 2, 3 and 6), $m(x) = 1000x_+^3(1-x)_+^3$ (case 4), and $m(x) = 10\sin(4\pi x)$ (case 5). Case 7 is same as case 1 except that X is a normalized chi-squared(4) random variable, and ϵ is generated as a Laplace random variable. Case 8 is same as case 1 except that $m(x) = H(100x) + H\{-100(x - .5)\}$, where $H(x) = \{1 + \exp(-x)\}^{-1}$. This function is poorly fit by a regression P-spline with 35 knots. "Naive" is the naive smoothing spline, "ICM" is the fully iterated ICM method, "Bayes" is the fully Bayesian method, and "Structural(m)" is the structural regression P-spline of Carroll et al. (1999) with m knots. In each column, the smallest MSE values is in boldface.

Case 4: Case 1 of Carroll et al. (1999), so that $m(x) = 1000x_+^3(1-x)_+^3$, $x_+ = xI(x > 0)$, with $n = 200$, $a = .1$, $b = .9$, $\sigma_\epsilon^2 = .0015^2$, $\sigma_u^2 = (3/7)\sigma_x^2$, $\mu_x = .5$, and $\sigma_x^2 = .25^2$.

Case 5: A modification of case 4 of Carroll et al. (1999), so that $m(x) = 10\sin(4\pi x)$, with $n = 500$, $a = .1$, $b = .9$, $\sigma_\epsilon^2 = .05^2$, $\sigma_u^2 = .141^2$, $\mu_x = .5$, and $\sigma_x^2 = .25^2$.

The methods compared were the following:

- Naive smoothing spline fit ignoring measurement error
- Fully iterated ICM approach
- Fully Bayesian approach
- Structural method (Carroll et al. 1999), 5 knots
- Structural method, 15 knots.

Table 1 presents summary results for mean squared bias and mean squared error (MSE). The SIMEX method discussed by Carroll et al. (1999) using a 40-knot quadratic P-spline and a quadratic extrapolant was also computed, with results better than the naive estimator but generally inferior to the others. The striking feature of this table is that our Bayesian estimator has at least as good *frequentist* properties as the frequentist methods. In cases 1 and 2 it clearly dominates, having less than half of the MSE of the other methods. In case 3, it MSE efficiency is 20% greater than the structural spline with 15 knots, whereas in case 5 it is only 25% less efficient. The improvement of the fully Bayesian method over the frequentist methods is especially large for smaller sample sizes, cases 1 and 6 for example, where $n = 100$.

Clearly, even this limited simulation suggests that our Bayesian method is at least competitive with other methods proposed previously in the literature.

4.2 Robustness to Priors

Our priors are proper yet not particularly informative. However, as suggested by a referee, it is interesting to compare our results when different priors are used. Here we focus on case 1 in the simulation, with the priors modified as follows: $\sigma_\epsilon^2 \sim \text{IG}(3, 1)$, $\sigma_u^2 \sim \text{IG}(3, 1)$, $\gamma \sim \text{G}(2, 2000)$, $\mu_x \sim \text{N}(0, 100^2)$, and $\sigma_x^2 \sim \text{IG}(3, 1)$. Compared with Table I, when we ran the simulation using these priors, the MSE of the Bayesian approach changed from 2.84 to 2.53, a minimal change. We have run selected exercises on datasets with different priors, and in all cases there were only minimal changes.

4.3 Distributional Robustness and Model Misspecification

The method that we have developed assumes that X and ϵ are normally distributed, and that the function $m(x)$ is adequately represented by a spline. We ran a limited number of simulations to study violations of these assumptions.

Case 6: The same as case 1 except that X is a normalized chi-squared(4) random variable. Squared bias and MSE are evaluated on $[-1.25, 2.00]$.

Case 7: The same as case 1 except that X is a normalized chi-squared(4) random variable and ϵ is generated as a Laplace random variable. Squared bias and MSE are evaluated on $[-1.25, 2.00]$.

Case 8: The same as case 1 except that $m(x) = H(100x) + H\{-100(x - .5)\}$, where $H(x) = \{1 + \exp(-x)\}^{-1}$. This function is poorly fit by a regression P-spline with 35 knots; see Figure 3. The results are also displayed in Table I.

In case 6, where the distribution from X is far from the normal distribution, the Bayes method is still more efficient than the other methods. For case 7, where in addition the distribution for ϵ is nonnormal, we see that the Bayes method is still best, although as expected with a small loss of efficiency. We are not sufficiently bold or naive to suggest that the Bayes method will retain distributional robustness in all cases, but the results are at least encouraging in the case of small model deviations.

In case 8, it is the spline representation of the function $m(x)$ that fails. Because all of the methods in Table 1 are based on splines, it was difficult to guess a priori what would happen in the simulation, although one perhaps would have expected that all the methods would be equally bad, although as it turns out the Bayes method had the smallest bias and MSE.

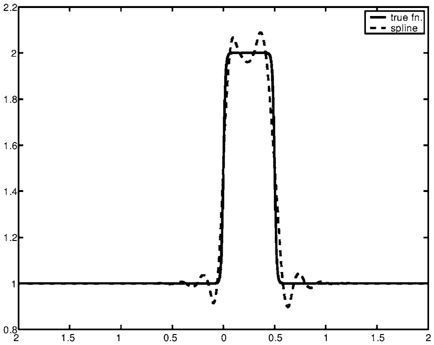


Figure 3. The Function $m(x) = H(100x) + H\{-100(x - .5)\}$, Where $H(x) = \{1 + \exp(-x)\}^{-1}$ (——), and the Best-Fitting Quadratic P-Spline With 35 Knots (- - - -).

5. EXAMPLE

This article was partially motivated by the analysis of a real dataset. Unfortunately, we do not have permission to discuss the study details here, or to make the data available to the public. The data that we have have been transformed and rescaled, and random noise has been added.

Essentially, there is a treatment group and a control group, which are evaluated using a scale at baseline (W) and at the end of the study (Y). Smaller values of both indicate a more severe disease. The scale itself is subject to considerable error, because it is based on a combination of self-report and clinic interview. The study investigators estimate that in their transformed and rescaled form, the measurement error variance is approximately $\sigma_u^2 = .35$.

A preliminary Wilcoxon test applied to the observed change from baseline, $Y - W$, indicated a highly statistically significant difference between the two groups.

In the notation of (3), the main interest focuses on the population mean change from baseline $\Delta(X) = m(X) - X$ for the two groups and, most importantly, on the difference between these two functions.

Preliminary nonparametric regression analysis of the data ignoring measurement error indicates possible nonlinearity in the data. A quadratic regression is marginally statistically significant in the control group $(p \approx .03)$ and marginally nonstatistically significant in the treated group $(\approx .07)$. When we corrected the quadratic fits for the measurement error (Cheng and Schneeweiss 1998) and bootstrapped the resulting parameter estimates, both p values exceeded .20, although the fitted functions had substantial curvature. Thus the evidence for a linear model is mixed. We are interested in understanding the nature of the statistically significant difference between the two groups as evidenced by the Wilcoxon test.

Figure 4 plots $\Delta(X)$ for the placebo and treatment group using the fully Bayesian method. The functions are fairly similar in shape, with the treated group having higher values, essentially uniform in the range $[-1.50, 2.25]$ and covering most of the distribution of X. Both fitted functions exhibit curvature, although for those with true baseline score exceeding 0, the fitted functions are fairly linear.

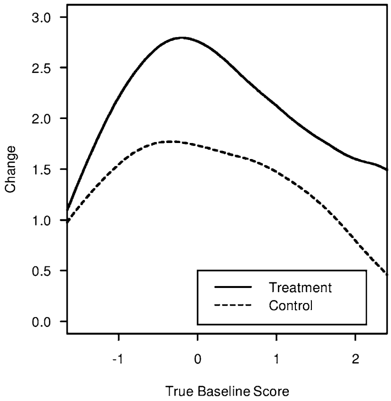


Figure 4. Estimate of the Function $\Delta(x) = m(x) - x$ for the Control Group (- - - -) and the Treatment Group (——) in the Example.

Figure 5 shows the differences between the two functions, along with the 90% pointwise credible interval using the fully Bayesian method. The upper part of this interval is below 0 from approximately -1 to 2, and thus is in agreement with the Wilcoxon analysis. Interestingly, there is no evidence that the treatment is particularly effective for those who have the most severe disease at baseline (i.e., those with a true baseline score less than -1).

6. DISCUSSION

The Bayesian approach to measurement error, modeling the mismeasured variables as latent random variables and integrating them out, is a powerful one. In this article we have developed a Bayesian method for nonparametric regression in the presence of measurement error. By modeling a smoothing spline from a Bayesian standpoint, we create algorithms to calculate the posterior distribution of the regression function. The resulting estimate accounts for the effects of measurement error both on the estimator and on the smoothing parameter. The resulting smoothing parameter selector appears to be the first to adjust for the effects of measurement error.

Two algorithms are presented. The first algorithm is a quick-and-dirty method to find a posterior mode. The technique, based on the ICM procedure, is easy to combine with a program that calculates splines and is very fast. The fully Bayesian procedure is based on an MCMC algorithm. The fully Bayesian procedure is computationally more difficult but benefits from the modeling of each unknown and exploring the posterior, rather than finding the mode.

The simulations demonstrate the flexibility of the fully Bayesian approach, and even its efficiency in the frequentist

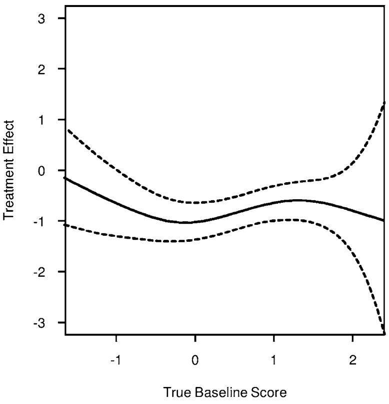


Figure 5. Estimate (solid line) of the Difference of the Function Δ(x) = m(x) − x Between the Treatment Group and the Control Group in the Example (control-treatment) With 90% Pointwise Credible Intervals (dashed lines).

sense, at least in the examples we have investigated. The fully Bayesian approach also enables inference on more than just the regression function.

We believe that the fully Bayesian approach works better than the previous proposals is because: often one can estimate the unknown X_i significantly more accurately using *all* information in the data about X_i rather than using just the $W_{ij}, j = 1, \ldots, n_i$. It is possible that a likelihood-based frequentist spline approach can also take advantage of this information, but such work is clearly outside the scope of this article. Many errors-in-variables techniques in parametric problems, and the technique of Carroll et al. (1999) in the nonparametric problems, and estimate X_i using only the W_{ij}. However, there is information in Y_i about X_i, and the fully Bayesian approach extracts this information. This fact can be seen in (9) for the conditional density of X_i given the other parameters.

As an illustration, we generated data from the following model. The sample size was $n = 201$, the X_i were normal(1, 1), $m_i \equiv 2$, and $m(x) = \sin(2x)$. Also $\sigma_\epsilon = .15$ and $\sigma_u = 1$. The fully Bayesian estimate of m is the spline fit with the imputed X's, averaged over the Gibbs sample. Thus the crucial quantity is how close the imputed $m(X_i)$ are, on average over the Gibbs samples, to the actual value of this quantity. In examples such as this one where m is nonmonotonic, for some cases the imputed X_i might not be close to the actual X_i but the imputed $m(X_i)$ might be close to the true $m(X_i)$. For estimation of $m(\cdot)$, the latter is good enough. Therefore, we compared various estimates of the X_i by using the norm $\|m(\mathbf{X}) - m(\widehat{\mathbf{X}})\|$, where $\mathbf{X} = (X_i, \ldots, X_n)^T$ is the vector of true X_i and similarly $\widehat{\mathbf{X}}$ is the vector of predicted X_i. Note that $m(\cdot)$ here is the true regression function. We are *not* comparing estimates of $m(\cdot)$, only estimates of the X_i. Consider two estimators of $\mathbf{X}$. The first estimator is the conditional expectation of X_i given $\overline{W}_i$. Because $(X_i, \overline{W}_i)$ is jointly normal, this is the optimal estimator given only $\overline{W}_i$. The second estimator uses the X_i from the Gibbs output and thus is a sample from the distribution of X_i given the Y_i and the W_{ij}. We calculated $\|m(\mathbf{X}) - m\{E(\mathbf{X}|\mathbf{W})\}\|$ and $\|m(\mathbf{X}) - \text{ave}\{m(\mathbf{X})\}\|$ where "ave" means average over the MCMC output and is thus a Monte Carlo estimate of conditional expectation given the Y_i and the W_{ij}. For six samples, the ratios of these norms were 3.1, 3.7, 3.8, 2.7, 2.1, and 4.2. These values are consistently well above 1, showing that the fully Bayesian approach is giving more information about X_i than what is available from $\overline{W}_i$ alone. The latter also uses the full structural model for the marginal distribution of the X_i, so it is not the structural assumption that is giving extra information about the X_i to the Bayesian estimator; rather, it is the regression model relating Y_i to X_i that provides this information. Clearly, this leaves open the possibility that with highly nonnormal errors, or highly heteroscedastic ones, the misspecified information from our simple model will lead to bias or other deleterious behavior in the Bayesian method.

We study the case where the measurement error and natural error are normally distributed. In most of the cases in the measurement error literature, the results are robust to the assumption of normality, once the additivity of errors in (4) is satisfied, possibly by transformation. Extending the normality of the ϵ's in (3), the natural error, to other distributions adds a level of complexity to the problem, because the normality of the complete conditional distribution for the spline will no longer hold. The methods presented in this article could be naturally combined with the work of Hastie and Tibshirani (1998) for modeling measurement error in semiparametric models.

While this manuscript was undergoing a final revision, an interesting unpublished manuscript by Ganguli, Staudenmayer, and Wand appeared. These authors extend our model by assuming multiple covariates and an additive regression function. They also estimate fixed effects and the variance components (of the random coefficients in the spline) by maximum likelihood. Because the smoothing parameters are ratios of variance components, these are also chosen by maximum likelihood. The likelihood involves a high-dimensional integral, so computation of the MLEs is not trivial, and the author uses the nesting EM algorithm of van Dyk (2000). Simulations indicate that the MLEs behave satisfactorily, but the MLEs were not compared to other estimators.

APPENDIX: BAYESIAN IMPLEMENTATION FOR REGRESSION P–SPLINES

For fixed-knot P-splines, $g(x) = \mathbf{B}^\mathsf{T}(x)\boldsymbol{\beta}$, $\mathbf{g} = \mathcal{B}\boldsymbol{\beta}$, and the penalty matrix is $\mathbf{D}$. Apportion $\mathbf{B}(x) = \{\mathbf{B}_1^\mathsf{T}(x), \mathbf{B}_2^\mathsf{T}(x)\}^\mathsf{T}$, where $\mathbf{B}_1^\mathsf{T}(x)$ is the first $p+1$ elements of $\mathbf{B}^\mathsf{T}(x)$. Apportion $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})^\mathsf{T}$ similarly. Then let the prior for $\boldsymbol{\beta}_1$ be normal$(0, \delta\Sigma)$ for a fixed covariance matrix Σ and δ "large," and the prior for $\boldsymbol{\beta}_2$ be normal$\{0, \gamma^{-1}\mathbf{I})$ ($\mathbf{I}$ is the identity matrix). Define the matrix $\mathbf{D}_* = \sigma_\epsilon^2 \text{diag}(\Sigma^{-1}/\delta, \gamma\mathbf{I})$. In the limit as $\delta \to \infty$, $\mathbf{D}_* \to \sigma_\epsilon^2 \gamma \mathbf{D}$. With these conventions, the joint posterior for $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\mathbf{W} = \text{normal}(\mathbf{QH},\mathbf{Q}),$$

$$\mathbf{H} = \sigma_\epsilon^{-2}\sum_{i=1}^{n}\mathbf{B}(X_i)Y_i = \sigma_\epsilon^{-2}\mathcal{B}^T\mathbf{Y},$$

and

$$\mathbf{Q} = \sigma_\epsilon^2\left\{\sum_{i=1}^{n}\mathbf{B}(X_i)\mathbf{B}^\mathsf{T}(X_i)+\mathbf{D}_*\right\}^{-1} = \sigma_\epsilon^2(\mathcal{B}^T\mathcal{B}+\mathbf{D}_*)^{-1}.$$

Similarly, the complete conditional for $\mathbf{X}$ is

$$X_i|Y_i, W_i \propto \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\{Y_i-\mathbf{B}^\mathsf{T}(X_i)\boldsymbol{\beta}\}^2 - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij}-X_i)^2 - \frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(X_i-\mu_x)^2\right\}.$$

As described in the text, Metropolis–Hastings steps can be used to generate new values of the X_i's.

The full conditional for γ given $\boldsymbol{\beta}$ is $\mathrm{G}(A_\gamma + k/2, \{B_\gamma^{-1} + \boldsymbol{\beta}_2^\mathsf{T}\beta_2/2\}^{-1})$. The full conditionals for σ_ϵ^2, σ_u^2, μ_x, and σ_x^2 are the same as for smoothing splines as given in Section 3.2.

Unlike smoothing splines, P-splines need not have their knots at the values of the X_i. Instead, we place the P-spline knots at fixed quantiles of the $\overline{W}_i$, so that the knots are fixed throughout the MCMC iterations. More specifically, if there are k knots, then we take $k+2$ values of p equally spaced on $[0,1]$, delete the first and last (0 and 1), and then place a knot at the pth quantile of the $\overline{W}_i$ for each of these k values of p.

[Received January 2000. Revised October 2000.]

REFERENCES

Amemiya, Y., and Fuller, W. A. (1988), "Estimation for the Nonlinear Functional Relationship," *The Annals of Statistics*, 16, 147–160.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Besag, J. (1986), "On the Statistical Analysis of Dirty Pictures" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 48, 259–279.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Regression With Errors in Covariates," *Biometrika*, 86, 541–554.

Carroll, R. J., Roeder, K., and Wasserman, L. (1999), "Flexible Parametric Measurement Error Models," *Biometrics*, 55, 44–54.

Cheng, C. L., and Schneeweiss, H. (1998),"Polynomial Regression With Errors in Variables," *Journal of the Royal Statistical Society*, Ser. B, 60, 189–200.

Cook, J. R., and Stefanski, L. A. (1994), "Simulation–Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-Splines and Penalties" (with discussion), *Statistical Science*, 11, 89–102.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing* (2nd ed.), New York: Marcel Dekker.

Fan, J., and Truong, Y. K. (1993), "Nonparametric Regression with Errors in Variables," *The Annals of Statistics*, 21, 1900–1925.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

——— (2000),"Bayesian Backfitting," (with discussion), *Statistical Science*, 15, 193–223.

Meng, X. L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Nychka, D. (1988), "Bayesian Confidence Intervals for a Smoothing Spline," *Journal of the American Statistical Association*, 83, 1134–1143.

——— (1990), "The Average Posterior Variance of a Smoothing Spline and a Consistent Estimate of the Average Squared Error," *The Annals of Statistics*, 18, 415–428.

Ruppert, D. (2000), "Selecting the Number of Knots for Penalized Splines," preprint (available at www.orie.cornell.edu/~davidr/papers).

Ruppert, D., and Carroll, R. J. (2000), "Spatially Adaptive Penalties for Spline Fitting," *Australia and New Zealand Journal of Statistics*, 42, 205–223.

van Dyk, D. A. (2000), "Nesting EM Algorithms for Computational Efficieny," *Statistica Sinica*, 10, 203–226.

Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B, 40, 364–372.

——— (1983),"Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Ser. B, 45, 133–150.

——— (1990), *Spline Models for Observational Data*, Providence: SIAM Press.

This article has been cited by:

1. Xin-Yuan Song, Zhao-Hua Lu. 2010. Semiparametric Latent Variable Models With Bayesian P-SplinesSemiparametric Latent Variable Models With Bayesian P-Splines. *Journal of Computational and Graphical Statistics* **19**:3, 590-608. [Abstract] [PDF] [PDF Plus] [Supplementary material]
2. Raymond J. Carroll, Aurore Delaigle, Peter Hall. 2009. Nonparametric Prediction in Measurement Error ModelsNonparametric Prediction in Measurement Error Models. *Journal of the American Statistical Association* **104**:487, 993-1003. [Abstract] [PDF] [PDF Plus] [Supplementary material]
3. Yu-Jen Cheng, Ciprian M. Crainiceanu. 2009. Cox Models With Smooth Functional Effect of Covariates Measured With ErrorCox Models With Smooth Functional Effect of Covariates Measured With Error. *Journal of the American Statistical Association* **104**:487, 1144-1154. [Abstract] [PDF] [PDF Plus] [Supplementary material]
4. Aurore Delaigle , , Jianqing Fan , , Raymond J. Carroll . 2009. A Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables ProblemA Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables Problem. *Journal of the American Statistical Association* **104**:485, 348-359. [Abstract] [PDF] [PDF Plus]
5. Duchwan Ryu , Debajyoti Sinha , Bani Mallick , Stuart R . Lipsitz , Steven E . Lipshultz . 2007. Longitudinal Studies With Outcome-Dependent Follow-upModels and Bayesian RegressionLongitudinal Studies With Outcome-Dependent Follow-up. *Journal of the American Statistical Association* **102**:479, 952-961. [Abstract] [PDF] [PDF Plus]
6. N. D . Pearce , M. P . Wand . 2006. Penalized Splines and Reproducing Kernel MethodsPenalized Splines and Reproducing Kernel Methods. *The American Statistician* **60**:3, 233-240. [Abstract] [PDF] [PDF Plus]
7. Panu Erästö , Lasse Holmström . 2005. Bayesian Multiscale Smoothing for Making Inferences About Features in ScatterplotsBayesian Multiscale Smoothing for Making Inferences About Features in Scatterplots. *Journal of Computational and Graphical Statistics* **14**:3, 569-589. [Abstract] [PDF] [PDF Plus]

Chapter 2
Transformation and Weighting

By David Ruppert

About the Author. David Ruppert is Andrew Schulz Jr. Professor of Engineering, School of Operations Research and Information Engineering, and Professor of Statistical Science, Cornell University. He received a PhD in Statistics and Probability from Michigan State University in 1977. He was Assistant and then Associate Professor of Statistics at the University of North Carolina, Chapel Hill, from 1977 to 1987 during which time his office was next to Ray's and they collaborated intensely. He is a Fellow of the ASA and IMS and received the Wilcoxon Prize in 1986 jointly with Ray. He has had 28 PhD students and three of them, Len Stefanski, David Giltinan, and Doug Simpson, were jointly advised with Ray. Professor Ruppert has written five books of which three, *Transformation and Weighting in Regression*, *Measurement Error in Nonlinear Models* (first and second editions), and *Semiparametric Regression* were coauthored with Ray. He and Ray have coauthored 37 papers.

Selected Papers on Transformation and Weighting

[TW-1]-[70] Carroll, R. J. and Ruppert, D. (1981). Prediction and the power transformation family. *Biometrika*, 68, 609–616.

[TW-2]-[57] Carroll, R. J. and Ruppert, D. (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, 77, 878–882.

[TW-3]-[150] Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79, 321–328.

[TW-4]-[422] Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1092.

By the early 1980s, regression with homoscedastic errors was well understood, but methodology for handling heteroscedastic noise was just being developed. There were two general approaches. In the first, studied by Carroll and Ruppert (1981 [TW-1], 1984 [TW-3]), the response is transformed to homoscedasticity. In the second, studied by Carroll and Ruppert (1982 [TW-2]) and Davidian and Carroll (1987 [TW-4]), one uses a variance function that specifies the conditional variance of the response given the covariates. Transformation has the added feature that it can also reduce skewness of the errors, but transformation is useful only when the conditional variance is of a special form and, in particular, is a function of the conditional mean; this is a common occurrence, but there are many applications where it does not occur. Transformation and variance functions can be combined into a very general methodology as described briefly below.

There are two important reasons for modeling the conditional variance. The first is that the regression parameters can be more precisely estimated if one weights by the reciprocals of the conditional variances. The second is that prediction and

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_2,

calibration intervals can be grossly inaccurate (true coverage probabilities far from nominal values) if one ignores the heteroscedasticity. As Davidian and Carroll (1987 [TW-4]) note, the second reason may be more important. A weighted analysis is significantly more efficient than an unweighted one only when there is substantial heteroscedasticity, but even a small amount of heteroscedasticity, say the conditional standard deviation varying by a factor of two, can cause prediction and calibration intervals to be seriously in error.

Transformation and the Box–Cox Controversy

Carroll and Ruppert (1981 [TW-1]) find a middle ground in a somewhat acrimonious controversy about the use of the Box–Cox transformation model in practice. Although the transformation of variables, e.g., replacing a variable by its logarithm, has had a long history in statistics, estimation of transformation parameters was not put on a firm theoretical footing until Box and Cox (1964). Their model is

$$y_i^{(\lambda)} = x_i\beta + \sigma\varepsilon_i, \tag{2.1}$$

where y_i is a nonnegative response for the ith case, x_i is a vector of predictors, β is a vector of regression coefficients, σ is the residual standard deviation, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are i.i.d. $N(0,1)$, or more generally i.i.d. F for some known F. Here,

$$\begin{aligned} y^{(\lambda)} &= (y^{\lambda}-1)/\lambda, \quad && \lambda \neq 0, \\ &= \log(y), && \lambda = 0, \end{aligned} \tag{2.2}$$

embeds the log transformation smoothly into the power transformation family. Model (2.1) states that, after transformation by an unknown parameter λ, the response follows a homoscedastic, Gaussian linear model.

The controversy was over whether inference about β should be conditional on the value of λ or not. Box and Cox (1982) recommend the conditional approach so that once λ is estimated, λ is treated as if it were known and equal to its estimator $\hat{\lambda}$. Bickel and Doksum (1981) disagree and study the sampling variability of β when λ is treated as unknown. Because the value of β is highly dependent on that of λ, the estimators $\hat{\beta}$ and $\hat{\lambda}$ are highly correlated, and the standard deviations of the components of $\hat{\beta}$ are much larger when λ is estimated compared to when λ is treated as known. In summary, Box and Cox argue that uncertainty about λ should be ignored when making inference about β, while Bickel and Doksum argue that this uncertainty should be acknowledged and has a large effect, so that inference about β is unstable.

Neither of these viewpoints seems entirely satisfactory. In a rebuttal to Bickel and Doksum, Box and Cox (1982) ask "how can it be sensible scientifically to state a conclusion as number measured on an unknown scale?" This is a reasonable question. On the other hand, there are few if any other estimation problems where ignoring the uncertainty in nuisance parameters is recommended in practice. Certainly, there must be some cost due to estimation of λ.

Carroll and Ruppert (1981 [TW-1]) study the problem of prediction about y on the original scale. That is, they study $f(\hat{\lambda}, x_0\beta^*)$ where $f(\lambda, \cdot)$ is the inverse of $y^{(\lambda)}$ so that $f(\lambda, y^{(\lambda)}) \equiv y$, x_0 is a value where prediction is to be made, and β^* is an estimator of β. Working on the original scale circumvents Box and Cox's objection to conclusions stated on an unknown scale.

Carroll and Ruppert (1981 [TW-1]) show that the high correlation between $\hat{\lambda}$ and β^* has effects that are similar to the effects of multicollinearity in multiple regression. Both have small, but non-ignorable, effects on prediction. Carroll and Ruppert first look at the case of simple linear regression with $\lambda = 0$ and prove a general result showing that the cost (inflation of the mean squared error) due to estimating λ cannot exceed 50 % and often is much smaller, e.g., at most 8 % in the balanced two-sample problem. Then, they look at the general case where the dimension of β is p and extend the model so that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are i.i.d. F for some known F. Asymptotic results are messy in the general case but simplify if one uses small-σ asymptotics where $\sigma \to 0$ as $n \to \infty$. Small-σ asymptotics were also used by Bickel and Doksum. In the small-σ case, the cost of estimating λ is $1/p$, exactly the same as the effect of adding an additional covariate in linear regression. In their last section, they look at the problem of predicting the mean response and show that the cost of adding r additional nuisance parameters when there are q parameters in the model is bounded by r/q.

Estimation of the mean on the original scale was studied further by Taylor (1986). Taylor (1988) studied the related problem of estimating event probabilities using binary regression where the link function contains an unknown parameter. Taylor, Siqueira, and Weiss (1996) propose a general framework that includes the Box–Cox model and binary regression with link parameters as special cases. In all three papers, it was found that the cost of estimating the unknown nuisance parameters is small but not ignorable.

Weighting in Regression

Carroll and Ruppert (1982 [TW-2]) address the question of whether one should use the generalized least squares estimator (GLSE) or the normal-theory maximum likelihood estimator (MLE) when fitting heteroscedastic models. The weighted least-squares estimator weights each squared residual by the reciprocal of its conditional variance, but is generally not available since the conditional variances typically are unknown. The GLSE replaces the unknown conditional variances by estimators. The MLE maximizes the likelihood under the working assumption that the errors are normally distributed. Of course, it is only a true maximum likelihood estimator when that assumption holds. The Carroll and Ruppert model is

$$Y_i = x_i^T \beta + \varepsilon_i \{f(x_i, \beta, \theta)\}^{-1/2}, \tag{2.3}$$

where Y_i is the response, x_i is a vector of covariates, β contains the regression coefficients, $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. with variance σ^2, f is unknown function that models the heteroscedasticity, and θ is a vector of parameters that specify the conditional variance of Y_i given x_i. A typical example of f is

$$\{f(x_i,\beta,\theta)\}^{-1/2} = (x_i^T\beta)^{\alpha}, \tag{2.4}$$

so that the conditional variance is proportional to a power of the conditional mean. (In this model, it is usually assumed that $x_i^T\beta$ is positive.)

There are two sources of information about β, the conditional mean $x_i^T\beta$ and the conditional standard deviation $[f(x_i,\beta,\theta)]^{-1/2}$. Maximum likelihood uses both sources and has the smallest possible asymptotic covariance matrix if the errors are Gaussian as the MLE assumes. The GLSE uses only the first source and, in general, is not fully efficient. However, variance models are often only approximations. Carroll and Ruppert (1982 [TW-2]) show that even a minor misspecification of the heteroscedasticity can degrade the performance of the MLE but has little effect on the GLSE.

More precisely, Carroll and Ruppert (1982 [TW-2]) assume that

$$Y_i = x_i^T\beta + \varepsilon_i[G_N(x_i,\beta,\theta)]^{-1/2}, \tag{2.5}$$

where N is the sample size, $G_N(x_i,\beta,\theta) = f(x_i,\beta,\theta)\{1+2BN^{-1/2}h(x_i,\beta,\theta)\}$, and $N^{-1}\sum_{i=1}^{N} h^2(x_i,\beta,\theta) \to \mu$ for some $0<\mu<\infty$. Thus, $2BN^{-1/2}h(x_i,\beta,\theta)$ represents the misspecification of the conditional standard deviation and, since it decays to 0 at rate $N^{-1/2}$, the model misspecification is too small to be detected with certainty even in the limit as $N\to\infty$. More formally, the true model is contiguous to the assumed model.

The asymptotic distribution of the GLSE assuming model (2.3) is the same under the models (2.3) and (2.5), so that the GLSE is not affected by contiguous misspecification. The asymptotic distribution of the MLE assuming model (2.3) has the same (fully efficient) asymptotic variance under models (2.3) and (2.5), but there is a bias under (2.5). Whether the MLE or the GLSE has the smaller asymptotic mean squared error (MSE) depends on the amount of model misspecification as determined by $B, h(x_1,\beta,\theta),\ldots,h(x_N,\beta,\theta)$, and how much information about β is contained in the conditional standard deviations. The latter is determined by $w_1,\ldots,w_N$ where, with $\hat\beta_M$ the MLE, we have

$$N^{1/2}(\hat\beta_M-\beta) = N^{-1/2}\sum_{i=1}^{N}\{v_i\varepsilon_i + w_i(\varepsilon_i^2-1)\} + o_P(1), \tag{2.6}$$

so that, roughly speaking, w_i, $i=1,\ldots,N$, determine how the second source of information about β is used and $v_1,\ldots,v_N$ do the same for the first source.

In summary, the asymptotic distribution of the GLSE is robust to misspecification of the conditional standard deviation, but this is not true of the MLE. If there is no misspecification, then the MLE has the smallest asymptotic mean squared error (MSE), but under misspecification either the MLE or the GLSE may have the smallest MSE.

Carroll and Ruppert (1982 [TW-2]) also discuss robustness to outliers. For the GLSE, (2.6) holds with $w_i\equiv 0$ so the GLSE depends linearly, not quadratically, on $\varepsilon_1,\ldots,\ \varepsilon_N$. Although neither the GLSE nor the MLE is robust to outliers, the

MLE is more seriously affected by outliers because it depends quadratically upon the errors. A robust M-estimator called ROBUST WEIGHTED is also considered in the paper and, in a Monte Carlo study, is the best performing estimator, even when the heteroscedasticity is correctly specified and the errors are normally distributed; in this case, it is tied with the MLE.

Carroll and Ruppert (1984 [TW-3]) propose a model that is at first glance superficially similar to, but ultimately rather different from, the Box–Cox (1964) transformation model. The Carroll–Ruppert model starts with a theoretical model

$$y_i = f(x_i, \theta_0),\ i = 1, \dots, N, \tag{2.7}$$

relating a response y_i to a covariate vector x_i. Here f is a known function that might have been derived from scientific theory, e.g., pharmacokinetics, and θ_0 is an unknown parameter vector. Model (2.7) will not hold exactly and in many cases there will be substantial variation of y_i about $f(x_i, \theta_0)$.

To estimate θ_0, one can expand (2.7) to the nonlinear regression model

$$y_i = f(x_i, \theta_0) + \varepsilon_i,\ i = 1, \dots, N, \tag{2.8}$$

where $\varepsilon_1, \dots, \varepsilon_N$ are i.i.d. errors and typically are assumed to be normally distributed. Carroll and Ruppert noted that (2.7) is equivalent to $h(y_i) = h\{f(x_i, \theta_0)\}$, for all i, where h is any invertible transformation. However, the noise model

$$h(y_i) = h\{f(x_i, \theta_0)\} + \varepsilon_i, \tag{2.9}$$

with $\varepsilon_1, \dots, \varepsilon_N$ i.i.d. Gaussian, can hold for at most one h. Therefore, there is no compelling reason to assume (2.8). Instead, Carroll and Ruppert (1984 [TW-3]) argue that (2.9) holds for some h in a parametric family of transformations, e.g., (2.2). As an example, if there are multiplicative lognormal errors so that $y_i = f(x_i, \theta_0)\exp(\varepsilon_i)$ where $\varepsilon_1, \dots, \varepsilon_N$ are i.i.d. normal, then (2.9) holds with $h(y) = \log(y)$.

Model (2.9) seeks a transformation h that induces additive, homoscedastic, and Gaussian errors. The Box–Cox transformation also has these goals, but the Box–Cox transformation model has a third goal, inducing a simple linear model. For example, $x_i\beta$ in (2.1) might be a no-interaction model and then one seeks a λ so that this no-interaction model holds; Box and Cox (1964) provide such an example. In other examples, $x_i = (1\ w_i)$ for a scalar covariate w_i and one seeks λ so that $E(y_i|w_i)$ is linear in w_i. In contrast, model (2.9) does *not* seek to simplify the regression model. Instead, it preserves the regression model by applying h to both y_i and $f(x_i, \theta_0)$. In practice, h will be monotonic and then (2.9) implies that the median of y_i is $f(x_i, \theta_0)$; this is the sense in which the model is preserved. Stated differently, the Carroll–Ruppert method is used when y_i already fits the regression model while the Box–Cox method is used when y_i must be transformed to fit the regression model.

Because $f(x_i, \theta_0)$ is the median of y_i, the problem of stating conclusions on an unknown scale is avoided. Conclusions can be stated about y_i itself. Therefore, the controversy discussed previously about inference for the Box–Cox model is avoided. Using small-σ asymptotics, Carroll and Ruppert show that the limit distribution of $\hat{\theta}$ is the same when the transformation parameter is unknown as when it is known. A more general result that does not use small-σ asymptotics is that the cost of not knowing the transformation parameter is at most $\pi/2 = 1.57$. This bound should be contrasted with the huge costs that Bickel and Doksum found for the Box–Cox model. Moreover, Carroll and Ruppert's Monte Carlo study shows that this bound is usually quite conservative.

Davidian and Carroll (1987 [TW-4]) provide a comprehensive study of variance function estimation and compare the many variance function estimators that have been proposed. They use the model

$$EY_i = \mu_i = f(x_i, \beta); \quad \text{var}(Y_i) = \sigma^2 g^2(z_i, \beta, \theta), \tag{2.10}$$

where Y_i is a response, x_i is a vector of covariates in the regression function f, z_i is the vector of covariates in the variance function g^2, β is a vector of regression parameters, θ is a vector of variance parameters, and $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d.. Typically, β is estimated by ordinary least squares and fixed. The residuals from this preliminary estimator of β can be used to estimate θ. For example, the squared residuals are estimators of g^2 though they are biased unless one corrects for the loss of degrees of freedom. Often, $\log(g)$ is linear in θ, and then it is tempting to use the logarithms of the absolute residuals as the responses, though Davidian and Carroll note that residuals near zero induce outliers when this is done. If the data come in groups where x_i and z_i are constant, then the sample variances of these groups are unbiased estimators of g^2 and can be used as the responses in a regression model with g^2 as the regression function.

Combining Transformation and Weighting

Transformation and weighting need to be combined in some applications. A generalization of (2.9) discussed in Chapter 5 of Carroll and Ruppert (1988) is

$$h(y_i) = h\{f(x_i, \theta_0)\} + \sigma g(z_i, \beta, \theta)\varepsilon_i. \tag{2.11}$$

One application of this model is to fitting the Michaelis–Menten equation of enzyme kinetics. A number of methods for estimating the Michaelis–Menten parameters have been proposed. Ruppert, Carroll, and Cressie (1989) show that all of these are special cases of a general transformation/weighting model, so each is efficient only for a certain error structure, that is, for particular values of the transformation and variance parameters. By using the general model, one can adapt to the error structure and obtain more accurate estimators.

References

Other publications by Ray Carroll cited in this chapter.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.

Ruppert, D., Carroll, R. J., and Cressie, N. (1989). A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics*, 45, 637–656.

Publications by other authors cited in this chapter.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296–311.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society, Series B*, 35, 473–479.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformation revised, rebutted. *Journal of the American Statistical Association*, 77, 209–210.

Taylor, J. M. G. (1986). The retransformed mean after fitting a power transformation, *Journal of the American Statistical Association*. **81**, 114–118.

Taylor, J. M. G. (1988). The cost of generalized logistic regression. *Journal of the American Statistical Association*, **83**, 1078–1083.

Taylor, J. M. G., Siqueira, A. L., and Weiss, R. T. (1996). The cost of adding parameters to a model. *Journal of the Royal Statistical Society, Series B*, 58, 593–607.

Biometrika (1981), 68, 3, *pp.* 609–15
Printed in Great Britain

On prediction and the power transformation family

BY R. J. CARROLL AND DAVID RUPPERT

Department of Statistics, University of North Carolina, Chapel Hill

SUMMARY

The power transformation family is often used for transforming to a normal linear model. The variance of the regression parameter estimators can be much larger when the transformation parameter is unknown and must be estimated, compared to when the transformation parameter is known. We consider prediction of future untransformed observations when the data can be transformed to a linear model. When the transformation must be estimated, the prediction error is not much larger than when the parameter is known.

Some key words: Asymptotic distribution; Box–Cox family; Maximum likelihood estimation; Monte-Carlo simulation; Prediction of conditional median; Robustness.

1. INTRODUCTION

The power transformation family studied by Box & Cox (1964) takes the following form: for some unknown λ and $i = 1, \ldots, n$,

$$y_i^{(\lambda)} = x_i\beta + \sigma\varepsilon_i, \quad x_i = (1, c_{i2}, \ldots, c_{ip}), \quad \beta' = (\beta_0, \ldots, \beta_{p-1}). \tag{1·1}$$

Here σ is the standard deviation; the ε_i are independently and identically distributed with mean zero, variance one and distribution F, and

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0), \\ \log y & (\lambda = 0). \end{cases}$$

Box & Cox propose maximum likelihood estimates for λ and β when F is the normal distribution. There are numerous alternative methods as well as proposals for testing hypotheses of the form H_0: $\lambda = \lambda_0$ (Hinkley, 1975; Andrews, 1971; Atkinson, 1973; Carroll, 1980). Carroll studied the testing problem via Monte-Carlo; by allowing F to be nonnormal he approximated a problem with outliers and found that the chance of mistakenly rejecting the null hypothesis can be very high indeed.

Bickel & Doksum (1981) develop an asymptotic theory for estimation. For technical reasons they assume that the design vectors $x_1, x_2, \ldots$ are independent and identically distributed according to G. If the maximum likelihood estimate of the regression parameter is $\hat\beta$ when λ is known, and $\beta^* = \hat\beta(\hat\lambda)$ when λ is unknown and estimated by $\hat\lambda$, they compute the asymptotic distributions of $n^{\frac{1}{2}}(\hat\beta - \beta)/\sigma$ and $n^{\frac{1}{2}}(\beta^* - \beta)/\sigma$ as $n \rightarrow \infty$ and $\sigma \rightarrow 0$. These distributions, which are given in the Appendix, are different, and as regards variances

> the cost of not knowing λ and estimating it . . . is generally severe. . . . The problem is that β^* and $\hat\lambda$ are highly correlated.

Their theoretical and Monte Carlo work indicate that $\hat{\lambda}$ and β^* are highly variable and highly correlated, and as discussed in § 2, the problem is similar in nature to that of multicollinearity. An example of the variability of β^* is given in the next section.

These results are somewhat controversial. One point of discussion concerns the scale on which inference is to be made: i.e. should one make unconditional inference about the regression parameter in the correct but unknown scale, as in Bickel & Doksum's theory, or a conditional inference for an appropriately defined 'regression parameter' in an estimated scale?

In order to eliminate such problems, we will study the cost of estimating λ when one wants to make inferences in the original scale of the observations. In the multicollinearity problem, reasonably good prediction is still possible if new vectors x arrive independently with the distribution G. Motivated by this fact, we focus our attention specifically on prediction, but we also discuss the two-sample problem and a somewhat more general estimation theory. Using Bickel & Doksum's asymptotic theory and Monte Carlo, we find that for prediction as well as other problems in the original scale there is a cost due to estimating λ, but it is generally not severe.

2. Predicting the conditional median in regression

2·1. *The general case*

Our model specifically includes an intercept, i.e. $x_i = (1, c_i)$; by suitable rescaling we assume the c_i have mean zero and identity covariance. From the sample we calculate $\hat{\lambda}$ and β^*, and we are given a new vector $x_0 = (1, c_0)$, which is independent of the other x's but still has the same distribution G. This formulation is simple but hardly necessary; the design vectors x_i could satisfy the usual regression assumptions, and x_0 can be thought of as chosen according to the design. Our predicted value in the transformed scale would be $x_0\,\beta^*$, so a natural predictor is $f(\hat{\lambda}, x_0\,\beta^*)$ where

$$f(\lambda, \theta) = \begin{cases} (1+\lambda\theta)^{1/\lambda} & (\lambda \neq 0), \\ e^{\theta} & (\lambda = 0). \end{cases}$$

Notice that if F has median equal to 0, then $f(\lambda, x_0\,\beta)$ is the median of the conditional distribution of y given x_0, even though it is not necessarily the conditional expectation. Calculation of conditional expectations would require the use of numerical integration and that F be known or an estimator of F be available. See § 3 for further discussion.

A Taylor expansion shows that

$$f(\hat{\lambda}, x_0\,\beta^*) - f(\lambda, x_0\,\beta)/g(\lambda, x_0\,\beta) \bumpeq x_0(\beta^* - \beta) + h(\lambda, x_0\,\beta)\,(\hat{\lambda} - \lambda) \qquad (2{\cdot}1)$$

where

$$g(\lambda, \theta) = f(\lambda, \theta)/(1+\lambda\theta), \quad h(\lambda, \theta) = \theta/\lambda - \{(1+\lambda\theta)\log(1+\lambda\theta)\}/\lambda^2.$$

Estimates $\hat{\lambda}$ and β^* are unstable and highly correlated, and expansion (2·1) shows that our problem as presently formulated is quite similar to a prediction problem in regression when there is multicollinearity.

2·2. *Case* 1

We now assume that F is a normal distribution, $\lambda = 0$, $\sigma = 1$, and the model is simple linear regression with slope β_1 and intercept β_0.

For this special case, likelihood calculations (Hinkley, 1975) can be made. Here the correct scale is the log scale and $E(c_i) = 0$, $E(c_i^2) = 1$, $E(c_i^3) = \mu_3$ and $E(c_i^4) = \mu_4$. Lengthy likelihood analysis shows

$$n \operatorname{cov}(\hat{\lambda}, \beta_0^*, \beta_1^*) \rightarrow \Sigma_0,$$

where

$$\Sigma_0 = 2\gamma^{-1} \begin{bmatrix} 1 & -c & \beta_0 \beta_1^* \\ -c & \tfrac{1}{2}\gamma + c^2 & -c\beta_0 \beta_1^* \\ \beta_0 \beta_1^* & -c\beta_0 \beta_1^* & \tfrac{1}{2}\gamma + \beta_0^2 \beta_1^{*2} \end{bmatrix}$$

and where

$$c = -\tfrac{1}{2}(1+\beta_0^2+\beta_1^2), \quad \beta_1^* = \beta_1 + \tfrac{1}{2}\beta_1^2 \mu_3/\beta_0, \quad \gamma = 3 + 4\beta_1^2 + \beta_1^4(\mu_4 - \mu_3^2 - 1).$$

Note that if λ were not estimated we would have had Σ_0 as the identity matrix, and in the next section we give an example which demonstrates the multicollinearity.

THEOREM 1. *Let* $\mathrm{MSE}(\lambda, x_0)$ *be the mean squared error for estimating the conditional median of Y given* x_0 *and* λ *known, while* $\mathrm{MSE}(\hat{\lambda}, x_0)$ *is the same quantity but with* λ *unknown. Then*

$$E_G\left(\| x_0 \|^2 \frac{\mathrm{MSE}(\hat{\lambda}, x_0)}{\mathrm{MSE}(\lambda, x_0)} \right) \Big/ E(\| x_0 \|^2) \rightarrow H(\beta_1), \tag{2·2}$$

where

$$H(\beta_1) = 1 + \tfrac{1}{2}\{1 + \beta_1^4(\mu_4 - 1 - \mu_3^2)\}\{6 + 8\beta_1^2 + \beta_1^4(\mu_4 - 1 - \mu_3^2)\}^{-1}.$$

Note that $\mu_4 - 1 - \mu_3^2 = E\{(c_1^2 - \mu_3 c_1 - 1)^2\} \geqslant 0$. The quantity (2·2) is a modified form of the average cost for prediction when λ is estimated. If one prefers to assume the design vectors are constants, then one might think of (2·2) as an average over the design. In either case the results are encouraging:

(i) there is a cost due to estimating λ, but it cannot exceed 50%;

(ii) for the balanced two-sample problem, $c_i = \pm 1$ with probability $\tfrac{1}{2}$, the cost is at most 8% and decreases to zero as $\beta_1 \rightarrow \infty$.

2·3. *Case 2: Symmetric errors*

We now allow λ and the number of regression parameters, p, to be arbitrary, but we assume that F is symmetric about zero.

Here we use the asymptotic theory of Bickel & Doksum, in which $n \rightarrow \infty$ and $\sigma \rightarrow 0$ simultaneously; see the Appendix for details. We report results only for the simplest case of an orthogonal design in which

$$n^{-1} \sum_{i=1}^{n} x_i' x_i \rightarrow I.$$

It then follows that $(\lambda, \hat{\beta}^*)$ is asymptotically normally distributed with mean (λ, β) and covariance $\sigma\Sigma_1/n$, where

$$\Sigma_1 = e^{-1} \begin{bmatrix} 1 & -D \\ -D' & eI + D'D \end{bmatrix},$$

and

$$x = (1, x_2, \ldots, x_p) = (x_1, \ldots, x_p), \quad H(a,\lambda) = \lambda^{-1} a - \lambda^{-2}(1+\lambda a)\log(1+\lambda a),$$

$$D = E\{H(x\beta,\lambda)x\}, \quad e = E[\{H(x\beta,\lambda)\}^2] - \sum_{j=1}^{p} [E\{x_j H(x\beta,\lambda)\}]^2.$$

It is interesting that in the case of simple linear regression $\lambda = 0$, Σ_1 is different from but of the same form as Σ_0. More precisely, c is replaced by $c_* = c + \frac{1}{2}$ and $\frac{1}{2}\gamma$ by $e = \beta_1^4(\mu_4 - \mu_3^2 - 1)/4$.

Theorem 2. *As $N \rightarrow \infty$ and $\sigma \rightarrow 0$ for any λ,*

$$E_G\left\{\|x_0\|^2 \frac{\mathrm{MSE}\,(\hat{\lambda}, x_0)}{\mathrm{MSE}\,(\lambda, x_0)}\right\} \Big/ E_G(\|x_0\|^2) \rightarrow 1 + 1/p,$$

where p is the dimension of the vector β.

The small σ asymptotics of Bickel & Doksum tell us that there is a positive but bounded cost due to estimating λ, with the cost decreasing as p increases. Note that Theorem 2 and Theorem 1 agree for simple linear regression, $\lambda = 0$, $\mu_4 - 1 - \mu_3^2 > 0$ and $\beta_1 \rightarrow \infty$.

Bickel & Doksum and Carroll also simultaneously introduced robust estimates of (λ, β) based on the ideas of Huber (1977). One can use Bickel & Doksum's small σ asymptotics to show that (i) the cost in robust estimation for estimating λ is still $1/p$ and (ii) Bickel & Doksum's and Carroll's methods have better robustness properties than does maximum likelihood.

We conducted a small Monte Carlo study to check small sample performance and to investigate the results of Theorems 1 and 2. The observations were generated according to $(1 + \beta_0 + \beta_1 c_i + \varepsilon_i)^{1/\lambda}$ for $\lambda = -1$, and $\exp(\beta_0 + \beta_1 c_i + \varepsilon_i)$ for $\lambda = 0$. Here $n = 20$, the ε_i are standard normal, $\beta_0 = 5$, $\beta_1 = 1$ and the c_i centred at zero, equally spaced, satisfy $\Sigma c_i^2 = n$ and range from $-1{\cdot}65$ to $1{\cdot}65$. Then $\mu_4 = 1{\cdot}79$ and $H(\beta_1) = 1{\cdot}06$, so that Theorems 1 and 2 lead us to expect very little cost due to estimating λ. There were 600 repetitions of the experiment. Likelihood calculations show that

$$\Sigma_0 = \begin{bmatrix} 0{\cdot}27 & 3{\cdot}65 & 1{\cdot}35 \\ \cdot & 50{\cdot}28 & 18{\cdot}25 \\ \cdot & \cdot & 7{\cdot}76 \end{bmatrix}$$

with correlation matrix

$$\begin{bmatrix} 1 & 0{\cdot}99 & 0{\cdot}93 \\ \cdot & 1 & 0{\cdot}92 \\ \cdot & \cdot & 1 \end{bmatrix},$$

which illustrates the multicollinearity quite well, for if λ were known then $n^{\frac{1}{2}}(\hat{\beta}_0 - \beta_0)$ and $n^{\frac{1}{2}}(\hat{\beta}_1 - \beta_0)$ would be uncorrelated with common variance 1.

In rows 1 to 4 of Table 1, we provide an analysis of the estimates β_0^* and β_1^* in the case that λ is estimated. The estimates are biased and have much larger mean squared errors than the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained for the case that λ is known.

The remaining rows of Table 1 give the results for the prediction problem. The last row corresponds to Theorems 1 and 2, although the actual mean squared errors are computed. It appears that, on the average, our asymptotic calculations are reasonable, and there is

Table 1. *Monte-Carlo results for the model* $y_i = \beta_0 + \beta_1 c_i + \sigma\varepsilon_i$, $\beta_0 = 5$, *and* $\beta_1 = 1$; E_K *and* E_U *denote expectation when* λ *is known and unknown, respectively*

	$\lambda = -1·0$	$\lambda = 0·0$
$\lvert E_U(\hat\beta_0) - \beta_0 \rvert$	0·44	0·60
$\lvert E_U(\hat\beta_1) - \beta_1 \rvert$	0·20	0·26
$[E_U\{(\hat\beta_0 - \beta_0)^2\}/E_K\{(\hat\beta_0 - \beta_0)^2\}]^{\frac{1}{2}}$	12·9	9·6
$[E_U\{(\hat\beta_1 - \beta_1)^2\}/E_K\{(\hat\beta_1 - \beta_1)^2\}]^{\frac{1}{2}}$	4·0	4·0
$\dfrac{E_U[\{f(\hat\lambda, \hat\beta_0) - f(\lambda, \beta_0)\}^2]}{E_K[\{f(\lambda, \hat\beta_0) - f(\lambda, \beta_0)\}^2]}$	—	1·35 1·27* 2·27†
$\dfrac{E_U[\{f(\hat\lambda, \hat\beta_0 - 1·65\hat\beta_1) - f(\lambda, \beta_0 - 1·65\beta_1)\}^2]}{E_K[\{f(\lambda, \hat\beta_0 - 1·65\hat\beta_1) - f(\lambda, \beta_0 - 1·65\beta_1)\}^2]}$	—	1·08 1·01* 2·00†
$\dfrac{E_U[\{f(\hat\lambda, \hat\beta_0 + \hat\beta_1 c_0) - f(\lambda, \beta_0 + \beta_1 c_0)\}^2]}{E_K[\{f(\lambda, \hat\beta_0 + \hat\beta_1 c_0) - f(\lambda, \beta_0 + \beta_1 c_0)\}^2]}$	1·02	1·06

* The value predicted by a likelihood analysis using Σ_0.
† The value predicted by the small σ analysis using Σ_1.
For the last entry, c_0 is randomly chosen from the design.

only a small cost involved in estimating λ for prediction. To read rows 5 and 6, we note that to this point we have defined the cost of estimating λ as an average over the distribution of the new value x_0. It is also of interest to study the costs conditional on a given value of x_0. For Case 1 when $x_0 = (1, c_0)$ and $\lambda = 0$ we find that

$$\frac{\text{MSE}\,(\hat\lambda, x_0)}{\text{MSE}\,(\lambda = 0, x_0)} \to \Upsilon_0(c_0, \beta),$$

while for Case 2 this limit is $\Upsilon_1(c_0, \beta)$, where

$$\Upsilon_j(c_0, \beta) = a\,\Sigma_j\,a^{\mathrm{T}} \quad (j = 1, 2), \quad a = [-\tfrac{1}{2}(\beta_0 + \beta_1 c_0)^2, 1, c_0).$$

Rows 5 and 6 of Table 1 give the ratios of the mean squared errors at two points, the centre and an extreme of the design. As expected from Theorems 1 and 2, there is only a slight cost due to estimating λ, and the small σ asymptotics of Bickel & Doksum are somewhat conservative.

3. Prediction of the conditional mean

The estimator in § 2 is the median of the conditional distribution of y given x_0. Our focus in this section is on estimating the conditional mean of y given x_0.

We sketch a general result which indicates that the cost of extra nuisance parameters, such as λ, is not large. We assume a regression model with (Y_i, X_i) having a joint density $g(y, x \mid \theta_0)$. As in normal theory regression we assume

$$g(y, x \mid \theta_0) = g_1(y \mid x, \theta_0)\, g_2(x).$$

Letting $L_n(\theta)$ denote the log likelihood, we make the usual assumptions:

$$\begin{gathered} E\{L'_n(\theta_0)\} = 0, \\ E\{L'_n(\theta_0)\, L'_n(\theta_0)^{\mathrm{T}}\} = -E\{L''_n(\theta_0)\} = I(\theta_0), \qquad (3·1) \\ n^{\frac{1}{2}}(\theta_n - \theta_0) \to N_q\{0, I^{-1}(\theta_0)\}, \end{gathered}$$

where θ_n is the maximum likelihood estimate, q is the dimension of the parameter θ_0 and the prime denotes differentiation with respect to θ at $\theta = \theta_0$. We are given a new value x_0 and wish to predict $E(Y\,|\,x_0)$; the natural estimate, which usually is only computable numerically, is

$$\hat{E}(Y\,|\,x_0) = \int y g_1(y\,|\,x_0, \theta_n)\, dy.$$

Taylor expansion shows that

$$\begin{aligned} A_n(\theta_0, x_0) &= n^{\frac{1}{2}}\{(\hat{E}(Y\,|\,x_0) - E(Y\,|\,x_0)\} \\ &\simeq \int \{y - E(y\,|\,x_0)\}\left\{\frac{d}{d\theta}\log g_1(y\,|\,x_0, \theta_0)\right\} n^{\frac{1}{2}}(\theta_n - \theta_0)\, g_1(y\,|\,x_0, \theta_0)\, dy \\ &= \int \{y - E(y\,|\,x_0)\}\left[\frac{d}{d\theta}\log g(y, x_0\,|\,\theta_0)\right] n^{\frac{1}{2}}(\theta_n - \theta_0)\, g_1(y\,|\,x_0, \theta_0)\, dy. \qquad (3{\cdot}2) \end{aligned}$$

An overall measure of the accuracy of the prediction is $E\{A_n^2(\theta_0, x_0)\}$; (3·1) and (3·2) and Schwarz's inequality show that for a sample $\mathscr{S}$

$$E\{A_n^2(\theta_0, x_0)\,|\,\mathscr{S}\} \leqslant \operatorname{var}\{y - E(y\,|\,x_0)\}\, n^{\frac{1}{2}}(\theta_n - \theta_0)^{\mathrm{T}} I(\theta_0)\, n^{\frac{1}{2}}(\theta_n - \theta_0).$$

Since $n^{\frac{1}{2}}(\theta_n - \theta_0)^{\mathrm{T}} I(\theta_0)\, n^{\frac{1}{2}}(\theta_n - \theta_0)$ converges in distribution to a chi-squared variable with q degrees of freedom, this suggests that

$$E\{A_n^2(\theta_0, x_0)\} \leqslant q \operatorname{var}\{y - E(y\,|\,x_0)\}. \qquad (3{\cdot}3)$$

Equation (3·3) shows that in prediction with q parameters the average squared prediction error is bounded, and this bound increases in relative magnitude by r/q when r additional nuisance parameters are added. A similar result holds for the two-sample problem.

Example. Consider the transformation model (1·1) but take $\lambda = 1$; this means one uses the Box–Cox model when transformation is unnecessary. If there are p regression parameters, then $q = p + 1$ when $\lambda = 1$ is known and

$$E\{A_n^2(\theta_0, x_0)\} = \operatorname{var}\{y - E(y\,|\,x_0)\}\, p.$$

When one estimates λ, (3·3) shows that

$$E\{A_n^2(\theta_0, x_0)\} \leqslant \operatorname{var}\{y - E(y\,|\,x_0)\}\, (p + 2).$$

Thus, the relative cost of estimating λ is at most $2/p$, which agrees qualitatively with Theorem 2.

We thank Professors Bickel and Doksum for providing a copy of their paper and the referee for his helpful comments.

APPENDIX

Some asymptotics

Suppose that the distribution function F is symmetric. In the theory of Bickel & Doksum (1981), it is assumed that $\sigma = r\eta$ where $r = r(n)$ is a known sequence tending to zero and η is unknown and fixed. Define

$$A = (x_1, \ldots, x_n)^{\mathrm{T}}, \quad P = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}, \quad Q = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}d^{\mathrm{T}}, \quad d = (d_1, \ldots, d_n).$$

$$d_i = \{\lambda^{-2}(v_i - 1) - v_i \log|v_i|\}, \quad v_i = 1 + \lambda x_i \beta, \quad e = dd^{\mathrm{T}} - dPd^{\mathrm{T}}.$$

Assuming that e converges to a positive limit, they prove after very detailed calculations that $n^{\frac{1}{2}}\{(\hat{\lambda}-\lambda)/\sigma, (\beta^*-\beta)/\sigma, (\hat{\eta}-\eta)/\eta\}$ is asymptotically normally distributed with mean zero and covariance

$$\lim_{n\to\infty} e^{-1}\begin{bmatrix} 1 & -Q & 0 \\ -Q^{\mathrm{T}} & (n^{-1}A^{\mathrm{T}}A)^{-1}e+QQ^{\mathrm{T}} & 0 \\ 0 & 0 & \frac{1}{2}e \end{bmatrix}.$$

Hence when λ is estimated one adds to the covariance of β^* the term $\lim(QQ^{\mathrm{T}}e^{-1})$, which is positive-semidefinite and, as the example shows, can often be much larger than the covariance of $\hat{\beta}$ when λ is known. It is this extra term which causes the instability of the regression estimate β^* when λ is estimated.

References

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika* **58**, 249–54.

Atkinson, A. C. (1973). Testing transformations to normality. *J. R. Statist. Soc.* B **35**, 473–9.

Bickel, P. J. & Doksum, K. A. (1981). An analysis of transformations revisited. *J. Am. Statist. Assoc.* **76**, 296–311.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* B **26**, 211–52.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *J. R. Statist. Soc.* B **42**, 71–8.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika* **62**, 101–11.

Huber, P. J. (1977). *Robust Statistical Procedures.* Philadelphia: Society of Industrial and Applied Mathematics.

[*Received February* 1980. *Revised May* 1981]

A Comparison Between Maximum Likelihood and Generalized Least Squares in a Heteroscedastic Linear Model

R.J. CARROLL and DAVID RUPPERT*

We consider a linear model with normally distributed but heteroscedastic errors. When the error variances are functionally related to the regression parameter, one can use either maximum likelihood or generalized least squares to estimate the regression parameter. We show that likelihood is more sensitive to small misspecifications in the functional relationship between the error variances and the regression parameter.

KEY WORDS: Linear models; Heteroscedasticity; Contiguity; Robustness; Weighted least squares; Maximum likelihood.

1. INTRODUCTION

There has been considerable recent interest in the heteroscedastic linear model, which we write as

$$Y_i = x'_i\beta + \epsilon_i[f(x_i, \beta, \theta)]^{-1/2}, \tag{1.1}$$

where $\beta(p \times 1)$ is the regression coefficient, $\{x_i(p \times 1)\}$ are the design vectors, $\{\epsilon_i\}$ are independent and identically distributed with distribution function F, and the function $f(x_i, \beta, \theta)$ expresses the possible heteroscedasticity. Bickel (1978) considers various tests of the hypothesis of homoscedasticity, that is, tests of

$$H_0 : f(x_i, \beta, \theta) \equiv \text{constant}. \tag{1.2}$$

His work has been extended by Carroll and Ruppert (1981), and the tests have been shown to be locally most powerful by Hammerstrom (1981). Other recent papers are Jobson and Fuller (1980), Carroll and Ruppert (1982), Box and Hill (1974), and Fuller and Rao (1978).

Box and Hill (1974), Carroll and Ruppert (1982), and Jobson and Fuller (1980) suggest various forms of generalized weighted least squares estimates (GLSE) of β. Basically, the suggestion is to obtain preliminary estimates $(\hat{\beta}_p, \hat{\theta})$ of (β, θ), estimate variances by $[f(x_i, \hat{\beta}_p, \hat{\theta})]^{-1}$, and then perform ordinary weighted least squares. Carroll and Ruppert (1982) emphasize robustness and develop methods that are robust against outliers and nonnormal distributions F; they prove that generalized M estimates of β, which include GLSE estimates as special cases, are just as good asymptotically as if the weights were really known. The same phenomenon has been found in other models of heteroscedasticity; see Williams (1975) for a review.

Jobson and Fuller (1980) suggest using the information about β in the function f to improve the GLSE. They state that their method is asymptotically equivalent to the MLE for β obtained by setting up the normal likelihood based on (1.1) and maximizing it; this likelihood is

$$\tfrac{1}{2}\sum_{i=1}^{N} \log(f(x_i, \beta, \theta)) - \tfrac{1}{2}\sum_{i=1}^{N} (Y_i - x'_i\beta)^2 f(x_i, \beta, \theta). \tag{1.3}$$

They have a very interesting result that suggests that as long as (1.1) is correct and F is normal the MLE will be preferred to the GLSE.

In this heteroscedasticity problem, we have an additional robustness consideration. Besides the usual goal (Huber 1981) of protecting ourselves against outliers and nonnormal error distributions, we also must protect ourselves against slight misspecifications in the functional relationship between $\operatorname{var}(Y_i)$ and (x_i, β, θ). Since this functional relationship expressed in (1.1) through f is typically at best an approximation, and since our primary interest is estimating β, we would prefer not to estimate β by a statistic that is adversely affected by slight misspecification of f.

In this note, we assume that the error distribution F is actually normal. We study the robustness of GLSE and MLE to small specification errors in f using simple contiguity techniques. We show that small mistakes in specifying f can easily make GLSE preferable to the MLE.

2. A CONTIGUOUS MODEL

We consider small deviations from (1.1) in the form of

$$Y_i = x'_i\beta + [g_N(x_i, \beta, \theta)]^{-1/2}\epsilon_i, \tag{2.1}$$

where for a scalar B and arbitrary unknown function h,

$$g_N(x_i, \beta, \theta) = f(x_i, \beta, \theta)\{1 + 2BN^{-1/2}h(x_i, \beta, \theta)\} \tag{2.2}$$

[*] R.J. Carroll is Associate Professor and David Ruppert is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. R.J. Carroll was supported by the U.S. Air Force Office of Scientific Research Contract AFOSR-80-0080. Part of his research was completed at the Universitat Heidelberg, with support from the Deutsche Forschunggemeinschaft. David Ruppert was supported by NSF Grant MCS78-01240. The authors wish to thank the editor and associate editor for many helpful comments.

© Journal of the American Statistical Association
December 1982, Volume 77, Number 380
Theory and Methods Section

$$N^{-1} \sum_{i=1}^{N} h^2(x_i, \beta, \theta) \to \mu \quad (0 < \mu < \infty)$$

$$\{\epsilon_i\} \quad \text{are iid standard normal.}$$

One should note that the model (2.1) is very close to the assumed model (1.1). Thus the model (2.1) fits our needs because the variance misspecification error is very small and decreases for larger sample sizes. An estimate of β that is robust against specification errors should have the same asymptotic properties under both models (1.1) and (2.1). Thus the question at hand is to study the sensitivity of the MLE and GLSE when (1.1) is assumed but (2.1) is true. If l_1 denotes the log-likelihood for (1.1), and l_2 is the log-likelihood for (2.1), it is quite simple to show that, when (1.1) is true, to order $o_p(1)$,

$$\begin{aligned} l_* &= l_2 - l_1 \\ &\doteq -B^2\mu - \sum_{i=1}^{N} (\epsilon_i^2 - 1)Bh(\underset{\sim}{x}_i, \underset{\sim}{\beta}, \underset{\sim}{\theta})N^{-1/2}, \end{aligned} \tag{2.3}$$

so that by the Central Limit Theorem,

$$\mathcal{L}(l_*) \xrightarrow{\mathcal{L}} N(-B^2\mu, 2B^2\mu) \quad \text{when model (1.1) holds,} \tag{2.4}$$

where $N(a, b)$ is the normal distribution with mean a and variance b. From Corollary 1.2 of Hájek and Sĭdák (1967, p. 204), this means that model (2.1) is contiguous to model (1.1).

3. LIMIT DISTRIBUTIONS FOR GLSE

Suppose that for some positive definite matrix S,

$$N^{-1} \sum_{i=1}^{N} x'_i x_i f(x_i, \beta, \theta) \to S. \tag{3.1}$$

Then, assuming normal errors and smoothness conditions on f, Carroll and Ruppert (1982) (as well as Jobson and Fuller 1980) show that when model (1.1) is true, the GLSE $\hat{\beta}_G$ satisfies

$$N^{1/2}(\beta_G - \beta) - N^{-1/2} \sum_{i=1}^{N} S^{-1} x'_i f^{1/2}(x_i, \beta, \theta)\epsilon_i \xrightarrow{p} 9, \tag{3.2}$$

$$N^{1/2}(\hat{\beta}_G - \beta) \xrightarrow{\mathcal{L}} N(0, S^{-1}). \tag{3.3}$$

A formal proof is possible as long as f is smooth, $\{f(x_i, \beta, \theta)\}$ is bounded away from ∞ uniformly in i, and $(\hat{\beta}_p, \hat{\theta})$ satisfy

$$N^{1/2}(\hat{\beta}_p - \beta) = 0_p(1)$$

and

$$N^{1/2}(\hat{\theta} - \theta) = 0_p(1). \tag{3.4}$$

Carroll and Ruppert (1982) and Jobson and Fuller (1980) verify (3.4) in the normal case under certain technical conditions.

Now, since $\{\epsilon_i\}$ are normal random variables, one uses (2.3) and (3.2) to show that $l_* = l_2 - l_1$ and $N^{1/2}(\hat{\beta}_G - \beta)$ are asymptotically independent, so that by LeCam's third lemma (Hájek and Sĭdák 1967, p. 208),

$$\mathcal{L}(N^{1/2}(\hat{\beta}_G - \beta)) \to N(0, S^{-1}), \tag{3.5}$$

and this *under either model* (1.1) *or* (2.1). This means that GLSE is robust against small specification errors of the variance function f. This encouraging result suggests that one will not go too wrong with GLSE as long as model (1.1) is reasonable. These results are easily extended to the robust estimates introduced by Carroll and Ruppert (1982).

4. LIMIT DISTRIBUTION FOR THE MLE

While GLSE is robust against minor errors in specifying the function f in model (1.1), the same cannot be said for the MLE. Denote this MLE by $\hat{\beta}_M$. Jobson and Fuller (1980) show that for a particular covariance matrix Σ, if the MLE is computed assuming (1.1), then under (1.1),

$$N^{1/2}(\hat{\beta}_M - \beta) \xrightarrow{\mathcal{L}} N(0, \Sigma). \tag{4.1}$$

The result of particular interest is that Σ is no larger than S^{-1} (see 3.1) and (3.3) in the sense that $S^{-1} - \Sigma$ is positive semi-definite under the model (1.1). In addition to (4.1), from (2.3) and the proof of Theorem 2 in Jobson and Fuller (1980), $N^{1/2}(\hat{\beta}_M - \beta)$ and l_* are jointly asymptotically normal with mean $(0, -B^2\mu)$, marginal variances $(\Sigma, 2B^2\mu)$, and covariances Bq computed below, that is,

$$(N^{1/2}(\hat{\beta}_M - \beta)', l_*) \xrightarrow{\mathcal{L}} N\left((0, -B^2\mu), \begin{bmatrix} \Sigma & Bq \\ q'B' & 2B^2\mu \end{bmatrix}\right). \tag{4.2}$$

We now indicate why it is true that the only cases in which the MLE can be expected to be robust against variance specification errors is when $S^{-1} = \Sigma$ and the MLE is asymptotically equivalent to GLSE. To see this, first consider model (1.1) to hold. Jobson and Fuller (1980) show that $\hat{\beta}_M$ is essentially a linear function of $\{\epsilon_i\}$ and $\{\epsilon_i^2 - 1\}$, that is, for vectors $\{v_i\}$ and $\{w_i\}$,

$$N^{1/2}(\hat{\beta}_M - \beta) = N^{-1/2} \sum_{i=1}^{N} \{v_i\epsilon_i + w_i(\epsilon_i^2 - 1)\} + o_p(1). \tag{4.3}$$

If we have $w_i \equiv 0$ $(i = 1, \ldots, N)$, then from (3.2), (4.3), and Gauss-Markov, we have that

$$N^{1/2}(\hat{\beta}_m - \hat{\beta}_G) \xrightarrow{P} 0,$$

and the estimates have the same limit distribution. Thus the only way for $\hat{\beta}_M$ to improve on $\hat{\beta}_G$ under (1.1) is for the $\{w_i\}$ to be nonzero. In this case, however, we can perform contiguity calculations based on (2.3) and (4.3), thus showing that under model (2.1), for the MLE computed assuming (1.1),

$$N^{1/2}(\hat{\beta}_M - \beta) \xrightarrow{\mathscr{L}} N(-2Bq, \Sigma), \tag{4.4}$$

$$q = \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} w_i h(x_i, \beta, \theta).$$

Of course, q will be nonzero in general if the $\{w_i\}$ are.

These results have important consequences for efficiency. Suppose we wish to estimate the linear combination $\alpha'\beta$. Then, under model (1.1),

$$N\,\mathrm{MSE}\,(\alpha'\hat{\beta}_G) \to \alpha' S^{-1}\alpha$$

$$N\,\mathrm{MSE}\,(\alpha'\hat{\beta}_M) \to \alpha\,\Sigma\,\alpha \le \alpha' S^{-1}\alpha. \tag{4.5}$$

However, under the model (2.1), when the GLSE and MLE are computed assuming (1.1),

$$N\,\mathrm{MSE}\,(\alpha'\hat{\beta}_G) \to \alpha'\,S^{-1}\alpha \quad \text{(no change)}$$

$$N\,\mathrm{MSE}\,(\alpha'\hat{\beta}_M) \to \alpha'\Sigma\,\alpha + 4B^2(\alpha' q)^2, \tag{4.6}$$

and of course $\alpha'\hat{\beta}_M$ will be a rather poor estimate if α is not orthogonal to q and B is large.

5. MONTE CARLO SPECIFICATIONS

We performed a small Monte Carlo study to illustrate the results given in the previous section, as well as to determine the effect of nonnormality; these are the two aspects of robustness discussed in this note, distributional robustness in heteroscedastic models as well as robustness against misspecification of the form of the variance function. All of the results are based on the following model ($\sigma_i = [f(x_i, \beta, \theta)]^{-1/2}$ in the previous notation):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \sigma_i \epsilon_i \quad (i = 1, \ldots, N).$$

Here $N = 40$ and the design $\{(x_{i1}, x_{i2})\}$ is as given in a similar experiment performed by Jobson and Fuller (1980). What varies in our experiments is the form of $\{\sigma_i\}$ and the distribution of the errors $\{\epsilon_i\}$. However, all weighted estimates were computed assuming the following model for variances:

$$\sigma_i^2 = \alpha_1 + \alpha_2 \tau_i^2, \tau_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}. \tag{5.1}$$

In context, (5.1) acts like (1.1) of the text. In all the experiments, we took $(\beta_0, \beta_1, \beta_2) = (10, -4, 2)$, as is done by Jobson and Fuller (1980). Normal random numbers were generated by the IMSL routine GGNPM. Contaminated normal random numbers were generated by first finding a normal deviate Z, and then multiplying Z by 3.0 if a uniform (0, 1) random number generated by IMSL's GGUBS exceeded .90. The starting seed was 325017, and the experiments were repeated 800 times.

The estimators we used included first of all the ordinary least squares estimate (LSE). We also attempted to study the estimator JLS of Jobson and Fuller (1980), which is a one-step version of the MLE; see their paper for details. Their estimate worked well at the normal error model and for their choice $(\alpha_1, \alpha_2) = (300.0, .2)$, but it was very bad at nonnormal distributions or even when the heteroscedasticity was severe. Consequently, the estimator JLS* studied here is a modification of Jobson and Fuller's. Basically, JLS* is JLS if both $\hat{\alpha}_{1\mathrm{JLS}} \ge 0$ and $\hat{\alpha}_{2\mathrm{JLS}} \ge 0$, where $\hat{\alpha}_{1\mathrm{JLS}}, \hat{\alpha}_{2\mathrm{JLS}})$ are the estimates of (α_1, α_2) using JLS. However, if either $\hat{\alpha}_{1\mathrm{JLS}} < 0$ or $\hat{\alpha}_{2\mathrm{JLS}} < 0$, we estimated (α_1, α_2) as in Equation (5.1) of Jobson and Fuller. The modified estimator JLS* appeared to us to be very much better than JLS in overall performance.

We also defined a GLSE called GLSE and a weighted robust estimate ROBUST WEIGHTED (Carroll and Ruppert 1982). In extensive trial and error work, we found that in small samples, the choice of method of estimating the weights has a very big effect on GLSE, although asymptotically there is no effect as long as consistent estimates are available; ROBUST WEIGHTED seems almost insensitive to the choice of weighting method even in small samples. Details will be reported in a future paper. We finally settled on the following somewhat complicated method.

First, for any function Ψ, define

$$\xi(\Psi) = (2\pi)^{-1/2} \int \Psi^2(v) \exp(-v^2/2)dv.$$

In general, Huber's Proposal 2 simultaneously solves

$$\Sigma\Psi((Y_i - X_i'\beta)/\sigma)\{X_i/\sigma\} = 0$$

and

$$\Sigma\Psi^2((Y_i - X_i'\beta)/\sigma) = (N - p)\,\xi\,(\Psi), \tag{5.4}$$

where p = dimension of β. Now define

$$\Psi_k(x) = \min\,(k, |\,x\,|\,) \quad \mathrm{sign} \quad (x).$$

The LSE solves (5.4) using $k = \infty$. A general algorithm for defining weighted estimates is based on k. Essentially, what we do is estimate α_2 robustly and α_1 consistently. The estimates of α_2 will also be consistent, although robust estimates of α_1 are apparently not feasible (see Carroll 1979, Sec. 3 for theoretical details). Estimating α_2 by any of the standard methods is not robust and results in poor overall performance of GLSE. For any given k, the algorithm we used is as follows.

1. Let $\hat{\beta}$ solve (5.4) using Ψ_2.
2. Define $\dot{r}_i = (Y_i - X_i'\hat{\beta})^2$, and P as in Jobson and Fuller (1980).
3. Form predicted values $t_i = x'_i\beta$.
4. Define H as an $(N \times 2)$ matrix, the first column of which consists of ones, the second the t_i^2.
5. Solve (5.4) for the regression model

$$E\dot{r} = PH\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$$

using Ψ_2 (this is much like (5.1) of Jobson and Fuller 1980). Define $Z_i = \dot{r}_i - \max\,(\hat{\alpha}_2, 0)t_i^2$.

6. Define $\hat{\alpha}_1 = N^{-1} \sum Z_i$.
7. Compute $\hat{\sigma}_i^2 = \max\,(\hat{\alpha}_1, 0) + \max\,(\hat{\alpha}_2, 0)\,t_i^2$.
8. Solve (5.4) using Ψ_2 with Y_i and X_i replaced by $Y_i/\hat{\sigma}_i$ and $X_i/\hat{\sigma}_i$. Call this estimate $\hat{\beta}$.

9. Repeat (2)–(8), except in Step (8) use Ψ_k instead of Ψ_2.

Our estimate GLSE uses $k = \infty$, while ROBUST WEIGHTED uses $k = 2$. The estimate HUBER was the Huber Proposal 2 computed in Step 1.

Finally, the mean squared error (MSE) of an estimator, as well as the standard error of this MSE, were calculated by the following simple device. Denote by WLS the weighted least squares estimate based on the weights σ_i^{-2}. This is not a real statistic since the weights are unknown in practice. Then, for JLS* as an example,

$$\begin{aligned}\text{MSE (JLS*)} &= E\{\hat{\beta}(\text{JLS*}) - \beta\}^2 \\ &= E[\{\hat{\beta}(\text{JLS*}) - \beta\}^2 - \{\hat{\beta}(\text{WLS}) - \beta\}^2] \\ &\quad + \text{MSE (WLS)}. \qquad (5.5)\end{aligned}$$

The second term on the right side of (5.5) is known exactly; the first term and its standard error are calculated by the Monte Carlo experiment. Because of the correlation between JLS* and WLS, this method produces better estimates of MSE(JLS*) than would the usual direct Monte Carlo calculation.

6. MONTE CARLO RESULTS

The first part of the study concerns the effect of nonnormality on the estimates and is reported in Table 1. In constructing this table, the assumed model (5.1) was actually true, with $(\alpha_1, \alpha_2) = (300, .2)$ as in Jobson and Fuller's work. For each estimator, the first line is the ratio of its MSE with that of WLS (the weighted estimator with known weights). Note that the Carroll-Ruppert ROBUST WEIGHTED is the best; it is quite competitive at the normal model and the clear winner at the contaminated normal model; this is in agreement with theory. Note too that, qualitatively at least, JLS* suffers the worst in the switch from normal to contaminated normal.

Table 1. Distributional Robustness When Model (5.1) Is Assumed and Is True, α_1 = 300 and α_2 = .20

	Standard Normal Errors			*Contaminated Normal Errors*		
	β_0	β_1	β_2	β_0	β_1	β_2
LSE	1.32	1.27	1.27	1.33	1.26	1.24
	.05	.04	.04	.06	.05	.05
	−.11	.03	−.01	.14	.00	.01
JLS*	1.18	1.17	1.12	1.25	1.21	1.21
	.04	.04	.04	.06	.06	.10
	−.35	.05	−.03	−.26	.05	−.02
GLSE	1.16	1.18	1.16	1.16	1.14	1.11
	.03	.04	.03	.05	.04	.04
	.16	.02	−.02	.39	.00	.00
Huber	1.27	1.27	1.27	0.96	0.94	.099
	.04	.04	.04	.04	.04	.05
	.07	.01	−.01	.32	−.01	.00
Robust Weighted	1.16	1.18	1.16	0.88	0.89	0.92
	.03	.03	.03	.04	.04	.05
	.23	.01	−.02	−.02	−.01	−.01
Actual MSE of WLS	196.9	1.08	.57	354.4	1.94	1.02

NOTE: The first row is the MSE ratio (MSE of indicated estimator/MSE of WLS), the second its standard error, and the third is the observed Monte Carlo bias.

The benefit of using our modification JLS* to Jobson and Fuller's JLS is dramatic here. Ordered as in Table 1, the MSE ratio values for JLS are 1.22, 1.27, 1.25, 2.72, 7.27, and 13.27.

Table 2 is designed to cover the problem of specification robustness discussed theoretically in Sections 1–4. Designing a Monte Carlo experiment that illustrated the theory was quite difficult because the theory is a local theory. We finally used heteroscedastic models that had fairly large inequalities in variances. The assumed model was (5.1), but with $\alpha_1 = 100$, $\alpha_2 = .20$. For the left side of Table 2, we do calculations when (5.1) is in fact true, the errors are normally distributed, and $\alpha_1 = 100$, $\alpha_2 = .20$. In the right side of Table 2 the model corresponding to (2.1) and (2.2) has

$$\sigma_i^2 = \alpha_1 \exp(2\alpha_2^2 \,|\, \tau_i \,|),$$

$$\tau_i = EY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2},$$

$$\alpha_1 = 100,$$

and

$$\alpha_2 = .128. \qquad (6.1)$$

The choice of $\alpha_2 = .128$ in Table 2 reflects a model whose variance behavior is close to that of (5.1) with $\alpha_1 = 100$, $\alpha_2 = .20$; the ratio of (6.1) to (5.1) over the range of the mean value is between .95 and 1.15. Further, the

Table 2. Specification Robustness When (5.1) Is Assumed. Small Specification Error

	Model (5.1) Is True *α_1 = 100, α_2 = .20* *(correct model)*			*Model (6.1) Is True* *α_1 = 100, α_2 = .128* *(misspecified model)*		
	β_0	β_1	β_2	β_0	β_1	β_2
LSE	1.62	1.59	1.60	1.63	1.63	1.58
	.07	.06	.06	.07	.07	.06
	−.04	.02	−.01	−.04	.02	−.01
JLS*	1.54	1.51	1.39	1.62	1.62	1.54
	.12	.12	.09	.15	.16	.16
	−.19	.04	−.03	−.22	.04	−.03
GLSE	1.27	1.29	1.28	1.28	1.31	1.27
	.05	.05	.05	.05	.05	.05
	.07	.02	−.02	.09	.02	−.02
Huber	1.62	1.59	1.60	1.63	1.63	1.58
	.07	.06	.06	.07	.07	.06
	−.04	.02	−.01	−.04	.02	−.01
Robust Weighted	1.27	1.29	1.28	1.27	1.29	1.26
	.05	.04	.04	.05	.05	.04
	.18	.01	−.02	.21	.01	−.01
Actual MSE of WLS	116.1	.72	.35	120.2	.74	.35

NOTE: The first row is the MSE ratio (MSE of indicated estimator/MSE of WLS), the second its standard error, and the third is the observed Monte Carlo bias.

difference in actual MSE for WLS (with known weights) is negligible, as can be seen in the last row of Table 2. As predicted by our theory, the only estimate that seems at all affected by the error in specifying the variances (assuming (5.1) when (6.1) is true) is JLS*, the MLE approximation. Note, too, how well our estimate ROBUST WEIGHTED performs; it is quite good at the normal error model and, as seen in Table 1, is superior for the contaminated normal model.

We also analyzed the case for (6.1) that $\alpha_2 = .131$, which illustrates a specification error resulting in overweighting the points with largest variance. This hardly affected WLS. Once again, the worst overall performance was turned in by JLS*; this MLE approximation was the weighted method most affected by the specification error.

One might ask how well our theory predicts the specific numbers in Table 2. As this section shows, the theory is at least qualitatively correct in predicting that the MLE approximation JLS* would be most sensitive to variance function misspecification, while GLSE and ROBUST WEIGHTED would be only slightly affected (see (4.6)). On the other hand, the result (4.5) that indicates that JLS* should be better than GLSE when the variances are correctly specified was not borne out. Another instance, which is really not too bad, in which asymptotic theory and Monte Carlo theory do not closely agree is that the GLSE had about 30 percent higher mean squared error than WLS. Because the approximation given by (4.5) is not close to the Monte Carlo results, evaluating the excess mean squared error $4B^2(\alpha' q)^2$ in (4.6) due to variance misspecification in unlikely to be too accurate. If

$$\sigma_i^2 = 1/f(x_i, \beta, \theta)$$

and

$$\delta_i^2 = 1/g_N(x_i, \beta, \theta),$$

then from (2.2) we have approximately

$$2Bh(x_i, \beta, \theta) \doteq N^{1/2}(\sigma_i^2/\delta_i^2 - 1). \qquad (6.2)$$

Using the theory of Jobson and Fuller, we can evaluate the terms $\{w_i\}$ of (4.3), which then enables us from (6.2) to approximate $2Bq$ of (4.4). When this is done, we are able to predict that in going from the correct model to the misspecified model in Table 2, the MSE for JLS* should increase by (4.7 percent, 5.8 percent, 4.7 percent) for estimating $(\beta_0, \beta_1, \beta_2)$. The actual Monte Carlo increases were (8.9 percent, 10.3 percent, 10.8 percent). In other words, in this example, the effect of variance function misspecification on the MLE approximation JLS* was more than that predicted by the asymptotic theory.

7. DISCUSSION

The theoretical work and the small Monte Carlo study presented here indicate that the maximum likelihood estimate (or approximations to it) in a heteroscedastic model is sensitive both to the normal error assumption and to small errors in specifying a functional form for the variances. Generalized least squares estimates are sensitive to the normal error assumption but, at least theoretically, are robust against small variance specification errors; a particular GLSE was constructed that, in a limited Monte Carlo study, had these properties in small samples. The robust weighted estimators of Carroll and Ruppert (1982) had the best theoretical and empirical robustness behavior, while at the same time giving up only very little when all assumptions about the variances and error distributions are true. For homoscedastic regression models, estimators with bounded influence functions have been defined and studied (Krasker and Welsch 1982). We did not consider the question but believe it is possible to develop bounded influence weighted estimators with appealing properties for heteroscedastic situations.

[*Received April 1981. Revised June 1982.*]

REFERENCES

BICKEL, P.J. (1978), "Using Residuals Robustly I: Tests for Heteroscedasticity, Nonlinearity," *Annals of Statistics*, 6, 266–291.

BOX, G.E.P., and HILL, W.J. (1974), "Correcting Inhomogeneity of Variances With Power Transformation Weighting," *Technometrics*, 16, 385–389.

CARROLL, R.J. (1979), "Estimating Variances of Robust Estimators When the Errors are Asymmetric," *Journal of the American Statistical Association*, 74, 674–679.

CARROLL, R.J., and RUPPERT, D. (1981), "On Robust Tests for Heteroscedasticity," *Annals of Statistics*, 9, 205–209.

——— (1982), "Robust Estimation in Heteroscedastic Linear Models," *Annals of Statistics*, 10, 429–441.

FULLER, W.A., and RAO, J.N.K. (1978), "Estimation for a Linear Regression Model With Unknown Diagonal Covariance Matrix," *Annals of Statistics*, 6, 1149–1158.

HÁJEK, J., and SĬDÁK, Z. (1967), *Theory of Rank Tests*, New York: Academic Press.

HAMMERSTROM, T. (1981), "Asymptotically Optimal Tests for Heteroscedasticity in the General Linear Model," *Annals of Statistics*, 9, 368–380.

HUBER, P.J. (1981), *Robust Statistics*, New York: John Wiley.

JOBSON, J.D., and FULLER, W.A. (1980), "Least Squares Estimation When the Covariance Matrix and Parameter Vector are Functionally Related," *Journal of the American Statistical Association*, 75, 176–181.

KRASKER, W.S., and WELSCH, R.E. (1982), "Efficient Bounded Influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595–604.

WILLIAMS, J.S. (1975), "Lower Bounds on Convergence Rates of Weighted Least Squares to Best Linear Unbiased Estimators," in *A Survey of Statistical Design and Linear Models*, ed. J.N. Srivastava, Amsterdam: North-Holland.

Power Transformations When Fitting Theoretical Models to Data

RAYMOND J. CARROLL and DAVID RUPPERT*

We investigate power transformations in nonlinear regression problems when there is a physical model for the response but little understanding of the underlying error structure. In such circumstances, and unlike the ordinary power transformation model, both the response and the model must be transformed simultaneously and in the same way. We show by an asymptotic theory and a small Monte Carlo study that for estimating the model parameters there is little cost for not knowing the correct transform a priori; this is in dramatic contrast to the results for the usual case where only the response is transformed. Possible applications of the theory are illustrated by examples.

KEY WORDS: Transformations; Box-Cox models; Theoretical models; Robustness; Nonlinear regression.

1. INTRODUCTION

Often in scientific work, an experimenter observes data y_i and $x_i' = (x_{1i} \cdots x_{pi})$ and postulates that these data follow a model

$$y_i = f(x_i, \theta_0), \quad i = 1, \ldots, N, \tag{1.1}$$

where θ_0 is a k-parameter vector. The function f may be derived, for example, from differential equations believed to govern the physical system that gave rise to the data. The deterministic model (1.1) is often inadequate since the data exhibit random variation, but whereas f was derived from theoretical considerations, there is really no firm understanding of the mechanism producing the randomness. In this case, the experimenter usually assumes that

$$y_i = f(x_i, \theta_0) + \epsilon_i, \tag{1.2}$$

where the $\{\epsilon_i\}$ are iid $N(0, \sigma_0^2)$. In those cases in which the data suggest that model (1.2) is also unsatisfactory, one might then, for example, assume that the errors are multiplicative and lognormal, so that

$$\log(y_i) = \log(f(x_i, \theta_0)) + \epsilon_i. \tag{1.3}$$

The point here is that model (1.1) is equivalent to the model

$$h(y_i) = h(f(x_i, \theta_0))$$

whenever $h(\cdot)$ is a monotonic transformation. Therefore (1.2) and (1.3) are based on the same theoretical model, but they allow variability to enter into the model in different fashions.

A more flexible approach is to take a sufficiently rich family of strictly monotonic transformations $h(y, \lambda)$, indexed by the m-vector parameter λ, and to assume that for some value λ_0,

$$h(y_i, \lambda_0) = h(f(x_i, \theta_0), \lambda_0) + \epsilon_i. \tag{1.4a}$$

Equation (1.1) could be understood to mean $Ey = f$ or $y = f$ when there is no error. We have in mind the latter meaning; the former is not possible under (1.4a). The model (1.4a) is in the spirit of Box and Cox (1964), who suggested the family of power transformations with $m = 1$ and

$$\begin{aligned} h(y, \lambda) = y^{(\lambda)} &= (y^\lambda - 1)/\lambda \quad \text{if } \lambda \neq 0 \\ &= \log(y) \qquad\quad \text{if } \lambda = 0. \end{aligned} \tag{1.4b}$$

However, as we will make clear, our proposed model (1.4) has greatly different ramifications than those usually associated with the power family. Box and Cox (1964) used their family in a study of the transformation model

$$h(y, \lambda_0) = x'\theta_0 + \epsilon. \tag{1.5}$$

Notice that here, unlike in (1.4), the regression function in (1.5) is *not* transformed. Box and Cox sought a transformation that achieves (a) a simple additive or linear model, (b) homoscedastic errors, and (c) normally distributed errors. Our model is different. Theoretical considerations already provide a regression function. We hope to transform the response *and* the regression function simultaneously to obtain homoscedasticity and normality.

There are two reasons for using model (1.4) instead of simply fitting (1.1) by least squares or some other method. First, estimation of θ_0 based on model (1.4) should be more efficient than other methods. Second, it may be necessary to estimate the entire conditional distribution of y given x; if the data clearly suggest that the distri-

[*] Raymond J. Carroll is Professor of Statistics and David Ruppert is Associate Professor of Statistics at the University of North Carolina, Chapel Hill, NC 27514. Research for this article was supported by the Air Force Office of Scientific Research Grant F49620-82-C-0009 and by National Science Foundation Grant MCS 8100748. Rod Reish kindly provided the authors with the menhaden data. He and Rick Deriso greatly aided our understanding of those data. The authors also thank a referee and an editor for their comments on an earlier version of this paper.

© Journal of the American Statistical Association
June 1984, Volume 79, Number 386
Theory and Methods Section

butions of $\{y_i - f(x_i, \theta_0)\}$ are not constant across i, one must go beyond standard regression methodology.

An example that motivated the research of this article is the relationship between egg production in a fish stock and subsequent recruitment into the stock. At least for some species, as egg production increases, the changes in the skewness and variance of recruitment are as large as the change in the median recruitment, and these changes in distributional shape may have important implications for management of the fishery. This example is discussed in more detail in Section 4.1.

Another possible reason for transformation is that often, for an appropriate h, $h(f(x_i, \theta))$ is a linear function of θ. Linearization was an accepted technique before the advent of nonlinear regression programs. Now, however, the statistician must decide whether to use linearization or nonlinear regression. As discussed later, our theory provides a method for deciding whether linearization is appropriate.

A natural question is, Which aspects of the data enable us to estimate λ_0? If we transform y_i by $h(\cdot, \lambda)$ and $\lambda \neq \lambda_0$, then information that $\lambda \neq \lambda_0$ is provided by both (a) nonnormality and (b) nonconstancy in i of the distribution of $h(y_i, \lambda) - h(f(x_i, \theta_0), \lambda)$. If the values of $f(x_i, \theta_0)$ are relatively constant, then (a) provides most of the information. On the other hand, if $\sigma^2 = \text{var}(\epsilon_i)$ is small, then most of the information is provided by heteroscedasticity. To see this last fact, suppose, for example, that (1.4b) holds and that we do not transform the data (i.e., we use $\lambda = 1$), but that the true value λ_0 is not 1. For each λ, let $g(\cdot, \lambda)$ be the inverse of the function $h(\cdot, \lambda)$, and define $g_y(y, \lambda) = (\partial/\partial y)\, g(y, \lambda)$. Then by (1.4) and a Taylor approximation, which is suitable if ϵ_i is small, we have

$$y_i = g[h(f(x_i, \theta_0), \lambda_0) + \epsilon_i, \lambda_0]$$
$$\approx f(x_i, \theta_0) + k_i\epsilon_i,$$

where $k_i = g_y[h(f(x_i, \theta_0), \lambda_0), \lambda_0]$; therefore y_i is approximately normally distributed with mean $f(x_i, \theta_0)$ and variance $k_i^2\sigma^2$.

When analyzing data, after we have determined estimates for θ, λ, and σ, we can estimate the density of y_i (or of $[y_i - f(x_i, \theta)]$, the residual from the median). By plotting this estimated density we can check for skewness and other signs of nonnormality on the original scale. By overlaying plots for several values of x_i we can also check for heterogeneity of the distribution of the untransformed data. Instead of graphing densities, we might graph quantiles against quantiles of the normal distribution; nonnormality would then be especially easy to detect. We use such a quantile-quantile plot in Example 4.1.

When we make inferences about θ, the issue arises whether $\hat{\lambda}$ should be treated as fixed or whether we should acknowledge that it is random. For example, there are at least two approaches to estimating the variance-covariance matrix of $\hat{\theta}$. The first is invert the estimated Fisher information matrix for $(\lambda, \sigma, \theta)$. The second is to transform the model and the response by $h(\cdot, \hat{\lambda})$ and then use standard nonlinear regression methodology. The second method is not strictly correct since it treats λ as known rather than estimated. However, it is convenient and expedient since existing nonlinear least squares software can be applied. In this article we report large-sample analysis and Monte Carlo results showing that the two methods tend to give similar results. The second method usually underestimates the variability of $\hat{\theta}$, but it does give a rough approximation to this variability. In the different model (1.5) of Box and Cox (1964), the two methods can give drastically different results, and this fact has led to considerable controversy; see Bickel and Doksum (1981), Carroll and Ruppert (1981), Hinkley and Runger (1984), and Box and Cox (1982).

Another major difference between our model and that of Box and Cox (1964) is that in our model the parameter θ has physical meaning even when λ_0 is unknown; $f(x_i, \theta_0)$ is the median of y_i regardless of the value of λ_0.

2. THEORETICAL ANALYSIS

To analyze the effect of treating $\hat{\lambda}$ as fixed (and equal to λ_0), we begin by computing the information matrices for $(\lambda_0, \theta_0, \sigma_0)$ and (θ_0, σ_0), the latter case assuming that λ_0 is known. The details quickly become intractable, so we resort to the approximation $\sigma_0 \approx 0$. The following theorems are proved in Appendix A.

Theorem 1. Under general conditions, if $N \to \infty$ and then $\sigma_0 \to 0$, the limit distribution of $\hat{\theta}$ is the same whether λ_0 is known or unknown. The limit distribution of $\hat{\sigma}$ depends on whether λ_0 is known or unknown.

Theorem 1 says that the effect of having to transform the problem to get homoscedastic, normal errors is small when σ_0 is small. However, we are not concerned only, or even primarily, with small σ_0. In fact, the need for transformation will probably be greater when σ_0 is large. When σ_0 is small, $\hat{\theta}$ from the untransformed data, $\hat{\theta}_{\lambda=1}$, will have a small bias because y_i will be approximately normally distributed. Moreover, although $\hat{\theta}_{\lambda=1}$ may be inefficient in terms of variance, there may be less need for an efficient estimate if σ_0 is small. The small σ_0 asymptotics do, however, lead to major simplifications, and the Monte Carlo results presented later agree with them.

Because we are interested in all values of σ_0, we looked at a second approach. This approach is outlined in Appendix A. Basically, we construct a third estimator of θ_0 and compute its efficiency with respect to $\hat{\theta}(\lambda_0)$, the estimator of θ_0 when λ_0 is known. This gives us a bound on the efficiency of the MLE.

Theorem 2. For any λ_0, σ_0, θ_0, f, or design $\{x_i\}$, as $N \to \infty$, the asymptotic relative efficiency of the MLE $\hat{\theta}(\hat{\lambda})$ compared to that estimate $\hat{\theta}(\hat{\lambda}_0)$ with λ_0 known is at least $2/\pi$, that is,

$$\text{ARE}(\hat{\theta}(\hat{\lambda}), \hat{\theta}(\lambda_0)) \geq 2/\pi.$$

This bound is very general, and if the Monte Carlo simulation in Section 3 is any guide, the bound is conservative. It follows that the practice of transforming and then using a standard errors for $\hat{\theta}(\hat{\lambda})$ the estimates output from a nonlinear least squares package will be only moderately in error.

3. MONTE CARLO

To study $\hat{\theta}$ when N is finite and σ_0 is not necessarily small, we undertook a small simulation of the model

$$h(y_i, \lambda_0) = h(\theta_1 + \theta_2 x_i, \lambda_0) + \sigma_0 \epsilon_i, \qquad (3.1)$$

where $h(\cdot)$ is the Box and Cox (1964) power family (1.4b). In our simulations, $N = 50$, the design points $\{x_i\}$ were equally spaced on $[-1, 1]$, the errors were normally distributed with mean zero and variance one, and $\theta_1 = 7$, $\theta_2 = 2$. We considered three estimators: (a) ML estimator, λ_0 known (KNOWN), (b) ML estimator, λ_0 unknown (MLE), and (c) The ordinary least squares estimator (LSE) without any transformation.

Since it is a rather frequent practice to use least squares estimation without transformation, we included the LSE in the study. The method of computation is outlined in Appendix B. We chose three values of σ_0: $\sigma_0 = .05, .10$, and .50. We present results in Tables 1 and 2 for $\lambda_0 = 0$ (lognormal data) and $\lambda_0 = .25$. There were 600 replications of the experiment for each (λ_0, σ_0) and each estimator, all generated from a common set of random numbers. The normal random deviates were generated from the IMSL routine GGNPM. Estimation of (θ_1, θ_2) for each λ was done by the IMSL routine ZXSSQ while ZXGSN was used to estimate λ_0.

The results for the ML estimator with λ_0 unknown (denoted by MLE) are very encouraging. The mean squared errors for MLE are reasonably close to those for KNOWN, the ML estimator with λ_0 known, especially for the slope θ_2. These results agree with our small σ theory and indicate the moderate cost of not knowing λ_0. The relative efficiencies of MLE to KNOWN are always well above the lower bound of $2/\pi$. To appreciate how well MLE does compared with KNOWN (line 2 of Tables 1 and 2), see Table 5 of Bickel and Doksum (1981); in their model, which we call (1.5), they have ratios MLE(λ_0 estimated)/KNOWN(λ_0 known) always at least 1.5 and as large as 211, while ours never exceed 1.2.

Table 1. Results of the Monte Carlo Study Described in the Text. (These results are for the INTERCEPT. The median response is linear with intercept = 7 and slope = 2.)

λ =	.00			.25		
σ =	.05	.10	.50	.05	.10	.50
Bias of KNOWN	.03	.06	.56	.01	.03	.23
MSE of KNOWN	2.41	9.67	24.87	.90	3.59	9.04
Bias of MLE	.02	.04	.60	.01	.02	.19
$\frac{\text{MSE of MLE}}{\text{MSE of KNOWN}}$	1.02	1.05	1.14	1.01	1.03	1.12
MSE of MLE − MSE of KNOWN	.05	.47	3.44	.01	.09	1.09
STD. ERROR of above difference	.02	.15	.77	.01	.04	.25
Bias of LSE	.11	.40	9.48	.04	.13	2.60
$\frac{\text{MSE of MLE}}{\text{MSE of LSE}}$	.97	.90	.22	1.00	.98	.63
MSE of MLE − MSE of LSE	−.06	−1.15	−96.62	.00	−.06	−6.07
STD. ERROR of above difference	.04	.33	4.71	.01	.06	.78

NOTE: Known = ML estimate with λ known, MLE = ML estimate with λ unknown, and LSE = ordinary least squares estimate. In these calculations, the mean squared error (MSE) and STD. ERROR of difference terms are multiplied by T**2. Here T = 10 if σ ≤ .10 and T = 1 if σ = .50.

Table 2. Results of the Monte Carlo Study Described in the Text. (These results are for the SLOPE. The median response is linear with intercept = 7 and slope = 2.)

λ =	.00			.25		
σ =	.05	.10	.50	.05	.10	.50
Bias of KNOWN	.01	.01	.03	.00	.01	.02
MSE of KNOWN	7.08	28.36	72.23	2.71	10.83	27.24
Bias of MLE	−.01	−.04	−.15	.00	−.02	−.16
$\frac{\text{MSE of MLE}}{\text{MSE of KNOWN}}$	1.06	1.06	1.01	1.06	1.06	1.03
MSE of MLE − MSE of KNOWN	.41	1.57	.95	.15	.60	.72
STD. ERROR of difference	.10	.40	.67	.04	.77	.27
Bias of LSE	.05	.15	2.97	.02	.04	.50
$\frac{\text{MSE of MLE}}{\text{MSE of LSE}}$	.98		.59	1.01	1.01	.91
MSE of MLE − MSE of LSE	−.16	−1.29	−50.54	.05	.13	−2.81
STD. ERROR of above difference	.18	.80	5.10	.06	.23	.74

NOTE: Known = ML estimate with λ known, MLE = ML estimate with λ unknown, and LSE = ordinary least squares estimate. In these calculations, the mean squared error (MSE) and STD. ERROR of difference terms are multiplied by T**2. Here T = 10 if σ ≤ .10 and T = 1 if σ = .50.

The other valuable point learned from Table 2 is that when we estimate the slope θ_2, the ML estimator with λ_0 unknown tends to dominate the LSE, especially for larger values of σ_0. In other words, for our model (1.4), there is real value to transformation when it is appropriate.

Finally, it should be noted that there is indeed a (moderate) cost for estimating θ_0 when λ_0 must also be estimated. The consequence of this moderate cost is that inference drawn in the "usual" way—treating $\hat{\lambda}$ as if it were preassigned—will be only moderately in error. (See Carroll and Ruppert 1981 and Carroll 1982a for details concerning the error in the usual inference for model (1.5), which tends to be moderate, on average, but which can be large for prediction at individual design points.)

4. EXAMPLES

4.1 Spawner-Recruit Data

This research was motivated by our study of the population dynamics of the Atlantic menhaden, which is, excluding shellfish, the third largest commerical U.S. fishery. The Atlantic menhaden fishery experienced a massive decline in the mid-1960's, and although there has been a slight recovery, present yields are only about half of those in the early 1960's. Our simulation study was an attempt to find strategies to reverse this decline in harvest; see Ruppert et al. (1983) for further details.

An important part of our study was the examination of the spawner-recruit (SR) relationship, in which we attempted to use the number of eggs E produced by mature menhaden (spawners) to predict the number R of juvenile menhaden recruited into the fishery (recruits). Estimates of E and R for the 21-year period 1955–1975 are given in Table 3.

An inspection of Table 3 or a plot of R against E shows that there is substantial variability. Note, for example, that 1958 has only the eighth-largest egg production, while it produced twice as many recruits as any other year. The year 1975 has the third-largest number of recruits but only the fourteenth largest egg production.

Two of the more usual ways to model the SR relationship are through the following approximations:

$$\text{(Beverton-Holt 1957)} \qquad R_i \approx (\alpha + \beta/E_i)^{-1}$$

$$\text{(Unnormalized Gamma)} \qquad R_i \approx \theta E_i^{\delta} \exp(\gamma E_i).$$

The Unnormalized Gamma (Gamma) is an extension of the Ricker (1954) equation, which allows only $\delta = 1$. Both the Beverton-Holt and the Ricker equations were derived from deterministic models. There appears to be no discussion in the fisheries literature on how these models should be interpreted for fish populations exhibiting highly variable SR relationships. The parameters are often estimated by using linearizing transformations. As stated in the Introduction, these two models can be thought of as part of a relationship driving the system, but they entail considerable variation. We wanted not only to decide upon one of the two models, but also, for our simulations, to do an adequate job of describing the nature of the variation in recruitment given egg production. The difference between the two models can have important effects on methods for managing the menhaden fishery. When, as is usual, $\gamma < 0$, the Gamma curve exhibits overcompensation; that is, eventually large egg production decreases recruitment, perhaps because of competition for food or perhaps because of a population explosion of a predator species. The Beverton-Holt model is much different, since it specifies that, except for random variation, large egg production will lead to an asymptote α^{-1} in recruitment. Since many strategies proposed for increasing the harvest depend on increasing egg production, perhaps beyond historically observed levels, the choice of the Gamma over the Beverton-Holt model could lead to a different management strategy. There has been no previous evidence for Atlantic menhaden supporting the Gamma curve, so a priori we would favor the Beverton-Holt curve, but it is obviously important for us to determine if the Beverton-Holt curve describes the present data as well as or better than the Gamma model.

Table 3. Spawner-Recruit Estimates

Year	*Egg Production E^a*	*Recruits R^b*
1955	2.42289	.85558
1956	1.77413	1.00935
1957	1.13816	.49287
1958	1.11338	2.10332
1959	1.32726	.31186
1960	1.88340	.41814
1961	2.62193	.30636
1962	1.63753	.30912
1963	.63302	.25417
1964	.33314	.29163
1965	.20943	.21642
1966	.16043	.30285
1967	.18389	.17046
1968	.23256	.24301
1969	.15267	.40457
1970	.22244	.20309
1971	.31532	.47767
1972	.33109	.37155
1973	.33011	.40746
1974	.27415	.52426
1975	.30154	.92933

[a] In units of 10^{14} eggs.
[b] In units of 10^{10} fish.

Linearization leads to the models

$$\text{(Beverton-Holt, Linear)} \quad R_i^{-1} = \alpha + \beta E_i^{-1} + \sigma_1\epsilon_i$$

$$\text{(Gamma, Linear)} \quad \log R_i = \delta \log E_i + \theta_* + \gamma E_i + \sigma_2\epsilon_i. \qquad (4.1)$$

From the point of view of meeting the assumption that $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, 1)$, the linearized Beverton-Holt is superior; the predictions of R_i are similar for the two models, but the residuals from the linearized Gamma are less normal-looking and somewhat more heteroscedastic. Thus, if we are constrained to admitting only the linearization models (4.1), the choice for simulation studies would be the Beverton-Holt.

There is, however, no reason why the variation about the Gamma model should be best explained by forcing linearization through logarithms. As argued in the Introduction, a more flexible model for determining the structure of the model variability is through our nonlinear Box-Cox models

$$\text{(Beverton-Holt)} \quad R_i^{(\lambda_B)} = \{(\alpha + \beta E_i^{-1})^{-1}\}^{(\lambda_B)} + \sigma_B\epsilon_i$$

$$\text{(Gamma)} \quad R_i^{(\lambda_G)} = \{\theta\, E_i \exp(\gamma E_i)\}^{(\lambda_G)} + \sigma_G\epsilon_i.$$

The MLE for λ_B is $\hat{\lambda}_B = -.72$, with a 90% confidence interval of $(-1.0, -0.17)$, and $\hat{\lambda}_B$ restricted to $[-1, 1]$. The likelihood ratio test for $H_0: \lambda_B = -1.0$ has value $\Lambda_B = .63$, indicating that the linearized Beverton-Holt model is at least reasonable. (Compare with X (1) quantiles.)

For the Gamma model, we obtained $\hat{\lambda}_G = -.71$, with a 90% confidence interval of $(-1.0, -.16)$. The likelihood ratio test for $H_0: \lambda_G = 0$ has value $\Lambda_G = 4.61$. This indicates that linearizing the Gamma model is probably not appropriate. In fact, having transformed by the power $\hat{\lambda}_G = -.71$, the residuals are essentially as normal looking and homoscedastic as those from the linearized Beverton-Holt.

The estimated Gamma curve reaches a maximum well above historically observed levels of egg production. In fact, the fitted Gamma and Beverton-Holt curves are quite similar over the observed range. However, our simulation experiments included allowing increased egg production where overcompensation would have an effect if the Gamma curve were used in the simulation model. We decided to base our simulations on the Beverton-Holt SR relationship, because there is no real evidence for overcompensation.

As this example makes clear, nonlinear models that can be linearized should not necessarily be linearized, since transformation analysis of response and predictor function can lead to a data scale with better distributional properties. In some cases, however, such as the Beverton-Holt model given here, the transformation analysis will provide added support for linearization.

Our theory predicts that the need to estimate λ is not costly in regard to estimation of α and β, and examination of the relevant Fisher information matrices suggests that this is, in fact, the case. If we fix $\lambda = \hat{\lambda}$, and (pretending that $\lambda = \hat{\lambda}$ was known a priori) invert the information matrix for α, β, and σ, then the estimated (asymptotic) variances are .2029, 2.0361, and .0258, respectively. If we invert the information matrix for α, β, σ, and λ, then the estimated (asymptotic) variances for α, β, and σ are .2213, 2.0394, and .1674, respectively. As our theory predicted, only the variance of $\hat{\sigma}$ increased substantially.

From our data analysis, we concluded that a realistic simulation model would need to be stochastic, and it was in the development of a stochastic model that power transformations proved to be most useful. In our simulation model we used

$$R = [(\hat{\alpha} + \hat{\beta}/E)^{-\hat{\lambda}} + \hat{\sigma}\epsilon]^{1/\hat{\lambda}}, \qquad (4.2)$$

where $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}$ are estimates on the $\hat{\lambda}$ scale, and ϵ is a standard normal pseudorandom number. With small probability the quantity in square brackets in (4.2) will be close to 0 or even negative, but in the model this quantity was truncated, so recruitment never exceeded twice the greatest recruitment observed in our data. In (4.2) one could use the MLE, $\hat{\lambda} = -.72$, but for simplicity, and because a likelihood ratio test indicated that $H_0\colon \lambda = -1.0$ was very credible, we used $\hat{\lambda} = -1.0$.

Model (4.2) with either $\hat{\lambda} = -1.0$ or $\hat{\lambda} = -.72$ is a particularly simple model that possesses these essential characteristics found in the data:

(i) Recruitment is highly variable and the variability increases with E.

(ii) Recruitment is positively skewed, and the skewness also increases with E. Therefore, except when E is small, the fishery has occasional dominant year classes.

The heteroscedasticity and variable skewness can be seen by examining the estimated distributions of recruitment with eggs set equal to the observed values during 1961 and 1969, the years with highest and lowest values of egg production, respectively, among all years for which we have data. In Figure 1, the quantiles of these estimated distributions are plotted against normal quantiles. The plots were obtained by plotting (4.2) with $\epsilon = \Phi^{-1}(i/70)$ on the horizontal axis and $\Phi^{-1}(i/70)$ on the vertical axis for $i = 2, \ldots, 68$, and interpolating these points with a spline. (Φ is the standard normal distribution function.) For the graphs, we used $\hat{\lambda} = -.72$ in (4.2), but $\hat{\lambda} = -1.0$ (the value used in simulations) would give similar plots.

With our model we were able to make a detailed simulation study of management policies for Atlantic menhaden. We found that management of a fishery with occasional, randomly occurring, dominant-year classes is a problem considerably different from managing a fishery with low variability.

In some situations, λ may be a nuisance parameter that is estimated only so that other parameters can be more efficiently estimated. However, as in this example, we may sometimes want to know the conditional distribution of the dependent variable, given the independent variables. λ then becomes a parameter equally as important as other parameters.

It is no coincidence that $\lambda_B \approx \lambda_G$. Since, for the range of E in the data, the Beverton-Holt and unnormalized Gamma curves with estimates substituted for the parameters are similar, their residuals from the estimated medians are also similar. $\hat{\lambda}$ is determined by the nonnor-

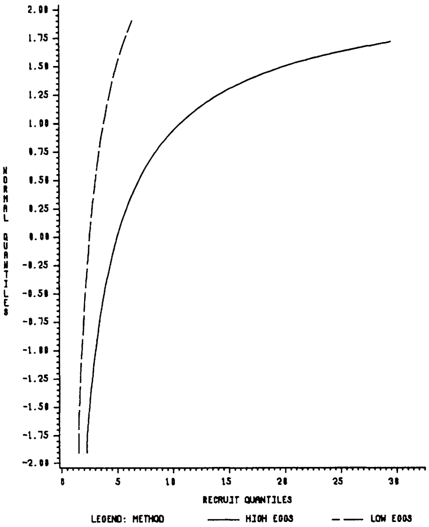


Figure 1. Estimated quantiles of recruitment plotted against standard normal quantiles. Recruitment is conditional on egg production being equal to the 1961 value (HIGH EGGS) *or the 1969 value* (LOW EGGS). *Recruitment is in units of 10^9* fish.

mality and heterogeneity of distribution that can be detected in these residuals.

As a final note, the analysis presented here was not merely an academic exercise; it formed a part of our study of the SR relationship, which itself was only a small (albeit important) component of a large study performed under time constraints. We welcome further analyses of the data, but we hope it is clear that we do not consider the reported analysis complete. In fact, we analyzed many other models under varying assumptions. For example, the inclusion of a quadratic time trend in the linearized Beverton-Holt model substantially improved the fit to the data. However, the time trend may be due to substantial overfishing in the 1960's, and the use of the trend for predicting future recruitments does not seem warranted. Another candidate for an explanatory variable in a more complex model is recruitment lagged one year.

4.2 Chemical Reaction Data

As a second sample, consider the data of Carr (1960) on the isomerization of pentane. For that data set, one proposed model is

$$y = \frac{\theta_0\theta_2(x_2 - x_3/1.632)}{1 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3}. \tag{4.3}$$

Box and Hill (1974) also list the data and discuss the model. They linearize (4.3) by taking inverses and then using a form of weighted least squares; without going into the full details, it suffices to state that their analysis suggests that $y^{(\lambda)}$ has constant variance, where $\lambda = .8$ (see also Pritchard, Downie, and Bacon 1977). We shall call the Box and Hill method power transformation (linearized) weight least squares (PTWLS).

Since the linearized model based on analyzing y^{-1} in (4.3) exhibits marked heteroscedasticity, it is interesting to see how our estimation method (based on (1.4a)–(1.4b)) performs; this method will be called PTBS for power transforming both sides. Based on Box and Hill's analysis, we should expect our PTBS to find $\lambda \approx .8$. As seen in Table 4, we estimated $\hat{\lambda} = .71$, which is definitely encouraging.

We applied PTBS to model (4.3), untransformed. See Table 4 for the results, which for θ are somewhat different from those obtained by Pritchard, Downie, and Bacon (1977), who used their algorithm DIRECT on the untransformed data. Possibly this difference is due to the presence of several local minima. When we applied unweighted nonlinear least squares to model (4.3), using Box and Hill's (1974) PTWLS solution as a starting value, other algorithms found a different solution with a smaller sum of squares than that reported by Pritchard, Downie, and Bacon (see Table 4).

Our aim in studying this example was to show that our PTBS gives reasonable results. We think our answers are perfectly sensible, and they correspond to PTWLS. For both, one obtains physically meaningful (positive) estimates of θ_1, θ_2, and θ_3, but unweighted linear least squares on the inverse scale gives negative estimates. We believe that PTWLS and PTBS can be recommended equally for this data set, although perhaps unweighted nonlinear least squares is just as effective and somewhat simpler.

A minor advantage of using the untransformed data is that on the inverse scale, Observation 6 of Box and Hill is highly influential even with power weighting (Carroll 1982b), while on the original scale no observation appears to have unusually high influence on the estimate of λ. Influence and diagnostics for inference in our model are questions that should be addressed in the future.

We used our transformation method successfully on other data sets, including the second data set mentioned by Pritchard, Downie, and Bacon.

APPENDIX A: PROOFS

Outline of Proof for Theorem 1

The likelihood analysis proceeds as follows. Define

$$z_i = dh(f_i(\theta_0), \lambda_0)/d\theta_0,$$

$$f_i(\theta) = f(x_i, \theta), f_i = f_i(\theta_0),$$

$$h_y(y) = h_y(y, \lambda) = dh(y, \lambda)/dy, \text{ and } h(y) = h(y, \lambda).$$

Let $h_\lambda(y)$ and $h_{\lambda\lambda}(y)$ be the gradient vector and Hessian of $h(y, \lambda)$ with respect to λ. By simple algebra we find

Table 4. Analysis of Carr's Data Using Unweighted, Least Squares, Power Transformation Weighted Least Squares (PTWLS), *and Power Transforming Both Sides* (PTBS)

Estimation Method	*Unweighted*	PTWLS	PTBS	*Unweighted*	*Unweighted*
Source	Pritchard et al.	Box and Hill	IMSL ZXSSQ[a] and ZXGSN	Pritchard et al.	BMDP3R[b]
Response Variable	y^{-1}	y^{-1}	y	y	y
λ	1	−.8	.71	1	1
Sum of Squares[c]	—	—	—	3.24397	3.23448
$\hat{\theta}_0$	16.3	40.00	39.2	35.9	35.9
$\hat{\theta}_1$	−.043	.75	.043	1.04	.071
$\hat{\theta}_2$	−.014	.35	.021	.55	.038
$\hat{\theta}_3$	−.098	1.85	.104	2.46	.167

[a] See Section 5.

[b] Same solution obtained with BMDPAR, SAS-NLIN with derivatives, and IMSL ZXSSQ.

[c] Used to compare the fits with $\lambda = 1$ and response 5 = y.

the joint information matrix of $(\theta_0, \sigma_0, \lambda_0)$ as (all summations are from 1 to N)

$$N^{-1}I = \begin{bmatrix} S/\sigma_0^2 & 0 & C_1/\sigma_0^2 \\ \cdot & 1/(2\sigma_0^4) & C_2/\sigma_0^4 \\ \cdot & \cdot & C_3/\sigma_0^2 \end{bmatrix},$$

where

$$S = N^{-1}\sum z_i z_i',$$

$$C_1 = -N^{-1}E\sum z_i[h_\lambda(y_i) - h_\lambda(f_i)]',$$

$$C_2 = -N^{-1}E\sum \epsilon_i[h_\lambda(y_i) - h_\lambda(f_i)]',$$

$$C_3 = N^{-1}E\sum\{[h_\lambda(y_i) - h_\lambda(f_i)][h_\lambda(y_i) - h_\lambda(f_i)]' + \epsilon_i[h_{\lambda\lambda}(y_i) - h_{\lambda\lambda}(f_i)] + (\partial/\partial\lambda)(\partial/\partial\lambda)'\log[h_y(y_i)]\}.$$

In general, C_1 and C_2 are not zero, and the asymptotic distribution of $(\hat\theta, \hat\sigma^2)$ when λ_0 is estimated differs from when λ_0 is known. The key question, of course, is whether C_1 and C_2 are sufficiently different from zero to seriously affect the distribution of $\hat\lambda$.

The expressions C_1, C_2, and C_3 are complex even when $f_i(\theta_0)$ has a nice form such as simple linear regression. To simplify matters sufficiently so that we can gain some insight about the difference between knowing and estimating λ_0, we follow Bickel and Doksum (1981) and others and let $\sigma_0 \to 0$.

Taylor expansions show that under mild regularity conditions $C_1 = 0(\sigma_0^2)$, $C_2 = 0(\sigma_0^2)$, and $C_3 = 0(\sigma_0^2)$ as $\sigma_0 \to 0$. Standard calculations show that when λ_0 is known,

$$N^{1/2}\text{covariance}\,[(\hat\theta - \theta_0)/\sigma_0, (\hat\sigma^2 - \sigma_0^2)/\sigma_0^2 \mid \lambda_0\,\text{known}]$$

$$\to A^{-1} = \begin{bmatrix} (\lim S)^{-1} & 0 \\ 0 & 2 \end{bmatrix}. \quad \text{(A.1)}$$

Let $D = \text{Diag}(\sigma_0, \ldots, \sigma_0, \sigma_0^2, 1, \ldots, 1)$. Then, to find this limiting covariance matrix when λ_0 is unknown, we must find the upper left $(k + 1) \times (k + 1)$ corner of

$$(\text{DID})^{-1} = \begin{bmatrix} S & 0 & C_1/\sigma_0 \\ \cdot & \frac{1}{2} & C_2/\sigma_0^2 \\ \cdot & \cdot & C_3/\sigma_0^2 \end{bmatrix}^{-1},$$

which by standard results on inverting partitioned matrices is $A^{-1} + FE^{-1}F'$, where A^{-1} is given in (A.1), $E = C_3/\sigma_0^2 - B'A\,B$, $F = A^{-1}B$, and $B' = (C_1/\sigma_0\; C_2/\sigma_0^2)$. Clearly,

$$F' = (S^{-1}C_1/\sigma_0 \quad 2C_2/\sigma_0)$$

and

$$E = C_3/\sigma_0^2 - C_1'S^{-1}C_1/\sigma_0^2 - 2C_2'C_2/\sigma_0^4.$$

To obtain simple asymptotics, we will assume that for σ_0 fixed, C_1/σ_0^2, C_2/σ_0^2, and C_3/σ_0^2 converge as $N \to \infty$, and that these, in turn, have limits D_1, D_2, and D_3, respectively, as $\sigma_0 \to 0$. We also assume that $S \to S_\infty$ (positive definite) as $N \to \infty$. If $D_3 - 2D_2'D_2$ is nonsingular, then

$$\lim_{\sigma_0\to 0}\lim_{N\to\infty} F\,E^{-1}\,F' = \begin{bmatrix} 0 & 0 \\ 0 & W \end{bmatrix},$$

where $W = 4D_2'[D_3 - D_2'D_2]^{-1}D_2$.

Outline of Proof of Theorem 2

Let $w_1, \ldots, w_N$ be positive numbers, and let $\hat\theta_1$ be any point that minimizes the expression

$$\sum w_i \mid y_i - f_i(\hat\theta_1) \mid.$$

Under (1.4), $f_i(\theta_0)$ is the unique median of y_i, so $\hat\theta_1$ will be consistent under some regularity conditions. The asymptotic distribution of $\hat\theta_1$ can be studied using techniques in Ruppert and Carroll (1980). A particularly simple asymptotic variance matrix is obtained if $w_i = h_y(f_i(\theta_0), \lambda_0)$, that is, if w_i is proportional to the density of $[y_i - f_i(\theta_0)]$ at its median, zero. Then

$$N^{1/2}(\hat\theta_1 - \theta_0)/\sigma_0 \xrightarrow{\mathcal{L}} N(0, (\pi/2)S^{-1}).$$

Although w_i depends on θ_0 and λ_0, the methods in Carroll and Ruppert (1982) can be used to show that the same limiting distribution holds if one substitutes $\sqrt{N}$-consistent estimates for θ_0 and λ_0.

Let $V(\lambda_0)$ and $V(\hat\lambda)$ be the asymptotic variance matrices of $\hat\theta(\lambda_0)$ and $\hat\theta(\hat\lambda)$, respectively. Since $V(\lambda_0) = S^{-1}$, the asymptotic optimality of the MLE shows that

$$S^{-1} \le V(\hat\lambda) \le (\pi/2)S^{-1},$$

where the inequalities are in the sense of positive definiteness.

APPENDIX B: COMPUTATION

Let $L(\theta, \sigma, \lambda)$ denote the log-likelihood for model (1.4). We do not recommend direct maximization of this likelihood by a canned routine for maximizing a function of many parameters. Rather, we adopt the usual practice for the Box-Cox (1964) model (1.5), which reduces the problem to maximizing a function of the scalar λ. Here are the general steps we used.

Step 1. Fix an initial scale $\lambda^{(1)}$. For the simulation and second example, $\lambda^{(1)} = 1.0$, while for the first example $\lambda^{(1)}$ was chosen to satisfy (4.1).

Step 2. Obtain preliminary estimates of θ, say $\theta^{(1)}$. For the simulation and first example, these were found by least squares, while for the second example the starting values are the last column of Table 4. The value $\sigma^{(1)}$ is simply the square root of the mean squared residual.

Step 3. Now begin the maximization of the log-likelihood. At the current value of λ, find $\theta(\lambda)$, $\sigma(\lambda)$ by using a nonlinear regression algorithm, starting from $\theta^{(1)}$, $\sigma^{(1)}$. After completion, update $\theta^{(1)} = \theta(\lambda)$, $\sigma^{(1)} = \sigma(\lambda)$. Define the one-parameter function $L^*(\lambda) = L(\theta(\lambda), \sigma(\lambda), \lambda)$.

Step 4. On the interval $\lambda \in [-1.0, 1.0]$, $L^*(\lambda)$ is often concave and can be maximized by a program specifically designed to maximize a concave function of one parameter. If $L^*(\lambda)$ is not concave, use a grid search.

For Steps 3 and 4, we used the IMSL subroutines ZXSSQ and XZGSN, respectively. The latter program includes a check for convexity of $-L^*(\lambda)$, which in the simulations was always satisifed.

[*Received November 1982. Revised October 1983.*]

REFERENCES

BEVERTON, R.J.H., and HOLT, S.J. (1957), *On the Dynamics of Exploited Fish Populations*, London: Her Majesty's Stationery Office.

BICKEL, P.J., and DOKSUM, K.A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311.

BOX, G.E.P., and COX, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

——— (1982), "An Analysis of Transformations Revisited, Rebutted," *Journal of the American Statistical Association*, 77, 209–210.

BOX, G.E.P., and HILL, W.J. (1974), "Correcting Inhomogeneity of Variance With Power Transformation Weighting," *Technometrics*, 16, 385–389.

CARR, N.L. (1960), "Kinetics of Catalytic Isomerization of *n*-Pentane," *Industrial and Engineering Chemistry*, 52, 391–396.

CARROLL, R.J. (1982a), "Prediction and the Power Transformation Family When Choice of Power Is Restricted to a Finite Set," *Journal of the American Statistical Association*, 77, 908–915.

——— (1982b), "Robust Estimation in Certain Heteroscedastic Linear Models When There Are Many Parameters," *Journal of Statistical Planning and Inference*, 7, 1–12.

CARROLL, R.J., and RUPPERT, D. (1981), "Prediction and the Power Transformation Family," *Biometrika*, 68, 609–617.

——— (1982), "Robust Estimation in Heteroscedastic Linear Models," *Annals of Statistics*, 10, 429–441.

HINKLEY, D.V., and RUNGER, G. (1984), "Analysis of Transformed Data," *Journal of the American Statistical Association*, 79, 302–309.

PRITCHARD, D.J., DOWNIE, J., and BACON, D.W. (1977), "Further Consideration of Heteroscedasticity in Fitting Kinetic Models," *Technometrics*, 19, 227–236.

RICKER, W.E. (1954), "Stock and Recruitment," *Journal of Fisheries Research Board of Canada*, 11, 559–623.

RUPPERT, D., and CARROLL, R.J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.

RUPPERT, D., REISH, R.L., DERISO, R.B., and CARROLL, R.J. (1983), "A Stochastic Population Model for Managing the Atlantic Menhaden Fishery and Assessing Managerial Risks," Mimeo Series No. 1532, Department of Statistics, University of North Carolina at Chapel Hill.

Variance Function Estimation

M. DAVIDIAN and R. J. CARROLL*

Heteroscedastic regression models are used in fields including economics, engineering, and the biological and physical sciences. Often, the heteroscedasticity is modeled as a function of the covariates or the regression and other structural parameters. Standard asymptotic theory implies that how one estimates the variance function, in particular the structural parameters, has no effect on the first-order properties of the regression parameter estimates; there is evidence, however, both in practice and higher-order theory to suggest that how one estimates the variance function does matter. Further, in some settings, estimation of the variance function is of independent interest or plays an important role in estimation of other quantities. In this article, we study variance function estimation in a unified way, focusing on common methods proposed in the statistical and other literature, to make both general observations and compare different estimation schemes. We show that there are significant differences in both efficiency and robustness for many common methods.

We develop a general theory for variance function estimation, focusing on estimation of the structural parameters and including most methods in common use in our development. The general qualitative conclusions are these. First, most variance function estimation procedures can be looked upon as regressions with "responses" being transformations of absolute residuals from a preliminary fit or sample standard deviations from replicates at a design point. Our conclusion is that the former is typically more efficient, but not uniformly so. Second, for variance function estimates based on transformations of absolute residuals, we show that efficiency is a monotone function of the efficiency of the fit from which the residuals are formed, at least for symmetric errors. Our conclusion is that one should iterate so that residuals are based on generalized least squares. Finally, robustness issues are of even more importance here than in estimation of a regression function for the mean. The loss of efficiency of the standard method away from the normal distribution is much more rapid than in the regression problem.

As an example of the type of model and estimation methods we consider, for observation-covariate pairs (Y_i, x_i), one may model the variance as proportional to a power of the mean response, for example,

$$E(Y_i) = f(x_i, \beta), \qquad \operatorname{var}(Y_i) = \sigma f(x_i, \beta)^\theta,$$
$$f(x_i, \beta) > 0,$$

where $f(x_i, \beta)$ is the possibly nonlinear mean function and θ is the structural parameter of interest. "Regression methods" for estimation of θ and σ based on residuals $r_i = Y_i - f(x_i, \hat{\beta}_*)$ for some regression fit $\hat{\beta}_*$ involve minimizing a sum of squares where some function T of the $|r_i|$ plays the role of the "responses" and an appropriate function of the variance plays the role of the "regression function." For example, if $T(x) = x^2$, the responses would be r_i^2, and the form of the regression function would be suggested by the aproximate fact $E(r_i^2) \approx \sigma^2 f(x_i, \hat{\beta}_*)^{2\theta}$. One could weight the sum of squares appropriately by considering the approximate variance of r_i^2. For the case of replication at each x_i, some methods suggest replacing the r_i in the function T by sample standard deviations at each x_i. Other functions T, such as $T(x) = x$ or $\log x$, have also been proposed.

KEY WORDS: Asymptotic efficiency; Heteroscedasticity; Regression; Variance estimation.

1. INTRODUCTION

Consider a heteroscedastic regression model for observable data Y:

$$EY_i = \mu_i = f(x_i, \beta); \qquad \operatorname{var}(Y_i) = \sigma^2 g^2(z_i, \beta, \theta). \quad (1.1)$$

Here, $\{x_i\}$ are the design vectors, $\beta(p \times 1)$ is the regression parameter, f is the mean response function, and the variance function g expresses the heteroscedasticity, where $\{z_i\}$ are known vectors, possibly the $\{x_i\}$, σ is an unknown scale parameter, and $\theta(r \times 1)$ is an unknown parameter. For example, the variance may be modeled as proportional to a power of the mean:

$$g(z_i, \beta, \theta) = f(x_i, \beta)^\theta, \qquad f(x_i, \beta) > 0. \quad (1.2)$$

One might also model the variance as quadratic in the predictors, that is,

$$\sigma g(z_i, \beta, \theta) = 1 + \theta_1 x_i + \theta_2 x_i^2,$$

or by an expanded power of the mean model, that is,

$$\sigma^2 g^2(z_i, \beta, \theta) = \theta_1 + \theta_2 f(x_i, \beta)^{\theta_3}. \quad (1.3)$$

Box and Meyer (1986) used

$$g(z_i, \beta, \theta) = \exp(z_i^t \theta).$$

An important feature of (1.1) is that no assumption about the distribution of the $\{Y_i\}$ has been made other than that of the form of the first two moments. Models that may be regarded as special cases of (1.1) are used in diverse fields, including radioimmunoassay, econometrics, pharmacokinetic modeling, enzyme kinetics, and chemical kinetics, among others. The usual emphasis is on estimation of β with estimation of the variances as an adjunct.

The most common method for estimating β is generalized least squares, in which one estimates $g(z_i, \beta, \theta)$ by using an estimate of θ and a preliminary estimate of β and then performs weighted least squares; see, for example, Carroll and Ruppert (1982a) and Box and Hill (1974). This might be iterated, with the preliminary estimate replaced by the current estimate of β, a new estimate of θ obtained, and the process repeated. Standard asymptotic theory as in Carroll and Ruppert (1982a) or Jobson and Fuller (1980) shows that as long as the preliminary estimators for the parameters of the variance function are consistent, all estimators of β obtained in this way will be asymptotically equivalent to the weighted least squares estimator with known weights.

There is evidence that for finite samples, the better one's estimate of θ, the better one's final estimate of β. Williams (1975) stated that "both analytic and empirical studies . . . indicate that . . . the ordering of efficiency (of estimates of β) . . . in small samples is in accordance with the ordering by efficiency (of estimates of θ)" (p. 563). Rothenberg (1984) showed via second-order calculations that if g does not depend on β, when the data are normally distributed the covariance matrix of the generalized least squares estimator of β is an increasing function of the covariance matrix of the estimator of θ.

* M. Davidian is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. R. J. Carroll is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. This work was supported by Air Force Office of Scientific Research Grant F-49620-85-C-0144.

Journal of the American Statistical Association
December 1987, Vol. 82, No. 400, Theory and Methods

Second-order asymptotics provide only a weak justification for studying the properties of variance function estimates. Instead, our thesis is that estimation of the structural variance parameter θ is of independent interest. In many engineering applications, an important goal is to estimate the error made in predicting a new observation; this can be obtained from the variance function once a suitable estimate of θ is available. In chemical and biological assay problems, issues of prediction and calibration arise. In such problems, the estimator of θ plays a central role. As motivation for the study of variance function estimation, in Section 2 we discuss the problem of calibration and prediction in the case of heteroscedasticity. For a formal investigation of how the statistical properties of prediction intervals and calibration constructs, such as the minimal detectable concentration, are highly dependent on how one estimates θ, see Davidian, Carroll, and Smith (1987). In off-line quality control, the emphasis is not only on the mean response but also on its variability; Box and Meyer (1986) stated that "one distinctive feature of Japanese quality control improvement techniques is the use of statistical experimental design to study the effect of a number of factors on variance as well as the mean" (p. 19). The goal is to adjust the levels of a set of experimental factors to bring the mean of the responses to some target value while minimizing standard deviation; the problem involves simultaneous consideration of both mean and variability, where the latter may be a function of the mean (see Box 1986; Box and Ramirez 1986). These authors advocated methods based on data transformations to account for the heteroscedasticity in separating the factors into those affecting dispersion but not location, those affecting location but not dispersion, and those affecting neither. Similarly, one might employ effective estimation of variance functions in this application. We briefly discuss the relationship between variance function estimation and one type of data transformation in Section 3.

It should be evident from this brief review that, far from being only a nuisance parameter, the structural variance parameter θ can be an important part of a statistical analysis. The foregoing discussion suggests the need for a unified investigation of estimation of variance functions, in particular, estimation of the structural parameter θ. Previous work in the literature tends to treat various special cases of (1.1) as different models with their own estimation methods. The intent of this article is to study parametric variance function estimation in a unified way. Nonparametric variance function estimation has also been studied (see, e.g., Carroll 1982); we will confine our study to the parametric setting.

Parametric variance function estimation may be thought of as a type of regression problem in which we try to understand variance as a function of known or estimable quantities and in which θ plays the part of a "regression" parameter. The major insight that allows for a unified study is that the absolute residuals from the current fit to the mean or the sample standard deviations from replicates are basic building blocks for analysis. At the graphical level, this means that transformations of the absolute residuals and sample standard deviations can be used to gain insight into the structure of the variability and to suggest parametric models. For estimation, a major contribution is to point out that most of the methods proposed in the literature are (possibly weighted) regressions of transformations of the basic building blocks on their expected values. Many exceptions to this are dealt with in this article as well.

Our study yields these major qualitative conclusions. As stated here, they apply strictly only to symmetric error distributions, but they are fairly definitive, and one is unlikely to be too successful ignoring them in practice.

1. Robustness plays a great role in the efficiency of variance function estimation, probably even greater than in estimation of a mean function. For example, if the variance does not depend on the mean response, the standard method will be normal theory maximum likelihood, as in Box and Meyer (1986). A weighted analysis of absolute residuals yields an estimator only 12% less efficient at the normal model, which rapidly gains efficiency over maximum likelihood for progressively heavier tailed distributions. This slope of improvement is much larger than is typical for estimation of the mean response. For a standard contaminated normal model for which the best robust estimators have efficiency 125% with respect to least squares, the absolute residual estimator of the variance function has efficiency 200%.

2. We obtain implications for fit to the means upon which the residuals are based. It has been our experience that unweighted least squares residuals yield unstable estimates of the variance function when the variances depend on the mean. This is confirmed in our study, in the sense that the asymptotic efficiency of the variance function estimators is an increasing function of the efficiency of the current fit to the means. Thus we suggest the use of iterative weighted fitting, so the variance function estimate is based on generalized least squares residuals. As far as we can tell, this part of our article is one of the first formal justifications for iteration in a generalized least squares context.

3. It is standard in many applied fields to take m replicates at each design point, where usually $m \leq 4$. Rather than using (transformations of) absolute residuals for estimating variance function parameters, one might use the sample standard deviations. We develop an asymptotic theory from which we obtain the efficiency of this substitution. The effect is typically, although not always, a loss of efficiency, at least when there are $m \leq 4$ replicates. The clearest results occur when the variance does not depend on the mean. Normal theory maximum likelihood is a weighted regression of squared residuals; the corresponding method would be a weighted regression based on sample variances. Using the latter entails a loss of efficiency, no matter what the underlying distribution. For normally distributed data, the efficiency is $(m - 1)/m$, thus being only 50% for duplicates. For other methods, using the replicate standard deviations can be more efficient. This is particularly true of a method due to Harvey (1976), which is based on the logarithm of absolute residuals. A small absolute residual, which seems always to occur in

practice, can wreak havoc with this method. This is consistent with our influence function calculations, so we suggest some trimming of the smallest absolute residuals before applying Harvey's method.

4. Our results indicate that the precision of estimates of θ is approximately independent of σ. In addition, in the power of the mean model (1.2), the efficiency of a regression estimator improves as the relative range of values of the mean response increases; efficiency depends on the spread of the logarithms of means, not their actual values. This helps explain why in assays, estimating variances is typically much harder than estimating means.

In Section 2 we discuss the prediction and calibration problems as a motivating example of a situation in which variance function estimation is of key importance. In Section 3 we describe a number of methods for estimation of θ. We do not discuss robust methods (see Giltinan, Carroll, and Ruppert 1986). In Section 4 we present an asymptotic theory for a general estimator of θ whose construction encompasses the methods of Section 3. Section 5 contains examples of specific applications of our theory and a discussion of the implications of our formulation. Sketches of proofs are presented in Appendix A.

2. AN EXAMPLE: THE ROLE OF VARIANCE ESTIMATION IN PREDICTION AND CALIBRATION PROBLEMS

One example in which heterogeneity of variation occurs is in calibration experiments in the physical and biological sciences, in which one fits a model such as (1.1) to a sample $\{Y_i, x_i\}$ $(i = 1, \ldots, N)$. The $\{x_i\}$ may be concentrations of a substance and the $\{Y_i\}$ may be counts or intensity levels that vary with concentration. The interest lies in using the estimated regression to make inference about a pair $\{Y_0, x_0\}$, which is independent of the original data set. One may wish to obtain point and interval predictors for Y_0 in the case in which x_0 is known (prediction) or estimate x_0 in the case in which Y_0 only is known (calibration) (see Rosenblatt and Speigelman 1981). As a motivating example for considering estimation of variance functions as an independent problem, we describe the primary role of form and estimation of the variance function in construction of prediction/calibration intervals in the case of heteroscedasticity.

Throughout this discussion assume that $x_i \equiv z_i$ so that we may write the variance function as $g(x_i, \beta, \theta)$, and assume that the data are approximately normally distributed. Given x_0, the standard point estimate of the response Y_0 is $f(x_0, \hat{\beta})$, where $\hat{\beta}$ is some estimate for β. For any consistent estimator $\hat{\beta}$ of β, under (1.1) the variance in the error made by the prediction is, for large sample sizes, $\text{var}\{Y_0 - f(x_0, \hat{\beta})\} \approx \sigma^2 g^2(x_0, \beta, \theta)$, so the error in prediction is determined mainly by the variance function $\sigma^2 g^2(x_0, \beta, \theta)$ and not the original data set itself. An approximate $(1 - \alpha)$ 100% confidence interval for Y_0 is $I(x_0) = \{\text{all } Y$ in the interval $f(x_0, \hat{\beta}) \pm t^{N-p}_{1-\alpha/2}\, \hat{\sigma} g(x_0, \hat{\beta}, \hat{\theta})\}$; here $t^{N-p}_{1-\alpha/2}$ is the $(1 - \alpha/2)$ percentage point of the t distribution with $(N - p)$ degrees of freedom and $\hat{\sigma}$ and $\hat{\beta}$ are estimates. If the parameters are estimated by a weighted analysis, such as generalized least squares assuming (1.1), all estimates are consistent and the prediction interval becomes

$$I(x_0) \approx \{\text{all } Y \text{ in the interval } f(x_0, \beta) \pm t^{N-p}_{1-\alpha/2}\sigma g(x_0, \beta, \theta)\}. \quad (2.1)$$

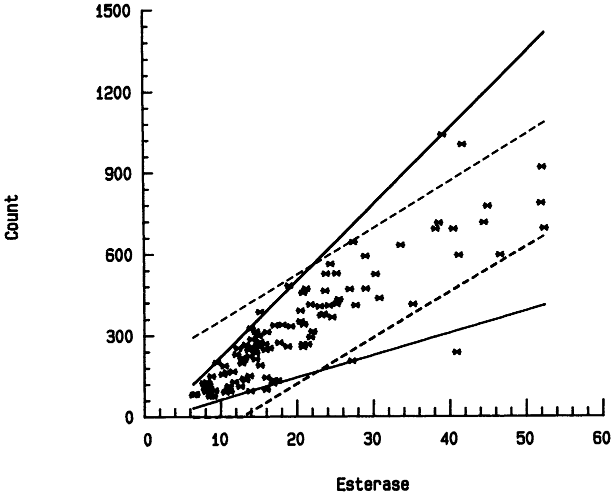


Figure 1. Approximate Form of Prediction Intervals for a Linear Mean Response Function Based on Unweighted (ignoring heteroscedasticity) and Weighted [as in (1.1)] Regression Fits. Esterase assay 95% prediction limits: dashed line—unweighted, solid line—weighted.

If one were to ignore the heterogeneity, the interval would be given by $I_U(x_0)$ = {all Y in the interval $f(x_0, \hat{\beta}) \pm t^{N-p}_{1-\alpha/2}\hat{\sigma}$}. For an unweighted analysis, however, σ^2 would be estimated by the unweighted mean squared error $\hat{\sigma}^2_U \approx \sigma^2 N^{-1} \sum g^2(x_i, \beta, \theta) = \sigma^2 g^2_N$ for large N. Thus the unweighted prediction interval satisfies

$$I_U(x_0) \approx \{\text{all } Y \text{ in the interval } f(x_0, \beta) \pm t^{N-p}_{1-\alpha/2}\sigma g_N\}. \quad (2.2)$$

Comparing (2.1) and (2.2), we see that where the variability is small, the unweighted interval will be too long and hence pessimistic, and conversely where the variance is large. Figure 1 illustrates this phenomenon for the results of an assay for the concentration of an enzyme esterase, where the responses are binding counts in the simple situation of an approximately linear mean response function where variability increases with mean response.

The situation is the same for calibration. For simplicity in discussing calibration, assume that $f(x, \beta)$ is strictly increasing or decreasing in x. Given Y_0, the usual estimate of x_0 is that value satisfying $Y_0 = f(x, \hat{\beta})$. The common confidence interval for x_0 is the set of all x values for which Y_0 falls in the prediction interval $I(x)$; this interval is actually a $(1 - \alpha)$ 100% confidence interval for the unknown x_0. Since the confidence interval for x_0 is thus an inversion of the intervals in Figure 1, again, the effect of not weighting is intervals that are too long for x_0 when the variance is small and the opposite when the variance is large. We are not familiar with any extensive investigation of calibration confidence intervals for heteroscedastic models, but see Watters, Carroll, and Spiegelman (1987).

The key point of this discussion is that when heterogeneity of variance is present, how well one models and estimates the variances will have substantial impact on prediction and calibration based on the estimated mean response, since the form of the intervals depends on the form of the variance function. Some theoretical work has been done verifying the implications of this discussion; for an investigation of how the statistical properties of estimators for calibration quantities depend on those of the estimator θ, see Davidian et al. (1987) and Carroll (1987).

3. ESTIMATION OF θ

We now discuss the form and motivation for several estimators of θ in (1.1). In what follows, let $\hat{\beta}_*$ be a preliminary estimator for β. This could be unweighted least squares or the current estimate in an iterative reweighted least squares calculation. Let $\varepsilon_i = \{Y_i - f(x_i, \beta)\}/\{\sigma g(z_i, \beta, \theta)\}$ denote the errors so that $E\varepsilon_i = 0$ and $E\varepsilon_i^2 = 1$, and denote the residuals by $r_i = Y_i - f(x_i, \hat{\beta}_*)$. We consider some methods requiring $m_i \geq 2$ replicates at each of M design points; for simplicity, we consider only the case of equal replication $m_i \equiv m$ and write in obvious fashion $\{Y_{ij}\}$ $(j = 1, \ldots, m)$ to denote the m observations at x_i where appropriate, so that $N = Mm$ is the total number of observations. In this case, let $\overline{Y}_{i\cdot}$ and s_i denote the sample mean and standard deviation at x_i. For consistency of exposition, however, we denote the sum over all observations as $\sum_{i=1}^N$ instead of $\sum_{i=1}^M \sum_{j=1}^m$. When we speak of replacing absolute residuals $\{|r_i|\}$ by sample deviations $\{s_i\}$ in the case of replication, $|r_i|$ or s_i appears m times in the sum.

3.1 Regression Methods

3.1.1. Pseudolikelihood. Given $\hat{\beta}_*$, the pseudolikelihood estimator maximizes the normal log-likelihood $l(\hat{\beta}_*, \theta, \sigma)$, where

$$l(\beta, \theta, \sigma) = -N \log \sigma - \sum_{i=1}^N \log\{g(z_i, \beta, \theta)\} - (2\sigma^2)^{-1} \sum_{i=1}^N \{Y_i - f(x_i, \beta)\}^2/g^2(z_i, \beta, \theta) \quad (3.1)$$

(see Carroll and Ruppert 1982a). Here the term "pseudolikelihood" is used as in Gong and Samaniego (1981). Generalizations of pseudolikelihood for robust estimation have been studied by Carroll and Ruppert (1982a) and Giltinan et al. (1986).

3.1.2. Least Squares on Squared Residuals. Besides pseudolikelihood, other methods using squared residuals have been proposed. The motivation for these methods is that the squared residuals have approximate expectation $\sigma^2 g^2(z_i, \beta, \theta)$ (see Amemiya 1977; Jobson and Fuller 1980). This suggests a nonlinear regression problem in which the "responses" are $\{r_i^2\}$ and the "regression function" is $\sigma^2 g^2(z_i, \hat{\beta}_*, \theta)$. The estimator $\hat{\theta}_{SR}$ minimizes in θ and σ,

$$\sum_{i=1}^N \{r_i^2 - \sigma^2 g^2(z_i, \hat{\beta}_*, \theta)\}^2.$$

For normal data the squared residuals have approximate variance $\sigma^4 g^4(z_i, \beta, \theta)$; in the spirit of generalized least squares, this suggests the weighted estimator that minimizes in θ and σ,

$$\sum_{i=1}^N \{r_i^2 - \sigma^2 g^2(z_i, \hat{\beta}_*, \theta)\}^2/g^4(z_i, \hat{\beta}_*, \hat{\theta}_*), \quad (3.2)$$

where $\hat{\theta}_*$ is a preliminary estimator for θ, $\hat{\theta}_{SR}$, for example. Full iteration, when it converges, would be equivalent to pseudolikelihood.

3.1.3. Accounting for the Effect of Leverage. One objection to methods such as pseudolikelihood and least squares based on squared residuals is that no compensation is made for the loss of degrees of freedom associated with preliminary estimation of β. For example, the effect of applying pseudolikelihood directly seems to be a bias depending on p/N. For settings such as fractional factorials, where p is large relative to N, this bias could be substantial.

Bayesian ideas have been used to account for loss of degrees of freedom (see Harville 1977; Patterson and Thompson 1971). When g does not depend on β, the restricted maximum likelihood approach of Patterson and Thompson suggests in our setting one estimate θ from the mode of the marginal posterior density for θ assuming normal data and a prior for the parameters proportional to σ^{-1}. When g depends on β, one may extend the Bayesian arguments and use a linear approximation as in Box and Hill (1974) and Beal and Sheiner (1987) to define a restricted maximum likelihood estimator.

Let Q be the $N \times p$ matrix with ith row $f_\beta(x_i, \beta)^t/g(z_i, \beta, \theta)$, where $f_\beta(x_i, \beta) = \partial/\partial\beta\{f(x_i, \beta)\}$, and let $H = Q(Q^tQ)^{-1}Q^t$ be the "hat" matrix with diagonal element $h_{ii} = h_{ii}(\beta, \theta)$; the values $\{h_{ii}\}$ are the leverage values. It turns out that the restricted maximum likelihood estimator is equivalent to an estimator obtained by modifying pseudolikelihood to account for the effect of leverage. This characterization, although not unexpected, is new; we derive this estimator and its equivalence to a modification of pseudolikelihood in Appendix B.

The least squares approach using squared residuals can also be modified to show the effect of leverage. Jobson and Fuller (1980) essentially noted that for nearly normally distributed data we have the approximations

$$Er_i^2 \approx \sigma^2(1 - h_{ii})g^2(z_i, \beta, \theta),$$

$$\text{var } r_i^2 \approx 2\sigma^4(1 - h_{ii})^2g^4(z_i, \beta, \theta).$$

To exploit these approximations modify (3.2) to minimize in θ and σ,

$$\sum_{i=1}^{N} \{r_i^2 - \sigma^2(1 - \hat{h}_{ii})g^2(z_i, \hat{\beta}_*, \theta)\}^2 \div \{(1 - \hat{h}_{ii})^2g^4(z_i, \hat{\beta}_*, \hat{\theta}_*)\}, \quad (3.3)$$

where $\hat{h}_{ii} = h_{ii}(\hat{\beta}_*, \hat{\theta}_*)$ and $\hat{\theta}_*$ is a preliminary estimator for θ. An asymptotically equivalent variation of this estimator in which one sets the derivatives of (3.3) with respect to θ and σ equal to 0 and then replaces $\hat{\theta}_*$ by θ can be seen to be equivalent to pseudolikelihood in which one replaces standardized residuals by studentized residuals. Although this estimator also takes into account the effect of leverage, it is different from restricted maximum likelihood.

3.1.4. Least Squares on Absolute Residuals. Squared residuals are skewed and long-tailed, which has led many authors to propose using absolute residuals to estimate θ (see Glejser 1969; Theil 1971). Assume that

$$E|Y_i - f(x_i, \beta)| = \eta g(z_i, \beta, \theta),$$

which is satisfied if the errors $\{\varepsilon_i\}$ are iid. Mimicking the least squares approach based on squared residuals, one obtains the estimator $\hat{\theta}_{AR}$ by minimizing in η and θ,

$$\sum_{i=1}^{N} \{|r_i| - \eta g(z_i, \hat{\beta}_*, \theta)\}^2.$$

In analogy to (3.2), the weighted version is obtained by minimizing

$$\sum_{i=1}^{N} \{|r_i| - \eta g(z_i, \hat{\beta}_*, \theta)\}^2/g^2(z_i, \hat{\beta}_*, \hat{\theta}_*),$$

where $\hat{\theta}_*$ is a preliminary estimator for θ, probably $\hat{\theta}_{AR}$. As for least squares estimation based on squared residuals, one presumably could modify this approach to account for the effect of leverage.

3.1.5. Logarithm Method. The suggestion of Harvey (1976) is to exploit the fact that the logarithm of the absolute residuals has approximate expectation $\log\{\sigma g(z_i, \beta, \theta)\}$. Estimate θ by ordinary least squares regression of $\log|r_i|$ on $\log\{\sigma g(z_i, \hat{\beta}_*, \theta)\}$, since if the errors are iid, the regression should be approximately homoscedastic. If one of the residuals is near 0, the regression could be adversely affected by a large "outlier"; hence in practice one might wish to delete a few of the smallest absolute residuals, perhaps trimming the smallest few percent.

3.2 Other Methods

Besides squares and logarithms of absolute residuals, other transformations could be used. For example, the square root and ⅔ root would typically be more normally distributed than the absolute residuals themselves. Such transformations appear to be useful, although they have not been used much to our knowledge. Our asymptotic theory applies to such transformations.

In a parametric model such as (1.1), joint maximum likelihood estimation is possible, where we use the term maximum likelihood to mean normal theory maximum likelihood. When the variance function does not depend on β, it can be easily shown that maximum likelihood is asymptotically equivalent to weighted least squares methods based on squared residuals. In the situation in which the variance function depends on β this is not the case. In this setting, it has been observed by Carroll and Ruppert (1982b) and McCullagh (1983) that, although maximum likelihood estimators enjoy asymptotic optimality when the model and distributional assumptions are correct, the maximum likelihood estimator of β can suffer problems under departures from these assumptions. This suggests that joint maximum likelihood estimation should not be applied blindly in practice. The theory of the next section shows the asymptotic equivalence of maximum likelihood with other methods in a simplifying special case. Based on this theory, we tend to prefer weighted regression methods even when the data are approximately normal for reasons of relative computational simplicity.

Although we have chosen to describe the methods of Section 3.1 as "regression methods," asymptotically equivalent versions of such methods may be derived by considering maximum likelihood assuming some underlying distribution. For example, the form of the weighted squared residuals method is that of normal theory maximum likelihood with β known and $\hat{\theta}_*$ replaced by θ (pseudolikelihood); the form of the weighted absolute residual method is that of maximum likelihood assuming β known and $\hat{\theta}_*$ replaced by θ under the double exponential distribution. Thus what we term a regression method may be viewed as an approximation to maximum likelihood assuming a particular distribution. We feel that the regression interpretation is a much more appealing and natural motivation, since no particular distribution need be considered

Table 1. Description of Some Methods for Variance Function Estimation

Maximum likelihood	Normal theory maximum likelihood in β, σ, θ.
Pseudolikelihood	Normal theory maximum likelihood when β is set to current value. When iterated, equivalent to maximum likelihood if the variance does not depend on β.
Weighted squared residuals	Regress squared residuals on the variance, function, weight inversely with squared current variance estimate.
Weighted absolute residuals	Regress absolute residuals on the standard deviation function, weight inversely with current variance estimate.
Logarithm method	Regress logarithm of absolute residuals on log of standard deviation function. Be wary of near-zero residuals.
Restricted maximum likelihood	Pseudolikelihood corrected for leverage. Maximizes marginal posterior for noninformative prior.

All of the preceding except restricted maximum likelihood have analogs formed by replacing absolute residuals by sample standard deviations in the case of replication. The following are based on the mean function or design being fully or partially unknown and are often used in assays.

Rodbard and Frazier	Regress log sample standard deviation on log sample mean, where the variance function depends on β only through the means.
Modified maximum likelihood	Modified functional maximum likelihood [Eq. (2.5)], where variance function depends on β only through means.
Sadler and Smith	Same as modified maximum likelihood, but means estimated by sample means.

to obtain the form of the estimators, only the mean-variance relationship.

Another joint estimation method is the extended quasi-likelihood of Nelder and Pregibon (1987) also described in McCullagh and Nelder (1983). This estimator is based on assuming a class of distributions "nearly" containing skewed distributions, such as the Poisson and gamma. Although it may be viewed as iteration between estimation of θ and σ and generalized least squares for β, technically this scheme does not fit in the general framework of the next section: an asymptotic theory was developed elsewhere (see Davidian and Carroll, in press). A related formulation was given by Efron (1986).

Methods requiring replicates at each design point have been proposed in the assay literature. These methods do not depend on the postulated form of the regression function; one reason that this may be advantageous is that in many assays, along with observed pairs (Y_{ij}, x_i), there will also be pairs in which only Y_{ij} is observed. A popular and widely used method is that of Rodbard and Frazier (1975). If we assume that

$$g(z_i, \beta, \theta) = g(\mu_i, z_i, \theta), \tag{3.4}$$

as in, for example, (1.2) or (1.3), the method is identical to the logarithm method previously discussed except that one replaces $|r_i|$ by the sample standard deviation s_i and $f(x_i, \hat{\beta}_*)$ in the "regression" function by the sample mean $\bar{Y}_{i\cdot}$. As a motivation for this and the method of Harvey, consider that under (1.2) θ is simply the slope parameter for a simple linear regression.

As an alternative, under the assumption of independence and (3.4), the modified maximum likelihood method of Raab (1981) estimates θ by joint maximization in the $(M + r + 1)$ parameters $\sigma^2, \theta, \mu_1, \ldots, \mu_M$ of the "modified" normal likelihood

$$\prod_{i=1}^{M} \{2\pi\sigma^2 g^2(\mu_i, z_i, \theta)\}^{(m-1)/2} \times \exp\left[-\sum_{j=1}^{m} (Y_{ij} - \mu_i)^2/\{2\sigma^2 g^2(\mu_i, z_i, \theta)\}\right]. \tag{3.5}$$

The modification serves to make the estimator of σ unbiased. The idea here is to improve upon the regression method of Rodbard by appealing to a maximum likelihood approach that, despite a parameter space increasing as the number of design points, is postulated to have reasonable properties. A related method is that in which θ and σ are estimated by maximizing (3.5) with μ_i replaced by $\bar{Y}_{i\cdot}$, the motivation being computational ease and evidence that this estimator may not be too different from that of Raab in practice (see Sadler and Smith 1985).

Table 1 contains a summary of some of the common methods for variance function estimation and their formulations.

4. AN ASYMPTOTIC THEORY OF VARIANCE FUNCTION ESTIMATION

In this section we construct an asymptotic theory for a general class of regression-type estimators for θ. Since our major interest lies in obtaining general insights, we do not state technical assumptions or details. In what follows, in the case of replication $N \to \infty$ in such a way that m remains fixed. The reader uninterested in this development may wish to review the definition of the form of the estimators in the first two paragraphs of Section 4.1 and then skip to Section 5, where conclusions and implications of the theory are presented.

4.1 Methods Based on Transformations of Absolute Residuals

Write $d_i(\beta) = |Y_i - f(x_i, \beta)|$. Let T be a smooth function and define M_i by

$$M_i(\eta, \theta, \beta) = E[T\{d_i(\beta)\}],$$

where η is a scale parameter that is usually a function of σ only. We consider estimation of the more general parameter η instead of σ itself for ease of exposition, and since σ is estimated jointly with θ in regression methods, our theory focuses on expansions for η and θ jointly. If $\hat{\eta}_*$, $\hat{\theta}_*$, and $\hat{\beta}_*$ are any preliminary estimators for η, θ, and β, define $\hat{\eta}$ and θ to be the solutions of

$$N^{-1/2} \sum_{i=1}^{N} H_i(\eta, \theta, \hat{\beta}_*)\{T\{d_i(\hat{\beta}_*)\} - M_i(\eta, \theta, \hat{\beta}_*)\} \div V_i(\hat{\eta}, \hat{\theta}_*, \hat{\beta}_*) = 0, \tag{4.1}$$

where $V_i(\eta, \theta, \beta)$ is a smooth function and H_i is a smooth function that for the estimators of Section 3 is the partial derivative of M_i with respect to (η, θ). In what follows, we suppress the arguments of the functions M_i, V_i, etcetera when they are evaluated at the true values η, θ, and β. Specific examples are considered in the next section.

The class of estimators solving (4.1) includes directly or includes an asymptotically equivalent version of the estimators of Section 3.1. For methods that account for the effect of leverage, M_i, V_i, and H_i will depend on the h_{ii}. In this case we need the additional assumption that if $h = \max\{h_{ii}\}$, then $N^{1/2}h$ converges to 0.

Theorem 4.1. Let $\hat{\eta}_*$, $\hat{\theta}_*$, and $\hat{\beta}_*$ be $N^{1/2}$ consistent for estimating η, θ, and β. Let $\dot{T}$ be the derivative of T, and define

$$C_i = H_i[T\{d_i(\beta)\} - M_i]/V_i,$$

$$B_{1,N} = N^{-1} \sum_{i=1}^{N} H_i H_i^t / V_i,$$

$$B_{2,N} = -N^{-1} \sum_{i=1}^{N} (H_i/V_i)\partial/\partial\beta\{M_i(\eta, \theta, \beta)\},$$

$$B_{3,N} = -N^{-1} \sum_{i=1}^{N} (H_i/V_i) f_\beta(x_i, \beta) E[\dot{T}\{d_i(\beta)\}\text{sign}(\varepsilon_i)].$$

Then, under regularity conditions as $N \to \infty$,

$$B_{1,N}N^{1/2}\begin{bmatrix} \hat{\eta} - \eta \\ \hat{\theta} - \theta \end{bmatrix} = N^{-1/2} \sum_{i=1}^{N} C_i + (B_{2,N} + B_{3,N})N^{1/2}(\hat{\beta}_* - \beta) + o_p(1). \quad (4.2)$$

We may immediately make some general observations about the estimator $\hat{\theta}$ solving (4.1). Note that if the variance function does not depend on β, then M_i does not depend on β and hence $B_{2,N} \equiv 0$. For the estimators of Section 2.1, $\dot{T}$ is an odd function. Thus, if the errors $\{\varepsilon_i\}$ are symmetrically distributed, $E[\dot{H}_i\{d_i(\beta)\}\text{sign}(\varepsilon_i)] = 0$ and hence $B_{3,N} \equiv 0$.

Corollary 4.1(a). Suppose that the variance function does not depend on β and the errors are symmetrically distributed. Then the asymptotic distributions of the regression estimators of Section 3.1 do not depend on the method used to obtain $\hat{\beta}_*$. If both of these conditions do not hold simultaneously, then the asymptotic distributions will depend in general on the method of estimating β.

The implication is that in the situation for which the variance function does not depend on β and the data are approximately symmetrically distributed, for large sample sizes the preliminary estimator for β will play little role in determining the properties of $\hat{\beta}$. Note also from (4.2) that for weighted methods, the effect of the preliminary estimator of θ is asymptotically negligible regardless of the underlying distributions.

The preliminary estimator $\hat{\beta}_*$ might be the unweighted least squares estimator, a generalized least squares estimator, or some robust estimator. See, for example, Huber (1981) and Giltinan et al. (1986) for examples of robust estimators for β. For some vectors $\{v_{N,i}\}$, these estimators admit an asymptotic expansion of the form

$$N^{1/2}(\hat{\beta}_* - \beta) = N^{-1/2} \sum_{i=1}^{N} \Psi(v_{N,i}, \varepsilon_i) + o_p(1). \quad (4.3)$$

Here Ψ is odd in the argument ε. In case the variance function depends on β, $B_{2,N} \neq 0$ in general; however, if the errors are symmetrically distributed and $\hat{\beta}_*$ has expansion of form (4.3), then the two terms on the right side of (4.2) are asymptotically independent. The following is then immediate.

Corollary 4.1(b). Suppose that the errors are symmetrically distributed and that $\hat{\beta}_*$ has an asymptotic expansion of the form (4.3). Then for the estimators of Section 3.1, the asymptotic covariance matrix of $\hat{\theta}$ is a monotone nondecreasing function of the asymptotic covariance matrix of $\hat{\beta}_*$.

By the Gauss–Markov theorem and the results of Jobson and Fuller (1980) and Carroll and Ruppert (1982a), the implication of Corollary 4.1(b) is that using unweighted least squares estimates of β will result in inefficient estimates of θ. This phenomenon is exhibited in small samples in a Monte Carlo study of Davidian et al. (1987). If one starts from the unweighted least squares estimate, one ought to iterate the process of estimating θ—use the current value $\hat{\beta}_*$ to estimate θ from (4.1), use these $\hat{\beta}_*$ and $\hat{\theta}$ to obtain an updated $\hat{\beta}_*$ by generalized least squares, and repeat the process $\mathcal{C} - 1$ more times. It is clear that the asymptotic distribution of $\hat{\theta}$ will be the same for $\mathcal{C} \geq 2$ with larger asymptotic covariance for $\mathcal{C} = 1$, so in principle one ought to iterate this process at least twice. See Carroll, Wu, and Ruppert (1987) for more on iterating generalized least squares.

4.2 Methods Based on Sample Standard Deviations

Assume replication, and as before let $\{s_i\}$ be the sample standard deviations at each x_i, which themselves have been proposed as estimators of the variance in generalized least squares estimation of β. This can be disastrous (see Jacquez, Mather, and Crawford 1968). When replication exists, however, practitioners feel comfortable with the notion that the $\{s_i\}$ may be used as a basis for estimating variances; thus one might reasonably seek to estimate θ by replacing $d_i(\hat{\beta}_*)$ by s_i in (4.1).

The following result is almost immediate from the proof of Theorem 4.1 in Appendix A.

Theorem 4.2. If $d_i(\hat{\beta}_*)$ is replaced by s_i in (4.1), then under the conditions of Theorem 4.1 the resulting estimator for θ satisfies (4.2) with $B_{3,N} \equiv 0$ and the redefinitions

$$C_i = (H_i/V_i)\{T(s_i) - M_i\}, \quad (4.4a)$$

$$M_i = E\{T(s_i)\} = M_i(\eta, \theta, \beta). \quad (4.4b)$$

If the errors are symmetrically distributed, then, from (4.2) and Theorem 4.2, whether one is better off using absolute residuals or sample standard deviations in the methods of Section 3.1 depends only on the differences between the expected values and variances of $T\{d_i(\beta)\}$ and $T(s_i)$. In Section 5 we exhibit such comparisons explicitly and show that absolute residuals can be preferred to sample standard deviations in situations of practical importance.

4.3 Methods Not Depending on the Regression Function

We assume throughout this discussion that the variance function has form (3.4) and replication is available. From Section 3.1 we see that the "regression function" part of the estimating equations depends on $f(x_i, \hat{\beta}_*)$, so in the general equation (4.1) M_i, V_i, and H_i all depend on $f(x_i, \hat{\beta}_*)$. In some settings, one may not postulate a form for the μ_i for estimating θ; the method of Rodbard and Frazier (1975), for example, uses s_i in place of $d_i(\hat{\beta}_*)$ as in Section 4.2 and replaces $f(x_i, \hat{\beta}_*)$ by the sample mean $\overline{Y}_{i\cdot}$. We now consider the effect of replacing predicted values by sample means for the general class (4.1).

The presence of the sample means in the variance function in (4.1) requires more complicated and restrictive assumptions than the usual large sample asymptotics applied heretofore. The method of Rodbard and Frazier and the general method (4.1) with sample means are nonlinear errors-in-variables problems as studied by Wolter and Fuller (1982) and Stefanski and Carroll (1985). Standard asymptotics for these problems correspond to letting σ go to 0 at rate $N^{-1/2}$. In Section 4.4 we discuss the practical implications of σ being small; for now, we state the following result.

Theorem 4.3. Suppose that we replace $f(x_i, \hat{\beta}_*)$ by $\overline{Y}_{i\cdot}$ in M_i, V_i, and H_i in (4.1) and adopt the assumptions of Theorems 4.1 and 4.2. Further, suppose that as $N \rightarrow \infty$, $\sigma \rightarrow 0$ simultaneously and

(i) $N^{1/2}\sigma \rightarrow \lambda$, $0 \leq \lambda < \infty$;

(ii) $N^{1/2} \sum_{i=1}^{N} C_i$ has a nontrivial asymptotic normal limit distribution;

(iii) The $\{\varepsilon_i\}$ are symmetric and iid;

(iv) $\{|\overline{Y}_{i\cdot} - \mu_i|/\sigma\}^2$ has uniformly bounded k moments, some $k > 2$.

Then the results of Theorems 4.1 and 4.2 hold with $B_{2,N} = B_{3,N} \equiv 0$.

This result shows that under certain restrictive assumptions, one may replace predicted values by sample means under replication; it is important to realize, however, that the assumption of small σ is not generally valid and hence the use of sample means may be disadvantageous in situations where these asymptotics do not apply. Further, relaxation of Assumption (iii) will result in an asymptotic bias in the asymptotic distribution of the estimator not present for estimators based on residuals regardless of the assumption of symmetry (see App. A).

The estimator of Raab (1981) discussed in Section 3.2 is also a functional nonlinear error-in-variables estimator, complicated by a parameter space with size of order N. Sadler and Smith (1985) observed that the Raab estimator is often indistinguishable from their estimator with μ_i replaced by $\overline{Y}_{i\cdot}$ in (3.5); such an estimator is contained in the general class (4.1). Davidian (1986) showed that under the asymptotics of Theorem 4.3 and additional regularity conditions the two estimators are asymptotically equivalent in an important special case. We may thus consider the result of Theorem 4.3 relevant to this estimator.

4.4 Small σ Asymptotics

In Section 4.3 technical considerations forced us to pursue an asymptotic theory in which σ is small. It turns out that in some situations of practical importance these asymptotics are relevant. In particular, in assay data we have observed values for σ that are quite small relative to the means. Such asymptotics are used in the study of data transformations in regression. It is thus worthwhile to consider the effect of small σ on the results of Sections 4.1 and 4.2 and to comment on some other implications of letting $\sigma \rightarrow 0$.

In the situation of Theorem 4.1, if the errors are symmetrically distributed, then for the estimators of Section 3.1, if $\sigma \rightarrow 0$ as $N \rightarrow \infty$, then there is no effect for estimating the regression parameter β. In the situation of Theorem 4.2, the errors need not even be symmetrically distributed. The major insight provided by these results is that in certain practical situations in which σ is small, the choice of $\hat{\beta}_*$ may not be too important even if the variance function depends on β.

Small σ asymptotics may be used to provide insight into the behavior of other estimators for θ that do not fit into the general framework of (4.1). It can be shown that the extended quasilikelihood estimator need not necessarily be consistent for fixed σ, but if one adopts the asymptotics of the previous section, this estimator is asymptotically equivalent to regression estimators based on squared residuals as long as the errors are symmetrically distributed. Otherwise, an asymptotic bias may result, which may have implications for inference for θ. For discussion see Davidian and Carroll (in press).

The small σ assumption also provides an illustration of the relationship between variance function estimation and data transformations. Let $l(y, \varphi) = (y^{\varphi} - 1)/\varphi$, and consider the model

$$E\{l(Y_i, \varphi)\} = l(f(x_i, \beta), \varphi\}, \qquad \operatorname{var}\{l(Y_i, \varphi)\} = \sigma; \tag{4.5}$$

such "transform both sides" models are proposed and motivated by Carroll and Ruppert (1984). For $\sigma \approx 0$, $E(Y_i) \approx f(x_i, \beta)$ and $\operatorname{var}(Y_i) \approx \sigma f(x_i, \beta)^{(1-\varphi)}$, so in (1.2) we have $\theta \approx 1 - \varphi$. Thus, when the small σ assumption is relevant, (4.5) and (1.1), (1.2) represent approximately the same model.

5. APPLICATIONS AND FURTHER RESULTS

In Section 4 we constructed an asymptotic theory for and stated some general characteristics of regression-type estimators of θ. In this section we use the theory to exhibit the specific forms for the various estimators of Section 3 and compare and contrast their properties. In our investigation we rely on the simplifying assumptions implied by the theory of Section 4, in particular the small σ asymptotic approach in which $\sigma \rightarrow 0$ and $N \rightarrow \infty$. Throughout, define $v(i, \beta, \theta) = \log g(z_i, \beta, \theta)$, let $v_\theta(i, \beta, \theta)$ be the column vector of partial derivatives of v with respect to θ, let $\xi(\beta, \theta)$ be the covariance matrix of $v_\theta(i, \beta, \theta)$, and let $\tau(i, \beta, \theta) = \{1, v_\theta^t(i, \beta, \theta)\}^t$. For simplicity, assume that the errors $\{\varepsilon_i\}$ are iid with kurtosis κ; $\kappa = 0$ for normality.

5.1 Maximum Likelihood, Pseudolikelihood, Restricted Maximum Likelihood, and Weighted Squared Residuals

Writing $\eta = \log \sigma$, we have $T(x) = x^2$, $M_i = \exp(2\eta)g^2(z_i, \beta, \theta)$, $V_i = M_i^2$, $H_i^t = \partial M_i/\partial(\eta, \theta^t)^t$, and $E[\dot{T}\{d_i(\beta)\}\text{sign}(\varepsilon_i)] = 2E[Y_i - f(x_i, \beta)] = 0$, so $B_{3,N} \equiv 0$ regardless of the underlying distributions. If $h \rightarrow 0$ such that $N^{1/2}h \rightarrow 0$ for methods accounting for the effect of leverage, then all of these methods admit an expansion of the form (4.2) with $B_{3,N} = 0$. The expansion will be different depending on whether $\hat{\beta}_*$ is a generalized least squares estimator for β or full maximum likelihood, since the maximum likelihood estimator has an expansion quadratic in the errors and that of the generalized least squares estimator is linear in the $\{\varepsilon_i\}$ (see Carroll and Ruppert 1982b). The implication is that regression methods based on iterated weighted squared residuals and full maximum likelihood are different in general asymptotically. Regardless of the underlying distributions, for fixed σ, Davidian (1986) showed that the asymptotic covariance matrix of the former methods increases without bound as a function of σ whereas that of maximum likelihood remains bounded for all σ. Further, a simple comparison of the two covariances reveals that under reasonable conditions maximum likelihood has smaller asymptotic covariance as long as $\kappa \leq 2$. Although these facts may suggest a preference for full maximum likelihood even away from normality, the computational and model robustness considerations mentioned earlier may make this preference tenuous. Generalized least squares and maximum likelihood estimators for β both satisfy $\hat{\beta}_* - \beta = O_p(\sigma N^{-1/2})$, so if $\sigma \rightarrow 0$ or g does not depend on β, then $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$(2 + \kappa)\{4N\xi(\beta, \theta)\}^{-1}. \tag{5.1}$$

As mentioned in Section 3, under the small σ asymptotics of Theorem 3.3, the extended quasilikelihood estimator of θ is asymptotically equivalent to the estimators here with asymptotic covariance matrix (5.1). Thus, if g does not depend on β or $\sigma \rightarrow 0$, pseudolikelihood, weighted squared residuals, restricted maximum likelihood, maximum likelihood and, if $\sigma \rightarrow 0$, extended quasilikelihood, are all asymptotically equivalent. In addition, all of these estimators have influence functions that are linear in the squared errors, indicating substantial nonrobustness.

We may also observe that these methods are preferable to unweighted regression on squared residuals. Write (5.1) as

$$(\tfrac{1}{2} + \kappa/4)(WV^{-1}W)^{-1}, \tag{5.2}$$

where V is the $N \times N$ diagonal matrix with elements V_i and W is the $N \times p$ matrix with ith row H_i^t. For the unweighted estimator based on squared residuals, calculations similar to those above show that the asymptotic covariance matrix when either g does depend on β or $\sigma \rightarrow 0$ is given by

$$(\tfrac{1}{2} + \kappa/4)(W^tW)^{-1}(W^tVW)(W^tW)^{-1}. \tag{5.3}$$

The comparison between (5.2) and (5.3) is simply that of the Gauss–Markov theorem, so (5.2) is no larger than (5.3).

5.2 Logarithms of Absolute Residuals and the Effect of Inliers

We do not consider deletion of the few smallest absolute residuals. Here $T(x) = \log x$, so $\dot{T}(x) = x^{-1}$. Letting $\eta = \log \sigma$ and assuming iid errors we have $M_i = \eta + v(i, \beta, \theta) + E \log|\varepsilon|$, $V_i \equiv 1$, and $H_i = \tau(i, \beta, \theta)$. Under the assumption of symmetry of the errors, with g not depending on β or $\sigma \rightarrow 0$, tedious algebra shows that $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$\text{var}\{\log(|\varepsilon|^2)\}\{4N\xi(\beta, \theta)\}^{-1}. \tag{5.4}$$

The influence function for this estimator is linear in the logarithm of the absolute errors. This indicates nonrobustness more for inliers than for outliers, which at the very least is an unusual phenomenon. If the errors are not symmetric, then there will be an additional effect due to estimating β not present for the methods of Section 5.1, even if g does not depend on β.

5.3 Weighted Absolute Residuals

Assume that the errors are iid, and let $\exp(\eta) = \sigma E|\varepsilon|$. Consider the weighted estimator. We have $T(x) = x$, $\dot{T}(x) = 1$, $M_i = \exp(\eta)g(z_i, \beta, \theta)$, and $V_i = M_i^2$. Thus, if the errors are symmetrically distributed and either g does not depend on β or $\sigma \rightarrow 0$, $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$\{\delta/(1 - \delta)\}\{N\xi(\beta, \theta)\}^{-1}, \tag{5.5}$$

where $\delta = \text{var}|\varepsilon|$. The influence function for this estimator is linear in the absolute errors. By an argument similar to that at the end of Section 5.1, we may conclude that when the effect of $\hat{\beta}_*$ is negligible one should use a weighted estimator and iterate the method.

5.4 General Transformations

One may also consider other power transformations of absolute residuals. If $\lambda \neq 0$ is the power of absolute residuals on which the regression is based, then define η by $\exp(\lambda\eta) = \sigma^\lambda E(|\varepsilon|^\lambda)$ and $T(x) = x^\lambda$. Then $M_i = \exp(\lambda\eta)g^\lambda(z_i, \beta, \theta)$, $V_i = M_i^2$. Straightforward calculations show that if the errors are symmetric and either g does not depend on β or $\sigma \rightarrow 0$, then $\hat{\theta}$ is asymptotically normally distributed with mean θ and asymptotic covariance matrix

$$[\text{var}(|\varepsilon|^\lambda)/\{E(|\varepsilon|^\lambda)\}^2]\{\lambda^2N\xi(\beta, \theta)\}^{-1}, \tag{5.6}$$

with influence function linear in $|\varepsilon|^\lambda$. Thus (5.6) yields (5.1) when $\lambda = 2$ and (5.5) when $\lambda = 1$. For square root transformations, for example, $\lambda = \frac{1}{2}$, and from (5.1) and (5.6), the asymptotic relative efficiency of the square root transformation relative to pseudolikelihood under normal errors is .693; from (5.5), the efficiency relative to weighted absolute residuals is .791.

At this point it is worthwhile to mention that under the simplifying assumptions of our discussion, the precision of general regression estimators does not depend on σ, since a general expression such as (5.6) is independent of η. Thus how well we estimate θ in many practical cases will be approximately independent of σ. Furthermore, when the power of the mean model for variance (1.2) holds, $v_\theta(i, \beta, \theta) = \log \mu_i$, so $\xi(\beta, \theta)$ is the limiting variance of the $\{\log \mu_i\}$. From the general expression (5.6), the precision with which one can estimate θ depends only on the relative spread of the mean responses, not their actual sizes, and clearly this spread must be fairly substantial so that the spread of the logarithms of the means will be so as well. The implications are that for (1.2), the design will play an important role in efficiency of estimation of θ, and in some practical situations we may not be able to estimate θ well no matter which estimator we employ.

5.5 Comparison of Methods Based on Residuals

We assume that the errors are symmetric and iid and that either g does not depend on β or σ is small. By (5.1), (5.4), and (5.5), the asymptotic relative efficiency of the three methods depends only on the distribution of the errors. For normal errors, using absolute residuals results in a 12% loss in efficiency, whereas for standard double exponential errors there is a 25% gain in efficiency for using absolute residuals. For normal errors, the logarithm method represents a 59% loss of efficiency with respect to pseudolikelihood.

In Table 2 we present asymptotic relative efficiencies for various contaminated normal distributions. The asymptotic efficiency of the weighted absolute residual method to pseudolikelihood is the same as the asymptotic relative efficiency of the mean absolute deviation with respect to the sample variance for a single sample (see Huber 1981, p. 3); the first column of the table is thus identical to that of Huber. The table shows that, although at normality neither the absolute residuals nor the logarithm methods are efficient, a very slight fraction of "bad" observations is enough to offset the superiority of squared residuals in a dramatic fashion. For example, just 2 bad observations in 1,000 negate the superiority of squared residuals. If 1% or 5% of the data are "bad," absolute residuals and the logarithm method, respectively, show substantial gains over squared residuals. The implication is that, although it is commonly perceived that methods based on squared residuals are to be preferred in general, these methods can be highly nonrobust. Our formulation includes this result for maximum likelihood, showing its inadequacy under slight departures from the assumed distributional structure. We also include asymptotic relative efficiencies for appropriately weighted residual methods based on square, cube, and $\frac{2}{3}$ roots to pseudolikelihood using (5.6) and observe that these methods also exhibit comparative robustness to contamination.

Table 2. Asymptotic Relative Efficiency of Appropriately Weighted Regression Methods Based on a Function T of Absolute Residuals and the Method Based on Logarithms of Absolute Residuals With Respect to Appropriately Weighted Regression Methods Based on Squared Residuals for Underlying Contaminated Normal Error Distributions With Distribution Function F(x) = (1 − α)Φ(x) + αΦ(x/3)

Contamination fraction α	$T(x)$				
	x	$x^{2/3}$	$x^{1/2}$	$x^{1/3}$	$\log x$
.000	.876	.772	.693	.606	.405
.001	.948	.841	.756	.662	.440
.002	1.016	.906	.816	.715	.480
.010	1.439	1.334	1.216	1.075	.720
.050	2.035	2.100	1.996	1.823	1.220

5.6 Methods Based on Sample Standard Deviations

Assume that $m \geq 2$ replicate observations are available at each design point. In practice, m is usually small (see Raab 1981). We compare using absolute residuals with using sample standard deviations in the estimators of Section 3.1. One advantage of sample standard deviations over absolute residuals is that, because they do not use the mean function, they will be robust to misspecification of the model for the mean response; absolute residuals will not. We assume that one is fairly confident in the postulated form of the model, thus viewing methods based on sample standard deviations as not taking full advantage of the information available. For simplicity, assume that the errors are iid and symmetrically distributed and that either g does not depend on β or σ is small. If the errors are not symmetric and σ is not small or the variance depends on β, using sample standard deviations presumably will be more efficient than in the discussion that follows. This issue deserves further attention.

Let s_m^2 be the sample variance of m errors $\{\varepsilon_1, \ldots, \varepsilon_m\}$. It is easily shown by calculations analogous to those of Section 5.1 that replacing absolute residuals by sample standard deviations has the effect of changing the asymptotic covariance matrices (5.1), (5.4), and (5.5) to

$$\text{Pseudolikelihood: } \{(2 + \kappa) + 2/(m - 1)\}\{4N\xi(\beta, \theta)\}^{-1};$$

$$\text{Logarithm method: } m \operatorname{var}\{\log(s_m^2)\}\{4N\xi(\beta, \theta)\}^{-1};$$

Weighted absolute residuals:

$$\{m\delta_* / (1 - \delta_*)\}\{N\xi(\beta, \theta)\}^{-1}, \quad (5.7)$$

where $\delta_* = \operatorname{var}(s_m)$. Table 3 compares the asymptotic relative efficiencies of using sample standard deviations with using transformations of absolute residuals for various values of m when the errors are standard normal. The values in the table for $T(x) = x^2$ and x indicate that if the data are approximately normally distributed, using sample standard deviations can entail a loss in efficiency with respect to using residuals if m is small. For substantial replication ($m \geq 10$), using sample standard deviations pro-

Table 3. Asymptotic Relative Efficiency of Regression Methods Based on a Function T of Sample Standard Deviations Relative to Using Regression Methods Based on a Function T of Absolute Residuals under Normality for T(x) (weighted methods)

	T(x)		
m	x^2	*log x*	*x*
2	.500	.500	.500
3	.667	1.000	.696
4	.750	1.320	.801
⋮	⋮	⋮	⋮
9	.889	1.932	.986
10	.900	1.984	1.001
∞	1.000	2.467	1.142

duces a slight edge in efficiency with respect to weighted absolute residuals for $T(x) = x$.

The second column of Table 3 shows that, for the logarithm method, using sample standard deviations surpasses using residuals in terms of efficiency except when $m = 2$ and is more than twice as efficient for large m. In its raw form, $\log|r_i|$ is very unstable because, at least occasionally, $|r_i| \approx 0$, producing a wild "outlier" in the regression. The effect of using sample standard deviations is to decrease the possibility of such inliers; the sample standard deviations will likely be more uniform, especially as m increases. The implication is that the logarithm method should not be based on residuals unless remedial measures are taken. The suggestion to trim a few of the smallest absolute residuals before using this method is clearly supported by the theory; presumably, such trimming would reduce or negate the theoretical superiority of using sample standard deviations.

Table 4 contains the asymptotic relative efficiencies of weighted squared sample standard deviations and logarithms of these to weighted squared residuals under normality of the errors. The first column is the efficiency of Raab's method to pseudolikelihood, and the second column is the efficiency of the Rodbard and Frazier method to pseudolikelihood. The results of the table imply that using the Raab and Rodbard and Frazier methods, which are popular in the analysis of radioimmunoassay data, can entail a loss of efficiency when compared with methods based on weighted squared residuals. Davidian (1986) showed that the Rodbard and Frazier estimator can have a slight edge in efficiency over the weighted squared residuals methods for some highly contaminated normal distributions. From (5.7), the squared residual methods will be more efficient than Raab's method in the limit. Also note that the entries for $T(x) = x$ and $\log x$ in Table 3 for $m = \infty$ are the reciprocals of the first row of Table 2 and that the entries for the last row of Table 4 are 1.0; thus if both N and m grow large all the methods yield the same results.

Table 4. Asymptotic Relative Efficiency of Regression Methods Based on a Function T of Sample Standard Deviations Relative to Regression Methods Based on Weighted Squared Residuals Under Normal Errors

	T(x)	
m	x^2	*log x*
2	.500	.203
3	.667	.405
4	.750	.535
⋮	⋮	⋮
9	.889	.783
10	.900	.804
∞	1.000	1.000

Table 4 also addresses the open question as to whether Raab's method is asymptotically more efficient than the Rodbard and Frazier method for normally distributed data. The answer is a general yes, thus agreeing with the Monte Carlo evidence available when the variance is a power of the mean. The results of this section suggest that, in the case of assay data containing pairs for which only Y_{ij} is observed, an estimator for θ combining estimation based on residuals for the observations for which x_i is known and on standard deviations otherwise in an appropriately weighted fashion would offer some improvement over the methods currently employed (see Davidian et al. 1987).

6. DISCUSSION

In Section 4 we constructed a general theory of regression-type estimation for θ in the heteroscedastic model (1.1). This theory includes as special cases common methods described in Section 3 and allows for the regression to be based on absolute residuals from the current regression fit as well as sample standard deviations in the event of replication at each design point. Under various restrictions such as symmetry or small σ, when the variance function g does not depend on β, we showed in Sections 4 and 5 that we can draw general conclusions about this class of estimators as well as make comparisons among the various methods.

When employing methods based on residuals, one should weight the residuals appropriately and iterate the process. There can be large relative differences among the methods in terms of efficiency. Under symmetry of the errors, squared residuals are preferable for approximately normally distributed data, but this preference is tenuous, since these can be highly nonrobust under only slight departures from normality; methods based on logarithms or the absolute residuals themselves exhibit relatively more robust behavior. For the small amount of replication found in practice, using sample standard deviations rather than residuals can entail a loss in efficiency if estimation is based on the squares of these quantities or the quantities themselves. For the logarithm method based on residuals, trimming the smallest few absolute residuals is essential, since for normal data using sample standard deviations is almost always more efficient than using residuals, even for a small number of replicates. Popular methods in applications such as radioimmunoassay based on sample means and sample standard deviations can be less efficient than methods based on weighted squared residuals. In some instances, the precision with which we can estimate θ depends on the relative range of values of the mean responses, not their actual values, so immediate implications for design are suggested.

Efficient variance function estimation in heteroscedastic regression analysis is an important problem in its own right. There are important differences in estimators for variance when it is modeled parametrically.

APPENDIX A: PROOFS OF MAJOR RESULTS

We now present sketches of the proofs of the theorems of Section 4. Our exposition is brief and nonrigorous, as our goal is to provide general insights. In what follows, we assume that

$$N^{1/2}\begin{bmatrix}\hat{\eta}-\eta\\ \hat{\theta}-\theta\end{bmatrix}=O_p(1); \qquad (A.1)$$

under sufficient regularity conditions it is possible to prove (A.1). Such a proof would be long, detailed, and essentially noninformative; see Carroll and Ruppert (1982a) for a proof of $N^{1/2}$ consistency in a special case.

Sketch of the Proof of Theorem 4.1. From (4.1), a Taylor series, the fact that $E[T\{d_i(\beta)\}]=M_i$ and laws of large numbers, we have

$$0=N^{-1/2}\sum_{i=1}^{N}(H_i/V_i)[T\{d_i(\hat{\beta}_*)\}-M_i(\hat{\eta},\hat{\theta},\hat{\beta}_*)]+o_p(1). \qquad (A.2)$$

By the arguments of Ruppert and Carroll (1980) or Carroll and Ruppert (1982a),

$$\begin{aligned}&N^{-1/2}\sum_{i=1}^{N}(H_i/V_i)[T\{d_i(\hat{\beta}_*)\}-T\{d_i(\beta)\}]\\ &\quad=N^{-1/2}\sum_{i=1}^{N}(H_i/V_i)\dot{T}\{d_i(\beta)\}\{d_i(\hat{\beta}_*)-d_i(\beta)\}+o_p(1)\\ &\quad=B_{3,N}N^{1/2}(\hat{\beta}_*-\beta)+o_p(1). \qquad (A.3)\end{aligned}$$

Applying this result to (A.2) along with a Taylor series in M_i gives

$$\begin{aligned}0&=N^{-1/2}\sum_{i=1}^{N}C_i+(B_{2,N}+B_{3,N})N^{1/2}(\hat{\beta}_*-\beta)\\ &\quad-B_{1,N}N^{1/2}\begin{bmatrix}\hat{\eta}-\eta\\ \hat{\theta}-\theta\end{bmatrix}+o_p(1),\end{aligned}$$

which is (4.2).

Theorem 4.2 follows by a similar argument; in this case the representation (A.3) is unnecessary.

Sketch of the Proof of Theorem 4.3. We consider Theorem 4.2; the proof for Theorem 4.1 is similar. Recall here that (3.4) holds. In the following, all derivatives are with respect to the mean μ_i and the definitions of C_i and M_i are as in (4.4).

Assumption (iv) implies that

$$N^{1/2}\max_{1\le i\le N}|\overline{Y}_{i\cdot}-\mu_i|\xrightarrow{p}0,$$

so a Taylor series in η, θ, and $\overline{Y}_{i\cdot}$ gives

$$\begin{aligned}&B_{1,N}N^{1/2}\begin{bmatrix}\hat{\eta}-\eta\\ \hat{\theta}-\theta\end{bmatrix}\\ &\quad=N^{-1/2}\sum_{i=1}^{N}C_i-N^{-1/2}\sum_{i=1}^{N}(\dot{M}_iH_i/V_i)(\overline{Y}_{i\cdot}-\mu_i)\\ &\qquad+N^{-1/2}\sum_{i=1}^{N}\{(\dot{H}_i/V_i)-(\dot{V}_i/V_i)\}(\overline{Y}_{i\cdot}-\mu_i)+o_p(1). \qquad (A.4)\end{aligned}$$

Since $\overline{Y}_{i\cdot}-\mu_i=\sigma g(\mu_i,z_i,\theta)\bar{\varepsilon}_{i\cdot}\approx\lambda N^{-1/2}g(\mu_i,z_i,\theta)\bar{\varepsilon}_{i\cdot}$, where $\bar{\varepsilon}_{i\cdot}$ is the mean of the errors at x_i, we can write the last two terms on the right side of (A.4) as

$$\lambda N^{-1}\sum_{i=1}^{N}\bar{\varepsilon}_{i\cdot}(q_{i,1}+q_{i,2}C_i) \qquad (A.5)$$

for constants $\{q_{i,j}\}$. Since $\bar{\varepsilon}_{i\cdot}$ has mean 0, (A.5) converges in probability to 0 if $E(\bar{\varepsilon}_{i\cdot}C_i)=0$, which holds under the assumption of symmetry. Thus (A.5) converges to 0, which from (A.4) completes the proof. Note that if we drop the assumption of symmetry, from (A.5) the asymptotic normal distribution of $N^{1/2}(\hat{\theta}-\theta)$ will have mean

$$p\text{-}\lim_{N\to\infty}\{\lambda B_{1,N}^{-1}N^{-1}\sum_{i=1}^{N}(\bar{\varepsilon}_{i\cdot}C_iq_{i,2})\}.$$

APPENDIX B: CHARACTERIZATION OF RESTRICTED MAXIMUM LIKELIHOOD

Let $\hat{\beta}_*$ be a generalized least squares estimator for β. Assume first that g does not depend on β. Let the prior distribution for the parameters $\pi(\beta,\theta,\sigma)$ be proportional to σ^{-1}. The marginal posterior for θ is hard to compute in closed form for nonlinear regression. Following Box and Hill (1974) and Beal and Sheiner (1987), we have the linear approximation

$$f(x_i,\beta)\approx f(x_i,\hat{\beta}_*)+f_\beta(x_i,\hat{\beta}_*)^t(\beta-\hat{\beta}_*).$$

Replacing $f(x_i,\beta)$ by its linear expansion, the marginal posterior for θ is proportional to

$$p(\theta)=\frac{\left\{\prod_{i=1}^{N}g_i^2(\theta)\right\}^{-1/2}}{\hat{\sigma}_G^{(N-p)}(\theta)\{\text{Det }S_G(\theta)\}^{1/2}}, \qquad (B.1)$$

where

$$\hat{\sigma}_G^2(\theta)=(N-p)^{-1}\sum_{i=1}^{N}r_i^2/g_i^2(\theta),$$

$$S_G(\theta)=N^{-1}\sum_{i=1}^{N}f_\beta(x_i,\hat{\beta}_*)f_\beta(x_i,\hat{\beta}_*)^t/g_i^2(\theta),$$

and where Det A = determinant of A. If the variances depend on β, we extend the Bayesian arguments by replacing $g_i(\theta)$ by $g(z_i,\hat{\beta}_*,\theta)$.

Let H be the hat matrix H evaluated at $\hat{\beta}_*$ and let $h_{ii}=h_{ii}(\hat{\beta}_*,\theta)$. From (3.1), pseudolikelihood solves in (θ,σ)

$$\begin{aligned}&\sum_{i=1}^{N}[r_i^2/\{\sigma^2g^2(z_i,\hat{\beta}_*,\theta)\}]\left[\frac{1}{v_\theta(i,\hat{\beta}_*,\theta)}\right]\\ &\quad=\sum_{i=1}^{N}\left[\frac{1}{v_\theta(i,\hat{\beta}_*,\theta)}\right]. \qquad (B.2)\end{aligned}$$

Since H is idempotent, the left side of (B.2) has approximate expectation

$$\sum_{i=1}^{N}\left[\frac{1-p/N}{v_\theta(i,\hat{\beta}_*,\theta)(1-h_{ii})}\right]. \qquad (B.3)$$

To modify pseudolikelihood to account for loss of degrees of freedom, equate the left side of (B.2) to (B.3). From matrix computations as in Nel (1980), this can be shown to be equivalent to restricted maximum likelihood.

[*Received July 1986. Revised April 1987.*]

REFERENCES

Amemiya, T. (1977), "A Note on a Heteroscedastic Model," *Journal of Econometrics*, 6, 365-370. [See also *Corrigenda*, Vol. 8, p. 265.]

Beal, S. L., and Sheiner, L. B. (1987), "Heteroscedastic Nonlinear Regression With Pharmokinetic Type Data," preprint.

Box, G. E. P. (1986), "Studies in Quality Improvement: Signal to Noise Ratios, Performance Criteria and Statistical Analysis: Part I," Report 11, University of Wisconsin–Madison, Center for Quality and Productivity Improvement.

Box, G. E. P., and Hill, W. J. (1974), "Correcting Inhomogeneity of Variance With Power Transformation Weighting," *Technometrics*, 16, 385–389.

Box, G. E. P., and Meyer, R. D. (1986), "Dispersion Effects From Fractional Designs," *Technometrics*, 28, 19–28.

Box, G. E. P., and Ramirez, J. (1986), "Studies in Quality Improvement: Signal to Noise Ratios, Performance Criteria and Statistical Analysis: Part II," Report 12, University of Wisconsin–Madison, Center for Quality and Productivity Improvement.

Carroll, R. J. (1982), "Adapting for Heteroscedasticity in Linear Models," *The Annals of Statistics*, 10, 1224–1233.

——— (1987), "The Effects of Variance Function Estimation on Prediction and Calibration: An Example," preprint.

Carroll, R. J., and Ruppert, D. (1982a), "Robust Estimation in Heteroscedastic Linear Models," *The Annals of Statistics*, 10, 429–441.

——— (1982b), "A Comparison Between Maximum Likelihood and Generalized Least Squares in a Heteroscedastic Linear Model," *Journal of the American Statistical Association*, 77, 878–882.

——— (1984), "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Association*, 79, 321–328.

Carroll, R. J., Wu, C. F. J., and Ruppert, D. (1987), "Variance Expansion and the Bootstrap in Generalized Least Squares," preprint.

Davidian, M. (1986), "Variance Function Estimation in Heteroscedastic Regression Models," unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill, Dept. of Statistics.

Davidian, M., and Carroll, R. J. (in press), "A Note on Extended Quasi-likelihood," *Journal of the Royal Statistical Society*, Ser. B, 50.

Davidian, M., Carroll, R. J., and Smith, W. (1987), "Variance Functions and the Minimum Detectable Concentration in Assays," unpublished manuscript.

Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association*, 81, 709–721.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986), "Some New Methods for Weighted Regression When There Are Possible Outliers," *Technometrics*, 28, 219–230.

Glejser, H. (1969), "A New Test for Heteroscedasticity," *Journal of the American Statistical Association*, 64, 316–323.

Gong, G., and Samaniego, F. J. (1981), "Pseudo-maximum Likelihood Estimation: Theory and Applications," *The Annals of Statistics*, 9, 861–869.

Harvey, A. C. (1976), "Estimating Regression Models With Multiplicative Heteroscedasticity," *Econometrica*, 44, 461–465.

Harville, D. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 79, 302–308.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley.

Jacquez, J. A., Mather, F. J., and Crawford, C. R. (1968), "Linear Regression With Non-constant, Unknown Error Variances: Sampling Experiments With Least Squares and Maximum Likelihood Estimators," *Biometrics*, 24, 607–626.

Jobson, J. D., and Fuller, W. A. (1980), "Least Squares Estimation When the Covariance Matrix and Parameter Vector Are Functionally Related," *Journal of the American Statistical Association*, 75, 176–181.

McCullagh, P. (1983), "Quasi-likelihood Functions," *The Annals of Statistics*, 11, 59–67.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, New York: Chapman & Hall.

Nel, D. G. (1980), "On Matrix Differentiation in Statistics," *South African Statistical Journal*, 14, 87–101.

Nelder, J. A., and Pregibon, D. (1987), "An Extended Quasi-likelihood Function," *Biometrika*, 74, 221–232.

Patterson, H. D., and Thompson, R. (1971), "Recovery of Inter-block Information When Block Sizes Are Unequal," *Biometrika*, 58, 545–554.

Raab, G. M. (1981), "Estimation of a Variance Function, With Application to Radioimmunoassay," *Applied Statistics*, 30, 32–40.

Rodbard, D., and Frazier, G. R. (1975), "Statistical Analysis of Radioligand Assay Data," *Methods of Enzymology*, 37, 3–22.

Rosenblatt, J. R., and Spiegelman, C. H. (1981), Discussion of "A Bayesian Analysis of the Linear Calibration Problem," by W. G. Hunter and W. F. Lamboy, *Technometrics*, 23, 329–333.

Rothenberg, T. J. (1984), "Approximate Normality of Generalized Least Squares Estimates," *Econometrica*, 52, 811–825.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Sqaures Estimation in the Linear Model," *Journal of the American Statistical Association*, 77, 828–838.

Sadler, W. A., and Smith, M. H. (1985), "Estimation of the Response–Error Relationship in Immunoassay," *Clinical Chemistry*, 31, No. 11, 1802–1805.

Stefanski, L. A., and Carroll, R. J. (1985), "Covariate Measurement Error in Logistic Regression," *The Annals of Statistics*, 13, 1335–1351.

Theil, H. (1971), *Principles of Econometrics*, New York: John Wiley.

Watters, R. L., Carroll, R. J., and Spiegelman, C. H. (1987), "Error Modeling and Confidence Interval Estimation for Inductively Coupled Plasma Calibration Curves," preprint.

Williams, J. S. (1975), "Lower Bounds on Convergence Rates of Weighted Least Squares to Best Linear Unbiased Estimators," in *A Survey of Statistical Design and Linear Models*, ed. J. N. Srivastava, Amsterdam: North-Holland, pp. 555–570.

Wolter, K. M., and Fuller, W. A. (1982), "Estimation of Nonlinear Errors-in-Variables Models," *The Annals of Statistics*, 10, 539–548.

Chapter 3
Epidemiology

By Laurence Freedman, Mitchell H. Gail, and Dale L. Preston

About the Authors. Laurence (Larry) Freedman is currently Director of the Biostatistics Unit at the Gertner Institute for Epidemiology at Tel Hashomer, Israel, where he directs a research and consulting program in biostatistics and advises the government on public health policy. He has published widely in the biostatistical and medical literature, was a founding coeditor of *Statistics in Medicine* and also served as coeditor of *Biometrics*. He founded the Eastern Mediterranean Region of the International Biometric Society and served as its first president. He first met Raymond Carroll in the late 1980s after taking a position as visiting scientist at the US National Cancer Institute (NCI). He eventually stayed at NCI for 10 years and during that time developed a strong working partnership with Ray, investigating statistical problems in nutrition research. The legacy was handed on, and after he left NCI to move to Israel, he and Ray continued to work with the NCI group. The collaboration continues to this day.

Mitchell H. Gail is a Senior Investigator in the Biostatistics Branch of the Division of Cancer Epidemiology and Genetics, NCI. His current interests include methodologic and applied research on the development and use of risk models, including models with genetic markers, for disease prevention and clinical epidemiology. He has published on statistical methods and their applications in clinical trials, epidemiology, and laboratory science. He served as President of the American Statistical Association and of the Eastern North American Region of the International Biometric Society and is a member of the Institute of Medicine. He first collaborated with Raymond Carroll in 1992 and has been a friend and collaborator since then. They have worked together on the method of back-calculation for projecting AIDS incidence, measurement error in case–control studies, the design of community intervention trials, measures of agreement based on categories defined by quantiles of marginal distributions, the kin-cohort method for estimating the penetrance of measured mutations, and meta-analytic analysis of surrogate outcomes in clinical trials.

Dale L. Preston spent almost 25 years at the Radiation Effects Research Foundation in Hiroshima, Japan, where he worked as chief of the Department of Statistics and led efforts to develop innovative methods for modeling the long-term health effects of radiation exposure in the atomic bomb survivors. He has had a long association with both the Radiation Epidemiology Branch and the Biostatistics Branch at NCI and is currently working as an independent consultant on studies of radiation effects in the atomic bomb survivors, populations exposed to radiation as a consequence of plutonium production in the Russian Southern Urals, and a large cohort of US radiation technologists. He is Fellow of the American Statistical Association, has served as an associate editor of *Radiation Research*, has been a member of the International Commission on Radiation Protection and has served as a consultant at the National Academy of Sciences committee on the Biological Effects of Ionizing Radiation and the United Nations Scientific Committee on the Effects of Atomic Radiation. Although he has never worked directly with Ray, he has spent considerable time studying Ray's work on measurement error and is currently involved in efforts to make use of complex Monte-Carlo-based dosimetry systems to allow for effects of dose measurement error on the uncertainty of radiation risk estimates.

Selected Papers on Epidemiology

[EPI-1]-[64] Carroll, R. J., Wang, C. Y., and Wang, S. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90, 157–169.

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_3,

[EPI-2]-[86] Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91, 722–732.

[EPI-3]-[135] Kipnis V., Midthune D., Freedman L.S., Bingham S., Schatzkin A., Subar A. and Carroll R.J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology*, 153, 394–403.

[EPI-4]-[39] Mallick, B., Hoffman, F. O., and Carroll, R. J. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. *Biometrics*, 58, 13–20.

[EPI-5]-[230] Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). Structure of dietary measurement error: results of the OPEN biomarker study. *American Journal of Epidemiology*, 158, 14–21.

[EPI-6]-[56] Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, 92, 399–418.

[EPI-7]-[56] Chatterjee, N., Kalaylioglu, Z. and Carroll, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, Interactions and joint effects. *Genetic Epidemiology*, 28, 138–156.

Raymond Carroll's work has had an important impact on epidemiologic research. This article reviews contributions to theory for the case–control design and to methods for nutritional and radiation epidemiology. Some of these contributions build on Ray's broad-ranging research on regression analysis, measurement error, and missing data problems. Ray has been a welcome visitor at the U. S. National Institutes of Health (NIH), first with the National Heart, Lung, and Blood Institute and later with the National Cancer Institute (NCI), both as a Visiting Scientist and Guest Researcher and as a friendly collaborator who drops by from time to time. At NIH, he has given valuable advice on a wide range of topics and collaborated on many projects not covered by this article, including the analysis of survival data with informative censoring (Wu and Carroll, 1988 [OW-2]), the design of community intervention trials (Gail et al., 1996), the design and analysis of the "kin-cohort" design for genetic epidemiology (Carroll et al., 2000; Gail et al., 1999), the meta-analysis of surrogate endpoints (Gail, 2000), and agreement of exposure assessments based on quantile groupings (Borkowf et al., 1997), among many others.

Theory of the Case–Control Study and Its Extensions

During a visit to NCI in 1990–91, Ray took an interest in the effects of measurement error on covariates measured in case–control studies (Carroll, Gail, and Lubin, 1993). He quickly mastered the theory of estimation for case–control studies and the subtleties of arguments (Prentice and Pyke, 1979), showing that one can analyze case–control data under a logistic model as if the data are from a prospective cohort design. Remarkably, the resulting estimates of log relative odds ratios are maximum likelihood under retrospective sampling, and their covariances, estimated as if the study were prospective, are also correct; only the variances of intercepts need adjustment. Wondering whether or not the prospective approach also applies to variants of the case–control design, Ray and his colleagues (Carroll, Wang and Wang, 1995 [EPI-1]) prove that the resulting estimates are consistent and give conditions under which covariances are consistently estimated, which is often the case. They prove that, if these conditions are violated, the prospective estimators of the variances of log odds ratios overestimate the true variance from the retrospective design, thus leading to conservative assessments. This work justifies using the prospective approach for case–control studies with differential and non-differential measurement error, with repeated exposure measurement designs and with missingness by design. Related work shows that the distribution of covariates in the general population is only identifiable from case–control data if the probability of disease is also known and that the results of Prentice and Pyke (Prentice and Pyke, 1979) only hold if the assumed distribution of covariates in the general population is unrestricted (Roeder, Carroll and Lindsay, 1996 [EPI-2]).

This latter point has important implications for genetic epidemiology, where one might wish to make assumptions on the distribution of covariates in the general population to gain efficiency in a case–control analysis. For example, Chatterjee and Carroll (Chatterjee and Carroll, 2005 [EPI-6]) assume that a genetic variant (G) is distributed independently of an environmental risk factor (E) in the population and develop methods of inference for the logistic model of risk. Although previous work exploits this assumption to test for G×E interactions, the work of Chatterjee and Carroll provides for a full logistic analysis of all risk factors and the G×E interactions. Their profile likelihood methods are also applicable to parametric assumptions on the covariate distribution, such as Hardy–Weinberg assumptions used to estimate diplotype distributions (Spinka, Carroll, and Chatterjee, 2005). This approach has greatly broadened the tools available for logistic analysis of case–control data whenever restrictive assumptions on the covariate distribution are justified; often those assumptions will increase the efficiency of the analysis.

Nutritional Epidemiology

It was David Byar who first introduced Ray to the field of nutritional epidemiology. Byar was the Chief of the Biometry Branch, the group of statisticians in Division of Cancer Prevention of NCI. In the mid- to late-1980s, that Division had as one of its main projects the investigation of the hypothesis that dietary fat intake is a major cause of breast cancer. This hypothesis was supported by animal studies, by ecologic studies and by a number of case–control studies, but the leaders of the Division wanted to conduct a very large randomized dietary intervention trial to prove the hypothesis. The projected cost of such a trial was very high, and the proposal met with strong opposition from some researchers, particularly those who were conducting large observational cohort studies that could potentially address the hypothesis. Byar, however, saw problems in the design of such nutritional cohort studies, not the least of which was the problem of dietary measurement error. He decided to highlight this issue and conduct a workshop on the statistical implications and methods of handling measurement error, to which he invited Ray, whom he recognized as one of the leaders in this field. Ray delivered a review paper at the workshop that was subsequently published (Carroll, 1989) together with others from the workshop in a special issue of *Statistics in Medicine*. This was Ray's introduction to a long and fruitful collaboration with the statisticians at the Biometry Branch of the Division of Cancer Prevention at NCI.

The watershed in this collaboration came in the mid-late 1990s. Up to that time, the accepted method of evaluating a dietary questionnaire of the type used in large cohort studies was to conduct a sub-study comparison of the intakes reported on the questionnaire with those reported from more detailed and shorter-term self-report instruments, such as multiple-day food diaries or 24-h recalls. Methods of statistical adjustment of the results from the cohort study were then based on the assumption that such self-report instruments were unbiased measures of true intake and had errors that were independent of true intake and, most importantly, independent of the errors in the cohort study questionnaire. However, many nutritionists in the field were dubious regarding this set of assumptions. Unfortunately, no definitive data were available to prove or disprove their doubts. In 1997, Ray participated in an informal meeting of statisticians at the Biometry Branch to discuss this problem. Together, they decided to investigate through calculation what would be the implications if the assumptions were contravened. It soon became clear that the crucial assumption was the independence of the errors of the questionnaire and the more detailed reference instrument. If these errors were positively correlated, then the questionnaires would be made to look less error-prone than they really were. We published this result (Kipnis et al., 1999), but there were still not sufficient data to say whether this was or was not a real concern.

By this time, some validation studies were being conducted using not only more detailed self-report instruments as reference measures but also biomarkers that had been shown to provide reliable measures of dietary intake (called "recovery" biomarkers), the foremost among them at that time being 24-h urinary nitrogen as a measure of protein intake. Availability of this extra measure, which could indeed be safely assumed to have errors independent of the errors in a questionnaire, in a study conducted in the UK, allowed empirical testing of the concerns explained above. The results, reported in Kipnis et al. (2001 [EPI-3]), confirmed the previous doubts regarding over-optimistic assessments of the questionnaire's accuracy. For protein intake, the correlation of the questionnaire report with true intake was estimated to be 0.28 using the biomarker, compared to the over-optimistic 0.43 using the more-detailed self-report. This result meant that under the previous erroneous assumptions, planned sample sizes of cohort studies were about half of that really required.

This preliminary work provided the impetus for NCI to approve a large validation study of dietary self-report instruments, which became known as the OPEN (Observing Protein and Energy Nutrition) Study, using as reference measures three recovery biomarkers for measuring energy intake, protein intake, and potassium intake, respectively. The results of the OPEN study (Kipnis et al., 2003 [EPI-5]) reconfirmed and extended the earlier results (Kipnis et al., 2001 [EPI-3]), and its design also allowed confirmation that the errors of the questionnaire and more-detailed self-report were substantially correlated. The success and impact of the study has inspired a new generation of validation studies using biomarkers.

More recently, Ray's insights have led to three further important developments in the field. The first concerns the use of so-called concentration biomarkers (those markers of dietary intake that do not enjoy the nice properties of recovery biomarkers). Ray had, early on in our work, insisted that these biomarkers could surely be used to advantage in nutritional epidemiology, but for a long time they were used only for informal checking on self-report instruments without any solid theoretical underpinning. Only recently has the use of such biomarkers been proposed for increasing the precision of relative risk estimates in nutritional epidemiology (Freedman et al., 2011; Prentice et al., 2009). The second concerns the idea of combining different types of self-report instruments, again for the purposes of increasing precision in estimating disease risk (Carroll et al., 2012). The third concerns his development of technology for multivariate analysis of many foods or nutrients (Zhang et al., 2011) that will have applications in dietary surveillance, dietary patterns, and other areas of nutritional epidemiology.

Ray's contributions to the work of statisticians and epidemiologists at the NCI and thus to nutritional epidemiology have been pivotal, as recognized by his being given the prestigious position of final author in the highly cited publications (Kipnis et al., 2001) and (Kipnis et al., 2003).

Radiation Epidemiology

In 1995, Ray, along with his colleagues David Ruppert and Len Stefanski, published the first edition of their landmark work on the effect of measurement error in nonlinear models (Carroll, Ruppert, and Stefanski, 1995). Around this time, statisticians such as Ethel Gilbert, Don Pierce, Duncan Thomas, and Dan Stram, among others, studied the health effects of radiation in atomic bomb survivors, in workers in nuclear facilities, and in persons exposed to fallout from US nuclear weapons tests. These researchers began to consider how to account for errors in radiation dosimetry to improve estimates of the effects of radiation on health outcomes. During this period Owen Hoffman, who had worked on a number of environmental dose reconstruction projects, also collaborated with Charles Land and others to develop empirical, Monte-Carlo-based methods to characterize uncertainty in estimates of the "probability of causation" (or assigned share) (National Research Council, 1984) for cancers occurring after radiation exposure. In 1997, Owen Hoffman and Elaine Ron organized a workshop on *Uncertainties in Radiation Dosimetry and Their Impact on Dose–Response Analysis* (Ron and Hoffman, 1999) that brought together scientists interested in this topic. Ray gave a keynote presentation on the limitations of replacement methods (such as regression calibration) and proposed the use of Bayesian methods with subjectively derived dose estimates (Carroll, 1999). His presentation was highly influenced by Owen Hoffman's work and, as a direct consequence of their interaction at this workshop, Ray began collaborations with Owen and others on topics related to the effect of dose uncertainty in radiation risk estimation.

Following the 1997 conference, Ray worked with Elaine Ron, Jay Lubin, Dan Schafer, and others on a reanalysis of data from an important study of scalp irradiation and thyroid cancer that lead to a presentation (Carroll et al., 2000b) and a pair of papers (Lubin et al., 2004; Schafer et al., 2001) that develop a generalized regression calibration method to estimate radiation dose–response on thyroid cancer risk. This method accounts for dose uncertainties that reflected a combination of assignment (Berkson) and classical measurement errors. A primary finding of this work is that adjustment for dosimetry errors had little impact on the magnitude of the risk estimate or its standard error in this study. This was believed to be because dose uncertainty in this study was largely the result of Berkson errors and the response was linear in dose.

Working with Owen Hoffman and others, Ray developed Bayesian methods to allow for the effects of dose uncertainty on thyroid cancer risk estimates in victims of radioactive fallout from nuclear weapons tests carried out at the Nevada test site. In the spirit of Ray's talk at the 1997 workshop, this work went beyond the use of simple replacement estimators. The resulting paper (Mallick, Hoffman and Carroll, 2002 [EPI-4]) develops Bayesian methods based on a latent variable model for the error structure that allows for both classical and Berkson error (Reeves et al., 1998). This novel method assumes that the total error variance is known for each individual but that the proportion of the total attributable to classical measurement or Berkson error is unknown. Priors are defined for the risk parameters (for parametric or monotone, semi-parametric dose–response models) and for the fraction of the total dose error arising due to Berkson error. A Markov chain Monte-Carlo (MCMC) method is used to obtain samples from the posterior densities. This paper ends with a brief discussion of the "open problem" of dealing with the effect of shared uncertainties on risk estimates. Uncertainties that are common to some or all individuals are said to be "shared uncertainties." This type of uncertainty is an important aspect of many complex radiation dosimetry systems. Examples of shared uncertainties include misspecification of the location of the hypocenters for the atomic bombs or misspecification of parameters in a model for plutonium clearance from the lung. In a subsequent paper (Li et al., 2007), Ray and his colleagues tackle the problem of shared uncertainties that arise as a result of the use of model-based dose surrogates. They show that simple replacement methods fail to account for the potentially complex effects of the correlated errors that arise as a consequence of shared uncertainties and developed both Bayesian (MCMC-based) and frequentist (Monte-Carlo EM) approaches to dealing with such shared errors.

In view of his longstanding interest in innovative research on characterizing and dealing with the effect of measurement error, it is not surprising that Ray Carroll has played an important role in highlighting the complex nature of the uncertainties in radiation dose estimation and the potential impact of these uncertainties on radiation risk estimates while helping to develop innovative statistical methods to address the problem. As radiation dosimetrists develop increasingly complex systems that often involve the construction of samples from the (possible) distribution of an individual's "true" dose given information on characteristics of the individual and the nature of the exposures, there is a need for additional statistical methodology to incorporate this information into risk estimation. While much remains to be done, Ray and his colleagues have taken important steps in this direction. Although not discussed in this section, other aspects of Ray's work are quite relevant to studies of radiation effects, including work on estimating response thresholds in the presence of dose-measurement error (Kuchenhoff and Carroll, 1997) and the use of SIMEX methods in dealing with dose uncertainty (Carroll et al., 2006; Kukush et al., 2011).

References

Other publications by Ray Carroll cited in this chapter.

Borkowf, C. B., Gail, M. H., Carroll, R. J., and Gill, R. D. (1997). Analyzing bivariate continuous data grouped into categories defined by empirical quantiles of marginal distributions. *Biometrics*, 53, 1054–1069.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8, 1075–1093.

Carroll, R. J. (1999). Risk assessment with subjectively derived doses. In *Uncertainties in Radiation Dosimetry and Their Impact on Dose-Response Analysis*, E. Ron and F. O. Hoffman (eds), 37–51. Bethesda, MD: National Cancer Institute Press.

Carroll, R. J., Gail, M. H., Benichou, J., and Pee, D. (2000). Score tests for familial correlation in genotyped-proband designs. *Genetic Epidemiology*, 18, 293–306.

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association*, 88, 185–199.

Carroll, R. J., Midthune, D., Subar, A. F., Shumakovich, M., Freedman, L. S., Thompson, F. E., and Kipnis, V. (2012). Taking advantage of the strengths of two different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *American Journal of Epidemiology*, 175, 340–347.

Carroll, R. J., Ruppert, D., and Stefanski, L. J. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D., Stefanski, L. J., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: a Modern Perspective, 2nd edition*. Boca Raton: CRC Press.

Carroll, R. J., Schafer, D. W., Lubin, J. H., Ron, E., and Stovall, M. (2000b). Thyroid cancer after scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Radiation Research*, 154, 721–722; discussion 723–724.

Freedman, L. S., Midthune, D., Carroll, R. J., Tasevska, N., Schatzkin, A., Mares, J., Tinker, L., Potischman, N., and Kipnis, V. (2011). Using regression calibration equations that combine self-reported intake and biomarker measures to obtain unbiased estimates and more powerful tests of dietary associations. *American Journal of Epidemiology*, 174, 1238–1245.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15, 1069–1092.

Gail, M. H., Pee, D., Benichou, J., and Carroll, R. (1999). Designing studies to estimate the penetrance of an identified autosomal dominant mutation: Cohort, case-control, and genotyped-proband designs. *Genetic Epidemiology*, 16, 15–39.

Gail M. H., Pfeiffer, R., Van Houwelingen, H. C., Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1, 231–246.

Kipnis, V., Carroll, R. J., Freedman, L. S., and Li, L. (1999). Implications of a new dietary measurement error model for estimation of relative risk: Application to four calibration studies. *American Journal of Epidemiology*, 150, 642–651.

Kuchenhoff, H. and Carroll, R. J. (1997). Segmented regression with errors in predictors: Semi-parametric and parametric methods. *Statistics in Medicine*, 16, 169–188.

Kukush, A., Shklyar, S., Masiuk, S., Likhtarov, I., Kovgan, L., Carroll, R. J., and Bouville, A. (2011). Methods for estimation of radiation risk in epidemiological studies accounting for classical and berkson errors in doses. *International Journal of Biostatistics*, 7, Article 15.

Li, Y., Guolo, A., Hoffman, F. O., and Carroll, R. J. (2007). Shared uncertainty in measurement error problems, with application to Nevada Test Site fallout data. *Biometrics*, 63, 1226–1236.

Lubin, J. H., Schafer, D. W., Ron, E., Stovall, M., and Carroll, R. J. (2004). A reanalysis of thyroid neoplasms in the Israeli tinea capitis study accounting for dose uncertainties. *Radiation Research*, 161, 359–368.

Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M., and Carroll, R. J. (2001). Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Biometrics*, 57, 689–697.

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology*, 29, 108–127.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175–188.

Zhang, S. J., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L. S., and Carroll, R. J. (2011). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Annals of Applied Statistics*, 5, 1456–1487.

Publications by other authors cited in this chapter.

National Research Council (1984). NAS/NRC Committee on Radioepidemiological Tables. Assigned Share for Radiation as a Cause of Cancer – Review of Radioepidemiological Tables. Assigning Probabilities of Causation (Final Report). Washington, DC: National Academies Press.

Prentice, R. L., Huang, Y., Tinker, L. F., Beresford, S. A., Lampe, J. W., and Neuhouser, M. L. (2009). Statistical aspects of the use of biomarkers in nutritional epidemiology research. *Statistics in Biosciences*, 1, 112–123.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.

Reeves, G. K., Cox, D. R., Darby, S. C., and Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*, 17, 2157–2177.

Ron, E. and Hoffman, F. O. (eds) (1999). *Uncertainties in Radiation Dosimetry and Their Impact on Dose-Response Analysis.* Bethesda, MD: National Cancer Institute.

Prospective Analysis of Logistic Case-Control Studies

R. J. Carroll, Suojin Wang, and C. Y. Wang*

In a classical case-control study, Prentice and Pyke proposed to ignore the study design and instead base estimation and inference on a random sampling (i.e., prospective) formulation. We generalize this prospective formulation of case-control studies to include multiplicative models, stratification, missing data, measurement error, robustness, and other examples. The resulting estimators, which ignore the case-control study aspect and instead are based on a random-sampling formulation, are typically consistent for nonintercept parameters and are asymptotically normally distributed. We derive the resulting asymptotic covariance matrix of the parameter estimates. The covariance matrix obtained by ignoring the case-control sampling scheme and using prospective formulas instead is shown to be at worst asymptotically conservative and asymptotically correct in a variety of problems; a simple sufficient condition guaranteeing the latter is obtained.

KEY WORDS: Asymptotics; Corrections for attenuation; Differential measurement error; Estimating equations; Measurement error; Missing data; Robust estimates.

1. INTRODUCTION

In a classical prospective logistic regression study, a random sample from a source population is taken and the status of a binary outcome D is ascertained, along with the values of covariates (Z, X), these being related via the logistic regression model

$$\mathrm{pr}(D = 1 \mid Z, X) = H(\theta_0^* + \theta_{11}' Z + \theta_{12}' X), \qquad (1)$$

where $H(\cdot)$ is the logistic distribution function. The classical case-control study (choice-based sample in econometrics) begins with the model (1), but instead uses retrospective sampling. Specifically, one first obtains a set of cases ($D = 1$) and controls ($D = 0$), and then samples from within the cases and controls to observe the covariates. The analysis of case-control studies of this type was described by Prentice and Pyke (1979), who showed that if one ignored the case-control sampling scheme and analyzed the data as if it came from a prospective sampling scheme, then the resulting estimates of $(\theta_{11}, \theta_{12})$ are consistent and the usual standard errors are asymptotically correct.

For prospective logistic regression studies, many other types of analyses and sampling schemes are possible. Here are a few examples:

- One might replace the classical logistic regression parameter estimates by robust methods of estimation (Copas 1988; Carroll and Pederson 1993, Künsch, Stefanski, and Carroll 1989).
- When X is measured with error, there is a large literature dealing with techniques for measurement error corrections in logistic regression (e.g., Carroll and Stefanski 1994; Rosner, Willett, and Spiegelman 1989; Satten and Kupper 1993; Stefanski and Carroll, 1987).
- In problems with partially missing data, one can use likelihood techniques (Little and Rubin 1987) or unbiased estimating equations due to Robins, Rotnitzky, and Zhao (1994).

Although the prospective analyses of these prospective techniques have been worked out, there is to date no corresponding general theory for whether they even lead to consistent estimates when applied to case-control studies and, if they do, whether these prospectively calculated standard errors are asymptotically correct in case-control studies. Our aim is to provide one version of such a theory, and in particular to answer the question: When can prospective analyses be used in case-control studies without having to adjust for the retrospective sampling structure?

We will show that, in general, using prospectively derived standard errors is at worst asymptotically conservative; that is, the standard errors are at worst too large. In addition, we derive a simple sufficient condition guaranteeing that prospective standard errors are asymptotically correct.

In the Appendix we sketch an informal argument derived from a semiparametric perspective that suggests that prospectively computed standard errors are retrospectively correct whenever the distribution of (Z, X) is left unrestricted. Much of this article is a formalization of this argument, along with consideration of cases that are not so easily categorized. The key feature of our analysis is that we start with a general class of unbiased estimating equations, instead of working with specific examples. The results allow for general patterns of missing data as well as for stratified studies. The asymptotic distribution theory is almost trivial to derive in this general framework, thus facilitating the identification of a simple sufficient condition for checking whether prospectively derived standard errors are asymptotically correct. Our results apply not only to the linear logistic model (1), but also to the multiplicative model of Weinberg and Wacholder (1993).

Here is an outline of the article. Section 2 reviews the known results on estimating equations, estimating functions, and sandwich covariance estimation in prospective studies. Using this background, we provide a simple argument show-

* R. J. Carroll is Professor, Suojin Wang is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. C. Y. Wang is Assistant Member, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98104. R. J. Carroll and C. Y. Wang's research was supported by Grant CA-57030 from the National Cancer Institute. Suojin Wang's research was supported by Texas Advanced Research Program Grant 160802 and National Science Foundation Grant DMS-9200610. The authors are grateful to Mitch Gail and Sholom Wacholder, whose questions led to the general formulation of this article, and to Jamie Robins, for help with the argument in Section A.1. They also thank two referees for their comments and an associate editor who read the article in painstaking detail and suggested many useful changes.

© 1995 American Statistical Association
Journal of the American Statistical Association
March 1995, Vol. 90, No. 429, Theory and Methods

ing why prospective standard errors are at worst asymptotically conservative when applied to case-control studies. Section 3 defines the general estimating equation framework allowing for missing and mismeasured data. Section 4 states the two main results.

The rest of the article considers important special cases, the results for which are new with one exception. Section 5 considers studies with no missing or mismeasured data. In work generalizing that of Weinberg and Wacholder (1993) for multiplicative models and Wang and Carroll (1993, 1995) for robust logistic estimation, we show that essentially any prospectively motivated estimator can be used retrospectively, with asymptotically correct standard errors.

Further sections deal with problems of missing and mismeasured data. Section 6 applies the general theory to a modification of the unbiased estimating equations proposed by Robins et al. (1994) for case-control studies with mismeasured data when there is a validation subsample, allowing for differential measurement error (formally defined in Sec. 6). Section 7 considers measurement error models with nondifferential measurement error in which a validation study can be done, using the simple prospective likelihood methods due to Satten and Kupper (1993). Section 8 discusses measurement error models when validation is not possible, and uses prospective correction-for-attenuation methods. In all three cases, prospective standard errors are asymptotically correct retrospectively.

Section 9 investigates the theory for the partial questionnaire design of Wacholder, Carroll, Pee, and Gail (1994), which has a nonmonotone pattern of missingness; it is shown that, in principle at least, prospective standard errors are asymptotically conservative. The results in Sections 5–9 are new. Section 10 studies the two-stage studies of Breslow and Cain (1988).

2. ESTIMATING EQUATIONS, SANDWICH ESTIMATORS, AND THE CLASSICAL MODEL

One of our main results is that prospectively derived standard errors are at worst asymptotically conservative. Justification for this result is easiest to understand in the classical simple logistic model, $\mathrm{pr}(D = 1 \mid X) = H(\theta_0^* + \theta_1 X)$. The argument uses nothing more than standard estimating equation theory; we will outline this theory and the nomenclature as we go along. Extensions to complex problems require little more than change in notation.

2.1 Prospective Sampling

We first consider prospective sampling, and write $\Theta^* = (\theta_0^*, \theta_1)^t$. The prospective ordinary logistic regression estimate is the solution to the equation

$$0 = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} \{D_i - H(\theta_0^* + \theta_1 X_i)\}$$

$$= \sum_{i=1}^{n} \psi(D_i, X_i, \Theta^*). \qquad (2)$$

The entire term on the right side of (2) is called an *estimating equation.* The arguments $\psi(D_i, X_i, \Theta^*)$ are called *estimating functions.* The prospective estimator is denoted by $\hat{\Theta}^*$.

Prospective theory requires that the estimating equation be *unbiased;* that is, it has mean zero when evaluated at the parameters, so that

$$0 = E\left\{ \sum_{i=1}^{n} \psi(D_i, X_i, \Theta^*) \right\}. \qquad (3)$$

For logistic regression, even more is true. The estimating functions are themselves unbiased, having mean zero at the parameters:

$$0 = E\{\psi(D_i, X_i, \Theta^*)\} \quad \text{for} \quad i = 1, \ldots, n. \qquad (4)$$

However, only Equation (3) is required for now.

By use of Taylor series, it is known that $\hat{\Theta}^*$ is asymptotically normally distributed (under regularity conditions), and we write the distribution as

$$n^{1/2}(\hat{\Theta}^* - \Theta^*)$$

$$\approx \mathrm{Normal}\{0, B^{-1}(\Theta^*)A(\Theta^*)B^{-t}(\Theta^*)\}, \quad \text{where} \qquad (5)$$

$$B(\Theta^*) = n^{-1} \sum_{i=1}^{n} E\left\{ \frac{\partial}{\partial \Theta^*} \psi(D_i, X_i, \Theta^*) \right\} \qquad (6)$$

and

$$A(\Theta^*) = n^{-1} \mathrm{cov}\left\{ \sum_{i=1}^{n} \psi(D_i, X_i, \Theta^*) \right\}. \qquad (7)$$

Formula (5) is often called the *sandwich formula,* because $A(\Theta^*)$ is sandwiched between inverses of $B(\Theta^*)$.

At this point, we may now use the fact that the estimating functions are unbiased—that is, use (4)—to conclude that if we define

$$n^{-1} \sum_{i=1}^{n} E\{\psi(D_i, X_i, \Theta^*)\psi^t(D_i, X_i, \Theta^*)\} = C(\Theta^*), \qquad (8)$$

then $A(\Theta^*) = C(\Theta^*)$ and $\hat{\Theta}^*$ is asymptotically normally distributed with mean Θ^* and covariance matrix $n^{-1}B^{-1}(\Theta^*)C(\Theta^*)B^{-t}(\Theta^*)$.

Of course, in ordinary logistic regression we know that $C(\Theta^*)$ and $B(\Theta^*)$ are equal and can be consistently estimated by the usual information formula. In general, though, a consistent nonparametric estimate of these terms can be based on the method of moments; that is, in (6) and (8) remove the expectations and replace Θ^* by $\hat{\Theta}^*$. The resulting covariance matrix estimator is sometimes called the *robust sandwich formula,* where in a misnomer the term "robust" is used as a replacement for "model-free" (Drum and McCullagh 1993). For example, the resulting model-free estimate of $B(\Theta^*)$ is just

$$\hat{B}(\hat{\Theta}^*) = n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta^*} \psi(D_i, X_i, \hat{\Theta}^*).$$

2.2 Retrospective Sampling

We now turn to retrospective sampling. The key point to notice in the preceding argument is that we used unbiasedness of the estimating *functions* only in showing that (7) equals (8).

In retrospective sampling, define $\theta_0 = \theta_0^* + \log(n_1/n_0)$

$-\log\{\mathrm{pr}(D=1)/\mathrm{pr}(D=0)\}$, where $\mathrm{pr}(D=1)$ is the unknown prospective rate. Prentice and Pyke (1979) showed that if $\Theta = (\theta_0, \theta_1)^t$ and we replace Θ^* by Θ, then the estimating function (2) is still unbiased; that is, (3) holds. But the estimating functions are not unbiased, so that (4) fails, and hence it is not true that $A(\Theta) = C(\Theta)$.

Asymptotically, the distribution (5) still remains the same, but with the prospective parameter Θ^* and prospective estimator $\hat{\Theta}^*$ replaced by the retrospective parameter Θ and retrospective estimator $\hat{\Theta}$, which of course is the solution to (2) under retrospective sampling.

Because the estimating equation is unbiased, we can rewrite (7) as follows:

$$\begin{aligned} A(\Theta) &= n^{-1}\mathrm{cov}\left[\sum_{i=1}^{n}\psi(D_i, X_i, \Theta) - E\left\{\sum_{i=1}^{n}\psi(D_i, X_i, \Theta)\right\}\right] \\ &= n^{-1}\sum_{i=1}^{n}\mathrm{cov}[\psi(D_i, X_i, \Theta) - E\{\psi(D_i, X_i, \Theta)\}] \\ &= n^{-1}\sum_{i=1}^{n}E\{\psi(D_i, X_i, \Theta)\psi^t(D_i, X_i, \Theta)\} \\ &\quad - n^{-1}\sum_{i=1}^{n}E\{\psi(D_i, X_i, \Theta)\}E\{\psi^t(D_i, X_i, \Theta)\} \\ &= C(\Theta) - D(\Theta). \end{aligned}$$

The main conclusion now follows, in a series of steps:

- Prospectively, the asymptotic covariance matrix is $n^{-1}B^{-1}(\Theta^*)C(\Theta^*)B^{-t}(\Theta^*)$.
- Applying prospective formula directly to a retrospective study is equivalent to basing estimation as if the correct covariance matrix were $n^{-1}B^{-1}(\Theta)C(\Theta)B^{-t}(\Theta)$.
- But the proper covariance is $n^{-1}B^{-1}(\Theta)\{C(\Theta) - D(\Theta)\}B^{-t}(\Theta)$.
- Because $D(\Theta)$ is positive semidefinite, prospective covariance formulas are at worst conservative.

2.3 Further Steps

The reasoning given previously is perfectly sound, but we have skipped over a few steps. For example, we have simply assumed that the actual covariance estimators derived from a prospective analysis estimate the corresponding quantities retrospectively, which is true but needs to be justified.

Our analysis shows that prospective covariance formulas are at worst conservative, but no insight is given as to when these formulas are asymptotically correct. Our second main contribution is to derive a simple sufficient condition for this asymptotic correctness. The condition is routine to check in the examples described in this article, as well as other examples that we have not included. Deriving the sufficient condition requires a more detailed examination of $D(\Theta)$. This task is relegated to the general theory.

3. PROSPECTIVE FORMULATION

3.1 Likelihood for Complete Data

The following conventions are used throughout. Disease status is denoted by D, observable covariates by Z, and covariates that may be partially missing by X. Anticipating the possibility that the study may be stratified, we use the stratum assignment variable S, taking on the values $1, \ldots, \mathcal{S}$. When considering measurement error problems, instead of observing X, a proxy W is typically observed for all study participants; for example, blood pressure measured at a single time point as a proxy for long-term blood pressure measurement. The vector of parameters of major interest is denoted by θ_1; for example, in (1), $\theta_1 = (\theta_{11}^t, \theta_{12}^t)^t$.

If there were no missing data, then we assumed a sampling mechanism of a classical case-control study within each stratum $S = s$, with n_{1s} cases, n_{0s} controls and $n_s = n_{0s} + n_{1s}$ observations. The total sample size is $n = \sum n_s$. We assume that the terms n_{js}/n converge to positive constants, so that our work does not apply to matched case-control studies.

In those cases where a proxy W exists, it is sometimes useful to allow for an error model for it. Thus we write the likelihood of W given (D, Z, X) and stratum $S = s$ as $f(w|z, x, d, s, \theta_2)$, where θ_2 is an unknown parameter. We will assume that the prospective model is of the form

$$\mathrm{pr}(D = 1|Z, X, S = s) = H\{\theta_{0s}^* + R_s(\theta_1, \theta_2, Z, X)\}, \qquad (9)$$

where $R_s(\theta_1, \theta_2, Z, X)$ is an arbitrary function. Although the vector θ_2 is in both the conditional likelihood for W and in (9), this is simply a convention; not all components of θ_2 must appear in both likelihoods. Model (9) includes the linear logistic model (1) and the multiplicative model of Weinberg and Wacholder (1993) as special cases.

From the usual odds ratio formulation of Prentice and Pyke (1979), the retrospective likelihood that $(Z, X, W) = (z, x, w)$ when $(D, S) = (d, s)$ is

$$(n_s/n_{ds})q_s(z, x)H_s^d(\cdot)\{1 - H_s(\cdot)\}^{1-d}f(w|z, x, d, s, \theta_2), \quad \text{where} \qquad (10)$$
$$H_s(\cdot) = H\{\theta_{0s} + R_s(\theta_1, \theta_2, z, x)\}.$$

In (10), $q_s(\cdot)$ is the marginal density of (Z, X) in stratum $S = s$ induced by the case-control sampling scheme and $\theta_{0s} = \theta_{0s}^* + \log(n_{1s}/n_{0s}) - \log\{\mathrm{pr}(D = 1|S = s)/\mathrm{pr}(D = 0|S = s)\}$, where $\mathrm{pr}(D = 1|S = s)$ is the prospective rate in stratum s. We write $\Theta = (\theta_{01}, \ldots, \theta_{0\mathcal{S}}, \theta_1^t, \theta_2^t)^t$, the retrospective parameter, and $\Theta^* = (\theta_{01}^*, \ldots, \theta_{0\mathcal{S}}^*, \theta_1^t, \theta_2^t)^t$, the prospective parameter.

3.2 Missing Data

The theory allows for the possibility that different components of X are missing in different subsets of the data. If there are J such possible patterns of missingness $\Delta = (\delta_1, \ldots, \delta_J)$ is a vector with a single nonzero component indicating which pattern is applicable. The only assumption is that the data are missing at random, and hence the missing data indicators and X are conditionally independent given (Z, W, S, D), with selection probabilities $\pi_j(Z, W, S, D) = \mathrm{pr}(\delta_j = 1|Z, W, S, D, X)$.

For example, suppose that X has two components, $X_{(1)}$ and $X_{(2)}$. There are four possible patterns of missingness here: (1) both components missing; (2) only $X_{(1)}$ missing;

(3) only $X_{(2)}$ missing; and (4) neither component missing. In this case, $\delta_1 = 1$ means that both components of X are missing, $\delta_2 = 1$ means that only $X_{(1)}$ is missing, and so on. Table 1 summarizes the notation.

Table 1. Notation Used in the Paper

Variable	*Explanation*
D	Response
Z	Fully observed covariates
X	Missing or mismeasured covariates
W	Proxy for X in measurement error problems
S	Stratum indicator variable
δ_j	Indicator that X is missing with pattern number j
$\pi_j(z, w, s, d)$	Probability of missing data pattern j
Θ	Retrospective parameter, including stratum intercepts
Θ^*	Prospective parameter, including stratum intercepts
θ_1	Non-intercept parameter in the prospective logistic model
θ_2	Error model parameter for the distribution of W

3.3 Prospective Estimating Equations

With the exception of the leading term, (10) is of the same general form as a prospective likelihood with stratum-specific intercepts. Thus a natural approach to estimation is to use prospective estimating equations. Let δ_{ijs} denote the value of δ_j for the ith individual in the sth stratum. The prospective estimating function defined for the jth pattern of missingness and the sth stratum is $\Psi_{js}(D, Z, X, W, s, \Theta)$, and the estimators are defined as solutions to

$$\begin{aligned} 0 &= n^{-1} \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{n_s} \sum_{j=1}^{J} \delta_{ijs} \Psi_{js}(D_{is}, Z_{is}, X_{is}, W_{is}, s, \Theta) \\ &= n^{-1} \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{n_s} \mathcal{L}_{is}(\Theta) = \mathcal{T}_n(\Theta). \end{aligned} \tag{11}$$

In effect, we are suggesting that one ignore the case-control study design and proceed as if the data arose from a prospective sample.

4. ASYMPTOTIC THEORY

4.1 Main Results

Readers who are interested mainly in the applications may skip this section without any loss.

In our analysis we make two basic assumptions. First, we assume that given $(D, S = s)$, the vectors $(Z_{is}, X_{is}, W_{is}, \Delta_{is})$ are independent and identically distributed as i varies. The individual components of these vectors are, of course, dependent. The assumption of independent and identically distributed data is only for simplicity in this analysis and is not always necessary, as we show in Section 10.

The second assumption is that Equation (11) is retrospectively unbiased, so that

$$0 = \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{n_s} E\{\mathcal{L}_{is}(\Theta) \mid D_{is}, s\}. \tag{12}$$

Assumption (12) is satisfied in all the cases that we have examined. As described in more detail in the Appendix, Section A.2, this appears to be a general phenomenon and not simply a matter of convenient example selection on our part.

To state the main result, we set the following definitions. Define $l_s(\cdot, \Theta) = H_s^d(\cdot)\{1 - H_s(\cdot)\}^{1-d} f(w|z, x, d, s, \theta_2) q_s(\cdot)$. The notation $d\mu(\cdot)$ means integration or summation with respect to the arguments of $\mu(\cdot)$. Let $\Psi_{js\Theta}$ be the matrix of partial derivatives of Ψ_{js} with respect to Θ. Also, define

$$\begin{aligned} \mathcal{T}_\Theta(\Theta) = \sum_{s=1}^{\mathcal{S}} (n_s/n) \sum_{d=0}^{1} \sum_{j=1}^{J} \int & \pi_j(\cdot) \Psi_{js\Theta}(\cdot, \Theta) \\ & \times l_s(\cdot, \Theta)\, d\mu(z, x, w), \end{aligned} \tag{13}$$

$$\begin{aligned} \mathbf{C}(\Theta) = \sum_{s=1}^{\mathcal{S}} (n_s/n) \sum_{d=0}^{1} \sum_{j=1}^{J} \int & \pi_j(\cdot) \Psi_{js}(\cdot, \Theta) \Psi_{js}^t(\cdot, \Theta) \\ & \times l_s(\cdot, \Theta)\, d\mu(z, x, w), \end{aligned} \tag{14}$$

and

$$\kappa_{ds} = \int \sum_{j=1}^{J} \pi_j(\cdot) \Psi_{js}(\cdot, \Theta) l_s(\cdot, \Theta)\, d\mu(z, x, w).$$

Theorem. Let $\hat{\Theta}$ be the solution to (11) under retrospective sampling, and let $\hat{\Theta}^*$ be the solution to (11) under prospective sampling. Under appropriate regularity conditions, *retrospectively,* $n^{1/2}(\hat{\Theta} - \Theta)$ is asymptotically normally distributed with mean zero and covariance matrix

$$\begin{aligned} \{\mathcal{T}_\Theta(\Theta)\}^{-1} \left[C(\Theta) - \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} \{n_s^2/(n n_{ds})\} \kappa_{ds} \kappa_{ds}^t \right] & \\ \times \{\mathcal{T}_\Theta(\Theta)\}^{-t}. & \end{aligned} \tag{15}$$

Prospectively, define $l_{s*}(\cdot, \Theta^*) = q_{s*}(z, x) H_{s*}^d(\cdot)\{1 - H_{s*}(\cdot)\}^{1-d} f(\cdot \mid \cdot, \theta_2)$, where $q_{s*}(\cdot)$ is the marginal of (Z, X) in the prospective sampling distribution in the sth stratum and $H_{s*}(\cdot)$ is the same as $H_s(\cdot)$ but with prospective stratum-specific intercepts. Let $C_*(\Theta^*)$ and $\mathcal{T}_{\Theta*}(\Theta^*)$ be defined similarly to $C(\Theta)$ and $\mathcal{T}_\Theta(\Theta)$, but with l_s and Θ replaced by l_{s*} and Θ^*. Then $n^{1/2}(\hat{\Theta}^* - \Theta^*)$ is asymptotically normally distributed with mean zero and covariance matrix

$$\{\mathcal{T}_{\Theta*}(\Theta^*)\}^{-1} C_*(\Theta^*) \{\mathcal{T}_{\Theta*}(\Theta^*)\}^{-t}. \tag{16}$$

The proof is sketched in the Appendix.

4.2 When are Prospective Standard Errors Asymptotically Correct?

For the moment, assume that prospectively derived covariance matrix estimates are consistent estimates of the quantity

$$\{\mathcal{T}_\Theta(\Theta)\}^{-1} C(\Theta) \{\mathcal{T}_\Theta(\Theta)\}^{-t}. \tag{17}$$

If this is true, then the foregoing theorem states that the prospective covariance matrix estimates are at worst conservative.

Here we state a simple sufficient condition that guarantees that prospectively derived standard errors are asymptotically correct. For most cases, $\kappa_{0s} = -\kappa_{1s}$ for each stratum, and we assume this here. This leads to the following simple result.

Corollary. Suppose that $\kappa_{1s} = -\kappa_{0s}$ and that κ_{1s} is proportional to the sth column of $\mathcal{T}_\Theta(\Theta)$ for $s = 1, \ldots, \mathcal{S}$. Then prospectively derived covariance formulas for (θ_1, θ_2) are asymptotically correct. More generally, the result holds if the rows of $\mathcal{T}_{\Theta}(\Theta)\kappa_{1s}$ corresponding to θ_1 all equal zero.

We show later that many examples satisfy the conditions of this corollary.

The reason that prospectively derived covariance matrix estimators actually estimate (17) is that they are in all circumstances derived from sums of functions of $\hat{\Theta}$ and the individual observations. For example, consider the model-free sandwich estimator from prospective formulas (Sec. 2), namely

$$\{\mathcal{T}_{n\Theta}(\hat{\Theta})\}^{-1}n^{-1}\sum_{s=1}^{\mathcal{S}}\sum_{i=1}^{n_s}\mathcal{L}_{is}(\hat{\Theta})\mathcal{L}_{is}^t(\hat{\Theta})\{\mathcal{T}_{n\Theta}(\hat{\Theta})\}^{-t},$$

where

$$\mathcal{T}_{n\Theta}(\hat{\Theta}) = n^{-1}\sum_{s=1}^{\mathcal{S}}\sum_{i=1}^{n_s}\frac{\partial}{\partial\Theta}\mathcal{L}_{is}(\hat{\Theta}).$$

Using the retrospective likelihood (10) and the fact that $\hat{\Theta}$ is a consistent estimator of Θ, it is easily seen that the model-free sandwich estimator consistently estimates (17).

For those cases for which prospective formulas are conservative, there are two ways to construct asymptotically correct covariance estimates. The preferred method is to begin with (15) and estimate $\mathcal{T}_{\Theta}(\Theta)$ and $C(\Theta)$ by prospective formulas; typically, one would not use the "model-free" estimates of these terms. For example, in the classical problem with no missing data, these matrices would be estimated by the observed information. To estimate κ_{ds} in (15), use $\hat{\kappa}_{ds} = n_s^{-1}\sum_{i=1}^{n_s} I(D_{is} = d)\mathcal{L}_{is}(\hat{\Theta})$, a model-free consistent estimate.

This hybrid approach, where κ_{ds} is estimated without a model and $\mathcal{T}_{\Theta}(\Theta)$ and $C(\Theta)$ typically being based on a prospective model, will work for most cases. But it need not yield a positive semidefinite covariance matrix estimate, because of the subtraction in (15). In such cases, a model-free sandwich covariance matrix estimate can be employed, namely $\{\mathcal{T}_{n\Theta}(\hat{\Theta})\}^{-1}B_n(\hat{\Theta})\{\mathcal{T}_{n\Theta}(\hat{\Theta})\}^{-t}$, where

$$\mathbf{B}_n(\Theta) = n^{-1}\sum_{s=1}^{\mathcal{S}}\sum_{d=0}^{1}\sum_{i=1}^{n_s} I(D_{is} = d)$$
$$\times [\mathcal{L}_{is}(\Theta) - \hat{m}(d, s, \Theta)][\cdots]^t,$$

where $\hat{m}(d,s,\Theta) = n_{ds}^{-1}\sum_{i=1}^{n_s} I(D_{is} = d)\mathcal{L}_{is}(\Theta)$ is an estimate of $E\{\mathcal{L}_{is}(\Theta)\,|\,D_{is} = d\}$.

5. CLASSICAL STUDIES

By a classical case-control study, we mean one with no missing data and a single stratum. Dropping the subscripts (j, s), which indicate missing data pattern and stratum number, from (9) we have $\mathrm{pr}(D = 1\,|\,X) = H\{\theta_0^* + R(\theta_1, X)\}$. In this section we show that in classical case-control studies, essentially any reasonable prospectively defined estimating equation yields consistent estimators, and the prospective standard errors are asymptotically correct. The work generalizes that of Weinberg and Wacholder (1993) on multiplicative models and Wang and Carroll (1993, 1995) on robust estimation.

To motivate the class of estimators, first consider simple linear logistic regression with $R(x, \theta_1) = \theta_1^t x$, and recall from (2) that the estimating function for the maximum likelihood estimator is $\psi(d, x, \Theta^*) = (1, x)^t\{d - H(\theta_0^* + \theta_1^t x)\}$. By assumption, prospectively

$$E\{\psi(D, X, \Theta^*)\,|\,X\} = 0, \qquad (18)$$

because $\mathrm{pr}(D = 1\,|\,X) = H(\theta_0^* + \theta_1^t X)$. For the prospective maximum likelihood estimator in the general model, the estimating equation for the maximum likelihood estimator is $\psi(d, x, \Theta^*) = \{1, (\partial/\partial\theta_1)R(\theta_1, x)\}[d - H\{\theta_0^* + R(\theta_1, x)\}]$, and (18) still holds. The same condition applies to all the robust estimators discussed by Carroll and Peterson (1993).

The fact then is that most estimators prospectively satisfy (18). We will say that an estimating function $\psi(D, X, \Theta^*)$ is (prospectively) conditionally unbiased if (18) holds prospectively for all Θ^*.

In the Appendix, we show the following result.

Lemma. Any conditionally unbiased estimating function leads to a retrospectively unbiased estimating equation, and prospectively derived standard errors are asymptotically correct.

The result is anticipated from the Appendix Section A.1, because in this context no restrictions have been made on the marginal distribution of X.

5.1 A Simulation

We performed a small simulation in simple linear logistic regression to illustrate the results. There were 75 cases and 75 controls. The predictor X was generated either as a normal random variable with mean zero and variance 1 or as a t-random variable with 3 degrees of freedom. We chose $\theta_0^* = -4.0$, $\theta_1 = -.4, -.6, -.8$. When X is normally distributed, the values of θ_1 were chosen so that the relative risks of moving from the 90th to the 10th percentile of the distribution of X equal 3, 5, and 8. There were 500 simulations for each case.

Two prospectively derived estimators were considered: (1) the ordinary linear logistic estimator, and (2) the robust leverage-downweighting estimators defined by Carroll and Pederson (1993, sec. 4.1). The results are given in Table 2. Note that in all cases, both the ordinary and the robust methods attain very nearly their nominal levels.

Table 2. Simulation of Ordinary and Robust Logistic Regression

		Ordinary			Robust		
Distribution	θ_1	*90%*	*95%*	*Median*	*90%*	*95%*	*Median*
Normal	−.4	.886	.944	−.435	.886	.940	−.434
	−.6	.878	.940	−.610	.884	.940	−.612
	−.8	.912	.964	−.816	.912	.964	−.806
$t(3)$	−.4	.912	.964	−.400	.914	.964	−.403
	−.6	.888	.936	−.618	.882	.924	−.619
	−.8	.906	.952	−.812	.898	.952	−.811

NOTE: In 500 simulations, the coverage rates are given for nominal 90% and 95% intervals. The median of the slope estimates is also listed.

6. MISMEASURED DATA: DIFFERENTIAL ERROR

6.1 Introduction

In most problems with missing data, and less frequently in problems with measurement error, X is observable in a subset of the study. A wide variety of parametric techniques have been developed for likelihood analysis of missing data, and the corresponding likelihoods for measurement error models are also well known. Recently, however, techniques have been developed which attempt to avoid strong parametric assumptions (see, for example, Carroll and Wand 1991, Pepe and Fleming 1991, and Reilly and Pepe 1995).

We will say that measurement error is *nondifferential* and that W is a *surrogate* for X if W is independent of D given (Z, X, S). Otherwise, measurement error is *differential.*

Robins et al. (1994) described a general class of prospectively unbiased estimating equations for missing and mismeasured data in a single stratum. We concentrate on this case in linear logistic regression and by modifying their approach slightly allow for differential measurement error. As a matter of interpretation, we take the view that interest lies in the effects of X on disease in the presence of the covariates Z measured without error, and not otherwise in W. Thus the interesting prospective logistic model is $H(\theta_0^* + \theta_{11}^t Z + \theta_{12}^t X)$. Our analysis requires a model for the error distribution of W given (Z, X, D).

6.2 Estimating Equations and Results

The estimating equations can be described as follows. Let $\psi(d, z, x, \Theta)$ be the usual logistic estimating function $M(z, x)\{d - H(\cdot)\}$, where $M(z, x) = (1, z^t, x^t)^t$. Write the conditional density or mass function for W as $f(w|z, x, d, \Theta) = f(w|z, x, d, \theta_2)$. Let $\chi(z, x, w, d, \Theta)$ be any unbiased estimating function for θ_2.

For any function $\xi = \xi(z, x, w, d)$, define

$$\mathcal{G}(z, x, w, \xi, \Theta) = \sum_{d=0}^{1} \xi(z, x, w, d) f(w|z, x, d, \Theta) H^d(\cdot)\{1 - H(\cdot)\}^{1-d}.$$

Then for an arbitrary function $\phi(d, z, w)$ (Robins et al. [1994] showed how one can choose ϕ prospectively; the same method applies retrospectively), with $j = 1$ being the case that all of (Z, X, W, D) are observed, we define

$$\Psi_1(\cdot, \Theta) = \begin{bmatrix} \psi(D, Z, X, \Theta) - \dfrac{\mathcal{G}(Z, X, W, \pi\psi, \Theta)}{\mathcal{G}(Z, X, W, \pi, \Theta)} - \dfrac{\mathcal{G}\{Z, X, W, (1 - \pi)\phi, \Theta\}}{\mathcal{G}(Z, X, W, \pi, \Theta)} \\ \chi(Z, X, W, D, \Theta) - \dfrac{\mathcal{G}(Z, X, W, \pi\chi, \Theta)}{\mathcal{G}(Z, X, W, \pi, \Theta)} \end{bmatrix}$$

and

$$\Psi_2(\cdot, \Theta) = \begin{bmatrix} \phi(D, Z, W) \\ 0 \end{bmatrix}.$$

Note that because there is a single stratum, we have dropped the index corresponding to stratum assignment. This estimating equation is prospectively unbiased and, as can be verified directly, also retrospectively unbiased (see the Appendix, Sec. A.5). In that section we also show that prospectively derived standard errors are asymptotically correct.

7. LIKELIHOOD AND NONDIFFERENTIAL MEASUREMENT ERROR

Satten and Kupper (1993) considered likelihood analysis for prospective studies with nondifferential measurement error. We study their easily computed "unconditional" method in the context of the logistic model (1), showing that it leads to consistent estimates in the retrospective model and that prospectively derived standard errors are retrospectively asymptotically correct.

Prospectively, Satten and Kupper formulated the problem as follows. For all subjects, (D, Z, W) is observed. But for the ith individual either X_i is also observed ($\delta_{i1} = 1$) or X_i is not observed ($\delta_{i2} = 1$). If $f_{X|Z,W,D}$ is the density or mass function of X given (Z, W, D), then the prospective likelihood can be written as

$$\prod_{i=1}^{n} [f^{\delta_{i1}}_{X|Z,W,D}(X_i|Z_i, W_i, D_i)\{\mathrm{pr}(D_i = 1|Z_i, W_i)\}^{D_i} \times \{1 - \mathrm{pr}(D_i = 1|Z_i, W_i)\}^{1-D_i}]. \quad (19)$$

The hard part is to compute each term. Satten and Kupper's approach is to model the distribution of X given $(Z, W, D = 0)$; that is, among the controls, depending on a parameter θ_2. Define

$$R(Z, W, \Theta) = \log[E\{\exp(\theta_{12}^t X)|Z, W, D = 0, \theta_2\}].$$

Prospectively, they showed that $\Pr(D = 1|Z, W) = H\{\theta_0^* + \theta_{11}^t Z + R(Z, W, \Theta)\}$, and further that the ratio of the density or mass functions is

$$\frac{f_{X|Z,W,D}(x|z, w, d = 1, \Theta)}{f_{X|Z,W,D}(x|z, w, d = 0, \Theta)} = \exp\{\theta_{12}^t x - R(z, w, \Theta)\},$$

thus writing the conditional density of X given $(Z, W, D = 1)$ in terms of that of $(Z, W, D = 0)$.

The prospective likelihood (19) is now specified, and the maximum likelihood estimator can be computed. In the Appendix, Section A.6, we show that maximizing this prospective likelihood leads to estimators that are retrospectively consistent and standard errors that are asymptotically correct.

8. MEASUREMENT ERROR AND REPLICATION

8.1 Introduction

The classical formulation of the measurement error problem (Fuller 1987) is one in which the true predictor X is not observable, and instead only an unbiased surrogate (defined in Sec. 6) W for X may be observed, possibly with replication on a subset of the data. If the variance of the measurement error is known or estimated from external data sources, then the standard linear regression method is the so-called "correction for attenuation." In nonlinear regression models, the same correction for attenuation often works extremely well. There are a variety of proposals based on the idea of a correction for attenuation (see Carroll and Stefanski 1990; Gleser 1990; Liu and Liang 1992; Rosner et al. 1989; Rosner, Spiegelman, and Willett 1990; and Schafer 1993). Carroll and Stefanski (1994) described an instrumental variables method.

These methods differ fundamentally from the moments methods of Section 6 in that they apply in the common case where X is never observable; for example, blood pressure or diet history.

The application of these ideas to case-control studies with nondifferential measurement error were briefly explored by Rosner et al. (1989), studied by Armstrong, Howe, and Whittemore (1989) and Buonaccorsi (1990) using discriminant analysis techniques and allowing for differential measurement error, and suggested as a general methodology with partially replicated data by Carroll, Gail, and Lubin (1993). Although all of these methods can be analyzed by our general theory, in this section we define and investigate a version of the correction for attenuation methodology based on prospective considerations but appropriate for case-control studies. The asymptotic distribution theory is most naturally studied using two strata.

8.2 Estimating Equations and Results

We will assume that W is a surrogate for X, that is, independent of D given (Z, X), and that the surrogate can be replicated with independent errors. Let $W = (W_1, W_2)$, where $W_j = X + U_j$ and U_1, U_2 are independent and identically distributed with mean zero and variance σ^2_u. To keep the analysis simple, we will ignore Z and study the prospective model $H(\theta^*_0 + \theta_1 X)$. There are two strata ($s = 1, 2$): one in which only W_1 is observed ($s = 1$), the other for which both (W_1, W_2) are observed ($s = 2$). Set $J = 1$ and $\pi_j \equiv 1$.

A good approximation (Carroll and Stefanski 1990; Gleser 1990; Rosner et al. 1989) to the probability of response given the observed surrogate is

$$\mathrm{pr}(D = 1 \mid W_1) \approx H\{\theta^*_0 + \theta_1 m_1(W_1)\}$$

and

$$\mathrm{pr}(D = 1 \mid \bar{W}) \approx H\{\theta^*_0 + \theta_1 m_2(\bar{W})\},$$

where $m_1(W_1) = E(X \mid W_1)$ and $m_2(\bar{W}) = E(X \mid \bar{W})$. The correction-for-attenuation methodology estimates the functions (m_1, m_2) and regresses the response on these estimated functions, with one intercept per stratum.

Of course the regression functions (m_1, m_2) are not estimable, because they depend on the underlying disease rates. But they can be approximated in the common case that the disease is rare, because they are approximately the same in the controls as they are in the source population, and hence $m_1(W_1) \approx E(X \mid W_1, D = 0)$ and $m_2(\bar{W}) \approx E(X \mid \bar{W}, D = 0)$, approximations that we will henceforth treat as exact. Let $\hat{\mu}_w$ be the mean of W_1 among the controls. Following Carroll and Stefanski (1990) and Gleser (1990), for $s = 1$, 2 estimates of the best linear approximations to these regressions are

$$g_1(W_1, \hat{\sigma}^2_w, \hat{\sigma}^2_u, \hat{\mu}_w) = \hat{\mu}_w + \{(\hat{\sigma}^2_w - \hat{\sigma}^2_u)/\hat{\sigma}^2_w\}(W_1 - \hat{\mu}_w)$$

and

$$g_2(\bar{W}, \hat{\sigma}^2_w, \hat{\sigma}^2_u, \hat{\mu}_w) = \hat{\mu}_w + \{(\hat{\sigma}^2_w - \hat{\sigma}^2_u)/(\hat{\sigma}^2_w - \hat{\sigma}^2_u/2)\}(\bar{W} - \hat{\mu}_w),$$

where $\hat{\sigma}^2_w$ is the sample variance of W_1 among all the controls and $\hat{\sigma}^2_u$ is the sample variance of $(W_1 - W_2)/2^{1/2}$ among the replicated data.

The algorithm then is as follows. Use the replicated data to construct $\hat{\sigma}^2_u$ and use the W_1's from all the control data to construct $\hat{\mu}_w$ and $\hat{\sigma}^2_w$. Then regress D on the functions (g_1, g_2) for $s = 1, 2$, with stratum specific intercepts.

Subject to the levels of approximation described, in the Appendix, Section A.7, we show that prospective covariance formulas may be used for the estimate of θ_1. The preceding analysis is readily extended to problems with vector predictors.

8.3 A Simulation

We performed a small simulation in simple linear logistic regression to illustrate the results. There were 300 cases and controls. The number of replicated observations in each case and control was 25. Prospectively, the variable X was generated as a normal random variable with mean zero and variance 1. We chose $\theta^*_0 = -4.0$, $\theta_1 = -.4, -.6, -.8$. With these values, prospectively the event rates are approximately 3%, the type of "rare-disease" situation one might expect. The values of θ_1 were chosen so that the relative risks of moving from the 90th to the 10th percentile of the distribution of X equal 3, 5, and 8.

The measurement error model was $W = X + U$, where U is independent of D and X and is generated as a normal random variable with mean zero and variance $\sigma^2_u = .25, .5, 1.0$, representing small, moderate, and large measurement error.

We estimated θ_1 using the foregoing algorithm. Standard errors were computed using a prospective method described in the Appendix, Section A.7. The results, given in Table 3, indicate that the prospectively derived confidence intervals achieve very nearly their nominal levels.

9. PARTIAL QUESTIONNAIRES

9.1 Introduction

The partial questionnaire design of Wacholder et al. (1994) is a single-stratum design where the covariate Z is of primary interest and the components of $X = (X_1, X_2)$ are partially missing by design. Such designs may be of considerable use when X_1 and X_2 are expensive or difficult to assess. The advantage of deliberately making components of (X_1, X_2) missing is a lesser burden on study subjects, possibly resulting

Table 3. Simulation of Correction for Attenuation

σ_u^2	θ_1	90%	95%	Mean	Median
.25	−.4	.89	.96	−.41	−.41
	−.6	.89	.96	−.61	−.60
	−.8	.91	.96	−.82	−.81
.50	−.4	.91	.96	−.41	−.41
	−.6	.89	.95	−.62	−.60
	−.8	.92	.96	−.82	−.81
1.00	−.4	.93	.95	−.43	−.40
	−.6	.90	.93	−.65	−.60
	−.8	.90	.94	−.87	−.80

NOTE: The measurement error variance is σ_u^2, the slope is θ_1, and the number of replicated cases and the number of replicated controls both equal 25. In 1,000 simulations, the coverage rates are given for nominal 90% and 95% intervals. The mean and median of the slope estimates are also listed.

in increased participation. Further details on the motivation of the study design were discussed by Wacholder et al. (1994).

The partial questionnaire design is under consideration for a study to be done by the National Cancer Institute. The study concerns the health effects of pesticide exposure (Z); diet and cooking practices (X_1) and level of physical activity (X_2) are also of interest. Measuring diet and cooking practices with any degree of accuracy is difficult, expensive, and time-consuming for both investigators and study participants; accurately measuring physical activity levels can be burdensome as well. Hence the investigators wish to minimize the number of subjects for whom both diet and activity are measured, because measuring both will affect compliance and accuracy.

The pattern of missingness here is nonmonotone in the sense of Little and Rubin (1987). We will show that the prospective formulas are not necessarily asymptotically correct and that in principle a correction needs to be made.

9.2 Estimating Equations and Theory

In this case, $J = 4$, $\pi_j = \pi_j(Z, D)$, and $\delta_j = 1$ if Z only is observed ($j = 1$), (X_1, Z) is observed ($j = 2$), (X_2, Z) is observed ($j = 3$), or the entire set (X_1, X_2, Z) is observed ($j = 4$). Wacholder et al. (1994) assumed that (X_1, X_2, Z) are discrete random variables.

The prospective model is $\mathrm{pr}(D = 1 \mid X_1, X_2, Z) = H(\theta_0^* + \theta_{11}X_1 + \theta_{12}X_2 + \theta_{13}Z)$. The marginal distribution of (X_1, X_2, Z) induced by the case-control sampling scheme before "covering up" (X_1, X_2) is written as $q(x_1, x_2, z, \theta_2)$, where we have included θ_2 as a parameter to allow various categorical data submodels; for example, the fully saturated model or the model in which (X_1, X_2) is independent of Z (Bishop, Fienberg, and Holland 1975). Thus $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})^t$ and $\Theta = (\theta_0, \theta_1^t, \theta_2^t)^t$.

There is no requirement in this formulation that (X_1, X_2, Z) be discrete. If they are not, then one must specify a model for the distribution of these random variables induced by the case-control sampling scheme.

In this case, we show in the Appendix, Section A.8, that the elements of $\mathcal{T}_\Theta(\Theta)\kappa_{11}$ corresponding to θ_1 need not all be zero, so that the prospective covariance formula for Θ_1 can be asymptotically conservative. When (X_1, X_2, Z) are all binary and their distribution is left unspecified, it can be shown that prospective covariance formula are correct, in accordance with the Appendix, Section A.1, but the prospective covariance formula is conservative when a logistic model applies.

10. TWO-STAGE STUDIES

Our estimating equation approach can be applied even when the missing-data indicators δ_{ijs} are dependent. As an illustration, consider the single-stratum two-stage study of Breslow and Cain (1988), which is based on the prospective model $\mathrm{pr}(D = 1 \mid X) = H(\theta_0^* + \theta_1^t X)$. The variable Z is a categorical surrogate with M levels. The assumption is that Z is conditionally independent of D given X, which might occur, for example, when Z is a categorical level of a continuous covariate X. In effect, the model is a linear logistic model, where the coefficient for Z is known to be zero.

At the first stage, we observe (D, Z), we note the number of observations in each (D, Z) category and then within each category select a further subsample of fixed size in which X is also observed.

In the Appendix, Section A.9, we show how to apply our estimating equation approach to rederive the Breslow and Cain result. (For a general discussion of two-stage designs, see Zhao and Lipsitz 1992.)

11. DISCUSSION

We have proposed a method for the analysis of prospective estimating equations in case-control studies. The major conclusions are that prospectively unbiased estimating equations are typically retrospectively unbiased and that the use of prospectively derived standard error estimates is asymptotically at worst conservative.

The examples we have considered allowed for multiplicative and linear logistic models, missing data, mismeasured data, and robust estimation. The techniques are applicable in general, and should prove useful in the consideration of other complex problems.

APPENDIX: TECHNICAL PROOFS

A.1 Semiparametric Perspective

Some insight into when the prospective standard errors are asymptotically correct can be gained by the following informal semiparametric argument. Suppose that the distributions of (Z, X) and $(W \mid Z, X, D)$ are not parameterized. Then they are completely unrestricted, or (semiparametrically) governed by infinite-dimensional parameters (ρ_1, ρ_2). The prospective likelihood can then be written as

$$f_{Z,X}(z, x \mid \rho_1) f_{D \mid Z,X}(d \mid z, x, \theta_0, \theta_1) f_{W \mid Z,X,D}(w \mid z, x, d, \rho_2).$$

From Prentice and Pyke (1979), the prospective likelihood also can be written as

$$f_{Z,X \mid D}(z, x \mid d, \theta_1, \rho_3) f_D(d, \rho_4) f_{W \mid Z,X,D}(w \mid z, x, d, \rho_2),$$

where (ρ_3, ρ_4) are unrestricted (infinite-dimensional) and θ_1 is characterized by the log-odds ratio. Because (ρ_2, ρ_3, ρ_4) are all unrestricted, $(D_1, \ldots, D_n)$ is retrospectively ancillary for θ_1, and hence the distributions of estimates of θ_1 should be the same prospectively and retrospectively, even with missing X's. The same argument applies even when the distribution of $(W \mid Z, X, D)$ is parameterized.

This informal argument is complementary to the results in Sections 5 and 6. It applies in Section 7 when the roles of X and W are interchanged. The result in Section 8 is not easily categorized. For the partial questionnaire design of Section 9, when the distribution of (Z, X) is not parameterized, the semiparametric argument also applies.

A.2 Retrospective Unbiasedness of Prospective Estimating Equations

We have no proof that prospective estimating equations are always retrospectively unbiased. But this is the case in every example we have examined, including the ones in this article. The following informal argument shows that retrospective unbiasedness is the rule, rather than the exception. This argument is a precise manifestation of the well-known fact that in a classical study, if we fictitiously "sampled" from a case-control study, case or control status would follow a logistic model.

We first show what it means for the estimating equation to be retrospectively unbiased. Define $l_s(\cdot, \Theta) = H_s^d(\cdot)\{1 - H_s(\cdot)\}^{1-d} f(w \mid z, x, d, s, \theta_2) q_s(\cdot)$, where as before $q_s(\cdot)$ is the marginal density or mass function of (Z, X) in the case-control sampling scheme. The notation $d\mu(\cdot)$ means integration or summation with respect to the arguments of $\mu(\cdot)$. Then, by (10), the estimating equation is retrospectively unbiased if

$$0 = \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} \sum_{j=1}^{J} n_s \int \pi_j(\cdot)\Psi_{js}(\cdot, \Theta) l_s(\cdot, \Theta)\, d\mu(z, x, w). \quad \text{(A.1)}$$

It will be useful in later work to note that the retrospective expectation (12) is given by

$$\sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} n_{ds} E\{\mathcal{L}_{is}(\Theta) \mid D_{is} = d, s\} = \sum_{s=1}^{\mathcal{S}} n_s \sum_{d=0}^{1} \kappa_{ds}. \quad \text{(A.2)}$$

Strictly speaking, retrospective unbiasedness of the estimating equation means that (20) holds for all Θ and $q_s(\cdot)$ in an appropriate class.

Now turn to the prospective formulation. For a prospective model, the likelihood of (D, X, Z, W) given $S = s$ is $l_{s*}(\cdot, \Theta^*) = q_{s*}(z, x) H_{s*}^d(\cdot)\{1 - H_{s*}(\cdot)\}^{1-d} f(\cdot \mid \cdot, \theta_2)$, where $\Theta^* = (\theta_{01}^*, \ldots, \theta_{0S}^*, \theta_1^t, \theta_2^t)^t$, $q_{s*}(\cdot)$ is the marginal of (Z, X) in the prospective sampling distribution in the sth stratum, and $H_{s*}(\cdot)$ is the same as $H_s(\cdot)$ but with prospective stratum-specific intercepts. Thus prospective unbiasedness means that for all Θ^* and $q_{s*}(\cdot)$ in an appropriate class,

$$0 = \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} \sum_{j=1}^{J} n_s \int \pi_j(\cdot)\Psi_{js}(\cdot, \Theta^*) l_{s*}(\cdot, \Theta^*)\, d\mu(z, x, w). \quad \text{(A.3)}$$

Note the similarity between (A.1) and (A.3). The equations are formally identical, with the only difference one of notation. Hence we can expect that prospectively unbiased estimating equations will also be retrospectively unbiased. In all the cases we have examined, the relationship between (A.1) and (A.3) trivially leads to retrospective unbiasedness of the estimating equation.

A.3 Sketch of Proof of the Main Theorem

Consider the retrospective formulation, where the parameter is Θ. By a Taylor series expansion, $n^{1/2}(\hat{\Theta} - \Theta) \approx -\{\mathcal{T}_{n\Theta}(\Theta)\}^{-1} n^{1/2}\mathcal{T}_n(\Theta)$. By a calculation similar to (A.2), $\mathcal{T}_{n\Theta}(\Theta)$ has expectation (13); suppressing the dependence on sample sizes, denote the result by $\mathcal{T}_\Theta(\Theta)$.

We next compute $\operatorname{cov}\{n^{1/2}\mathcal{T}_n(\Theta)\}$ (conditioned on all the D's of course). Let the notation $[\cdots]$ indicate a repeat of the preceding term. Using (12), we have

$$\begin{aligned}
&\operatorname{cov}\{n^{1/2}\mathcal{T}_n(\Theta)\} \\
&\quad = n^{-1} \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{n_s} E([\mathcal{L}_{is}(\Theta) - E\{\mathcal{L}_{is}(\Theta) \mid D_{is}, s\}][\cdots]^t \mid D_{is}, s) \\
&\quad = n^{-1} \sum_{s=1}^{\mathcal{S}} \sum_{i=1}^{n_s} (E\{\mathcal{L}_{is}(\Theta)\mathcal{L}_{is}^t(\Theta) \mid D_{is}, s\} \\
&\qquad\qquad - [E\{\mathcal{L}_{is}(\Theta) \mid D_{is}, s\}][\cdots]^t) \\
&\quad = C(\Theta) - \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} (n_{ds}/n)[E\{\mathcal{L}_{is}(\Theta) \mid D_{is} = d, s\}][\cdots]^t.
\end{aligned}$$

It is easily seen that $E\{\mathcal{L}_{is}(\Theta) \mid D_{is} = d, s\} = (n_s/n_{ds})\kappa_{ds}$, thus showing that

$$\operatorname{cov}\{n^{1/2}\mathcal{T}_n(\Theta)\} = C(\Theta) - \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} \{n_s^2/(n n_{ds})\}\kappa_{ds}\kappa_{ds}^t. \quad \text{(A.4)}$$

Now, using (10), we have

$$\begin{aligned}
C(\Theta) &= \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} (n_{ds}/n) E\{\mathcal{L}_{is}(\Theta)\mathcal{L}_{is}^t(\Theta) \mid D_{is} = d, s\} \\
&= \sum_{s=1}^{\mathcal{S}} \sum_{d=0}^{1} (n_s/n) \sum_{j=1}^{J} \int \pi_j(\cdot)\Psi_{js}(\cdot, \Theta)\Psi_{js}^t(\cdot, \Theta) l_s(\cdot, \Theta) \\
&\quad \times d\mu(z, x, w). \qquad \text{(A.5)}
\end{aligned}$$

This is identical to (14), as required.

A.4 Theory for the Classical Model

We have defined conditional unbiasedness to mean that (18) holds for all Θ^*. Write $\Theta = (\theta_0, \theta_1^t)^t$, $H(x, \Theta) = H\{\theta_0 + R(\theta_1, x)\}$, and $H^{(1)}(x, \Theta) = H(x, \Theta)\{1 - H(x, \Theta)\}$. Then conditional unbiasedness means that for any (x, Θ),

$$0 = \sum_{d=0}^{1} \psi(d, x, \Theta) H^d(x, \Theta)\{1 - H(x, \Theta)\}^{1-d}. \quad \text{(A.6)}$$

In our notation, $\kappa_d = \int \psi(d, x, \Theta) H^d(x, \Theta)\{1 - H(x, \Theta)\}^{1-d} q(x)\, dx$, so that $\kappa_0 + \kappa_1 = 0$ by (A.6), and hence prospective estimating equations that are conditionally unbiased are also unbiased retrospectively. It also follows that

$$\mathcal{T}_\Theta(\Theta) = \int \sum_{d=0}^{1} \psi_\Theta(d, x, \Theta) H^d(x, \Theta)\{1 - H(x, \Theta)\}^{1-d} q(x)\, dx.$$

Differentiating the right side of (A.6) with respect to Θ and then integrating with respect to $q(x)dx$, we find that the first column of $-\mathcal{T}_\Theta(\Theta)$ is

$$\int \sum_{d=0}^{1} (2d - 1)\psi(d, x, \Theta) H(x, \Theta)\{1 - H(x, \Theta)\} q(x)\, dx. \quad \text{(A.7)}$$

It follows then that κ_1 equals the first column (A.7) of $-\mathcal{T}_\Theta(\Theta)$ if

$$\psi(1, x, \Theta) = \{1 - H(x, \Theta)\}\{\psi(1, x, \Theta) - \psi(0, x, \Theta)\},$$

which follows directly from (A.6).

Based on the lemma in Section 4, we have thus shown that prospectively defined standard errors for slope parameters are retrospectively asymptotically correct for the general class of conditionally unbiased prospective estimating equations.

A.5 Theory for Differential Error

We first briefly sketch an argument showing that the estimating equations of Section 6 are retrospectively unbiased. Because there is only a single stratum, we drop the stratum indicators and, as in Section 6, write $\kappa_d = [\{\kappa_d(\psi) + \kappa_d(\phi)\}^t, \kappa_d^t(\chi)]^t$. We will show that $\kappa_0(\chi) + \kappa_1(\chi) = 0$, the other cases being similar. By definition,

$$\begin{aligned}
&\sum_{d=0}^{1} \kappa_d(\chi) \\
&\quad = \sum_{d=0}^{1} \int \pi(d, z, w) H^d(\cdot)\{1 - H(\cdot)\}^{1-d} \\
&\qquad \times f(w|z, x, d, \Theta) q(z, x) \\
&\qquad \times \left\{\chi(z, x, w, d, \Theta) - \frac{\mathcal{G}(z, x, w, \pi\chi, \Theta)}{\mathcal{G}(z, x, w, \pi, \Theta)}\right\} d\mu(z, x, w) \\
&\quad = \int \left\{\mathcal{G}(\cdot, \pi\chi, \Theta) - \frac{\mathcal{G}(\cdot, \pi\chi, \Theta)}{\mathcal{G}(\cdot, \pi, \Theta)} \sum_{d=0}^{1} \pi(\cdot) H^d(\cdot)\right. \\
&\qquad \left. \times \{1 - H(\cdot)\}^{1-d} f(\cdot, \Theta)\right\} q(\cdot)\, d\mu(z, x, w).
\end{aligned}$$

Because this integrand is zero, this yields the desired result.

We now show that except for the intercept, prospective covariance formulas are asymptotically correct. To do this, we must show that κ_1 is proportional to the first column of $\mathcal{T}_\Theta(\Theta)$. Let $H^{(1)}(\cdot) = H(\cdot)\{1 - H(\cdot)\}$ be the derivative of $H(\cdot)$. Because $\mathcal{G}\{\cdot, \xi M(z, x), \Theta\} = M(z, x)\mathcal{G}(\cdot, \xi, \Theta)$, direct calculations indicate that $\kappa_1 = [\{\kappa_1(\psi) + \kappa_1(\phi)\}^t, \kappa_1^t(\chi)]^t$, where

$$\begin{aligned}
\kappa_1(\psi) &= \int \pi(1, z, w) M(z, x) f(w|z, x, d = 1, \Theta) q(z, x) \\
&\qquad \times \left[H^{(1)}(\cdot) - \frac{H(\cdot)\mathcal{G}\{\cdot, \pi(1 - H), \Theta\}}{\mathcal{G}(\cdot, \pi, \Theta)}\right] d\mu(z, x, w), \\
\kappa_1(\phi) &= \int H(\cdot) f(w|z, x, d = 1, \Theta) q(z, x) \\
&\qquad \times \left[\{1 - \pi(1, z, w)\}\phi(1, z, w) - \pi(1, z, w)\right. \\
&\qquad\quad \left. \times \frac{\mathcal{G}\{\cdot, (1 - \pi)\phi, \Theta\}}{\mathcal{G}(\cdot, \pi, \Theta)}\right] d\mu(z, x, w),
\end{aligned}$$

and

$$\begin{aligned}
\kappa_1(\chi) &= \int \pi(1, z, w) H(\cdot) f(w|z, x, d = 1, \Theta) q(z, x) \\
&\qquad \times \left[\chi(w, z, x, d = 1, \Theta) - \frac{\mathcal{G}(\cdot, \pi\chi, \Theta)}{\mathcal{G}(\cdot, \pi, \Theta)}\right] d\mu(z, x, w).
\end{aligned}$$

With θ_0 being the intercept, $(\partial/\partial\theta_0)\Psi_2 = 0$, $(\partial/\partial\theta_0)\psi = M(z, x)H^{(1)}$, $(\partial/\partial\theta_0)H^d(1 - H)^{1-d} = (2d - 1)H^{(1)}$, and for any function $\xi(\cdot)$, $(\partial/\partial\theta_0)\mathcal{G}(\cdot, \xi, \Theta) = \mathcal{G}(\cdot, \xi_{\theta_0}, \Theta) + \mathcal{G}\{\cdot, (d - H)\xi, \Theta\}$. Writing $\mathcal{G}(\cdot, \xi, \Theta) = \mathcal{G}(\xi)$, by direct but tedious algebra, we find that if $\Psi_1 = (\psi_{1a}^t, \psi_{1b}^t)^t$, then

$$\begin{aligned}
&\mathcal{G}(\pi)(\partial/\partial\theta_0)\psi_{1a} \\
&\quad = -M(z, x)\left[\mathcal{G}\{\pi(d - H)^2\} - \frac{\mathcal{G}^2\{\pi(d - H)\}}{\mathcal{G}(\pi)}\right] \\
&\qquad - \left[\mathcal{G}\{(1 - \pi)(d - H)\phi\} - \frac{\mathcal{G}\{(1 - \pi)\phi\}\mathcal{G}\{\pi(d - H)\}}{\mathcal{G}(\pi)}\right].
\end{aligned} \tag{A.8}$$

Clearly, $(\partial/\partial\theta_0)\psi_{1a}$ does not depend on d. Remembering that $(\partial/\partial\theta_0)\Psi_2 = 0$, the part of the first column of $\mathcal{T}_\Theta(\Theta)$ corresponding to ψ_{1a} is

$$\begin{aligned}
&\sum_{d=0}^{1} \int \pi(\cdot)(\partial/\partial\theta_0)\psi_{1a}(\cdot) H^d(\cdot) \\
&\qquad \times \{1 - H(\cdot)\}^{1-d} f(\cdot\,|\,\cdot, \Theta) q(\cdot)\, d\mu(z, x, w) \\
&\quad = \int \mathcal{G}(\cdot, \pi, \Theta)(\partial/\partial\theta_0)\psi_{1a}(\cdot) q(\cdot)\, d\mu(z, x, w).
\end{aligned} \tag{A.9}$$

We now substitute the right side of (A.8) into (A.9), noting that it factors naturally into components depending on ϕ and ψ, the latter through M. We thus rewrite (A.9) as

$$-\int \{G_1(\psi) + G_2(\phi)\} q(\cdot)\, d\mu(z, x, w). \tag{A.10}$$

Direct but tedious algebra shows that the integrands corresponding to ψ and ϕ in (A.10) exactly equal the integrands in $\kappa_1(\psi)$ and $\kappa_1(\phi)$.

Similarly, the part of the first column of $\mathcal{T}_\Theta(\Theta)$ corresponding to ψ_{1b} is

$$\begin{aligned}
&-\int \left[\mathcal{G}\{\cdot, \pi\chi(d - H), \Theta\}\right. \\
&\qquad \left. - \frac{\mathcal{G}\{\cdot, \pi\chi, \Theta\}\mathcal{G}\{\cdot, \pi(d - H), \Theta\}}{\mathcal{G}(\cdot, \pi, \Theta)}\right] q(\cdot)\, d\mu(z, x, w).
\end{aligned}$$

It can be shown that the integrand of this expression equals the integrand of $\kappa_1(\chi)$.

We have thus shown that κ_1 is proportional to the first column of $\mathcal{T}_\Theta(\Theta)$, and hence that prospective covariance formulas may be used.

A.6 Theory for Nondifferential Measurement Error

The analysis requires a small notational change, namely to interchange the roles of x and w. Dropping the stratum indicators, (10) then becomes

$$(n/n_d) q(z, w) H_R^d(\cdot)\{1 - H_R(\cdot)\}^{1-d} f_{X|Z,W,D}(x|z, w, d, \theta_2);$$
$$H_R(\cdot) = H\{\theta_0 + \theta_{11}^t z + R(z, w, \Theta)\}.$$

In our theory, $H_R(\cdot)$ replaces $H(\cdot)$ and $f_{X|Z,W,D}(\cdot)$ replaces $f(\cdot)$.

Dropping stratum indicators, the estimating equations for maximum likelihood then fit into our notation with $J = 2$:

$$\Psi_1(\cdot) = (\mathcal{M}^t(\cdot)\{D - H_R(\cdot)\}, S_{\theta_{12}}^t, S_{\theta_2}^t)^t$$

and

$$\Psi_2(\cdot) = (\mathcal{M}^t(\cdot)\{D - H_R(\cdot)\}, 0^t, 0^t)^t,$$

where $S_{\theta_2} = (\partial/\partial\theta_2)\log\{f_{X|Z,W,D}(X|Z, W, D, \Theta)\}$ and $S_{\theta_{12}}$ is defined similarly. In addition, $\mathcal{M}(Z, W) = (1, Z^t, U_{\theta_{12}}, U_{\theta_2}^t)^t$, where $U_{\theta_2} = (\partial/\partial\theta_2)R(Z, W, \Theta)$ and similarly for $U_{\theta_{12}}$.

It is easy to show that these prospective estimating equations are retrospectively unbiased. With the redefinition of $l(\cdot)$, the first column of $\mathcal{T}_\Theta(\Theta)$ is

$$\begin{aligned}
\mathcal{T}_{\Theta 1}(\Theta) &= -\left(\sum_{d=0}^{1} \int \mathcal{M}^t(z, w) H_R^{(1)}(\cdot) l(\cdot, \Theta)\, d\mu(z, x, w), 0^t, 0^t\right)^t \\
&= -\left(\sum_{d=0}^{1} \int \mathcal{M}^t(z, w) H_R^{(1)}(\cdot) H_R^d(\cdot)\{1 - H_R(\cdot)\}^{1-d}\right. \\
&\qquad \left. \times q(z, w) f_{X|Z,W,D}(x|z, w, d, \Theta)\, d\mu(z, x, w), 0^t, 0^t\right)^t \\
&= -\left(\int \mathcal{M}^t(z, w) H_R^{(1)}(\cdot) q(z, w)\, d\mu(z, w), 0^t, 0^t\right)^t,
\end{aligned} \tag{A.11}$$

the last step following because the only term depending on x is $f_{X|Z,W,D}(x|z, w, d, \Theta)$, which is a density and hence integrates to 1. That (A.11) equals $-\kappa_1$ is immediate.

We have thus shown that the Satten and Kupper (1993) "unconditional" method for prospective studies can be applied without change to retrospective studies.

A.7 Theory for Corrections for Attenuation

Let the mean of W_1 among the controls be μ_w and the mean of $W_1 - W_2$ among the controls be $\mu_e = 0$. For technical reasons having to do with the fact that σ_u^2 is being estimated by the sample variance $\hat{\sigma}_u^2$, we must include an estimating equation for μ_e even though it is known; this estimating equation has no effect on the standard error estimates. The estimating equations for this algorithm are, with $\theta_2 = (\mu_w, \mu_e, \sigma_w^2, \sigma_u^2)^t$,

$$\Psi_{11} = \begin{bmatrix} \begin{pmatrix} 1 \\ 0 \\ g_1(W_1, \sigma_w^2, \sigma_u^2, \mu_w) \end{pmatrix} [d - H\{\theta_{01} + \theta_1 g_1(W_1, \sigma_w^2, \sigma_u^2)\}] \\ (W_1 - \mu_w) I(D = 0) \\ \{(W_1 - \mu_w)^2 - \sigma_w^2\} I(D = 0) \\ 0 \\ 0 \end{bmatrix}$$

and

$$\Psi_{11} = \begin{bmatrix} \begin{pmatrix} 0 \\ 1 \\ g_2(\bar{W}, \sigma_w^2, \sigma_u^2, \mu_w) \end{pmatrix} [d - H\{\theta_{02} + \theta_1 g_2(\bar{W}, \sigma_w^2, \sigma_u^2)\}] \\ (W_1 - \mu_w) I(D = 0) \\ \{(W_1 - \mu_w)^2 - \sigma_w^2\} I(D = 0) \\ \{(W_1 - W_2)/2^{1/2} - \mu_e\} \\ [\{(W_1 - W_2)/2^{1/2} - \mu_e\}^2 - \sigma_u^2] \end{bmatrix}.$$

By treating the approximations we have made to be exact, it is easily shown that the estimating equations are unbiased. If $f(w|x)$ is the density function of $W = (W_1, W_2)$ given X, because κ_{1s} is based on the case $d = 1$ it follows trivially that

$$\kappa_{1s} = \int \{a, b, g_s(\cdot), 0, 0, 0, 0\}^t H_s^{(1)}(\cdot) q(\cdot) f(w|x)\, d\mu(x, w),$$

where $(a, b) = (1, 0)$ and $(0, 1)$ for $s = 1, 2$. It is also easily verified that the first column of $\mathcal{T}_\Theta(\Theta)$, which is based on the expectations of the derivative of Ψ_{11} with respect to θ_{01}, equals $-\kappa_{11}$, whereas the second column of $\mathcal{T}_\Theta(\Theta)$ is $-\kappa_{12}$. Hence, subject to the levels of approximation described, the conclusion is that prospective covariance formulas may be used for the estimate of θ_1.

We estimated standard errors using a prospective formulation. The covariances of the terms Ψ_{11} and Ψ_{12} were originally computed using the "model-free" method, with the usual exception that in the upper 3 × 3 matrix corresponding to the logistic parameters, we replaced the model-free terms by the usual information contributions.

For the terms corresponding to the derivatives of Ψ_{11} and Ψ_{12}, we again started with the model-free method and again modified the upper 3 × 3 matrix by substituting information contributions. For the other terms in the first three rows of Ψ_{11} and Ψ_{12}, we explicitly used the prospective result that eliminates the contributions of the derivatives of the terms g_1 and g_2 when they are outside of $H(\cdot)$, because prospectively terms such as

$$\left\{\frac{\partial}{\partial \sigma_u^2} g_1(W_1, \sigma_w^2, \sigma_u^2, \mu_w)\right\} [D - H\{\theta_{01} + \theta_1 g_1(W_1, \sigma_w^2, \sigma_u^2, \mu_w)\}]$$

have mean zero. Generalization to vector predictors is immediate.

A.8 Theory for Partial Questionnaires

We will use the estimating equations from prospective maximum likelihood, which can be described as follows. Define

$$G_4 = Hq, \qquad G_3 = \int Hq\, d\mu(x_1),$$

$$G_2 = \int Hq\, d\mu(x_2), \qquad G_1 = \int Hq\, d\mu(x_1, x_2),$$

$$A_4 = q, \qquad A_3 = \int q\, d\mu(x_1),$$

$$A_2 = \int q\, d\mu(x_2), \qquad A_1 = \int q\, d\mu(x_1, x_2),$$

$$\mathcal{C}_j(\cdot) = \{\partial/\partial(\theta_0, \theta_{11}, \theta_{12}, \theta_{13})\} G_j(\cdot),$$

$$L_j(\cdot) = (\partial/\partial\theta_2) G_j)(\cdot),$$

$$\mathcal{M}_j(\cdot) = (\partial/\partial\theta_2) A_j(\cdot), \qquad M(\cdot) = (1, x_1, x_2, z)^t.$$

The prospective estimating equations are in our general form with

$$\Psi_{j1} = \begin{pmatrix} \mathcal{C}_j(\cdot)\{DA_j(\cdot) - G_j(\cdot)\}/[G_j(\cdot)\{A_j(\cdot) - G_j(\cdot)\}] \\ DL_j(\cdot)/G_j(\cdot) + (1 - D)\{\mathcal{M}_j(\cdot) - L_j(\cdot)\}/\{A_j(\cdot) - G_j(\cdot)\} \end{pmatrix}.$$

The estimating equation is retrospectively unbiased, and by definition we write

$$\begin{aligned} \kappa_{11} &= (\kappa_{11a}^t, \kappa_{11b}^t)^t \\ &= \sum_{j=1}^{4} \int \pi_j(z, 1) \Psi_{j1}(1, z, x_1, x_2, \Theta) H(\cdot) q(\cdot)\, d\mu(x_1, x_2, z). \end{aligned}$$

Define $d\mu_4(\cdot) = d\mu(x_1, x_2, z)$, $d\mu_3(\cdot) = d\mu(x_2, z)$, $d\mu_2(\cdot) = d\mu(x_1, z)$, and $d\mu_1(\cdot) = d\mu(z)$. Recall that $(\partial/\partial v)H(v) = H^{(1)}(v) = H(v)\{1 - H(v)\}$ and define $R(\cdot) = (\partial/\partial\theta_2) q(\cdot)$. Then

$$\kappa_{11a} = \sum_{j=1}^{4} \int \pi_j(z, 1)\{\mathcal{C}_j(\cdot)/G_j(\cdot)\} H(\cdot) q(\cdot)\, d\mu(x_1, x_2, z).$$

Note, however, that $\mathcal{C}_j(\cdot)$ and $G_j(\cdot)$ depend only on z, (x_1, z), (x_2, z), and (x_1, x_2, z) for $j = 1, 2, 3, 4$, so that

$$\begin{aligned} \kappa_{11a} &= \sum_{j=1}^{4} \int \pi_j(z, 1) \mathcal{C}_j(\cdot)\, d\mu_j(\cdot) \\ &= \sum_{j=1}^{4} \int \pi_j(z, 1) \mathcal{C}_4(\cdot)\, d\mu(x_1, x_2, z) \\ &= \int \mathcal{C}_4(\cdot)\, d\mu(x_1, x_2, z) \\ &= \int M(\cdot) H^{(1)}(\cdot) q(\cdot)\, d\mu(x_1, x_2, z). \end{aligned}$$

Similarly,

$$\begin{aligned} \kappa_{11b} &= \sum_{j=1}^{4} \int \pi_j(z, 1)\{L_j(\cdot)/G_j(\cdot)\} H(\cdot) q(\cdot)\, d\mu(x_1, x_2, z) \\ &= \int L_4(\cdot)\, d\mu(x_1, x_2, z) = \int H(\cdot) R(\cdot)\, d\mu(x_1, x_2, z). \end{aligned}$$

Because $\mathcal{T}_\Theta(\Theta)$ and $C(\Theta)$ of (13)-(14) are the same as in the prospective case, and the estimating equations are obtained from maximum likelihood, it follows that $\mathcal{T}_\Theta(\Theta) = -C(\Theta)$; this may be verified directly by algebra. It is easier to compute $C(\Theta)$, which is given by

$$C(\Theta) = \sum_{j=1}^{4} \int \pi_j(z, 1)\begin{pmatrix} \mathcal{C}_j(\cdot) \\ L_j(\cdot) \end{pmatrix}\begin{pmatrix} \mathcal{C}_j(\cdot) \\ L_j(\cdot) \end{pmatrix}^t \{G_j(\cdot)\}^{-1}\, d\mu_j(\cdot)$$

$$+ \sum_{j=1}^{4} \int \pi_j(z, 1)\begin{pmatrix} -\mathcal{C}_j(\cdot) \\ \mathcal{M}_j(\cdot) - L_j(\cdot) \end{pmatrix}\begin{pmatrix} -\mathcal{C}_j(\cdot) \\ \mathcal{M}_j(\cdot) - L_j(\cdot) \end{pmatrix}^t$$

$$\times \{A_j(\cdot) - G_j(\cdot)\}^{-1}\, d\mu_j(\cdot).$$

When (Z, X_1, X_2) are all binary and their distribution is left unspecified, detailed considerations show that prospective covariance formulas are asymptotically correct.

A.9 Theory for Two-Stage Studies

Let n_d be the number of observations with $D = d$, let n_{md} be the random number of observations with $(D = d, Z = m)$, and let n^*_{md} be the fixed number of observations in the second stage within each (D, Z) category. Define $\theta_{2,md} = \mathrm{pr}(Z = m \mid D = d)$. Note that $(n_{1d}, \ldots, n_{Md})$ is a multinomial random variable with probabilities $(\theta_{2,1d}, \ldots, \theta_{2,Md})$.

Because there is only a single stratum, we will drop the stratum assignment indicators. In (11), $j = 1$ refers to observations selected into the second-stage sample, and $j = 2$ denotes those which are not so selected. Define $\psi_{21}(\cdot, \Theta) = 0$ and

$$\psi_{11}(\cdot, \Theta) = (1, X^t)^t[d - H\{\theta_0 + \theta_1^t X + \log(n_0/n_1)$$

$$+ \log(n^*_{Z1}/n^*_{Z0}) + \log(\theta_{2,Z0}/\theta_{2,Z1})\}]. \quad \text{(A.12)}$$

Let $\psi_{12} = \psi_{22}(\cdot, \Theta)$ be the vector of size $2M$ whose $(dM + m)$th element equals $I(D = d)\{I(Z = m) - \theta_{2,md}\}$.

Let $\Psi_j = (\psi^t_{j1}, \psi^t_{j2})^t$. Denote the logistic argument in (A.12) by $H_*(Z, X, \Theta)$ and write $H^{(1)}_* = H_*(1 - H_*)$.

At the end of this section, the estimating equation is shown to be unbiased, the particular method being to condition on all $n_{md} \geq n^*_{md}$ or, equivalently, on all the δ's. In addition, with $(\tilde{D}, \tilde{Z})$ denoting the collection of all (D, Z)'s, we later show that

$$0 = E\left\{n^{-1} \sum_{i=1}^{n} \delta_{i1}\psi_{11}(\cdot, \Theta) \mid \tilde{D}, \tilde{Z}\right\}. \quad \text{(A.13)}$$

Next we show that the estimating equations for (θ_0, θ_1) are uncorrelated with those for $(\theta_{2,10}, \ldots, \theta_{2,M1})$, so that the covariance matrix $A(\Theta) = C(\Theta) - D(\Theta)$ is block diagonal. To see this, first note that the off-diagonal term in the covariance matrix is

$$n^{-1}E\left[\left\{\sum_{i=1}^{n} \delta_{i1}\psi_{11}(\cdot, \Theta)\right\}\right.$$

$$\left.\times \left\{\sum_{i=1}^{n} \delta_{i1}\psi_{12}(\cdot, \Theta) + n^{-1}\sum_{i=1}^{n} \delta_{i2}\psi_{22}(\cdot, \Theta)\right\}^t \mid \tilde{D}\right].$$

The terms associated with $\psi_{12} = \psi_{22}$ depend only on $(\tilde{D}, \tilde{Z})$. So, if we condition on this term and apply (A.13), then we have the desired result.

Breslow and Cain did not use our estimating equation approach, but their results are equivalent to ours, except that they worked with the parameterization $\xi_{md} = \log(\theta_{2,md})$. In our notation we have shown that

$$\mathcal{T}_{\Theta}(\Theta) = \begin{bmatrix} \mathcal{T}_{\Theta 11} & \mathcal{T}_{\Theta 12} \\ 0 & \mathcal{T}_{\Theta 22} \end{bmatrix}; \qquad A(\Theta) = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}.$$

Except for the change in parameterization and the fact that our asymptotics are based on the total sample size n rather than on the total $n_* = \sum_{m,d} n^*_{md}$ (a notational difference of no effect on the final results), their term H corresponds to our $\mathcal{T}_{\Theta 11}$, their term A to our $\mathcal{T}_{\Theta 12}$, their term **B** to our A_{22}, and, as we show at the end of this section, their term G to our A_{11}.

We now turn to filling in the main technical steps. We first show (A.13). The notation is that $H_*(\cdot, \Theta)$ refers to the logistic argument in (A.12). Because the conditional density or probability mass function $f(x \mid z, d) = f(x, z \mid d)/\theta_{2,zd}$, then

$$E\left[\sum_{i=1}^{n} \delta_{i1}(1, X_i^t)^t\{D_i - H_*(\cdot, \Theta)\} \mid \tilde{D}, \tilde{Z}, \tilde{\delta}\right]$$

$$= \sum_{m=1}^{M}\sum_{d=0}^{1} n^*_{md}E[(1, X^t)^t\{d - H_*(\cdot, \Theta)\} \mid \tilde{D}, \tilde{Z}, \tilde{\delta}]$$

$$= \sum_{m=1}^{M}\sum_{d=0}^{1} n^*_{md}E[(1, X^t)^t\{d - H_*(\cdot, \Theta)\} \mid \tilde{D}, \tilde{Z}]$$

$$= \sum_{m=1}^{M} \int (1, x^t)^t q(x, m)\left[\sum_{d=0}^{1} \frac{n^*_{md}}{n_d\theta_{2,md}/n}\{d - H_*(\cdot)\}\right.$$

$$\left.\times H^d(\cdot)\{1 - H(\cdot)\}^{1-d}\right] d\mu(x).$$

(A.14)

However, the term in square brackets in (A.14) equals zero, proving (A.13).

Next we prove that the estimating equation is unbiased. The part corresponding to $\psi_{11}(\cdot, \Theta)$ is unbiased by (A.13). For the other part, for specificity consider the estimating equation corresponding to $\theta_{2,m1}$. This has expectation

$$(n_1/n) \int \{I(z = m) - \theta_{2,m1}\}(n/n_1)H(\cdot)q(x, z)\, d\mu(x, z).$$

Because $(n/n_1)H(\cdot)q(x, z)$ is the distribution of (X, Z) given $D = 1$, the estimating equation has expectation $(n_1/n)\{\mathrm{pr}(Z = m \mid D = 1) - \theta_{2,m1}\} = 0$.

Now we show that our A_{11} is the same as Breslow and Cain's matrix **G**, except for the replacement in our calculations of n^{-1} by their n_*^{-1}, where $n_* = \sum_{z,d} n^*_{zd}$. Define $\xi(\cdot, \Theta) = E\{\psi_{11}(\cdot, \Theta) \mid \tilde{D}, \tilde{Z}\}$. We have that

$$A_{11}(\Theta) = n^{-1}\mathrm{cov}\left\{\sum_{i=1}^{n} \delta_{i1}\psi_{11}(\cdot, \Theta) \mid \tilde{D}\right\}$$

$$= n^{-1}E\left[\mathrm{cov}\left\{\sum_{i=1}^{n} \delta_{i1}\psi_{11}(\cdot, \Theta) \mid \tilde{D}, \tilde{Z}, \tilde{\delta}\right\} \mid \tilde{D}\right]$$

$$+ n^{-1}\mathrm{cov}\left[E\left\{\sum_{i=1}^{n} \delta_{i1}\psi_{11}(\cdot, \Theta) \mid \tilde{D}, \tilde{Z}, \tilde{\delta}\right\} \mid \tilde{D}\right]$$

$$= n^{-1}E\left[\mathrm{cov}\left\{\sum_{i=1}^{n} \delta_{i1}\psi_{11}(\cdot, \Theta) \mid \tilde{D}, \tilde{Z}, \tilde{\delta}\right\} \mid \tilde{D}\right],$$

the last step following from (A.14). Thus

$$A_{11}(\Theta) = n_*^{-1}E\left(E\left[\sum_{i=1}^{n} \delta_{i1}\{\psi_{11}(\cdot, \Theta)\psi^t_{11}(\cdot, \Theta)\right.\right.$$

$$\left.\left. - \xi(\cdot, \Theta)\xi^t(\cdot, \Theta)\} \mid \tilde{D}, \tilde{Z}, \tilde{\delta}\right] \mid \tilde{D}\right)$$

$$= \sum_{z=1}^{M}\sum_{d=0}^{1} (n^*_{zd}/n_*)E\{\psi_{11}(\cdot, \Theta)\psi^t_{11}(\cdot, \Theta)$$

$$- \xi(\cdot, \Theta)\xi^t(\cdot, \Theta) \mid \mathbf{D} = d, Z = z\}.$$

This last term is Breslow and Cain's matrix G.

[Received September 1992. Revised March 1994.]

REFERENCES

Armstrong, B., Howe, G., and Whittemore, A. S. (1988), "Correcting for Measurement Error in a Nutrition Study," *Statistics in Medicine,* 8, 1151–1163.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis,* Cambridge, MA: MIT Press.

Breslow, N. E., and Cain, K. C. (1988), "Logistic Regression for Two-Stage Case-Control Data," *Biometrika,* 75, 11–20.

Buonaccorsi, J. P. (1990), "Double Sampling For Exact Values in the Normal Discriminant Model With Application to Binary Regression," *Communications in Statistics, Part A—Theory and Methods,* 19, 4569–4586.

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993), "Case-Control Studies With Errors in Covariates," *Journal of the American Statistical Association,* 88, 185–199.

Carroll, R. J., and Pederson, S. (1993), "Further Remarks on Robustness in the Logistic Regression Model," *Journal of the Royal Statistical Society,* Ser. B, 80, 461–465.

Carroll, R. J., and Stefanski, L. A. (1990), "Approximate Quasi-likelihood Estimation in Models With Surrogate Predictors," *Journal of the American Statistical Association,* 85, 652–663.

——— (1994), "Measurement Error, Instrumental Variables, and Corrections for Attenuation With Applications to Meta-Analysis," *Statistics in Medicine,* 13, 1265–1282.

Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society,* Ser. B, 53, 573–585.

Copas, J. B. (1988), "Binary Regression Models for Contaminated Data (with discussion)," *Journal of the Royal Statistical Society,* Ser. B, 50, 225–265.

Drum, M., and McCullagh, P. (1993), Comment on "Regression Models for Discrete Longitudinal Response," by Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. *Statistical Science,* 8, 300–301.

Fuller, W. A. (1987), *Measurement Error Models,* New York: John Wiley.

Gleser, L. J. (1990), "Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-Variables Regression Models," in *Statistical Analysis of Measurement Error Models and Application,* eds. P. J. Brown and W. A. Fuller, Providence: American Mathematics Society.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989), "Conditionally Unbiased Bounded Influence Estimation in General Regression Models, With Applications to Generalized Linear Models," *Journal of the American Statistical Association,* 84, 460–466.

Little, R. J. A., and Rubin, D. B. (1987), *Analysis of Missing Data,* New York: John Wiley.

Liu, X., and Liang, K. Y. (1992), "Efficacy of Repeated Measures in Regression Models With Measurement Error," *Biometrics,* 48, 645–654.

Pepe, M. S., and Fleming, T. R. (1991), "A General Nonparametric Method for Dealing With Errors in Missing or Surrogate Covariate Data," *Journal of the American Statistical Association,* 86, 108–113.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika,* 66, 403–411.

Reilly, M., and Pepe, M. S. (1995), "The Mean Score Method for Mismeasured and Auxiliary Covariate Data in Regression Models," *Biometrika,* to appear.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association,* 89, 846–866.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Measurement Error: The Case of Multiple Covariates Measured With Error," *American Journal of Epidemiology,* 132, 734–745.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error," *Statistics in Medicine,* 8, 1051–1070.

Satten, G. A., and Kupper, L. L. (1993), "Inferences about Exposure–Disease Association Using Probability of Exposure Information," *Journal of the American Statistical Association,* 88, 200–208.

Schafer, D. (1993), "Replacement Methods for Measurement Error Models," preprint.

Stefanski, L. A., and Carroll, R. J. (1987), "Conditional Scores and Optimal Scores in Generalized Linear Measurement Error Models," *Biometrika,* 74, 703–716.

Wacholder, S., Carroll, R. J., Pee, D. Y., and Gail, M. H. (1994), "The Partial Questionnaire Design for Case-Control Studies," *Statistics in Medicine,* 13, 623–634.

Wang, C. Y., and Carroll, R. J. (1993), "Robust Estimation in Case-Control Studies," *Biometrika,* 80, 237–241.

——— (1995), "On Robust Logistic Case-Control Studies With Response-Dependent Weights," *Journal of Statistical Planning & Inference,* 43, 331–340.

Weinberg, C. R., and Wacholder, S. (1993), "Prospective Analysis of Case-Control Data Under General Multiplicative Intercept Risk Models," *Biometrika,* 80, 461–465.

Zhao, L. P., and Lipsitz, S. (1992), "Designs and Analysis of Two-Stage Studies," *Statistics in Medicine,* 11, 769–782.

A Semiparametric Mixture Approach to Case-Control Studies With Errors in Covariables

Kathryn Roeder, Raymond J. Carroll, and Bruce G. Lindsay

Methods are devised for estimating the parameters of a prospective logistic model in a case-control study with dichotomous response D that depends on a covariate X. For a portion of the sample, both the gold standard X and a surrogate covariate W are available; however, for the greater portion of the data, only the surrogate covariate W is available. By using a mixture model, the relationship between the true covariate and the response can be modeled appropriately for both types of data. The likelihood depends on the marginal distribution of X and the measurement error density $(W|X, D)$. The latter is modeled parametrically based on the validation sample. The marginal distribution of the true covariate is modeled using a nonparametric mixture distribution. In this way we can improve the efficiency and reduce the bias of the parameter estimates. The results also apply when there is no validation data provided the error distribution is known or estimated from an independent data source. Many of the results also apply to the easier case of prospective sampling.

1. INTRODUCTION

In this article we examine logistic case-control studies with errors in covariables, using a semiparametric mixture model approach. Although we use case-control studies to illustrate our points, many of the results and methods apply more generally.

To study the relationship between disease status and exposure level to a suspected disease causing agent (X), epidemiologists often use retrospective sampling in which the diseased population (the "cases," with $D = 1$) and the disease free population (the "controls," with $D = 0$) are sampled separately to determine their levels of exposure X. Viewed from the perspective of the joint population of (X, D), we are taking observations from the conditional distributions of $X|D$.

Suppose that the probability of disease in the source population can be described by the prospective logistic model, $\Pr(D = 1|X = x) = \mathcal{K}(x) = [1 + \exp(-\beta_0 - \beta_1^t x)]^{-1}$, and the marginal density of X in the population is $g(x)$, which will be modeled as an unknown density. Then the population is described by parameters (β_0, β_1, g). It is well known that standard logistic regression, performed as if D were the dependent variable and the covariates X were fixed, leads to the maximum likelihood (ML) estimate of β_1 for retrospective sampling (Prentice and Pyke 1979). Unless the prevalence of disease in the source population (ϕ) is known, β_0 must be viewed as a nuisance parameter. In Section 2 we shed new light on this well-known result by demonstrating that the retrospective model is identifiable up to a specific equivalence class. If ϕ is unknown, then only β_1 is identifiable. The remaining parameters, β_0 and g, are linked by ϕ.

Epidemiologists often use a validation design, in which some of the data are "complete" in that the covariates are measured both directly (without error) and indirectly (with error), whereas for the remainder of the data, the covariates are measured only indirectly. The latter sample is called the "reduced" or "incomplete" sample. Let W and X denote the covariates measured with and without error. Carroll, Gail, and Lubin (1993) extended the use of the prospective logistic model to account for this errors-in-variables design. Other papers in the area include the work of Armstrong, Whittemore, and Howe (1989), Buonaccorsi (1990), and Satten and Kupper (1993).

A motivating example that we analyze in Section 8 concerns the effect of low-density lipoprotein (LDL) cholesterol (X) on the probability of heart disease (D). Consider a design in which the case-control sample is split into a group of size n_C and another group of size n_R, independent of LDL level. Suppose that total cholesterol (W) and LDL are measured for n_C individuals, of which n_{C1} are cases and n_{C0} are controls. Because LDL is expensive to measure relative to total cholesterol, only total cholesterol is measured in the remaining n_R individuals. The complete and reduced data sets are $\{X_i, W_i, D_i, i = 1, \ldots, n_C\}$ and $\{W_j, D_j, j = 1, \ldots, n_R\}$.

Following Carroll et al. (1993), we assume a parametric conditional distribution for W, given X and D, denoted by $f_{W|X,D}(w|x, d; \alpha)$. We choose a parametric model for the measurement error distribution for two reasons: (a) there is opportunity to test this assumption through the validation data, and (b) in many problems, other studies will also have assessed this assumption, and the error distribution may be transportable from study-to-study. On the other hand, G, the distribution of X corresponding to density g, is left unspecified, largely because this distribution depends on the source population.

A natural way to model errors-in-variables data is with a semiparametric mixture model (Kiefer and Wolfowitz 1956). Let $h_j(\xi; \alpha, \beta)$ denote the conditional likelihood of $(W_j, D_j|X = \xi)$. For the reduced data, it is immediately obvious that the joint likelihood of (W_j, D_j) can be written in the form of a mixture model, $L_j(\alpha, \beta, G)$

Kathryn Roeder is Associate Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. Raymond J. Carroll is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Bruce G. Lindsay is Professor, Department of Statistics, Penn State University, University Park, PA 16802. The authors are grateful to D. Bennett for providing the code for the simulations. Roeder's and Lindsay's research was supported by the National Science Foundation. Carroll's research was supported by National Cancer Institute Grant CA-57030.

Journal of the American Statistical Association
June 1996, Vol. 91, No. 434, Theory and Methods

$= \int h_j(\xi;\alpha,\beta)\,dG(\xi)$. The joint likelihood of the complete data (X_i, W_i, D_i) can also be written in the form of a mixture model, $L_i(\alpha,\beta,G) = \int h_i(\xi;\alpha,\beta)I(x_i = \xi)\,dG(\xi)$, with the use of an indicator function.

But the terms in the conditional (retrospective) likelihood are not of the form of a mixture model. In Section 3 we show that if θ maximizes the joint likelihood, then it lies in an equivalence class of parameters that also maximize the retrospective likelihood. Hence maximizing the joint likelihood is equivalent to maximizing the retrospective likelihood. Thus a semiparametric mixture model approach can be taken to obtain ML estimates after all. Specifically, the retrospective ML estimate of the parameter of interest, β_1, can be obtained by maximizing the joint likelihood over the parametric (α,β) and nonparametric (G) components of the likelihood. A confidence interval for β_1 is obtained from the profile likelihood. Assuming the likelihood satisfies the regularity conditions specified by Kiefer and Wolfowitz (1956), the profile ML estimator, $\hat{\beta}_1$, is consistent. In certain models with conditional structure, the estimates of the parametric component are known to be asymptotically efficient (Lindsay, Clogg, and Grego 1991; van der Vaart, in press). Although many simulation studies have shown high efficiency in other models (e.g., Lesperance 1989), a general theory of efficiency has not yet been developed.

Because X was not measured in the reduced sample, the problem can be viewed as a missing-data problem. In Section 4 we present an EM algorithm to obtain the ML estimate of G and also note gradient-based algorithms appropriate for semiparametric mixture models.

In Section 5 we partially extend the results to models for which the probability of disease depends on additional covariates. In Section 6 we develop an extension to a study design without direct validation data. Suppose that no complete data are available in the sample, but that an independent study has been conducted in which both X and W have been measured. Although none of the other measurement error models can handle this situation, the semiparametric mixture method can analyze data of this form. This model requires that the measurement error be nondifferential (i.e., the distribution of $W|X, D$ does not depend on D).

When validation data are available, a natural competitor to the mixture method is the pseudolikelihood method proposed by Carroll et al. (1993). They estimated the marginal distribution of X using a weighted average of the empirical distributions of $X|D = d$ obtained from the complete data. This estimate is plugged into the likelihood, from which the maximum pseudolikelihood estimates of the remaining parameters is obtained. By modeling the relationship between W and D, using a rough-and-ready estimate of the unobserved distribution of X, the information about β_1 contained in the reduced data can be partially recovered. But because the distribution of W depends on X, there is some additional information in the reduced sample about the distribution of X. Consequently, jointly maximizing the full likelihood should yield more information about β_1 than is available when only the complete data are used to estimate the distribution of X.

In Section 7 we present a simulation experiment that evaluates the performance of the mixture method for various sample sizes and amounts of measurement error. The mixture method always performs as well or better than the pseudolikelihood method. The amount of improvement depends on the sample size and the amount of error in the measurements. In Section 8 we analyze an epidemiological data set and present profile likelihoods.

2. IDENTIFIABILITY

Throughout this section we treat X as a discrete random variable for ease of exposition; however, all of the results extend directly to the continuous case. Assume that the probability of disease in the source population for a given level of exposure can be modeled by the prospective logistic model $\mathcal{K}(x)$. Then it is readily determined that

$$f_D(d) = \left\{\sum_x \mathcal{K}(x)g(x)\right\}^d \left\{\sum_x \bar{\mathcal{K}}(x)g(x)\right\}^{1-d}, \quad (1)$$

where $\bar{\mathcal{K}}(x) = 1 - \mathcal{K}(x)$. It follows that

$$\Pr(X = x|D = d) = \frac{\mathcal{K}(x)^d \bar{\mathcal{K}}(x)^{1-d} g(x)}{\{\sum_x \mathcal{K}(x)g(x)\}^d \{\sum_x \bar{\mathcal{K}}(x)g(x)\}^{1-d}}. \quad (2)$$

For the p-dimensional parameter β_1 to be identifiable from this distribution, it will be assumed that $g(x_j) > 0$ for some set $x_0, x_1, \ldots, x_p$ of affinely independent vectors. The following lemma provides useful information for the consideration of identifiability issues and, later, for maximization of the likelihood.

Lemma 1. Suppose that (β_0, β_1, g) and $(\beta_0^*, \beta_1^*, g^*)$ are two logistic regression models satisfying

$$\Pr(D = 1; \beta_0, \beta_1, g) = \phi$$

and

$$\Pr(D = 1; \beta_0^*, \beta_1^*, g^*) = \phi^*.$$

Then

$$\Pr(X = x|D = d; \beta_0, \beta_1, g) = \Pr(X = x|D = d; \beta_0^*, \beta_1^*, g^*) \quad (3)$$

if and only if

a. $\beta_1 = \beta_1^*$;
b. $\beta_0^* = \beta_0 + \log(\phi^*\bar{\phi})/(\bar{\phi}^*\phi)$; and
c. $g^*(x) = [1 + \exp(\beta_0^* + \beta_1^t x)]/[1 + \exp(\beta_0 + \beta_1^t x)]\, g(x)/\sum_x\{[1 + \exp(\beta_0^* + \beta_1^t x)]/[1 + \exp(\beta_0 + \beta_1^t x)]\}g(x)$,

where $\bar{\phi} = 1 - \phi$ and $\bar{\phi}^* = 1 - \phi^*$.

Proof. See the Appendix.

Thus we see that from the retrospective sample, only the parameter β_1 is fully identifiable. The marginal distribution of X can be determined only up to an equivalence class of functions. But if the true population probability of disease, ϕ, is otherwise known, then the foregoing formulas show that β_0 and g are identifiable.

3. THE ERRORS-IN-VARIABLES MODEL

We assume that the true covariates and the error-prone measurements are available in a validation study consisting of a random sample of n_{C1} cases and n_{C0} controls. Thus the complete data consist of $n_C = n_{C0} + n_{C1}$ observations, $\{X_i, W_i, D_i; i = 1, \ldots, n_C\}$. In addition, we have $n_R = n_{R0} + n_{R1}$ incomplete or reduced observations, $\{W_j, D_j; j = 1, \ldots, n_R\}$, obtained from a random sample of n_{R1} cases and n_{R0} controls. The number of cases overall is $n_1 = n_{C1} + n_{R1}$, and the number of controls is $n_0 = n_{C0} + n_{R0}$. It is assumed that the data are missing X at random (Little and Rubin 1987).

Because we are assuming that the division of the cases and controls into the reduced and complete subsamples is done conditionally on D but independently of the values of X, the conditional distribution of $X = x$ given both $D = d$ and the subsample information is still just the conditional distribution of $X = x$ given $D = d$ that was described by (2) in the previous section: $\Pr(X = x|D = d) = f_{X,D}(x,d)/f_D(d)$, where $f_{X,D}(x,d) = \mathcal{K}(x)^d \bar{\mathcal{K}}(x)^{1-d} g(x)$. The marginal distribution of D, given in (1), can be reexpressed in the form of a mixture,

$$f_D(d) = \left\{ \int \mathcal{K}(x)\, dG(x) \right\}^d \left\{ \int \bar{\mathcal{K}}(x)\, dG(x) \right\}^{1-d}, \qquad (4)$$

where G may be discrete or continuous. Recall from Lemma 1 that there is an equivalence class of choices for (β_0, g) for which the same retrospective model is obtained. Now $\phi = f_D(1)$ identifies a particular pair. If it were necessary to evaluate $f_D(1)$ for each subsample (reduced and complete) separately, then it would be necessary to incorporate a differential shift for each β_0 and to estimate a different g for each subsample. The following argument shows why this is not necessary.

In the reduced sample, the likelihood contribution from an observation (W, D) is $f_{W|D}(w|d) = f_{W,D}(w,d)/f_D(d)$, where

$$f_{W,D}(w,d) = \int h(\xi)\, dG(\xi), \qquad (5)$$

and

$$h(\xi) = \mathcal{K}(\xi)^d \bar{\mathcal{K}}(\xi)^{1-d} f_{W|X,D}(w|\xi, d). \qquad (6)$$

In the complete subsample, the likelihood contribution from an observation (W, X, D) is $f_{W,X|D}(w, x|d) = f_{W,X,D}(w,x,d)/f_D(d)$, where

$$f_{W,X,D}(w,x,d) = \int h(\xi) I\{x = \xi\}\, dG(\xi). \qquad (7)$$

Note the device of using an indicator function so that both (5) and (7) are written as integrals $dG(\xi)$.

Putting these together, we get that the likelihood to maximize is of the form

$$L_C(\alpha, \beta, G) = \frac{L_J(\alpha, \beta, G)}{L_M(\beta, G)},$$

where

$$L_J(\alpha, \beta, G) = \prod_{j=1}^{n_R} f_{W,D}(w_j, d_j) \prod_{i=1}^{n_C} f_{W,X,D}(w_i, x_i, d_i) \qquad (8)$$

and

$$L_M(\beta, G) = \{f_D(1)\}^{n_1} \{f_D(0)\}^{n_0}.$$

In the foregoing, C, J, and M stand for conditional, joint, and marginal. Note that as functions of G, both L_J and L_M have the product-of-integrals form $\prod_i \int t_i(\xi)\, dG(\xi)$, for some nonnegative functions $t_i(\xi)$. It follows that the conditional likelihood $L_C(\alpha, \beta, G)$ is a ratio of two products of integrals. Consequently, we cannot apply the theory of nonparametric mixture models to the retrospective likelihood. Here now we run into some good fortune, as we can simply estimate the parameters from $L_J(\alpha, \beta, G)$, which does have mixture form. The parameter estimates that we so obtain will be in an equivalence class of estimators that maximize $L_C(\alpha, \beta, G)$, *being that member of that equivalence class that fits the observed fractions of cases and controls.* Moreover, the profile likelihood for β_1 from L_J is the same as from L_C.

We state the result in a general manner, as it has applications in other problems (see, e.g., Lindsay et al. 1991). Suppose that for a two-parameter model (θ, ϕ), we have a likelihood decomposition of the form

$$L_C(\theta, \phi) = \frac{L_J(\theta, \phi)}{L_M(\theta, \phi)},$$

where $L_M(\theta, \phi)$ is of multinomial form: $L_M(\theta, \phi) = \prod_k^K [p_k(\theta, \phi)]^{m_k}$, where $m = m_1 + \cdots + m_K$ and $p_1(\theta, \phi) + \cdots + p_K(\theta, \phi) = 1$. In this setting we have the following simple theorem.

Theorem 2. Let Ω be the set of all (θ, ϕ) in the parameter space such that there exists ϕ^* depending on (θ, ϕ) satisfying (a) $L_C(\theta, \phi) = L_C(\theta, \phi^*)$; and (b) $p_k(\theta, \phi^*) = m_k/m$. Suppose that in the set of (θ, ϕ) maximizing $L_C(\theta, \phi)$, there exists $(\hat{\theta}_C, \hat{\phi}_C)$ in Ω. Then the following hold:

1. $(\hat{\theta}_C, \hat{\phi}_C^*)$ maximizes $L_J(\theta, \phi)$.
2. If $(\hat{\theta}_J, \hat{\phi}_J)$ is any maximizer of $L_J(\theta, \phi)$, then it satisfies (b) and also maximizes $L_C(\theta, \phi)$.

Moreover, if the entire parameter space is in Ω, then L_J and L_C generate the same profile likelihoods for θ.

Proof. See the Appendix.

Corollary 3. If $(\hat{\alpha}, \hat{\beta}, \hat{G})$ maximizes $L_J(\alpha, \beta, G)$, then $(\hat{\alpha}, \hat{\beta}, \hat{G})$ also maximizes $L_C(\alpha, \beta, G)$ and satisfies the equation $\Pr(D = 1; \hat{\alpha}, \hat{\beta}, \hat{G}) = n_1/n$. The profile likelihoods for β_1 from L_C and L_J are identical.

Proof. See the Appendix.

Besides extending the result of Prentice and Pyke (1979) to the situation with errors in covariables, this manner of proof considerably clarifies the kind of structural features necessary for an equivalence between a retrospective and prospective likelihood analysis. One needs sufficiently rich structure to fit perfectly the marginal distribution of D without diminishing the ability to fit the conditional distributions of the covariates given D. In particular, it is important to note that if we had modeled the marginal distribution G parametrically, then exact equivalence between prospective and retrospective inferences would not necessarily follow.

4. ALGORITHMS

In this section we present two algorithms for maximizing the likelihood with respect to the mixing distributions for a fixed value of (α, β). Algorithms for estimating (α, β) given G have been discussed by Carroll et al. (1993). To obtain the joint ML estimates for (α, β) and G, we alternate between the two estimation problems. Throughout the discussion of estimates for G, we suppress the dependence of the likelihood on the parametric component of the model (α, β).

4.1 Geometric Results About Mixture Models

We first summarize a number of results about nonparametric mixture estimators due to Lindsay (1983). For a fixed value of (α, β), the problem reduces to one of maximizing a concave functional, $l(G) = sum_{i=1}^{n_C} \log L_i(G) + \sum_{j=1}^{n_R} \log L_j(G)$, over a convex set. The likelihood vector, evaluated at the ML estimate $(L_1(\hat{G}), \ldots, L_i(\hat{G}), \ldots, L_j(\hat{G}), \ldots, L_{n_R}(\hat{G}))$, is unique. Furthermore, the ML estimate $\hat{G}$ is known to be a discrete distribution that has a fixed upper bound, K, on the number of support points: K equals the number of distinct terms in the likelihood vector. Because the maximization problem has these convenient properties, it is not difficult to construct an algorithm that walks up the likelihood surface and converges to the ML estimate regardless of the starting value of G.

The gradient method described later can be applied directly to obtain the ML estimate of G. But the EM algorithm can be conveniently applied only if the problem is simplified somewhat. We presuppose that G has support on a fixed grid, $\xi = (\xi_1, \ldots, \xi_M)$. The problem of finding the ML estimate on this grid inherits all of the convenient properties of the full maximization problem. In particular, algorithms can be constructed that are guaranteed to converge regardless of the starting value.

4.2 EM Algorithms

Because the reduced data are missing X, the problem is suited to the EM algorithm, which estimates missing data and then maximizes the likelihood, given these estimates. The missing data can be thought of as a membership indicator variable for the M possible values of the covariate $\xi = (\xi_1, \ldots, \xi_M)$. For convenience, we assume throughout that ξ is known, although it is possible to implement a more general version of the EM algorithm to estimate the best grid for G. This assumption is justified shortly. The group membership for the complete data is obvious because X is observed; in other words, for the ith observation, the posterior probability of group membership is one for $\xi = X_i$. Consequently, the set of support points, ξ, must include all distinct observed X's. The group membership for reduced data can be estimated in the usual say (see, e.g., Titterington, Smith, and Makov 1985). Let $L_j(G^{(m)}) = \sum_t g^{(m)}(\xi_t) h_j(\xi_t)$. The EM algorithm, at the $(m+1)$st step, puts mass $g^{(m+1)}(\xi_k) = \{A_{Ck} + A_{Rk}\}/(n_C + n_R)$, at ξ_k, where

$$A_{Rk} = \sum_{j=1}^{n_R} \frac{g^{(m)}(\xi_k) h_j(\xi_k)}{L_j(G^{(m)})}$$

and

$$A_{Ck} = \sum_{i=1}^{n_C} I(x_i = \xi_k).$$

For a fixed (α, β), $\lim_{m \to \infty} G^{(m)}$ maximizes the likelihood, provided that the support points are known.

Remark. Although the EM algorithm is typically used to estimate both the location of the support ξ and the mass associated with each point of support, we chose to select a fixed grid of support points and then estimate the probability associated with each point of support. Selection of ξ can be based on the observed X's, W's, and the distributional relationship between X and W. We found it sufficient to include each distinct X from the complete data set as well as a grid of points separated by at most $1/5\hat{\sigma}_{W|X}$, where $\sigma_{W|X}$ may depend on X. The range of the grid can be determined by estimating $E[X|W]$ for the minimum and maximum value of W observed in the experiment; call the minimum predicted value x_l and the maximum predicted value x_h. Define the grid on the interval $[x_l - 2\hat{\sigma}_{W|X=x_l}, x_h + 2\hat{\sigma}_{W|X=x_h}]$. Provided that a sufficiently dense grid is used, the solution will not differ measurably from the exact ML estimate. Time required to perform a single iteration of the algorithm increases directly with the number of grid points. We chose to use a fixed grid in the interest of computational feasibility; the EM algorithm is notoriously slow, especially when both the location and the weights must be estimated.

4.3 Gradient Methods

Because the optimization problem reduces to one of maximizing a concave function over a convex set, gradient-based algorithms are ideally suited to this maximization problem.

The gradient of the log-likelihood at G toward $\delta(\xi)$, a point mass at ξ, is

$$D(G, \xi) = \sum_i [I(x_i = \xi) - 1] + \sum_j \left[\frac{h_j(\xi)}{L_j(G)} - 1 \right]. \quad (9)$$

The algorithms are based on the following characterization due to Lindsay (1983). The mixing distribution $\hat{G}$ maximizes the likelihood if and only if

a. $D(\hat{G},\xi) \le 0$ for all ξ; and
b. $D(\hat{G},\xi) = 0$ in the support of $\hat{G}$.

This result is the basis of the gradient methods. We present a simple but somewhat inefficient algorithm known as the VDM (Federov 1972; Wynn 1970). At the mth step, let $G^{(m)}$ be the current estimator. At each step, find ξ^m to maximize $D(G^{(m)})$, and then find ε_m to maximize the likelihood evaluated at $(1-\varepsilon)G^{(m)} + \varepsilon\delta(\xi^m)$; set $G^{(m+1)} = (1-\varepsilon_m)G^{(m)} + \varepsilon\delta(\xi^m)$. Iterate until condition (a) is virtually satisfied. Other, more efficient versions of this sort of algorithm have been developed (see Boehning 1985, Lesperance and Kalbfleisch 1992, and Wu 1978a,b).

4.4 Combining Algorithms for the Parametric and Nonparametric Components of the Model

Although the likelihood is concave as a function of G, there are no guarantees that it will even be unimodal when viewed as a function of (α,β). Therefore, it is essential that good starting values be obtained for the parametric component of the model. We used the following algorithm with good success.

1. Estimate β from the complete observations (X,D) using a standard logistic regression algorithm. Call this β_C.
2. Estimate α from the complete observations (X,W,D). Call this α_C.
3. From the complete data, estimate G independently from (α,β) using a weighted average of the empirical distribution of X for cases and controls. Call it G_{PL}: this is the estimator used by Carroll et al. (1993).
4. Set $G = G_{\mathrm{PL}}$, and estimate (α,β), using an algorithm such as the modified Newton–Raphson to maximize the likelihood. The complete-data estimators, (α_C,β_C), are natural choices for a starting value. The resulting estimators are the partial likelihood estimators of Carroll et al. (1993). Call these $(\alpha_{\mathrm{PL}},\beta_{\mathrm{PL}})$.
5. Fix the parametric components at $(\alpha_{\mathrm{PL}},\beta_{\mathrm{PL}})$. Using either the gradient algorithm or the EM algorithm, maximize the likelihood over G. If a gradient algorithm is used, then any starting value will work; a natural choice is G_{PL}. If the EM algorithm is used, then any starting value will suffice, provided that positive mass is associated with each grid point. We used a starting value equal to $.9 * G_{\mathrm{PL}} + .1 * \mathrm{Uniform}(\xi)$, where the discrete uniform had equal mass on the grid points ξ.
6. Maximize the likelihood over (α,β), fixing G at the maximum obtained in the previous iteration.
7. Maximize the likelihood over G, fixing (α,β) at the maximum obtained in the previous iteration.
8. Repeat Steps 6 and 7 until convergence.

Because the steps involved in estimating G for a fixed value of (α,β) are guaranteed to converge, all of the difficulties encountered in this algorithm are shared by other parametric ML estimation problems. For example, if the Newton–Raphson algorithm is not modified, then it can overstep the maximum and fail to converge. All of the usual warnings appropriate to maximizing a function apply. By using multiple starting values and checking the likelihood at each step of the parametric estimation scheme, one can ensure convergence eventually. There is no need to monitor the algorithms that estimate the nonparametric component of the model, as they are guaranteed to walk up the likelihood surface.

5. ADDITIONAL COVARIATES MEASURED WITHOUT ERROR

Frequently, additional covariates measured without error will be available. For instance, indicator variables such as sex and smoking status may have been recorded. In this section we extend our methodology to incorporate this extra information. Let Z denote an arbitrary set of covariables measured without error. Model the probability of disease, given all the covariates, as a prospective logistic regression model, where

$$\Pr(D=1|X=\xi,Z=z) = \{1+\exp(-\beta_0-\beta_1^t\xi-\beta_2 z)\}^{-1} \equiv \mathcal{K}(\xi,z).$$

Mimicking the development of the likelihood given in (5)–(7), let

$$h_j(\xi) = \mathcal{K}^{d_j}(\xi,z_j)\bar{\mathcal{K}}(\xi,z_j)^{1-d_j} f_{W|X,Z,D}(w_j|\xi,z_j,d_j). \quad (10)$$

Let $dG_{X,Z}(x,z) \equiv dF_Z(z) \times G_{X|Z}(x|z) \equiv dG_X(x) \times F_{Z|X}(z|x)$ represent the joint distribution of (X,Z) in the case-control population. The contributions to the likelihood from the reduced and complete observations are then

$$L_j(\alpha,\beta,G) = \int_{\mathcal{X}} h_j(\xi;\alpha,\beta)\,dG_{X,Z}(\xi,z) \quad (11)$$

and

$$L_i(\alpha,\beta,G) = \int_{\mathcal{X}} h_i(\xi;\alpha,\beta) I(x_i=\xi)\,dG_{X,Z}(\xi,z). \quad (12)$$

We now arrive at a curious situation where ML procedures for estimating the joint distribution $G_{X,Z}$ will break down. To simplify the issues, consider the case in which we observe variables (W,Z), where W is X observed with error but Z is observed directly. Suppose that the distribution of W given $X=x$ is symmetric and unimodal about x. If the observed Z_i are all distinct, then by writing $G_{X,Z} = G_Z G_{X|Z}$, one can show that the ML estimator for the joint distribution $G_{X,Z}$ is $n^{-1}\sum\delta(w_i,z_i)$, where δ is the point mass distribution. This follows because w_i is the value of x maximizing the error density $f_{W|X}(w_i|x)$. Thus the ML estimator converges to $G_{W,Z}$, not $G_{X,Z}$. Curiously, regardless of whether the variables are both observed without error or both observed with error, ML still produces appropriate estimates. The problem seems to be that the sharpness of the Z observations prevents the pooling together of information over i that is necessary to obtain good estimates of conditional distribution of X given Z. For example, if Z is discrete with a finite number of values, then this problem does not arise, as then the conditional distribution for each $Z=z$ can be consistently estimated and will simply be the mixture estimator of G_X obtained from the set of W's that have $Z=z$.

With this in mind, we must consider alternative strategies for this case. One strategy would be to attempt to model the Z given X distribution. If $f_{Z|X}$ were known, then for the reduced observations, (11) becomes

$$L_j = \int_{\mathcal{X}} h_j(\xi;\alpha,\beta) f_{Z|X}(z|x)\, dG_X(\xi),$$

and the problem is of the same form as that solved previously. Furthermore, if $f_{Z|X}$ is known to follow some parametric form depending on η and $W|X, D, Z$ depends on α, then the results of the previous sections apply with parametric component (η, α, β). The drawback of this approach is that if the form of $Z|X$ is unknown, then the results will depend on parametric modeling assumptions. This problem is especially difficult if Z is multivariate with discrete and continuous components.

Another possibility that retains the nonparametric flavor of our approach and that has the same mathematical structure would be to model $G_{X,Z}$ as a discrete mixture of continuous densities, such as $\sum \pi_j N(\mu_j, hI)$, where h would have to be chosen so as to balance the criteria of providing a flexible family of distributions (h small) and providing sufficient smoothing to alleviate the foregoing problem (h large). This approach has been studied in a slightly different context by Magder and Zeger (in press). Further research is necessary to resolve these issues.

6. NO COMPLETE DATA OBSERVED

Consider a situation in which only W, D have been measured (no validation data). If the measurement error distribution, $f_{W|X,D}$, is known, then the likelihood is of the form $L_J(\beta, G) = \prod_{j=1}^{n_R} f_{W,D}(w_j, d_j)$, where $f_{W,D}$ is given by (5). Notice that this is equivalent to (8) for $n_C = 0$ and α known. Provided that the model satisfies the identifiability constraints of Kiefer and Wolfowitz (1956), β and G can be consistently estimated by the ML estimates even though no complete data are available. In fact this is the usual form of a semiparametric mixture model (see, e.g., Butler and Louis 1992; Lindsay 1995).

If the measurement error distribution is unknown, then it clearly cannot be estimated from the reduced sample. But if an independent data set is available in which both X and W have been measured, then it is possible to proceed as indicated earlier, provided that the measurement error distribution is nondifferential (i.e., does not depend on D). The idea is to use these data to obtain an estimate of the distribution of $W|X$ in the case-control population. The assumption of nondifferential error is necessary, because otherwise $f_{W|X}$ is a function of the study population. In particular,

$$f_{W|X}(w|x) = f_{W|X,D}(w|x, D=0) f_D(0) + f_{W|X,D}(w|x, D=1) f_D(1),$$

which depends on the prevalence of diseased individuals in the population unless $f_{W|X,D} = f_{W|X}$. Consequently, the measurement error distribution is transportable only if it is nondifferentiable. We conclude that the ML methods presented so far are just as applicable in this situation as they are in the usual validation study, provided that the measurement error is nondifferential.

Contrast this scenario to the situation encountered when using methods that require a model for $X|W$ (see, e.g., Satten and Kupper 1993). Clearly, by the same argument, the distribution of $X|W$ is not transportable unless the effect is null; hence such methods cannot be applied in the no-validation situation. The pseudolikelihood method proposed by Carroll et al. (1993) clearly is not applicable either, as it requires an empirical estimate of the distribution of X.

The only other literature that allows for consistent estimation when there is no validation and the error model is estimated independently is the paper by Stefanski and Carroll (1987), who assumed normally distributed measurement error. Our results allow for any error distribution.

To fit the model in this situation, use the following modification of the algorithm given in Section 4.4:

1. Obtain an ad hoc estimate of $E[X|W = w_i], i = 1, \ldots, n$ using the known distribution of $W|X$. Call it $\hat{X}_i$, and refer to $\hat{X}, D, W$ as the pseudo-complete data.
2. Estimate β from the pseudo-complete data $(\hat{X}_i, D_i, i = 1, \ldots, n)$ using a standard logistic regression algorithm. Call this β_{PC}.
3. From the pseudo-complete data, estimate G independently from β using a weighted average of the empirical distribution of $\hat{X}$ for cases and controls. Call it G_{PL}.
4. Set $G = G_{\text{PL}}$, and estimate β using an algorithm such as the modified Newton–Raphson to maximize the likelihood. The pseudo-complete estimator, (β_{PC}), is a natural choice for a starting value. Call the result β_{PL}.
5. Fix the parametric component at β_{PL}. Using either the gradient algorithm or the EM algorithm, maximize the likelihood over G. If a gradient algorithm is used, then any starting value will work; a natural choice is G_{PL}. If the EM algorithm is used, then any starting value will suffice, provided that positive mass is associated with each grid point. We used a starting value equal to $.9 * G_{\text{PL}} + .1 * \text{Uniform}(\xi)$, where the discrete uniform had equal mass on the grid points ξ which are chosen as in Section 4.
6. Maximize the likelihood over β, fixing G at the maximum obtained in the previous iteration.
7. Maximize the likelihood over G, fixing β at the maximum obtained in the previous iteration.
8. Repeat Steps 6 and 7 until convergence.

7. A SIMULATION EXPERIMENT

To examine the performance of our estimator, we simulated data with measurement error under two scenarios: with and without a validation study. We chose a lognormal distribution for both X and $W|X, D$, because this distribution often arises in practice (see, e.g., Nero, Schwehr, and Nazaroff 1986). Distributions and parameter values were chosen to be similar to or identical with those simulated by Carroll et al. (1993). Data were generated prospectively, and then a case-control sample was obtained from this simulated source sample. The prospective logisticmodel was

Table 1. Known Measurement Error Model With No Complete Data, $\beta_1 = .5$

n_{R0}	n_{R1}		$\sigma = 0$	$\sigma = .25$	$\sigma = .50$	$\sigma = .75$
40	40	Mean	.53	.54	.57	.59
		MSE	.061	.078	.111	.180
		RMSE	1.00	1.28	1.82	2.95
120	120	Mean	.51	.50	.53	.55
		MSE	.012	.022	.026	.054
		RMSE	1.00	1.83	2.16	4.50

NOTE: Mean and MSE of $\hat{\beta}_1$ are calculated based on 50 repetitions of the simulation experiment. RMSE is the ratio of the MSE of the model of interest divided by the MSE of the model with $\sigma = 0$.

given by $\Pr(D = 1|X = x) = \mathcal{K}(x)$, with $\beta_0 = -3.09$ and $\beta_1 = .5$. The true covariate, X, was generated as a lognormal random variable so that $\log(X)$ had mean $-1/2\sigma_x^2$ and variance σ_x^2, where $\sigma_x = 1.08$. The surrogate predictor was also lognormal, with $\log(W)$, given (X, D), having mean $\log(X)$ and standard deviation σ, which varied. We repeated each experiment 50 times.

The general form of the measurement error model that we fit is $\log(W) = \alpha_0 + \alpha_1 \log(X) + \sigma\varepsilon$, with $\varepsilon \sim N(0,1)$; the simulated data fell into this class, with $\alpha_0 = 0, \alpha_1 = 1$, and σ varying. In addition to the unknown mixing distribution G, the model also has five free parameters, $(\alpha_0, \alpha_1, \sigma, \beta_0, \beta_1)$; hence we call this the five-parameter model. In the following subsection we assume that $(\alpha_0, \alpha_1, \sigma)$ are known. This leaves only two parametric unknowns, (β_0, β_1); hence we call this the two-parameter model.

We used the EM algorithm to estimate the mixing distribution. We chose sample sizes and parameters to provide valid comparisons for realistic sample sizes and to illustrate the limits of the method. A value of $\sigma = 1$ represents a large amount of measurement error: $\sigma_x^2 \approx \sigma^2$. The method failed to converge occasionally with this amount of measurement error. In practice one could fit this model when $\sigma = 1$, provided that a number of starting values could be used to find the ML estimate. Alternatively, $\sigma < .25$ represents a rather small measurement error. In principle, this method will perform well when $\sigma < .25$; however, the EM grid would have to be quite dense to yield an estimate similar to the actual ML estimate for G. The computations are prohibitively slow for simulations in this situation. Consequently, we report only the performance of the method for $\sigma = .25, .50, .75$.

7.1 Known Measurement Error

In this section we assume that the measurement error distribution is known and examine the performance of the two-parameter model with a small, medium, and large amount of measurement errors ($\sigma = .25, .50, .75$). Samples $\{X_i, W_i, D_i, i = 1, \dots n\}$ of size 80 and 240 were generated, each consisting of half cases and half controls. First, using $\{X_i, D_i, i = 1, \dots n\}$, we estimated β using the prospective logistic model (equivalent to $\sigma = 0$). Next, omitting the gold standard variables X_i and using only $\{W_i, D_i, i = 1, \dots n\}$, we estimated (β, G) using the mixture model (the MIX method). Results are presented in Table 1. Notice that as the variance of the measurement error distribution increased, the mean squared error (MSE) of the MIX method increased rapidly. In fact, when $\sigma = 1$, the reduced data appeared to have almost no information remaining relative to when $\sigma = 0$ (results not reported). As the sample size increased, the MSE decreased substantially, as would be expected. Of greater interest is the increase in the disparity between the performance with and without measurement error. This is especially apparent for models with larger measurement error.

For sample sizes smaller than 80, it is unlikely that the semiparametric model provides any advantages over a model that assumes a parametric form for the unobserved covariate. For smaller sample sizes, the data simply do not provide sufficient information from which to estimate the distribution of this unobservable accurately in a nonparametric setting.

Next, we simulate data from a validation study. To see the effect of increasing amount of complete data, first $\frac{1}{8}$ and then $\frac{1}{4}$ of the data were considered as complete. The performance of the MIX method is compared with the pseudolikelihood (PL) method proposed by Carroll et al. (1993) in Table 2. We found that if a substantial proportion of the data are complete ($\frac{1}{4}$ or more), then the PL and MIX methods perform similarly. When only $\frac{1}{8}$ of the data are complete, the pattern of results is markedly different. The MSE of both methods increased dramatically; however, the MIX method performed much better than the PL method for this scenario. The difference in performance was greatest when the variance of the measurement error was greatest.

For the most part, we found that the MIX method either provided the same estimate as the PL method or provided one that was better. When the number of reduced observations was not large relative to the number of complete

Table 2. Known Measurement Error Model With Some Complete Data, $\hat{\beta}_1 = .5$

					$\sigma = .25$		$\sigma = .50$		$\sigma = .75$	
n_{C0}	n_{C1}	n_{R0}	n_{R1}		PL	MIX	PL	MIX	PL	MIX
30	30	90	90	Mean	.496	.496	.500	.502	.512	.505
				MSE	.016	.016	.021	.020	.038	.032
				RMSE		1.00		1.05		1.19
15	15	105	105	Mean	.512	.501	.568	.515	.611	.522
				MSE	.040	.017	.089	.033	.143	.038
				RMSE		2.35		2.70		3.76

NOTE: Mean and MSE of β_1 are calculated based on 50 repetitions of the simulation experiment. RMSE is the ratio of the MSE of the PL estimator to the MSE of the MIX estimator.

Table 3. Unknown Measurement Error Model With Some Complete Data, $\beta_1 = .5$

					$\sigma = .25$		$\sigma = .50$		$\sigma = .75$	
n_{C0}	n_{C1}	n_{R0}	n_{R1}		*PL*	*MIX*	*PL*	*MIX*	*PL*	*MIX*
30	30	30	30	Mean	.501	.507	.500	.503	.518	.522
				MSE	.038	.034	.046	.045	.063	.060
				RMSE		1.12		1.15		1.05
30	30	90	90	Mean	.481	.478	.471	.468	.493	.489
				MSE	.025	.022	.034	.027	.067	.048
				RMSE		1.14		1.26		1.40

NOTE: Mean and MSE of β_1 are calculated based on 50 repetitions of the simulation experiment. RMSE is the ratio of the MSE of the PL estimator to the MSE of the MIX estimator.

observations, it was often the case that no information beyond that provided by the PL estimate was available in the data about the distribution of X. In these cases the algorithm essentially stopped at Step 4 (Sec. 4.4) and produced the PL estimate. Thus it is apparent from these simulations that the MIX method outperforms the PL method only when there are substantially more reduced observations than complete observations; however, as n_R grows, the discrepancy between the performance will continue to increase.

When the measurement error distribution is known (or obtained from another study), there is no lower bound on the number of complete cases required. Comparing Table 1 to Table 2, one can see the reduction in the MSE when at least some of the observations are complete (compare simulations with total sample size of 240). As expected, when $\sigma = .25$, the improvement is not substantial: .022 versus .016 and .017. When $\sigma = .75$, the difference is more notable: .054 versus .032 and .038.

7.2 Unknown Measurement Error

In this section we assume that the measurement error distribution is known to be lognormal, but the three parameters of the lognormal model are unknown. This puts us in the framework of the five-parameter model. To compare the MIX method and the PL method, we simulated validation data sets with two different sample sizes and three different measurement errors: $n_{C0} = n_{C1} = 30, n_{R0} = n_{R1} = 30; n_{C0} = n_{C1} = 30, n_{R0} = n_{R1} = 90$; and $\sigma = 25, .50, .75$.

For the first choice of sample sizes, the MIX method and the PL method performed almost identically, suggesting that for data like those simulated, little extra information can be extracted by maximizing over the mixing distribution. But again, the MIX method outperforms the PL method when the number of complete observations is small relative to the number of reduced observations.

The disparity between the methods increased with the variance in the measurement error distribution. The PL method also tends to perform poorly when the measurement error is small, relative to the number of complete observations (not reported). We believe that PL performs poorly in this setting, because when W is not close to any X in the sample, the contribution to the pseudolikelihood is nearly zero and hence the efficiency of the method is reduced.

Comparing the two-parameter and five-parameter results (Tables 2 and 3), it is clear that the variance of the estimator increases substantially when the parameters of the measurement error distribution must be estimated.

We discovered that the algorithm was not stable with less than 60 complete observations, when the five-parameter model was fit. Again, we found that the MIX method tended to provide the same estimate as the PL method, or one that was better. Comparing the two-parameter and five-parameter results with sample sizes 30, 30, 90, 90, we see that the MIX method outperformed the PL method more often when the measurement error distribution was not known; (for example, for $\sigma = .75$, the root mean squared error (RMSE) was 1.40 for the five-parameter model, but only 1.19 for the two-parameter model.)

The measurement error model can be extended to allow for differential error by incorporating two extra parameters: $\log(w) = \alpha_0 + \alpha_1 \log(x) + \alpha_2 d + \sigma_d \varepsilon$. Carroll et al. (1993) performed extensive simulations, comparing the seven-parameter PL model to competing methods in the validation study setting, and found that incorporating the two extra parameters did not significantly reduce the efficiency of the PL model, and, moreover, it provided a substantial increase in the robustness of the model. Our methods demonstrated a serious bias in the estimates of the β_1 when the nondifferential error was ignored. We found that for the sample sizes and parameter values given in Table 3, the seven-parameter MIX method performed equal to or better than the seven-parameter PL method (not reported). In toto, these results suggest that it is worthwhile to use a richer measurement error model unless there is evidence that the measurement error is nondifferential.

Table 4. Coverage of Nominal 95% Confidence Intervals

n_{C0}	n_{C1}	n_{R0}	n_{R1}	*Model type*	*Coverage*
0	0	40	40	2*p*	92
0	0	120	120	2*p*	94
30	30	90	90	2*p*	95
15	15	105	105	2*p*	96
30	30	30	30	5*p*	92
30	30	90	90	2*p*	93

NOTE: In each case, $\sigma = .5$. The experiment was repeated 250 times to obtain estimated coverage values. The standard error of the coverage is sqrt(p(1 − p)/250), where p is the true coverage probability.

7.3 Profile Likelihoods

There is currently no standard methodology for computing the variance of a profile ML estimate in a semi-

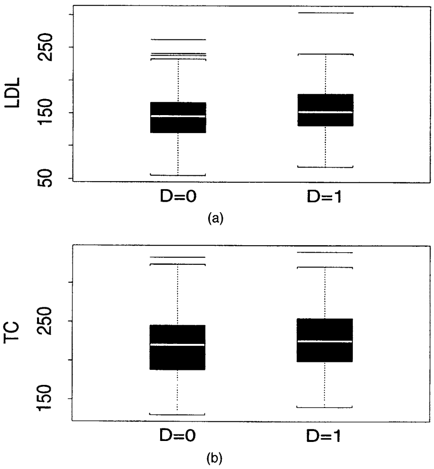


Figure 1. Boxplot of LDL Cholesterol (LDL) and Total Cholesterol (TC) for Individuals With (D = 1) and Without (D = 0) Heart Disease. Here (a) shows the differences in distribution for LDL cholesterol across cases and controls, while (b) shows a weaker, but similar pattern for total cholesterol.

parametric mixture model. Here we let ϕ denote all nuisance parameters, including the mixing distribution, and let β_1 denote the parameter of interest. For a real-valued parameter of interest, Lesperance (1989) proposed inverting the semiparametric generalized likelihood ratio test $\Lambda(\beta_1) = 2\log\sup_{\beta_1,\phi} L(\beta_1,\phi)/\sup_{\phi} L(\beta_1,\phi)$ to obtain a $(1-\alpha)100\%$ profile confidence interval $\{\beta_1\colon \Lambda(\beta_1) < \mathcal{X}_1^2(\alpha)\}$. She observed good coverage properties when using this method for a number of standard mixture problems. We also found good coverage using this method (Table 4). As expected, the results were slightly anticonservative. Based on Lesperance's simulations and ours, we recommend this approach for finding a confidence interval for β_1.

8. A CHOLESTEROL STUDY

In this example we analyze a data set concerning the risk of coronary heart disease (CHD) as a function of blood cholesterol level. These data were extracted from the Lipids Research Clinics study, which was previously discussed by Satten and Kupper (1993). We use a portion of these data involving men age 60–70 who do not smoke (256 records: 4 outliers were removed). A subject is recorded as having CHD ($D = 1$) if they have had a previous heart attack, an abnormal exercise electrocardiogram, history of angina pectoris, and so forth. The measured covariables are low-density lipoprotein (LDL) cholesterol level and total cholesterol (TC) level. Direct measurement of LDL levels is time-consuming and requires costly special equipment. For this reason, we are interested in whether TC serves as a useful surrogate for LDL. Note that the measurement error of TC is not the source of error of primary interest; rather, the unknown quantity of the other components of TC (triglycerides and high density lipoproteins) lead to the "measurement error." Henceforth CHD, LDL/100, and TC/100 play the roles of D, X, and W.

In this data set, both X and W have been recorded for each subject. In the full data set there are 113 cases, of which 47 had LDL levels higher than 160. Among the 143 controls, 43 had elevated LDL levels. Figure 1 presents boxplots of the data. Using X as the predictor, the prospective logistic regression estimate for β_1 was .656 with a standard error of .336. Contrast this with the attenuated estimate (.540) obtained when measurement error was ignored and W was used as the predictor.

A nondifferential lognormal measurement error model provided a good fit to the data (Fig. 2), with the exception of a slight increase in the variance of $W|X$ for small values of X. The differential measurement error model fit significantly better but did not change the parameters enough to have a practical impact on the estimation procedure. Consequently, the measurement error was modeled using a nondifferential error model.

To illustrate the information present in the reduced data, we analyzed a sample of data with and without the reduced observations. From the 113 cases and 143 controls, 32 cases and 40 controls were randomly selected to serve as complete data. The remaining observations were treated as re-

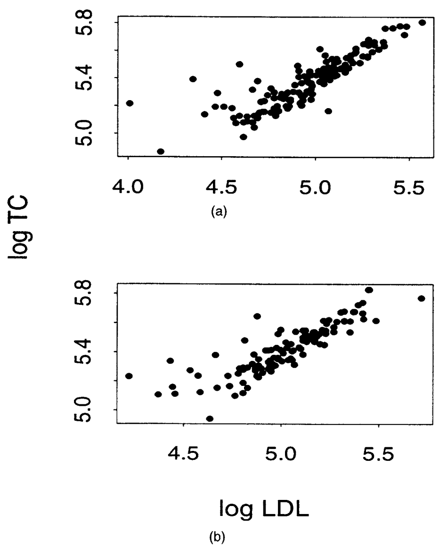


Figure 2. Relationship Between Total Cholesterol (TC) and LDL Cholesterol (LDL). Here (a) illustrates the relationship for controls (D = 0), and (b) illustrates cases (D = 1).

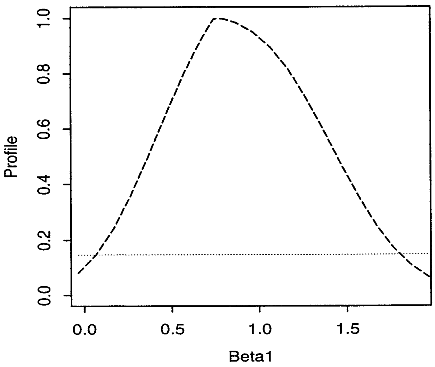


Figure 3. Profile Likelihood of β_1 for the Cholesterol Data. The horizontal bar marks the 95% profile confidence interval.

duced observations. Using only the 72 complete observations, we obtained $\hat{\beta}_1 = .943$ with standard error of .62. Figure 3 illustrates the profile likelihood for the nondifferential model when both complete and reduced data are used: $\hat{\beta}_1 = .765$. To compare, notice that a 95% confidence interval using all of the data as complete had length of 1.34, a 95% profile interval for the 72 complete observations and 184 reduced observations had length 1.75, and a 95% confidence interval for the 72 complete observations only had length 2.48. We conclude that the precision of the estimate increases substantially when both complete and reduced data are used.

9. CONCLUSIONS

In this article, we have suggested using nonparametric mixture methods to estimate regression parameters when one or more of the regression parameters are measured with error. The theoretical results and implementation of the MIX method have both concentrated on the logistic case-control study, allowing for differential measurement error. We have considered situations where the predictor X is observed in a subset of the study (Secs. 3 and 4) or cannot be observed at all (Sec. 6). Simulations (Sec. 7) and an example (Sec. 8) indicate the feasibility of the methodology.

The use of mixture methods, and the MIX method itself, is not restricted to logistic case-control studies but also applies to any prospective (as opposed to case-control) likelihood problem. In principle, one needs only a likelihood for the distribution of the response Y given the predictor X, as well as a parametric error model relating the observed predictor W to (Y, X) (differential error) or relating W to X (nondifferential error). When X is partially observed in this context, there are many other competing techniques (see Robins, Hsieh, and Newey 1995 and references therein). When X is unobserved, as occurs in the classical measurement error problem, nondifferential measurement error is required. Mixture methods apply as discussed in Section 6, as long as there is sufficient information to identify the parameters of the relevant distributions, especially the error distribution of W given X. We discussed in Section 6 the case where there is an independent experiment that estimates the distribution of W given X. It is possible to extend the results to a second case, where there are replicates of W that are sufficient in themselves to identify the error distribution.

APPENDIX: PROOF OF RESULTS FROM SECTIONS 2 AND 3

Proof of Lemma 1

a. $\beta_1 = \beta_1^*$, because the log odds ratio is proportional to β_1 for either retrospective model.
b. Follows from Bayes's theorem.
c. When $d = 1$, (3) implies

$$\mathcal{K}(x)g(x)/\phi = \mathcal{K}^*(x)g^*(x)/\phi^*.$$

It follows that

$$g^*(x) = g(x)\,\frac{\phi^*[1+\exp(\beta_0^* + \beta_1^t x)]}{\phi[1+\exp(\beta_0 + \beta_1^t x)]}\ .$$

For $d = 0$, the same relationship holds. The result follows from noting that $\sum_x g^*(x) = 1$.

Proof of Theorem 2

The key to the proof is that any (θ, ϕ^*) satisfying (b) necessarily maximizes $L_{\mathrm{M}}(\theta, \phi^*)$, as this term is maximized over all possible multinomial probabilities $p_1, \ldots, p_K$. Thus given any $(\hat{\theta}_{\mathrm{C}}, \hat{\phi})$ maximizing $L_{\mathrm{C}}(\theta, \phi)$, the corresponding $(\hat{\theta}_{\mathrm{C}}, \hat{\phi}^*)$ satisfying (a) and (b) simultaneously maximizes both terms in the product $L_{\mathrm{C}}(\theta, \phi)L_{\mathrm{M}}(\theta, \phi)$. If the entire parameter space is in Ω, then for any θ_0 there is a $\hat{\phi}_0^*$ that maximizes $L_{\mathrm{C}}(\theta_0, \phi)$ over ϕ and satisfies $p_k(\theta_0, \hat{\phi}_0^*) = m_k/m$. Consequently, the joint and conditional profile likelihoods for θ are equal.

Proof of Corollary 3

From Lemma 1, for any fixed set of parameters (α, β, G), we can find a set (α^*, β, G^*) giving the same conditional distributions for X given D but having the prespecified value of $\Pr\{D = 1; \alpha^*, \beta, G^*\} = n_1/n$. But it follows that the conditional distributions of W given D are also identical for (α^*, β, G^*) and (α, β, G); hence the values of L_{C} are identical, and so the hypothesis of the theorem is met.

[Received May 1994. Revised July 1995.]

REFERENCES

Armstrong, B. G., Whittemore, A. S., and Howe, G. R. (1989), "Analysis of Case-Control Data With Covariate Measurement Error: Application to Diet and Colon Cancer," *Statistics in Medicine*, 8, 1151–1163.

Boehning, D. (1985), "Numerical Estimation of a Probability Measure," *Journal of Statistical Planning and Inference*, 11, 57–69.

Buonaccorsi, J. P. (1990), "Double Sampling for Exact Values in the Normal Discriminant Model With Application to Binary Regression," *Communications in Statistics, Part A—Theory and Methods*, 19, 4569–4586.

Butler, S. M., and Louis, T. A. (1992), "Random Effects Models With Nonparametric Priors," *Statistics in Medicine*, 11, 1981–2000.

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993), "Case-Control Studies With Errors in Predictors," *Journal of the American Statistical Association*, 88, 185–199.

Federov, V. V. (1972), *Theory of Optimal Experiments*, New York: Academic Press.

Kiefer, J., and Wolfowitz, J. (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 886–906.

Lesperance, M. L. (1989), "Mixture Models as Applied to Models Involving Many Incidental Parameters," unpublished doctoral dissertation, University of Waterloo, Dept. of Statistics and Actuarial Science.

Lesperance, M. L., and Kalbfleisch, J. D. (1992), "An Algorithm for Computing the Nonparametric MLE of a Mixing Distribution," *Journal of the American Statistical Association*, 87, 120–126.

Lindsay, B. G. (1983), "The Geometry of Mixture Likelihoods, Part I: A General Theory," *The Annals of Statistics*, 11, 86–94.

——— (1995), *Mixture Models: Theory, Geometry and Applications*, IMS Monograph Series, Hayward, CA: Institute of Mathematical Statistics.

Lindsay, B. G., Clogg, C. C., and Grego, J. (1991), "Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis," *Journal of the American Statistical Association*, 86, 96–107.

Little, R. J. A., and Rubin, D. B. (1987), *Analysis of Missing Data*, New York: John Wiley.

Magder, L., and Zeger, S. (in press), "A Smooth Nonparametric Estimate of a Mixing Distribution Using Mixtures of Gaussians," *Journal of the American Statistical Association.*

Nero, A. V., Schwehr, M. B., and Nazaroff, W. M. (1986), "Distribution of Airborne Radon-222 Concentration in U.S. Homes," *Science*, 234, 992–997.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.

Robins, J. M., Hsieh, F., and Newey, W. (1995), "Semiparametric Efficient Estimation of a Conditional Density With Missing or Mismeasured Covariates," *Journal of the Royal Statistical Society*, Ser. B, 57, 409–424.

Satten, G. A., and Kupper, L. L. (1993), "Inferences About Exposure–Disease Association Using Probability of Exposure Information," *Journal of the American Statistical Association*, 88, 200–208.

Stefanski, L. A., and Carroll, R. J. (1987), "Conditional Scores and Optimal Scores in Generalized Linear Measurement Error Models," *Biometrika*, 74, 703–716.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley.

Wu, C. F. J. (1978a), "Some Algorithmic Aspects of the Theory of Optimal Designs," *The Annals of Statistics*, 6, 1286–1301.

——— (1978b), "Some Iterative Procedures for Generation of Nonsingular Optimal Designs," *Communications in Statistics, Part A—Theory and Methods*, 7, 1399–1412.

Wynn, H. P. (1970), "The Sequential Generation of D-Optimum Experimental Designs," *The Annals of Mathematical Statistics*, 41, 1655–1664.

van der Vaart, A. (in press), "Efficient MLE in Semiparametric Mixture Models," *The Annals of Statistics*.

American Journal of Epidemiology
Copyright © 2001 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved

Vol. 153, No. 4
Printed in U.S.A.

Empirical Evidence of Correlated Biases in Dietary Assessment Instruments and Its Implications

Victor Kipnis,[1] Douglas Midthune,[1] Laurence S. Freedman,[2] Sheila Bingham,[3] Arthur Schatzkin,[4] Amy Subar,[5] and Raymond J. Carroll[6]

Multiple-day food records or 24-hour recalls are currently used as "reference" instruments to calibrate food frequency questionnaires (FFQs) and to adjust findings from nutritional epidemiologic studies for measurement error. The common adjustment is based on the critical requirements that errors in the reference instrument be independent of those in the FFQ and of true intake. When data on urinary nitrogen level, a valid reference biomarker for nitrogen intake, are used, evidence suggests that a dietary report reference instrument does not meet these requirements. In this paper, the authors introduce a new model that includes, for both the FFQ and the dietary report reference instrument, group-specific biases related to true intake and correlated person-specific biases. Data were obtained from a dietary assessment validation study carried out among 160 women at the Dunn Clinical Nutrition Center, Cambridge, United Kingdom, in 1988–1990. Using the biomarker measurements and dietary report measurements from this study, the authors compare the new model with alternative measurement error models proposed in the literature and demonstrate that it provides the best fit to the data. The new model suggests that, for these data, measurement error in the FFQ could lead to a 51% greater attenuation of true nutrient effect and the need for a 2.3 times larger study than would be estimated by the standard approach. The implications of the results for the ability of FFQ-based epidemiologic studies to detect important diet-disease associations are discussed. *Am J Epidemiol* 2001;153:394–403.

biological markers; dietary assessment methods; epidemiologic methods; measurement error; models, statistical; model selection; regression analysis; research design

Scientists have long sought a connection between diet and cancer. A number of large prospective studies have now challenged conventional wisdom, which was derived in large part from international correlation studies and animal experiments, in reporting no association between dietary fat and breast cancer (1) and, most recently, no association between dietary fiber and colorectal cancer (2). These null epidemiologic findings may ultimately be shown to reflect the truth about these diet-cancer hypotheses. Alternatively, however, the studies themselves may have serious methodological deficiencies.

Usually, in large studies, a relatively inexpensive method of measurement, such as a food frequency questionnaire (FFQ), is employed. Investigators now recognize that errors in the values reported on FFQs can profoundly affect the results and interpretation of nutritional epidemiologic studies (3–5). Dietary measurement error often attenuates (biases toward 1) the estimated disease relative risk and reduces the statistical power to detect an effect. An important relation between diet and disease may therefore be obscured.

Realization of this problem has prompted the integration into large epidemiologic investigations of calibration substudies that involve a more intensive but presumably more accurate dietary reporting method, called the "reference" instrument. Typically, the instruments chosen for reference measurements have been multiple-day food records, sometimes with weighed quantities instead of estimated portion sizes, or multiple 24-hour recalls. FFQs have been "validated" against such instruments, and correlations between FFQs and reference instruments, sometimes adjusted for within-person random error in the reference instrument, have been quoted as evidence of FFQ validity (6, 7). Additionally, on the basis of such studies, statistical methods have been employed to adjust FFQ-based relative risks

Received for publication January 31, 2000, and accepted for publication November 28, 2000.

Abbreviations: AIC, Akaike Information Criterion; BIC, Bayes Information Criterion; FFQ, food frequency questionnaire.

[1] Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD.

[2] Department of Mathematics, Statistics and Computer Science, Bar Ilan University, Ramat Gan, Israel.

[3] Dunn Human Nutrition Unit, Medical Research Council, Cambridge, United Kingdom.

[4] Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

[5] Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD.

[6] Department of Statistics, Texas A&M University, College Station, TX.

Reprint requests to Dr. Victor Kipnis, National Cancer Institute, Executive Plaza North, Room 344, 6130 Executive Blvd., MSC 7354, Bethesda, MD 20892-7354 (e-mail: victor_kipnis@nih.gov).

for measurement error (8), using the regression calibration approach.

The correct application of the regression calibration approach relies on the assumptions that errors in the reference instrument are uncorrelated with 1) true intake and 2) errors in the FFQ (9). Throughout this paper, we take these two conditions as requirements for a valid reference instrument.

Recent evidence suggests that these assumptions may be unwarranted for dietary report reference instruments. Studies involving biomarkers, such as doubly labeled water for measuring energy intake and urinary nitrogen for measuring protein intake (10–16), suggest that reports using food records or recalls are biased (on average, towards underreporting) and that individuals may systematically differ in their reporting accuracy. This could mean that all dietary report instruments involve bias at the individual level, although direct evidence for individual macronutrients other than protein is not yet available. Part of the bias may depend on true intake (which manifests itself in what we call group-specific bias), therefore violating the first assumption for a reference instrument. Part of the bias may also be person-specific (defined below in detail) and may correlate with its counterpart in the FFQ, thereby violating the second assumption.

For this reason, Kipnis et al. (9) proposed a new measurement error model that allows for person-specific bias in the dietary report reference instrument as well as in the FFQ. Using sensitivity analysis, they showed that if the correlation between person-specific biases in the FFQ and the reference instrument was 0.3 or greater, the usual adjustment for measurement error in the FFQ would be seriously incorrect. However, the paper presented no empirical evidence that such correlations exist.

In this paper, we present results of a reanalysis of a calibration study conducted in Cambridge, United Kingdom (17–19) that employed urinary nitrogen excretion as a biomarker for assessing nitrogen intake (20) in addition to the conventional dietary instruments. The biomarker measurements allowed us to generalize the model by Kipnis et al. (9) and further explore the structure of measurement error in dietary assessment instruments and its implications for nutritional epidemiology.

MODELS AND METHODS

Effect of measurement error

Consider the disease model

$$R(D|T) = \alpha_0 + \alpha_1 T, \qquad (1)$$

where $R(D|T)$ denotes the risk of disease D on an appropriate scale (e.g., logistic) and T is the true long term usual intake of a given nutrient, also measured on an appropriate scale. In this analysis, all nutrients were measured on the logarithmic scale. The slope α_1 represents an association between nutrient intake and disease. Let $Q = T + e_Q$ denote the nutrient intake obtained from an FFQ (also on a logarithmic scale), where the difference between the reported and true intakes, e_Q, defines measurement error. Note that short term variation in diet is included in e_Q, as well as systematic and/or random error components resulting from the instrument itself. We assume throughout that error e_Q is nondifferential with respect to disease D; that is, reported intake contributes no additional information about disease risk beyond that provided by true intake.

Fitting model 1 to observed intake Q instead of true intake T yields a biased estimate $\tilde{\alpha}_1$ of the exposure effect. To an excellent approximation (21), the expected observed effect is expressed as

$$E(\tilde{\alpha}_1) = \lambda_1\alpha_1, \qquad (2)$$

where the bias factor λ_1 is the slope in the linear regression calibration model

$$T = \lambda_0 + \lambda_1 Q + \xi, \qquad (3)$$

where ξ denotes random error.

Although, in principle, when measurement error e_Q is correlated with true exposure T, λ_1 could be negative or greater than 1 in magnitude, in nutritional studies λ_1 usually lies between 0 and 1 (22) and can be thought of as an attenuation of the true effect α_1.

Measurement error also leads to loss of statistical power for testing the significance of the disease-exposure association. Assuming that the exposure is approximately normally distributed, the sample size required to reach the requested statistical power for a given exposure effect is proportional to (22)

$$N \propto 1/[\rho^2(Q,T)\sigma_T^2] = 1/(\lambda_1^2\sigma_Q^2), \qquad (4)$$

where $\rho(Q,T)$ is the correlation between the reported and true intakes, σ_Q^2 is the variance of the questionnaire-reported intake, and σ_T^2 is the variance of true intake. Thus, the asymptotic relative efficiency of the "naive" significance test, compared with one based on true intake, is equal to the squared correlation coefficient $\rho^2(Q,T)$.

Commonly used measurement error adjustment

Following equations 2 and 3, the unbiased (adjusted) effect can be calculated as $\hat{\lambda}_1^{-1}\tilde{\alpha}_1$, where $\hat{\lambda}_1$ is the estimated attenuation factor. Estimation of λ_1 usually requires simultaneous evaluation of additional dietary intake measurements made by the reference instrument in a calibration substudy. The common approach in nutritional epidemiology, introduced and made popular by Rosner et al. (8), uses food records/recalls as reference measurements (F), assuming that they are unbiased instruments for true long term nutrient intake at the personal level. For person i and repeat measurement j, the common model can be expressed as

$$Q_i = T_i + e_{Qi}, \qquad (5)$$

$$F_{ij} = T_i + e_{Fij}, \tag{6}$$

where it is assumed that errors e_{Qi} and e_{Fij} satisfy

$$E(e_{Fij}|T_i) = 0, \tag{7}$$

$$\mathrm{Cov}(e_{Fij}, e_{Fij'}) = 0, j \neq j', \tag{8}$$

$$\mathrm{Cov}(e_{Fij}, e_{Qi}) = 0. \tag{9}$$

Note that the assumption in equation 7 assures that $\mathrm{Cov}(e_{Fij},T_i) = 0$.

The calibration data

The data were obtained from a dietary assessment validation study carried out at the Medical Research Council's Dunn Clinical Nutrition Center, Cambridge, United Kingdom (17). One hundred and sixty women aged 50–65 years were recruited through two general medical practices in Cambridge. Subjects from practice 1 (group 1) were studied from October 1988 to September 1989, and those from practice 2 (group 2) were studied from October 1989 to September 1990. The principal measures for this study were a 4-day weighed food record and two 24-hour urine collections obtained on each of four occasions (seasons) over the course of 1 year. Season 1 was October–January; season 2, February–March; season 3, April–June; and season 4, July–September.

The weighed food record was the primary dietary report instrument of interest. The weighed records were obtained using portable electronic tape-recorded automatic scales that automatically record verbal descriptions and weights of food without revealing the weight to the subject. Each 4-day period included different days chosen to ensure that all days of the week were studied over the year, with an appropriate ratio of weekend days to weekdays.

Urine specimens were checked for completeness with *p*-aminobenzoic acid and were used to calculate urinary nitrogen excretion (23). Since it is estimated that approximately 81 percent of nitrogen intake is excreted through the urine (20), the urinary nitrogen values were adjusted, dividing by 81 percent, to estimate the total nitrogen intake of each individual. Subjects were asked to collect the first 24-hour urine sample on the third or fourth day of their food record procedure and the second sample 3–4 days later.

In this analysis, we studied nitrogen intake (g/day) and analyzed the Oxford FFQ, which is based on the widely used FFQ of Willett et al. (24), modified to accommodate the characteristics of a British diet. Nitrogen in foods is analyzed directly and then converted to dietary protein content using established factors of 5.18–6.38 (25). The FFQ was administered 1 day before the start of the weighed food record in season 3. We used the weighed food record as the dietary report reference instrument and the adjusted urinary nitrogen measurements as the biomarker. Urinary nitrogen has long been used as a critical measure of protein nutriture in nitrogen balance studies (20, 26–39), and adjusted urinary nitrogen appears to provide a marker for nitrogen intake that is valid as a reference instrument, as defined in the Introduction. (See the Appendix for more details.)

Note that both weighed food records and urinary nitrogen measure intake over a short period of time, while the FFQ assesses diet during the previous year. Therefore, errors in weighed food records and urinary nitrogen may reflect seasonal patterns in food consumption, but FFQ errors should not, in principle, contain seasonality.

In all of our analyses, we applied logarithmic transformation to the data to better approximate normality. Table I lists the mean values and variances of the transformed data according to instrument and season.

Check of standard reference instrument assumptions

As we noted above, it is a requirement that the reference instrument in a calibration study contain only error that is unrelated to true nutrient intake and is independent of error in the FFQ. Here we demonstrate an indirect check of these assumptions for the weighed food record in the Medical Research Council data. A critical assumption in our analysis is that adjusted urinary nitrogen meets the above requirements of a reference instrument for nitrogen intake.

Suppose that the common assumptions (equations 5–9) for a reference instrument hold for the weighed food record. We would then expect that using the common approach (8) with the weighed food record as the reference instrument should lead to nearly the same estimated attenuation as using the urinary nitrogen as the reference instru-

TABLE 1. Numbers of individuals, mean values, and variances of log-transformed nitrogen intake measurements in the Medical Research Council study*

Instrument	No.	Mean	Variance
Food frequency questionnaire	137	2.544	0.0709
Weighed food record			
Season 1	160	2.371	0.0584
Season 2	160	2.380	0.0490
Season 3	160	2.354	0.0502
Season 4	156	2.321	0.0450
Adjusted urinary nitrogen level			
Season 1			
First measurement	117	2.497	0.0547
Second measurement	112	2.476	0.0606
Season 2			
First measurement	112	2.538	0.0425
Second measurement	111	2.523	0.0466
Season 3			
First measurement	116	2.507	0.0517
Second measurement	110	2.483	0.0515
Season 4			
First measurement	122	2.446	0.0469
Second measurement	116	2.446	0.0530

* Data were obtained from a dietary assessment validation study (17) carried out at the Dunn Human Nutrition Unit, Cambridge, United Kingdom, 1988–1990.

ment. Figures 1 and 2 display scatterplots of averaged weighed food record data versus FFQ data and averaged urinary nitrogen data versus FFQ data, respectively; the slopes of the regression lines give the estimates of the respective attenuation factors. The former method yielded an estimated attenuation factor of 0.282, while the latter estimated it as 0.187; using a statistical test based on their bootstrap distributions, the difference between these two estimates is statistically significant ($p = 0.022$). This important finding means that the attenuation caused by measurement error in the FFQ is in fact more severe than it would appear when using the weighed food record as the reference instrument. If we accept the previously stated assumptions concerning urinary nitrogen, this result suggests that the weighed food record does not satisfy at least one of the two major requirements for a reference instrument—namely, that its error be unrelated to true intake and independent of error in the FFQ.

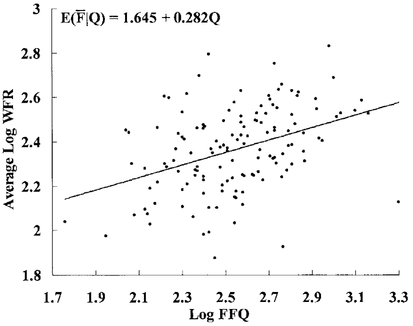


FIGURE 1. Scatterplot of log nitrogen intake as measured by averaged values from a dietary report reference instrument $\bar{F}$ (weighed food record (WFR)) versus a food frequency questionnaire (FFQ) Q, with an estimated linear regression line. Data were obtained from a dietary assessment validation study (17) carried out at the Dunn Clinical Nutrition Center, Cambridge, United Kingdom, 1988–1990.

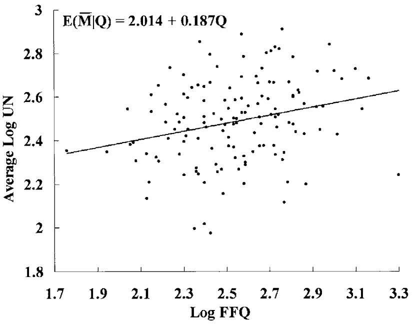


FIGURE 2. Scatterplot of log nitrogen intake as measured by the averaged biomarker $\bar{M}$ (adjusted urinary nitrogen (UN) excretion) versus a food frequency questionnaire (FFQ) Q, with an estimated linear regression line. Data were obtained from a dietary assessment validation study (17) carried out at the Dunn Clinical Nutrition Center, Cambridge, United Kingdom, 1988–1990.

A new dietary measurement error model

Model for the FFQ. The error in an FFQ is thought likely to include a systematic within-person bias b that may depend on the individual's true intake T, as well as within-person variation ε (19, 21, 40), so that

$$Q = T + e_Q = T + b + \varepsilon.$$

We approximate the relation between bias b and true intake T as the linear regression

$$b = \beta_{Q0} + \beta^*_{Q1}T + r,$$

where r has zero mean and variance σ_r^2 and is independent of T. T itself has mean μ_T and variance σ_T^2. The component $\beta_{Q0} + \beta^*_{Q1}T$ is common to all persons with the same true intake and may be called group-specific bias. The second term $\beta^*_{Q1}T$ can be thought of as arising from correlation between error and true intake. For example, given the social/cultural pressure to follow the "correct" dietary pattern, persons with a low intake of supposedly healthy food may be tempted to overreport their intake, and those with a high intake of supposedly unhealthy food may be tempted to underreport. In this case, as in many other instances, β^*_{Q1} is negative, giving rise to the flattened slope phenomenon in the regression of reported intake on true intake $E(Q|T) = \beta_{Q0} + (\beta^*_{Q1} + 1)T$.

The difference r between within-person bias and its group-specific component varies from person to person and may be determined by personality characteristics such as susceptibility to social/cultural influences. We will call it person-specific bias. Note that this error component is part of within-person systematic error and will be reproduced in repeated measurements on the same individual.

Gathering all of the error components together, we model the FFQ intake Q_{ij} for individual i and repeat measurement (season) j as

$$Q_{ij} = \mu_{Qj} + \beta_{Q0} + \beta_{Q1}T_i + r_i + \varepsilon_{ij}, \qquad (10)$$

where $\beta_{Q1} = \beta^*_{Q1} + 1$. The term μ_{Qj} represents a possible seasonal effect at the population level, a factor that usually improves model fit (41). Similarly, below we use the symbols μ_{Fj} and μ_{Mj} to represent seasonal effects in reference instrument reports and in marker levels, respectively. Within-person random error ε_{ij} has variance σ_ε^2 and is independent of other terms in model (equation) 10.

Model for the dietary report reference instrument. As we have argued, we need to allow for systematic group-specific and person-specific biases in dietary report reference instruments. Thus, we now make the same assumptions regarding the error structure for the reference instrument as for the FFQ and use a model which is analogous to that of model 10.

In the Medical Research Council study, each individual i was requested to provide the weighed food record in each (j) of the four seasons. We model these data as

$$F_{ij} = \mu_{Fj} + \beta_{F0} + \beta_{F1}T_i + s_i + u_{ij},$$
$$i = 1, \ldots, n,\ j = 1, 2, 3, 4, \qquad (11)$$

where $\beta_{F0} + \beta_{F1}T_i$ represents group-specific bias and where s_i and u_{ij} denote person-specific bias and within-person random error, with variances $\sigma_s{}^2$ and $\sigma_u{}^2$, respectively, and are assumed to be independent of each other and of true intake T_i. As before, μ_{Fj} represents a seasonal effect at the population level.

Note that the term s_i in equation 11 is parallel to the term r_i in equation 10 for the FFQ. Since the same personality traits can influence both person-specific biases, one may anticipate that the two will have a nonzero correlation $\rho(r,s)$.

Because there was only one application of the FFQ in the Medical Research Council study (17), we cannot estimate $\sigma_\varepsilon{}^2$ and $\sigma_r{}^2$ separately, only their sum. Thus, we can estimate the covariance between r and s and the correlation between $r + \varepsilon$ and s, but not the correlation $\rho(r,s)$. The correlation between $r + \varepsilon$ and s will be smaller than $\rho(r,s)$, because ε is independent of s.

Model for the biomarker. As we mentioned above, it is reasonable to assume that adjusted urinary nitrogen has errors that are unrelated to true intake and to errors in dietary assessment instruments. The Medical Research Council study included two repeat urinary nitrogen measurements in each of the four seasons. Letting j denote season (j = 1, 2, 3, 4), as before, and k denote the repeat measurement within the season (k = 1, 2), we write this model as

$$M_{ijk} = \mu_{Mj} + T_i + w_i + \nu_{ijk}, \qquad (12)$$

where 1) M_{ijk} denotes the kth repeat of the urinary nitrogen measurement of person i in season j; 2) w_i and ν_{ijk} denote person-specific bias and within-person random error, with variances $\sigma_w{}^2$ and $\sigma_\nu{}^2$, respectively, and are assumed to be independent of each other and of true intake T_i; and 3) μ_{Mj} represents a seasonal effect at the population level. It is critical that w_i is independent of true intake T_i and of all error components in the dietary report instruments Q and F.

As we explain in the Appendix, external evidence suggests that the variance of the person-specific bias, w_i, is very small relative to the variance of other terms in the model. Therefore, we assume in our main analysis that its variance is actually zero, and we show in the Appendix that our results do not change appreciably when other reasonable values of the variance are used.

Unlike model 10–11 for dietary assessment methods, which is not identifiable without biomarker data (9), model 12 with a specified value for the variance of w_i, such as zero, is identifiable on its own. Fitting it to the Medical Research Council data supports the assumption that the within-person random errors ν_{ijk} are mutually independent (i.e., they are not correlated within season) and have constant variances within seasons but not between seasons. In particular, season 2 has a different error variance than the other three seasons, which have similar variances, so that, denoting the variance of ν_{ijk} by $\sigma^2_{\nu j}$, $\sigma^2_{\nu 1} = \sigma^2_{\nu 3} = \sigma^2_{\nu 4} \neq \sigma^2_{\nu 2}$.

In contrast, the variances of ε_{ij} and u_{ij} are assumed to be constant for all i and j; this assumption is supported by examination of plots of residuals after fitting model 10–12 to the data. The within-person random errors ε_{ij}, u_{ij}, and ν_{ijk} are assumed to be mutually independent, except when the instruments are administered in the same season, in which case seasonal fluctuations in diet are assumed to produce nonzero correlation between u_{ij} and ν_{ijk}. To verify that FFQ errors were not affected by seasonality, we initially allowed for nonzero correlations between ε_{ij} and each of the errors u_{ij} and ν_{ijk} in season 3. As we expected, these correlations were found to be very small and statistically nonsignificant, and we did not include them in the final model.

Model 10–12 involves 20 unknown parameters. From the data, we can estimate 19 unique variances and covariances. These, together with an assumed value for the variance of w_i, allow us to estimate all of the parameters of the model. In practice, we use the method of maximum likelihood for estimation, which increases efficiency when there are missing values in the data.

Alternative measurement error models

Several alternatives to measurement model 10–12 have been proposed in the literature. In table 2, we list six models that are special cases of (and nested within) the more general model 10–12. These include the common model of

TABLE 2. Six alternative models that are special cases of the new model, model 10–12

Model	Parameter restrictions
Common model (Rosner et al. (8))	$\beta_{F1} = 1$; $\sigma_s{}^2 = 0$; $\rho(r,s) = 0$; $\rho(\varepsilon,u) = 0$
Freedman et al. (42)	$\beta_{F1} = 1$; $\sigma_s{}^2 = 0$; $\rho(r,s) = 0$; $\rho(\varepsilon,u)$ is nonzero for contemporaneous measures
Kaaks et al. (40)	$\beta_{F1} = 1$; $\rho(r,s) = 0$
Spiegelman et al. (43)	$\beta_{F1} = 1$; $\sigma_r{}^2 = \sigma_s{}^2 = 0$; $\rho(\varepsilon,u)$ is nonzero for all measures
Kipnis et al. (9)	$\beta_{F1} = 1$; $\rho(\varepsilon,u)$ is nonzero for contemporaneous measures
New (restricted)	$\rho(r,s) = 0$; $\rho(\varepsilon,u)$ is nonzero for contemporaneous measures

Rosner et al. (8) and models proposed by Freedman et al. (42), Kaaks et al. (40), Spiegelman et al. (43), and Kipnis et al. (9). The defining manner in which each model departs from model 10–12 is given in the table. To test the significance of the correlation between person-specific biases in the FFQ and the weighed food record, we also included in the comparison a version of model 10–12 with $\rho(r,s) = 0$.

For comparison purposes, we slightly modified the literature-based models by adding the term μ_{Fj} to represent a possible seasonal effect in the weighed food record. We also included the urinary nitrogen measurements that were modeled by equation 12.

Plummer and Clayton (19) suggest a quite general model (their model II(c)) that includes our model as a special case. They do not consider person-specific biases but allow group-specific biases to vary in repeat administrations of the same instrument. In addition, within-person random errors are assumed to be correlated, both across repeat administrations of the same instrument and across instruments, with the exception of errors in the biomarker. These are assumed to be correlated across repeat administrations within the same season and with errors in dietary report instruments in the same season but to be independent of measurements taken in different seasons. Moreover, all of the correlations and variances that are assumed to exist are allowed to differ from one another.

Prentice (44) suggested a model similar to that presented by Kipnis et al. (9), except that he explicitly assumed that $\rho(r,s) = \sigma_s/\sigma_r$ (9). However, all model parameters are allowed to depend on body mass index, and we do not include his model in this comparison.

MODEL COMPARISON USING MEDICAL RESEARCH COUNCIL DATA

Model comparison criteria

All models mentioned above were fitted to the Medical Research Council data by the method of maximum likelihood under multivariate normality—a reasonable assumption after the logarithmic transformation—and compared using three criteria. We first tested the models' goodness of fit by comparing each model with the unstructured (i.e., fully saturated) model using the likelihood ratio χ^2 test. A model that fits the data should produce a nonsignificant p value, thereby indicating that it does not explain the data significantly worse than the most general model possible. We also applied the likelihood ratio test to evaluate differences in model fit for nested models. In addition, all models were compared using two standard model selection criteria, the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) (29). These are defined as

$$\text{AIC} = \log(\text{likelihood}) - d$$

and

$$\text{BIC} = \log(\text{likelihood}) - \log(n) \times d/2,$$

where d is the number of parameters and n is the sample size. Larger values of AIC and BIC are desirable. Both AIC and BIC penalize more complex models: The "best" models chosen by the BIC tend to be simpler than those chosen by the AIC.

Model comparison results

The results of model comparison are given in table 3. Ideally, one aims to find a model that passes the goodness-of-fit test, is not significantly different from any more complex model, provides a significantly better fit than all models nested within it, and has the highest AIC and BIC scores among all models. For the Medical Research Council data, model 10–12 emerges as best by these criteria. First, it is one of only four models, together with its two simplified versions and the model of Plummer and Clayton (19), to pass the goodness-of-fit test. Second, it does not fit the data significantly worse than the more general model of Plummer and Clayton. The likelihood ratio χ^2 statistic comparing the two models is 38.8 (38.8 = 1,173.2 – 1,134.4), based on 37 degrees of freedom (37 = 56 – 19) ($p = 0.39$). Third, model 10–12 provides a significantly better fit ($p \leq 0.0011$) than

TABLE 3. Results of a model comparison using the Medical Research Council data*

Model	–2 × log likelihood	Degrees of freedom†	*p* value‡	Akaike Information Criterion	Bayes Information Criterion
Unstructured (fully saturated)	–1,222.3	104		507.2	224.7
Plummer and Clayton (19) model II(c)	–1,173.2	56	0.426	530.6	378.5
New model (equations 10–12)	–1,134.4	19	0.393	548.2	496.6
New model restricted to ρ(*r*,*s*) = 0	–1,123.7	18	0.167	543.9	495.0
Kipnis et al. (9)	–1,122.2	18	0.142	543.1	494.2
Kaaks et al. (40)	–1,112.4	17	0.049	539.2	493.0
Spiegelman et al. (43)	–1,058.0	17	<0.001	512.0	465.8
Freedman et al. (42)	–1,050.1	16	<0.001	509.1	465.6
Common model (Rosner et al. (8))	–1,050.1	15	<0.001	510.1	469.3

* Data were obtained from a dietary assessment validation study (17) carried out at the Dunn Human Nutrition Unit, Cambridge, United Kingdom, 1988–1990.

† Number of parameters.

‡ *p* value for goodness-of-fit test relative to the unstructured model.

any model nested in it. For example, comparing it with its version with uncorrelated person-specific biases, the likelihood ratio χ^2 statistic is 10.7 (10.7 = 1,134.4 – 1,123.7), based on 1 degree of freedom (1 = 19 – 18), with a p value of 0.0011.

These results suggest that group- and person-specific biases exist in both the FFQ and the weighed food record, and that these person-specific biases are indeed correlated. With only one FFQ measurement, this correlation cannot be estimated directly, but it is at least 0.35 (the low bound for $\rho(r,s)$ corresponding to $\sigma_\varepsilon^2 = 0$) and may be considerably higher. For example, if the variance of the person-specific bias is the same for the FFQ and the weighed food record, this correlation is estimated as 0.81.

Attenuation of estimated effect and statistical power

Table 4 displays the estimates of the most interesting parameters for model 10–12 and the common model. They include the attenuation factor λ_1, the variance of true intake σ_T^2, the correlation $\rho(Q,T)$ between the FFQ and true usual intake, and the slopes β_{Q1} and β_{F1} that represent group-specific biases in the FFQ and weighed food record, respectively. For all parameters, except σ_T^2, there are major differences between model 10–12 and the common approach. First, the slope of the regression of the weighed food record on true intake, β_{F1}, assumed to be 1 in the common approach, is estimated as 0.766 in our model, thereby demonstrating the flattened slope phenomenon in the reference instrument. In addition, the common approach suggests that the slope in the regression of the FFQ on true intake, β_{Q1}, is 0.661 and the correlation $\rho(Q,T)$ between the FFQ and true usual intake is 0.432, while our model estimates them as 0.430 and 0.284, respectively, indicating much less accuracy.

The major parameter controlling the ability to detect disease-nutrient relations using an FFQ is the attenuation factor λ_1. The common approach yields the attenuation factor of 0.282, while our model estimates it as 0.187. Since the true effect of an exposure is calculated as the observed effect divided by the attenuation factor, our model suggests that the true effect would be 51 percent greater than the one estimated by the common approach. There is also a much greater impact on the design of epidemiologic studies. As follows from equation 4, for any two models, the ratio of the sample sizes for the same required statistical power is the same as the squared ratio of their attenuation factors. Thus, our model suggests that the study size based on the common model should be increased by the factor 2.3 ($(0.282/0.187)^2 = 2.3$); that is, studies would have to be more than twice as large as suggested by the common model in order to maintain nominal power.

DISCUSSION

Our purpose has been to propose a statistical framework (model 10–12) for evaluating common dietary assessment reference instruments (multiple-day food records, 24-hour recalls) and to employ this framework to evaluate the weighed food record as a reference instrument for nitrogen intake (which is essentially equivalent to protein intake) using data from the Medical Research Council study (17). We have demonstrated that our model produces the best fit to these data when compared with several other models proposed in the literature:

- Its fit is not significantly different from that of the more complex models that we studied.
- It provides a significantly better fit than the simpler models, which are special cases of it.
- It has the highest values of AIC and BIC, two numerical measures of model fit.

Our statistical framework allows evaluation of two major common assumptions about a dietary report reference instrument: 1) there is no correlation between its measurement error and true intake; and 2) there is no correlation between its measurement error and that of the FFQ. Our results using the Medical Research Council data suggest that both assumptions are violated because of the presence of both group- and person-specific biases in the weighed food record and the correlation of the person-specific bias with that in the FFQ.

The statistical model we used rests on the requirement that the urinary nitrogen marker for nitrogen intake does itself satisfy assumptions 1 and 2 above. Assumption 1 is supported by several studies, documented in the Appendix, that have examined urinary nitrogen under various controlled feeding situations. Assumption 2 is based on the strong intuition that discrepancies between this biomarker measurement and true intake are caused by physiologic factors and therefore will be unrelated to errors in a dietary report instrument.

We have thus demonstrated that, at least for these data, the weighed food record may well be a flawed reference instrument. There still remains the question, Do these flaws

TABLE 4. Estimated parameters for the new and common models using the Medical Research Council data*

Model	Attenuation factor (λ_1)	σ_T^2	$\rho(Q,T)$	β_{Q1}	β_{F1}
New	0.187 (0.056)†	0.031 (0.004)	0.284 (0.082)	0.430 (0.129)	0.766 (0.066)
Common	0.282 (0.054)	0.030 (0.004)	0.432 (0.076)	0.661 (0.131)	1

* Data were obtained from a dietary assessment validation study (17) carried out at the Dunn Human Nutrition Unit, Cambridge, United Kingdom, 1988–1990.

† Numbers in parentheses, standard error.

translate into anything of importance? We believe that they do. As was shown above, using the common approach yields the estimated attenuation factor of 0.282, but it is estimated as 0.187 when using the new model. In addition, the estimated correlation between the FFQ-based nitrogen intake and true intake is 0.432 by the common approach but only 0.284 by the new model. This correlation is used as a measure of the FFQ validity, and its squared value represents the loss in statistical power to test the significance of a disease-exposure association. Thus, for these data, the real effect of measurement error in the FFQ is a greater attenuation (51 percent) and a greater loss of power (52 percent) for testing the true effect than would be estimated by the common approach.

Our estimates of the attenuation factor also indicate that the common approach may lead to unexpectedly underpowered studies. For the Medical Research Council data, our model suggests the need for a study 2.3 times larger than would have been designed had the common approach been used.

In summary, our results suggest that the impact of measurement error in dietary assessment instruments on the design, analysis, and interpretation of nutritional studies may be much greater than has been previously suspected, at least regarding protein intake. Both the attenuation of diet effect and the loss of statistical power in FFQ-based epidemiologic studies may be greater than previously estimated, because of the use of dietary reporting methods as reference instruments. This means that current and past studies may be underpowered and may explain some of the null results that have been found in nutritional epidemiology. There is a need to confirm our results by conducting further studies with biomarkers.

Our paper covers only the analysis of protein intake unadjusted for total energy intake. Further work is needed on the effects of measurement error on the analysis of protein density or energy-adjusted protein intake (6), an approach that is often used in nutrition analyses. This will require simultaneous consideration of both energy intake, using a biomarker such as doubly labeled water (10), and protein intake, using urinary nitrogen excretion. Black et al. (16) reported results from a small study with such data that supported a correlation between underreporting of protein and underreporting of energy, but also higher rates of underreporting of energy than of protein. As was reported previously (45), the effect of measurement error in energy-adjusted models can be more complex than in univariate analysis. Therefore, further studies are needed in which data from questionnaires, dietary report reference instruments, and biomarkers for protein and energy intakes are all collected and analyzed simultaneously to investigate the effects of measurement error on protein density or energy-adjusted protein intake.

ACKNOWLEDGMENTS

Dr. Raymond Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-E509106).

REFERENCES

1. Hunter DJ, Spiegelman D, Adami H-O, et al. Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. N Engl J Med 1996;334:356–61.
2. Fuchs CS, Giovannucci EL, Colditz GA, et al. Dietary fiber and the risk of colorectal cancer and adenoma in women. N Engl J Med 1999;340:169–76.
3. Beaton GH, Milner J, Corey P, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Am J Clin Nutr 1979;32:2546–59.
4. Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. Nutr Cancer 1988;11:243–50.
5. Freedman LS, Schatzkin A, Wax J. The impact of dietary measurement error on planning a sample size required in a cohort study. Am J Epidemiol 1990;132:1185–95.
6. Willett WC. Nutritional epidemiology. New York, NY: Oxford University Press, 1990:69–91.
7. Rosner B, Willett WC. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. Am J Epidemiol 1988;127: 377–86.
8. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8: 1051–69.
9. Kipnis V, Carroll RJ, Freedman LS, et al. A new dietary measurement error model and its implications for the estimation of relative risk: application to four calibration studies. Am J Epidemiol 1999;150:642–51.
10. Bandini LG, Schoeller DA, Cyr HN, et al. Validity of reported energy intake in obese and nonobese adolescents. Am J Clin Nutr 1990;52:421–5.
11. Livingstone MB, Prentice AM, Strain JJ, et al. Accuracy of weighed dietary records in studies of diet and health. Br Med J 1990;300:708–12.
12. Heitmann BL. The influence of fatness, weight change, slimming history and other lifestyle variables on diet reporting in Danish men and women aged 35–65 years. Int J Obes 1993;17: 329–36.
13. Heitmann BL, Lissner L. Dietary underreporting by obese individuals—is it specific or non-specific? Br Med J 1995; 311:986–9.
14. Martin LJ, Su W, Jones PJ, et al. Comparison of energy intakes determined by food records and doubly labeled water in women participating in a dietary intervention trial. Am J Clin Nutr 1996;63:483–90.
15. Sawaya AL, Tucker K, Tsay R, et al. Evaluation of four methods for determining energy intake in young and older women: comparison with doubly labeled water measurements of total energy expenditure. Am J Clin Nutr 1996;63:491–9.
16. Black AE, Bingham SA, Johansson G, et al. Validation of dietary intakes of protein and energy against 24 urinary N and DLW energy expenditure in middle-aged women, retired men and post-obese subjects: comparisons with validation against presumed energy requirements. Eur J Clin Nutr 1997;51:405–13.
17. Bingham SA, Gill C, Welch A, et al. Comparison of dietary assessment methods in nutritional epidemiology: weighed food records v. 24 h recalls, food frequency questionnaires and estimated diet records. Br J Nutr 1994;72:619–43.
18. Bingham SA, Cassidy A, Cole TJ, et al. Validation of weighed records and other methods of dietary assessment using the 24 h nitrogen technique and other biological markers. Br J Nutr 1995;73:531–50.

19. Plummer M, Clayton D. Measurement error in dietary assessment: an investigation using covariance structure models. (Parts I and II). Stat Med 1993;12:925–48.
20. Bingham SA, Cummings JH. Urine nitrogen as an independent validatory measure of dietary intake: a study of nitrogen balance in individuals consuming their normal diet. Am J Clin Nutr 1985;42:1276–89.
21. Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London, United Kingdom: Chapman and Hall Ltd, 1995.
22. Kaaks R, Riboli E, van Staveren W. Calibration of dietary intake measurements in prospective cohort studies. Am J Epidemiol 1995;142:548–56.
23. Bingham S, Cummings JH. The use of 4-aminobenzoic acid as a marker to validate the completeness of 24 h urine collections in man. Clin Sci 1983;64:629–35.
24. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 1985;122:51–65.
25. Matthews DE. Proteins and amino acids. In: Shils ME, Olson JA, Shike M, et al, eds. Modern nutrition in health and disease. 9th ed. Baltimore, MD: Williams and Wilkins Company, 1999: 11–48.
26. Campbell WW, Crim MC, Dallal GE, et al. Increased protein requirements in elderly people: new data and retrospective reassessments. Am J Clin Nutr 1994;60:501–9.
27. Zanni E, Calloway DH, Zezulka AY. Protein requirements of elderly men. J Nutr 1979;109:513–24.
28. Oddoye EA, Margen S. Nitrogen balance studies in humans: long term effect of high nitrogen intake and nitrogen accretion. J Nutr 1979;109:363–77.
29. Weller LA, Calloway DH, Margen S. Nitrogen balance of men fed amino acid mixtures based on Rose's requirements, egg white protein, and serum free amino acid patterns. J Nutr 1971; 101:1499–508.
30. Bunker VW, Lawson MS, Stansfield MF, et al. Nitrogen balance studies in apparently healthy elderly people and those who are housebound. Br J Nutr 1987;57:211–21.
31. Uauy R, Scrimshaw NS, Young VR. Human protein requirements: nitrogen balance response to graded levels of egg protein in elderly men and women. Am J Clin Nutr 1978;31:779–85.
32. Castaneda C, Charnley JM, Evans WJ, et al. Elderly women accommodate to a low-protein diet with losses of body cell mass, muscle function, and immune response. Am J Clin Nutr 1995;62:30–9.
33. Cheng AH, Gomez A, Bergan JG, et al. Comparative nitrogen balance study between young and aged adults using three levels of protein intake from a combination wheat-soy-milk mixture. Am J Clin Nutr 1978;31:12–22.
34. Atinmo T, Mbofung CM, Egun G, et al. Nitrogen balance study in young Nigerian adult males using four levels of protein intake. Br J Nutr 1988;60:451–8.
35. Rand WM, Scrimshaw NS, Young VR. Retrospective analysis of data from five long term, metabolic balance studies: implications for understanding dietary nitrogen and energy utilization. Am J Clin Nutr 1985;42:1339–50.
36. Tarnopolsky MA, Atkinson SA, MacDougall JD, et al. Evaluation of protein requirements for trained strength athletes. J Appl Physiol 1992;73:1986–95.
37. Pannemans DL, Wagenmakers AJ, Westerterp KR, et al. Effect of protein source and quantity metabolism in elderly women. Am J Clin Nutr 1998;68:1228–35.
38. Wayler A, Queiroz E, Scrimshaw NS, et al. Nitrogen balance studies in young men to assess the protein quality of an isolated soy protein in relation to meat proteins. J Nutr 1983;113: 2485–91.
39. Young VR, Wayler A, Garza C, et al. A long term metabolic balance study in young men to assess the nutritional quality of an isolated soy protein and beef proteins. Am J Clin Nutr 1984; 39:8–15.
40. Kaaks R, Riboli E, Esteve J, et al. Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. Stat Med 1994;13:127–42.
41. Landin R, Carroll RJ, Freedman LS. Adjusting for time trends when estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. Biometrics 1995;51:169–81.
42. Freedman LS, Carroll RJ, Wax Y. Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. Am J Epidemiol 1991;134: 310–20.
43. Spiegelman D, Schneeweiss S. McDermott A. Measurement error correction for logistic regression models with an alloyed gold standard. Am J Epidemiol 1997;145:184–96.
44. Prentice R. Measurement error and results from analytic epidemiology: dietary fat and breast cancer. J Natl Cancer Inst 1996;88:1738–47.
45. Kipnis V, Freedman LS, Brown CC, et al. Effect of measurement error on energy-adjustment models in nutritional epidemiology. Am J Epidemiol 1997;146:842–55.
46. Millward DJ. Urine nitrogen as an independent validatory measure of dietary intake: potential errors due to variation in magnitude and type of protein intake. Br J Nutr 1997;77:141–3.
47. Bingham SA. Urine nitrogen as an independent validatory measure of protein intake. Br J Nutr 1997;77:144–8.
48. Millward DJ, Bowtell JL, Pacy P, et al. Physical activity, protein metabolism and protein requirements. Proc Nutr Soc 1994;53:223–40.
49. Millward DJ. The nutritional value of plant-based diets in relation to human amino acid and protein requirements. Proc Nutr Soc 1999;58:249–60.

APPENDIX

Nitrogen balance studies require known levels of protein/nitrogen intakes and complete urine collections in addition to either estimation or collection of fecal, sweat, or other miscellaneous losses in order to be valid, and this has been done with varying levels of rigor and oversight. Generally, the goals of such studies have been to assess protein requirements and protein sources.

Among studies with varying levels of controlled conditions in which protein intakes were provided at levels necessary to maintain a positive nitrogen balance (a near-given in diets in developed countries), the long term ratio of urinary nitrogen to dietary nitrogen among individuals is generally within a range of 70–90 percent (20, 26–39). Bingham and Cummings (20) specifically addressed the question of nitrogen output and validation of dietary intakes in a rigorously controlled feeding study of eight adults adhering to their regular diets and found that the mean ratio of urinary nitrogen to dietary nitrogen was 81 percent, with a standard error of 2 percent (range, 78–83 percent). In other well-controlled studies, group means have ranged from 77 percent to 88 percent (26–32). Generally, urinary nitrogen is robust in free-living adults, except when there is inadequate total energy and/or protein intake, inadequate essential amino acid intake, a very high fiber intake, or profuse sweating (46–49). None of these conditions are prevalent in adequately nourished populations, and a range of 70–90 percent represents a realistic range for biologic variability in the ratio of urinary nitrogen to dietary nitrogen that does not depend on age, gender, and source of protein, as long as subjects maintain a positive nitrogen balance. This is supported by different studies that measured this range in old and young participants and in men and women with soy, egg, meat, or mixed sources of protein in their diets (20, 26–39).

Nevertheless, the ratio of urinary nitrogen to dietary nitrogen does not represent an exact biologic constant and may still include interperson variability, or person-specific bias. Three studies described by Bingham and Cummings (20), Oddoye and Margen (28), and Castaneda et al. (32) and two studies described by Young et al. (39) provided information on within-person variation in the ratio (R) of urinary nitrogen to dietary nitrogen and therefore can be analyzed by analysis of variance to estimate and/or test the presence of person-specific bias in the urinary nitrogen biomarker. These five studies represent a valuable subsample of the controlled feeding studies and include both men (20, 28, 39) and women (32), young (28, 39), middle-aged (20), and elderly (32) participants, and a variety of protein sources, including soy protein (39), meat-free protein (32), formula diets (26), beef protein (39), and usual diet (20).

We carried out a meta-analysis of these five studies using a random effects model for ratio R that included both a random study effect and, nested in it, a random person effect (person-specific bias). The study effect η was very small (variance $\sigma_\eta^2 = 0.0006$) and not statistically significant ($p = 0.21$), while the person effect w was also relatively small (variance $\sigma_w^2 = 0.0021$) but highly statistically significant ($p = 0.0008$). These results provide some evidence that although ratio R does not seem to depend on age, gender, and source of protein intake, it does contain a small person-specific bias. After we pooled all of the participants from the five studies and fitted a random effects model with a random effect representing person-specific bias, the variance of this bias was estimated as 0.0027 (standard deviation 5.2 percent). The mean long term ratio of urinary nitrogen to dietary nitrogen was estimated as 83.5 percent (standard error 2.3 percent), which agrees well with the calibration constant of 81 percent suggested by Bingham and Cummings (20). The mean ratio (83.5 percent) and the standard deviation of its person-specific bias (5.2 percent) agree well with the general observation that individual ratios fall between 70 percent and 90 percent.

These results suggest that urinary nitrogen level satisfies both requirements for a reference instrument. The stability of the urinary nitrogen:dietary nitrogen ratio and the relatively low person-specific bias support the essential absence of correlation between errors in adjusted urinary nitrogen and true nitrogen. The relatively low person-specific bias and the fact that the bias is probably physiologically based rather than psychologically based also support the essential absence of correlation between errors in adjusted urinary nitrogen and errors in dietary report instruments.

It is interesting to note that the estimated variation due to person-specific bias in the urinary biomarker for protein intake constitutes only about 10 percent of the estimated variation of true protein intake. Nevertheless, to investigate how this person-specific bias might change the result of our model fit, we conducted a sensitivity analysis by including person-specific bias in the biomarker model and changing its value from $\sigma_w^2 = 0$ (the value assumed in the main text) to $\sigma_w^2 = 0.0027$ (the value estimated in this appendix). The results are reported in appendix table 1. The estimated attenuation factor was not affected by the presence of person-specific bias in the biomarker, since this bias does not violate the two major requirements for the reference instrument. Other parameters in the model changed slightly. The estimated variance of true intake was reduced by the variation due to person-specific bias. The estimated correlation between true intake and its FFQ measure was increased by 4.5 percent, and the estimated slopes in the regressions of FFQ and weighed food record on true intake were increased by approximately 10 percent each. However, the general conclusions reached in the paper remain the same.

APPENDIX TABLE 1. Estimated parameters for the new model, with and without person-specific biases in the urinary biomarker, using the Medical Research Council data*

Model	Attenuation factor (λ_1)	σ_T^2	$\rho(Q,T)$	β_{Q1}	β_{F1}
$\sigma_w^2 = 0$	0.187 (0.056)†	0.031 (0.004)	0.284 (0.082)	0.430 (0.129)	0.766 (0.066)
$\sigma_w^2 = 0.0027$	0.187 (0.056)	0.028 (0.004)	0.297 (0.085)	0.472 (0.142)	0.839 (0.074)

* Data were obtained from a dietary assessment validation study (17) carried out at the Dunn Human Nutrition Unit, Cambridge, United Kingdom, 1988–1990.

† Numbers in parentheses, standard error.

BIOMETRICS 58, 13–20
March 2002

Semiparametric Regression Modeling with Mixtures of Berkson and Classical Error, with Application to Fallout from the Nevada Test Site

Bani Mallick,[1,*] F. Owen Hoffman,[2,] and Raymond J. Carroll[1,***]**

[1]Department of Statistics, Texas A&M University,
College Station, Texas 77843-3143, U.S.A.
[2]SENES Oak Ridge, Center for Risk Analysis,
102 Donner Drive, Oak Ridge, Tennessee 37830, U.S.A.
**email:* bmallick@stat.tamu.edu
***email:* fohoff3084@aol.com
****email:* carroll@stat.tamu.edu

SUMMARY. We construct Bayesian methods for semiparametric modeling of a monotonic regression function when the predictors are measured with classical error, Berkson error, or a mixture of the two. Such methods require a distribution for the unobserved (latent) predictor, a distribution we also model semiparametrically. Such combinations of semiparametric methods for the dose–response as well as the latent variable distribution have not been considered in the measurement error literature for any form of measurement error. In addition, our methods represent a new approach to those problems where the measurement error combines Berkson and classical components. While the methods are general, we develop them around a specific application, namely, the study of thyroid disease in relation to radiation fallout from the Nevada test site. We use this data to illustrate our methods, which suggest a point estimate (posterior mean) of relative risk at high doses nearly double that of previous analyses but that also suggest much greater uncertainty in the relative risk.

KEY WORDS: Bayes; Berkson error; Classical error; Dose–response; Latent variables; Likelihood; Measurement error; Pólya trees; Radiation epidemiology; Semiparametric; Thyroid cancer.

1. Introduction

This article develops semiparametric Bayesian methods for regression problems where a predictor is measured with either classical error, Berkson error, or a combination of classical and Berkson measurement error. We allow the regression function and the distribution of the unobservable (latent) covariate to be modeled either parametrically or nonparametrically. Our methods are applied to a study of thyroid cancer induced by fallout from nuclear testing (Stevens et al., 1992).

There is of course an enormous literature on regression problems where the latent covariate is measured either entirely with classical error or entirely with Berkson error (Carroll, Ruppert, and Stefanski, 1995). There have been numerous articles that model the latent variable semiparametrically (Roeder, Carroll, and Lindsay, 1996; Müller and Roeder, 1997; Carroll, Roeder, and Wasserman, 1999; Schafer, 2001; Richardson et al., unpublished manuscript). There are also articles that model the regression function semiparametrically (Carroll, Maca, and Ruppert, 1999). However, to date, no one has exhibited methods that are semiparametric both in the model and in the latent variable distribution. This article exhibits such methods. We focus for specificity on radiation epidemiology, where the latent variable is the dose to an individual, typically measured with a combination of classical and Berkson errors. The methods developed to date for these models (cf., Reeves et al., 1998; Schafer et al., 2002) rely on approximations to the regression function given the observed data and typically use as the predictor an estimate of its conditional expectation given the observed dose, the so-called regression calibration approach.

This article takes a Bayesian approach. In the problem of interest, the regression function is reasonably thought to be monotone in the latent variable, so we allow either a parametric form or a semiparametric monotone form. In addition, the likelihood of the mixed Berkson-classical model depends on the distribution of the latent variable; this distribution we model either parametrically or flexibly semiparametrically.

In our example and in other such exercises in radiation dosimetry, the estimation of an individual's dose is the result of a complex modeling process including physical transport systems, biological processes, and direct measurements. It is typical to assign to the dose a total uncertainty, which is in effect the sum of the Berkson error variance and the classical error variance. This total uncertainty is known nominally at the individual level, but the relative contribution of Berkson and classical errors is unknown. However, in these cases, is it reasonable to suppose that the proportion p of the error variance that is due to Berkson error lies within a defined interval on $[0, 1]$. Our Bayesian methods place a prior distribution on the relative contribution p, being uniformly distributed on the predefined interval.

This article is structured as follows. In Section 2, we describe the Nevada test-site data in detail. In Section 3, we describe parametric and semiparametric models for the dose–response. Of particular note in this section is that we develop a semiparametric approach that makes the dose–response monotonic, using a mixtures-of-beta cumulative distribution functions (CDFs) approach.

In both these sections, we show that the likelihood function depends on the distribution of a latent variable and may thus be sensitive to misspecification of this distribution. Indeed, in our example, we present evidence that the latent variable in the natural log-dose scale is far from normally distributed. Section 4 describes our Bayesian approach in detail. Of particular note here is that, instead of specifying a distribution for the latent variable, we model the latent variable semiparametrically, using Pólya trees. Section 5 contains the reanalysis of the Nevada test-site data. Section 6 contains the results of a small simulation study. Section 7 has concluding remarks. An appendix gives brief details of the priors and Metropolis–Hastings proposals used in our calculations.

2. Berkson/Classical Errors

Stevens et al. (1992) describe a study of thyroid disease in relation to fallout from the Nevada test site (NTS). Similar statistical issues arise in the Hanford Thyroid Disease Study (Davis et al., 1998) and the Oak Ridge Radiation Study (Ostrouchov, Frome, and Kerr, 1998). In the Nevada study, 2473 individuals who were exposed to radiation as children were examined for thyroid disease. The primary radiation exposure came from milk and vegetables. Dosimetry calculations were based on age at exposure, gender, residence history, x-ray history, whether as a child the individual was breast fed, and a diet questionnaire filled out by the parent focusing on milk consumption and vegetables. The data were then fed into a complex model and, for each individual, the point estimate of thyroid dose and an associated standard error were reported. Unfortunately, only the summary statistics are available in the data file.

A statistically significant relationship between dose and neoplasms developed was obtained when fitting a logistic regression model with stratum-specific intercepts, adjustments for confounders, and a term for dose of the form $\log(1+\theta\text{dose})$ (see below for more details). In one such analysis (Stevens et al., p. 208), the estimate of θ more than doubled after accounting for dose uncertainty by assuming a classical error model, i.e., if all the error were classical and error is ignored, relative risks are underestimated. In Section 3.1, we show that assuming that all the error is Berkson and ignoring classical error overestimates relative risk.

It is helpful to consider the model used to calculate dose to the thyroid of a specified individual from a single milk source contaminated by a single Nevada test-site event. This model has the following form (Stevens et al., p. 85):

$$W = C \times DCF \times I \times TD \times FP, \tag{1}$$

where W = reported dose to thyroid of the subject; C = time-integrated radioiodine concentration of milk; DCF = ingestion dose conversion factor; I = individual milk intake rate in liters per day, measured by a food frequency questionnaire; FP = frequency of purchase correction factor; and TD = time-delay factor. A detailed elicitation of the error structure for each component is not possible because of space limitations. The following is a brief summary. Milk intake (I), information on frequency of purchase (FP), and the sources of milk used to compute the time-delay factor (TD) come from a food frequency questionnaire (FFQ) filled out by the parent. As such, the error here is probably best thought of as mainly classical. The ingestion dose conversion factors (DCF) are specific for age and isotope. Uncertainties associated with DCF are probably best modeled as a mixture of Berkson and classical types. The time-integrated radioiodine concentration of milk (C) is specific not to individuals but to producers. One would ordinarily think of C as of Berkson type, but there is a major component of it that is classical, namely the deposition of I^{131} (Kerber et al., 1993; Simon et al., 1995) across the regions under study. Thus, the error structure for estimated dose to the thyroid has a mixture of classical and Berkson error.

This brief outline is simplistic. For example, the mass interception of I^{131} on vegetation and the transfer of iodine from feed to cow's milk are important components of the DCF. Their distribution is estimated by a combination of data from a literature review and expert judgment, thus combining classical and Berkson error in complex ways.

Reeves et al. (1998) consider data with a mixture of Berkson and classical error, although in a context far different from ours. At a formal mathematical level, their model is applicable to the Utah study; our approaches to analysis are far different. Let Y be the indicator of disease and let Z denote a vector of covariates measured without error, e.g., age, sex, and state. We will take logarithms in (1) and assign Berkson and classical error formulas to the pieces as described above. Denote true dose by X and observed dose by W. Then there is a latent variable L, which we call the latent intermediate variable, such that

$$\log(X) = \log(L) + U_b, \tag{2}$$

$$\log(W) = \log(L) + U_c. \tag{3}$$

The terminology latent intermediate variable is suggestive because L is intermediate between X and W.

We assume that (W, L) are conditionally independent of the response Y given (X, Z) and that the Berkson and classical errors are independent. Here U_b is Berkson error with variance σ_b^2, U_c is classical error with variance σ_c^2, and $\log(L)$ has mean μ_L and variance σ_L^2. With a change in notation, these models are the same as model (4) in Reeves et al. There are covariates Z measured without error, so as is standard in the measurement error problem, e.g., μ_L may be allowed to depend on a linear function of Z and σ_L^2 is understood to be a conditional variance given Z.

The Utah study data file provides the sum of the Berkson and classical error variances for each individual but does not provide the relative contribution of each to the sum. Our approach is to allocate the total error variance across the two types, where we allow for a fixed proportion p of an individual's error to come from each source, and vary this source in our Bayesian analysis by placing an informative prior on the fixed proportion.

3. Dose–Response Modeling

In this section, we provide a discussion of model fitting when the distribution of the latent intermediate variable L in (2)–(3) is specified. For convenience, for now we assume that $\log(L)$ is normally distributed conditional on Z and $X > 0$, where Z consists of the patient age at exposure, sex, and state of residence (Utah, Nevada, Arizona). In state s,

$$[\log(L) \mid Z,\ \text{state} = \text{s}] \sim \text{normal}\left(\alpha_{s0} + Z^{\mathrm{T}}\alpha_{s1}, \sigma_s^2\right). \quad (4)$$

We denote by $\mathcal{A}$ the collection of these parameters.

In general, we consider four types of modeling efforts: (a) all dose uncertainties are ignored; (b) error purely of Berkson type; (c) error purely of classical type; and (d) error a mixture of Berkson and classical errors, with the fraction of variance due to the Berkson part being p, i.e., $\sigma_b^2/(\sigma_b^2 + \sigma_c^2) = p$. In the latter case, we face an identifiability issue: while the total uncertainty $\sigma_b^2 + \sigma_c^2$ is given in the data base, p itself is not identifiable. For our Bayesian analysis, we handle this issue by using an informative prior for p. Based on previous considerations, it seems reasonable to balance the classical and Berkson errors, with a substantial fraction being of each type. Thus, we gave p a uniform prior on the interval $[0.2, 0.8]$, creating a form of model mixing.

3.1 *Parametric Dose-Response Models*

The model used by Stevens et al. (1992) in their dose–response was defined as follows. For numerical convenience, we rescaled dose to be dose/max(dose) = dose/0.461 Gy (Gray). The model is

$$\text{logit}\,\{\text{pr}(Y = 1 \mid Z, X)\} = \beta_0 + Z^{\mathrm{T}}\beta_1 + \log(1 + \theta X). \quad (5)$$

Let $\mu_{L|Z}$ and $\sigma_{L|Z}^2$ be the conditional mean and variance of $\log(L)$ given Z. Define $\lambda_{x|w,\ell} = \sigma_{L|Z}^2/(\sigma_{L|Z}^2 + \sigma_c^2)$. Using standard calculations, assuming that the nonzero L's are log normal, and making the usual exponential approximation to the logistic function appropriate for rare events, it can be shown that, for the observed data,

$$\begin{aligned} &\text{logit}\,\{\text{pr}(Y = 1 \mid Z, W)\} \\ &\qquad \approx \beta_0 + Z^{\mathrm{T}}\beta_1 + \log(1 + \theta\gamma W^{\lambda_{x|w,\ell}}), \quad (6) \\ &\quad \gamma = \exp\{(1 - \lambda_{x|w,\ell})(\mu_{L|Z} + \sigma_{L|Z}^2/2) + \sigma_b^2/2\}. \end{aligned}$$

In the Berkson case, $\lambda_{x|w,\ell} = 1$ and $\gamma = \exp(\sigma_b^2/2) > 1$, so that the right-hand side of (6) reduces to $\beta_0 + Z^{\mathrm{T}}\beta_1 + \log(1 + \theta\gamma W)$, meaning that an analysis that ignores Berkson errors overestimates the dose–response parameter by the factor γ, thus falsely inflating the effect of dose. Indeed, when regressing Y against (Z, W), if one assumes Berkson error, then W should be replaced by γW, where γ varies among individuals; essentially, such an approach was used by Stevens et al. (1992).

3.1.1 *MCMC calculations.* In terms of our MCMC calculations, we make the following comments. When the measurement error in the dose is incorporated, (X, L) are treated as latent variables, i.e., augmented data, and observations are sampled from their complete conditionals. Let $f(L \mid Z, \mathcal{A})$ be the density of L depending on Z and the parameter $\mathcal{A}$. Remember that the data base gives us the value of $\sigma_b^2 + \sigma_c^2$, with the possible unknown being $p = \sigma_b^2/(\sigma_b^2 + \sigma_c^2)$. Let the prior density be $\pi(\beta_0, \beta_1, \theta, \mathcal{A}, p)$. Then the complete distribution for an observation is written as

$$\begin{aligned} &f(Y \mid X, Z, \beta_0, \beta_1, \theta) f(X, W \mid L, \sigma_b^2, \sigma_c^2, p) f(L \mid Z, \mathcal{A}) \\ &\qquad \times \pi(\beta_0, \beta_1, \theta, \mathcal{A}, p). \quad (7) \end{aligned}$$

In (7), when $\sigma_b^2 + \sigma_c^2 = 0$, we have $L = X = W = 0$.

All priors were chosen to be proper but noninformative, with the exception of that for θ. For θ, the prior was chosen to be normal, truncated at zero, with prior mean being the empirical Bayes estimator ignoring measurement error but with a large variance.

Due to the logistic model framework, the complete conditional distributions for $(\beta_0, \beta_1, \theta, X)$ are nonstandard. We used a Metropolis step to generate observations from these nonstandard distributions. The complete conditionals for L and $\mathcal{A}$ are standard.

3.2 *Monotonic, Semiparametric Dose-Response*

Here we replace the term $\log(1+\theta X)$ in (5) by a more flexible semiparametric form, namely

$$\text{logit}\,\{\text{pr}(Y = 1 \mid Z, X)\} = \beta_0 + Z^{\mathrm{T}}\beta_1 + g(X). \quad (8)$$

Following (5), in (8) it makes sense to have $g(\cdot)$ be an unknown but strictly monotone function, with the property that $g(0) = 0$. For a general function $g(x)$, modeling in such a circumstance has been considered previously by many authors, e.g., using regression splines. These methods do not guarantee monotonicity of the dose–response. We thus use instead the approach of Mallick and Gelfand (1994), which has three steps: (a) monotonically transform the range of the function to the unit interval, (b) note that then modeling g is equivalent to modeling an unknown distribution function, and (c) model this distribution function as a mixture of beta distribution functions.

Thus, for some function $\mathcal{T}(\cdot)$, we assume that $g(\cdot)$ satisfies

$$\mathcal{T}\{g(x)\} = \sum_{\ell=1}^{r} \omega_\ell \text{IB}[\mathcal{T}\{g_0(x)\}, c_\ell, d_\ell]. \quad (9)$$

In (9), $\mathcal{T}$ is a monotonic transformation from the real line to $[0, 1]$. In addition, $\text{IB}(u; c, d)$ denotes the incomplete beta function, associated with a beta density in standard form having parameters c and d but evaluated at u. In (9), r denotes the number of mixands and ω_ℓ denotes the mixing weights, with the constraints that $\omega_\ell \geq 0$ and $\Sigma_{\ell=1}^r \omega_\ell = 1$. Finally, g_0 is a centering function for g. The data will revise g_0 to an estimator of the unknown function g revealing the extent of departure from g_0. In our application, it is natural to set $g_0(x) = 1 + \hat{\theta}x$, where in our example $\hat{\theta}$ is the posterior mean estimate of the corresponding parametric model. Because X and hence $G(x)$ are nonnegative, we slightly modify Mallick and Gelfand's suggestion by choosing $\mathcal{T}(v) = v/(1 + v)$.

In viewing g as unknown. we might think of r, $(\omega_1, \ldots, \omega_r)$. and the (c_ℓ, d_ℓ) as unknown. In practice, we have found that assuming r is unknown gains little compared with, say, $r = 6$.

Given r, it is mathematically easier to assume that the component beta densities are specified but that the weights are unknown. Following Mallick and Gelfand (1994), we take $c_\ell = \ell$, $d_\ell = r + 1 - \ell$, providing a collection of densities that blanket the unit interval. Hence, specification of g is equivalent to specification of the ω's. In addition to the constraints that $\omega_\ell \geq 0$ and $\Sigma_{\ell=1}^{r} \omega_\ell = 1$, (9) and the condition $g(0) = 0$ implies that $k_1/(1+k_1) = \Sigma_{\ell=1}^{r} \omega_\ell \mathrm{IB}\{k_1/(1+k_1), c_\ell, d_\ell\}$, i.e., the ω_ℓ satisfy an additional linear constraint.

For the Bayesian analysis, we need to specify a prior distribution for the ω's, noting this is a distribution on the r-dimensional simplex. We chose for this distribution the Dirichlet($\gamma = 1$) (Berger, 1985, p. 561). The intuition behind this choice is as follows. If g_0 is a baseline function for g, then we might choose $f(\omega)$ such that, *a priori*, g is centered around g_0. The data would then revise this prior in terms of the support for g_0. Centering g around g_0 corresponds to centering $\Sigma_{l=1}^{r} w_l \mathrm{IB}(u; c_l, d_l)$ around u. If we center using the mean, as is typically done in the case of Dirichlet processes, we obtain

$$\sum_{\ell=1}^{r} \mathrm{E}(w_\ell)\mathrm{IB}(u; c_\ell, d_\ell) = u. \tag{10}$$

Then (9) requires $r^{-1}\Sigma\,\mathrm{IB}(u; c_l, d_l) = u$. If we use c_l and d_l as in previous the paragraph and take r even, expansion of the terms in this summation about 1/2 yields, to a first0order approximation, an average that is u.

4. Intermediate Variable Distribution

We next propose a flexible parametric model for $\log(L)$. There are many ways to specify a flexible, skewed, heavy-tailed distribution for $\log(L)$. Possibilities include the skewed-normal distribution, the mixture of normals distribution, or models such as those used by Davidian and Gallant (1993). These methods are easy to write down, but the MCMC calculations involving them are not entirely straightforward since they require Metropolis steps.

In contrast, our method is to assume $\log(L)$ has an unknown distribution and impose a Pólya-tree prior (Lavine, 1992; Walker and Mallick, 1996). The method allows considerable flexibility in the model for $\log(L)$ as well as great ease of calculation. The flexibility and ease of calculation are bought at the price of difficult notation.

We give here a brief description of the methodology used. Within each state s, we assumed that the distribution function of $\log(L)$ for nonzero doses was $F_s(x - \alpha_{s0} - Z^{\mathrm{T}}\alpha_{s1})$, where $F_s(\cdot)$ is the realization of a random distribution function. The prior for $F_s(\cdot)$ is a Pólya tree distribution, defined as follows.

We start with a base distribution function G, the normal distribution function (with a large standard deviation, in this case 40). We then partition the real line. At stage $m = 1$, the first partition is (B_0, B_1), where $B_0 = (-\infty, G^{-1}(1/2))$. At stage $m = 2$, we partition B_0 and B_1 separately into (B_{00}, B_{01}) and (B_{10}, B_{11}), respectively, where $B_{00} = (-\infty, G^{-1}(1/4))$ and $B_{10} = [G^{-1}(1/2), G^{-1}(3/4))$. We continue in this way so that, at stage $m+1$, we partition $B_{i_1,\ldots,i_m}$ into $B_{i_1,\ldots,i_m,0}$ and $B_{i_1,\ldots,i_m,1}$. At any stage m, order the $j = 1, \ldots, 2^m$ partitions into B_j^* and note that $B_j^* = [G^{-1}\{(j-1)/2^m]\}, G^{-1}(j/2^m))$. In our calculations, we continued with $m = 1, \ldots, M = 8$ levels of partitioning.

The Pólya tree prior for F_s is defined on the sets B_j^* for $j = 1, \ldots, 2^M$ as follows. At stage $m = 1$, let C_0 be the realization of a beta random variable with indices (γ_0, ζ_0). Then $F_s(B_0) = C_0$, and of course $F_s(B_1) = C_1 = 1 - C_0$. At stage $m = 2$, let C_{00} and C_{10} be realizations of beta random variables with indices $(\gamma_{00}, \zeta_{00})$ and $(\gamma_{10}, \zeta_{10})$, respectively. Then $F_s(B_{00}) = C_0 C_{00}$, $F_s(B_{01}) = C_0(1 - C_{00})$, $F_s(B_{10}) = C_1 C_{10}$, $F_s(B_{11}) = C_1(1 - C_{10})$. We continue in this way for $m = 3, \ldots, M$, thus defining F_s on the sets B_j^* for $j = 1, \ldots, 2^M$. This defines a Pólya tree distribution with partition $\Omega = (B_j^*)_{j=1}^{2^M}$ and parameters $\mathcal{A} = (\gamma_0, \zeta_0, \gamma_{00}, \zeta_{00}, \ldots)$, which we denote as $\mathrm{PT}(\Omega, \mathcal{A})$. For our prior, at stage m, we set the γ's and the $\zeta's$ all equal to $c_{\text{polya}} m^2$, where $c_{\text{polya}} = 0.5$, although we experimented with different values $0.1 \leq c_{\text{polya}} \leq 1.0$ and the results changed hardly at all.

We have now defined the Pólya tree prior for F_s. Given observations L_{is} from state s, the posterior of F_s is also a Pólya tree distribution with the same set partition Ω. The parameters are updated as follows. First, at stage $m = 1$, γ_0 is updated to $\gamma_0 + n_{0s}$, where n_{0s} is the number of L's in state s that fall into the set B_0. At stage $m = 2$, γ_{00} and γ_{10} are updated to $\gamma_{00} + n_{00}$ and $\gamma_{10} + n_{10}$, where n_{j0} is the number of L's in state s that fall in B_{j0}. Further levels of the γ's are generated in the same way.

In the MCMC calculations, suppose that the complete conditional for F_s is $\mathrm{PT}(\Omega, \mathcal{A}_{s*})$. We generate observations from state s as follows. First generate F_s. In state s, the distribution function for the L's is $F_s(x - \alpha_{s0} - Z^{\mathrm{T}}\alpha_{s1})$. Observations from this distribution function are easily generated by a Metropolis–Hastings step. Conditioned on F, the regression parameters α_{s0} and α_{s1} are also generated by a Metropolis–Hastings step. See the Appendix for details.

5. Analysis of the Nevada Test-Site Data

5.1 *Model Fitting*

This section provides our reanalysis of the Nevada test-site data, where we illustrate the methods we have developed. In what follows, we will refer to the parametric dose–response model (5) and the semiparametric dose–response model (8). We will also refer to four error structures: (a) none, i.e., ignoring measurement error; (b) Berkson, i.e., when all measurement error is Berkson; (c) classical, i.e., when all measurement error is classical; and (d) mixture, i.e., when the fraction p of the measurement error variance is Berkson and p is uniformly distributed on the interval [0.2, 0.8]. We will also refer to models for the latent intermediate variable L, namely, the parametric normal model (4) and the semiparametric latent intermediate variable model described in Section 4.

In our analyses, the response Y was the period prevalence (1985–1986) of thyroid neoplasms. There were only 19 such neoplasms in the data set, although the effect of dose is statistically significant when ignoring measurement error and performing a likelihood ratio test.

Table 1 gives results for the parametric dose–response model (5) for the cases that measurement error is ignored, is purely Berkson, is purely classical, or is a mixture of

Table 1
Posterior means and credible sets for the parameter θ in model (5) *and for the relative risk at true dose* 1 *Gy* (100 *rad*)

Error model	Latent variable	Posterior mean θ	Lower 95% credible bound	Upper 95% credible bound	RR at dose = 1 Gy	Lower 95% credible bound	Upper 95% credible bound
No error		38.90	16.28	58.98	9.43	4.53	13.79
Classical	Normal	74.06	34.52	108.55	17.06	8.48	24.54
Classical	Semi	68.19	30.91	102.15	15.79	7.70	23.16
Berkson		31.90	13.09	48.00	7.92	3.84	11.41
Mixture	Normal	56.11	18.58	101.98	13.17	5.03	23.12
Mixture	Semi	45.60	12.15	94.99	10.89	2.63	22.60

classical and Berkson error. In the first two cases, no latent intermediate variable model is assumed, while for the other cases, we allow for the parametric or semiparametric latent intermediate variable model. This table gives results both for the estimate of θ as well as for the relative risk at true dose 1 Gy = 100 rad.

Note that, as expected from the theory, ignoring the measurement error leads to a slight overestimate of the dose–response rate as compared with a pure Berkson error analysis. In contrast, if all the measurement error were classical, ignoring measurement error would lead to a substantial underestimate of risk. This is in agreement with the calculations of Stevens et al. (1992). In results not reported here, we computed the maximum likelihood estimate for θ via numerical integration, the estimated value being almost the same as the posterior mean. As might be expected from these considerations, the mixture error model gives risk estimates between the no-error and 100%-classical error estimates.

Figure 1 illustrates the lack of normality of $\log(L)$ in Utah. Specifically, we computed a posterior mean Pólya tree distribution by averaging the MCMC probability values for each partition. We then generated 5000 observations from this posterior mean Pólya tree. As seen in Figure 1, the result is skew, pointing out the need for more flexible latent intermediate variable modeling in the log scale. Coupled with this plot indicating the need for a flexible distribution for the latent intermediate variable, we will present evidence in Section 5.2 in support of the need for a flexible dose–response function.

Table 2 gives the general results and compares the parametric and semiparametric dose–response models (5) and (8). Here we restrict attention to estimating the relative risk at true dose 1 Gy = 100 rad. In our discussion, we specifically want to contrast two analyses: (a) Berkson error model with the dose–response function (5), an analysis fairly close to that done in Stevens et al. (1992), and (b) the mixture of Berkson and classical errors with semiparametric dose–response and latent intermediate variable functions. Note that the latter model suggests a near doubling of the posterior mean relative risk from 7.92 to 14.23.

Perhaps the more interesting result is the comparison between the uncertainties in these posterior means as exhibited through 95% credible intervals. It is well known in measurement error models that correction for measurement error affects both parameter estimation and precision of inference. In our case, the Berkson error model suggests a lower bound on the relative risk of 3.84, while the mixture semiparametric model suggests a lower bound of 1.68. Corresponding large differences are seen in the upper 95% credible bounds, not too surprising given the extra flexibility in our modeling approach.

5.2 *Model Selection*

In selecting among the models described in Section 5.1, customary Bayesian model screening selects the model with the largest value of the marginal density of the data evaluated at the observations. In the present case, we will use the deviance information criterion (DIC) as in Spiegelhalter, Best, and Carlin (unpublished manuscript) to do this calculation. Let $\bar{D}$ be the posterior expectation of the deviance of the model and P_D be the effective number of parameters in the model, defined as $P_D = \bar{D} - D(\bar{\eta})$, where η contains all the parameters of the model and $\bar{\eta}$ is its posterior expectation. Then $DIC = \bar{D} + P_D$ and can be calculated easily using the MCMC samples and using the sample means of the simulated values of D and the plug-in estimates of the deviance using the sample means of the simulated values of all the parameters η.

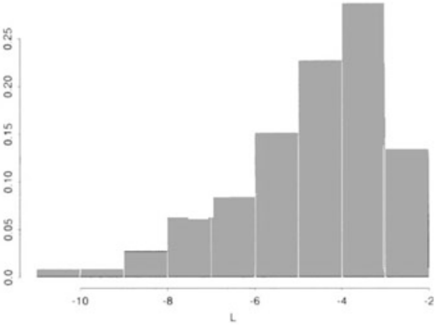


Figure 1. A histogram illustrating the posterior distribution of nonzero values of $\log(L)$ for Utah when modeled by a Pólya tree distribution.

Table 2

Posterior means and credible sets for the relative risk at true dose 1 Gy (100 rad)

Error model	Regression model	Latent variable model	Relative risk at dose 1 Gy	Lower 95% credible interval	Upper 95% credible interval
No error	Parametric	—	9.43	4.53	13.79
	Semiparametric	—	13.77	2.46	18.91
Berkson	Parametric	—	7.92	3.84	11.41
	Semiparametric	—	9.95	3.10	13.20
Classical	Parametric	Parametric	17.06	8.48	24.54
		Semiparametric	15.79	7.70	23.16
	Semiparametric	Parametric	21.54	9.41	36.66
		Semiparametric	18.98	7.98	32.69
Mixture	Parametric	Parametric	13.17	5.03	23.12
		Semiparametric	10.89	2.63	22.60
	Semiparametric	Parametric	16.42	2.19	34.75
		Semiparametric	14.23	1.68	33.61

For logistic regression with probabilities p_i, the deviance is $D = 2\Sigma_i Y_i \log(Y_i/p_i) + (Y_i - 1)\log\{(1 - Y_i)/(1 - p_i)\}$. DIC for no dose effect is 229.2. DIC for parametric models without error, with Berkson errors, with classical errors, and with mixture errors are 218.5, 214.6, 213.9, and 211.4, respectively. DIC for semiparametric models without error, with Berkson errors, with classical errors, and with mixture errors are 216.3, 211.7, 210.4, and 207.2, respectively. This gives some support for the need to use the semiparametric regression model (8) coupled with the semiparametric dose–response model (Section 4).

6. Simulations

We performed a small simulation study to understand the relative performance of our methods. The sample size was the same as in the data set, with two logistic functions in true dose X: (a) $\text{logit}\{\text{pr}(Y = 1 \mid X)\} = \log(1 + 0.6X)$ and (b) $\text{logit}\{\text{pr}(Y = 1 \mid X)\} = 2 - 1./(1 + X^2)$. We assumed that half the measurement error was classical and half was Berkson and that the measurement error was relatively large. Specifically, $\log(L) = \text{normal}(-0.3466, 0.8408^2)$, $\log(X) = \log(L) + \text{normal}(0, 0.8408^2)$, and $\log(W) = \log(L)+\text{normal}(0, 0.8408^2)$. These specifications means that $X < 0.10$ with probability 0.05 and $X < 5$ with probability 0.95, not too far from what appears to actually happen in the actual data with dose divided by 0.461Gy. We evaluated the relative risk on the interval 0 to 5. The main difference between the simulation and the data is that the former has many more observations with $Y = 1$.

We generated a single data set. Figures 2 and 3 compare the true relative risk function (thicker solid line), the estimated relative risk when error is ignored (thin solid line), and the fit via our semiparametric dose–response and latent intermediate variable model (dashed line). The figures demonstrate the superiority of our methods in these two cases.

In addition, we computed the DIC for these two simulated data sets. As expected, in the first simulation, the parametric model (dose–response and latent intermediate variable) had the lowest DIC, while in the second, the semiparametric model had the lowest DIC.

7. Discussion

7.1 *Summary Comments*

We have constructed Bayesian methods for analysis of data when predictors have a combination of classical and Berkson measurement error. We applied our methods to an important data set in radiation epidemiology. The methods allow for a semiparametric yet monotonic regression function along with a semiparametric latent intermediate variable model. The methods are easily extended to any generalized linear model. In our example, the combination of the two semiparametric approaches yielded the smallest deviance information criterion. It also yielded a much larger relative risk at relatively high doses than suggested by a Berkson error model with parametric dose–response function, albeit with much wider uncertainties in the estimate of this relative risk.

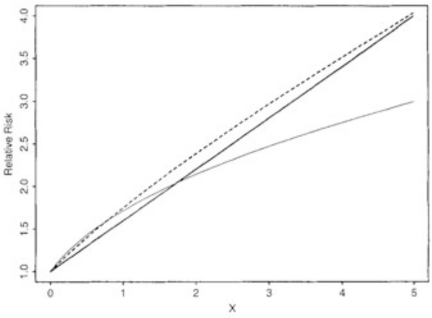


Figure 2. Results for the simulated data set with $\text{logit}\{\text{pr}(Y = 1 \mid X)\} = \log(1 + 0.6X)$. The true relative risk is the thicker solid line, the estimated relative risk ignoring measurement error is the thin solid line, and our semiparametric estimate is the dashed line.

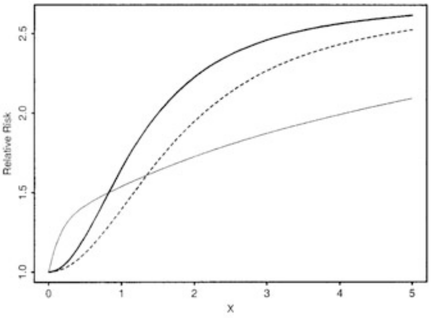


Figure 3. Results for the simulated data set with $\text{logit}\{\text{pr}(Y = 1 \mid X)\} = 2 - 1./\{1 + (X^2)\}$. The true relative risk is the thicker solid line, the estimated relative risk ignoring measurement error is the thin solid line, and our semiparametric estimate is the dashed line.

In our example, the total error variance, i.e., the sum of the Berkson and classical error variances, was assumed to be known for each individual but the individual error variances was unknown. In the context of the example, it was impractical for us to redo the dosimetry construction and to thus untangle the relative contributions to the total error variance. Our solution to this dilemma was to assume that a fraction p of each variance was of Berkson type, and we placed a uniform distribution on p within a well-defined interval that is reasonable in the context of the example. Of course, if the dosimetry could have been redone, then this information could be incorporated naturally into the Bayesian framework.

The preceding paragraph makes clear, and it is worth reemphasizing, that we have been forced to assume that the data set accurately specified the total error variance. This is clearly a major limitation of any analysis of dose uncertainties. It is also probable that the percentage p of total error that is classical varies from individual to individual; we have chosen to make p fixed across individuals, although at least in principle we could have allowed it to vary according to some specified distribution. These two pieces of unavoidable roughness in the data base means that the Nevada test-site data should best be thought of as an illustration of the general methodology.

In other examples, the total error variance may not be known. Our methods are, in principle, easily extended to this case, although issues of identifiability become more complex.

Whenever there is a component of classical error, a distribution for a latent intermediate variable must be specified. At least in principle, it is possible that misspecifying the distribution of the latent intermediate variable could cause biases in regression modeling. We assumed separate regression models for a categorical variable and considered both fully parametric and flexible semiparametric distributions, the latter based on the Pólya tree distribution. Clearly, there is nothing magical about the Pólya tree distribution, and other flexible semiparametric distributions could be used.

We also examined the form of the dose–response function, allowing the dose to be modeled semiparametrically. In the context of the example, it was reasonable to assume a monotonic function, and our semiparametric approach incorporates the monotonicity naturally.

7.2 *Shared Uncertainties*

Finally, we comment on our assumption that the Berkson and classical errors are independent across individuals. This is almost certainly not the case, and thus our data analysis may thus be best thought of as an illustration of methodology. The radioiodine concentration of milk, C, in (1) includes the deposition of I^{131} by region, its mass interception on vegetation, the effective half-life of I^{131} in the vegetation, the consumption of vegetation by cows, and the milk transfer coefficient (abbreviated here as MTC). While similar issues apply to the mass interception and the dose conversion factor, consider for example the MTC for a child in a particular region whose milk comes from a backyard cow: the problem we now discuss is probably even greater for children consuming milk from commercial dairies. As we understand it, as part of the modeling process, the Utah study generated a distribution for the MTCs as log-normally distributed with mean μ_{MTC} and variance σ^2_{MTC}. If these parameters were known, then the error structure for the MTC would be primarily Berkson. However, these parameters are not known and are instead estimated by a combination of historical data and literature review. This means that the error in estimating the coefficients is the same, hence shared, by all the children in the region with a backyard cow.

Understanding how such shared uncertainties affect parameter estimation and inference is an open problem worth considerable study. We have performed one preliminary calculation. We consider the parametric dose–response model, the parametric (normal) latent intermediate variable (L) model, and the mixture of Berkson and classical error structure. We allowed for shared uncertainties in the Berkson error model (2). Specifically, for the six groups formed by the combinations of states and genders, the Berkson errors for individuals within each group were assumed to have common correlation ρ, which we varied from 0.0, 0.2, 0.4, and 0.6. The posterior mean estimates of the parameter θ for each of these situations were 56.11, 65.85, 84.12, and 95.06, respectively. The credible intervals were $(18.58, 101.98)$, $(21.44, 120.21)$, $(30.41, 142.95)$, and $(38.23, 151.69)$, respectively. The fairly large changes in parameter estimates and credible intervals suggest the need in future for data to be gathered that can account for the possibility of shared uncertainties.

Acknowledgements

This research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We are deeply grateful to Duncan Thomas and Richard Kerber and to a referee for many useful suggestions.

Résumé

Nous construisons des méthodes Bayésiennes pour une modelisation semi-paramétrique d'une fonction de régression

monotone quand les prédicteurs sont mesurés avec une erreur classique, une erreur de Berkson ou les deux. De telles méthodes demandent une distribution pour le prédicteur non observé (variable latente), distribution que nous modélisons aussi de façon semi-paramétrique. De telles combinaisons de méthodes semi-paramétriques pour la réponse à des doses aussi bien que pour la distribution de la variable latente n'ont pas été étudiées dans la littérature sur les erreurs de mesure, quelle que soit la forme de l'erreur de mesure. De plus, nos méthodes proposent une nouvelle approche des problèmes où l'erreur de mesure combine une composante de Berkson et une composante classique. Ces méthodes sont générales mais nous les développons autour d'une application particulière qui est l'étude des maladies de la thyroïde en relation avec les radiations venant du site de test du Nevada. Nous utilisons ces données pour illustrer nos méthodes, lesquelles suggèrent une estimation ponctuelle (moyenne a posteriori) du risque relatif à des doses près de deux fois plus élevée que celle obtenue par les analyses précédentes, mais suggère aussi une bien plus grande incertitude sur le risque relatif.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis,* 2nd edition. New York: Springer.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* London: Chapman and Hall.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression with errors in covariates. *Biometrika* **86,** 541–554.

Carroll, R. J., Roeder, K., and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics* **55,** 44–54.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80,** 475–488.

Davis, S., Kopecky, K. J., Hamilton, T. E., and Amundson, B. (1998). *Hanford Thyroid Disease Study Draft Final Report.* Fred Hutchinson Cancer Research Center, Seattle, Washington.

Kerber, R. L., Till, J. E., Simon, S. L., Lyon, J. L., Thomas, D. C., Preston-Martin, S., Rollison, M. L., Lloyd, R. D., and Stevens, W. (1993). A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *Journal of the American Medical Association* **270,** 2076–2083.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *The Annals of Statistics* **20,** 1222–1235.

Mallick, B. K. and Gelfand, A. E. (1994). Generalized linear models with unknown link functions. *Biometrika* **81,** 237–245.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case–control studies with errors in variables. *Biometrika* **84,** 523–537.

Ostrouchov, G., Frome, E. L., and Kerr, G. D. (1998). *Dose estimation from daily and weekly dosimetry data: Final draft.* Technical Report. Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Reeves, G. K., Cox, D. R., Darby, S. C., and Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* **17,** 2157–2177.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A nonparametric mixture approach to case–control studies with errors in covariables. *Journal of the American Statistical Association* **91,** 722–732.

Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics* **57,** 53–61.

Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M., and Carroll, R. J. (2002). Thyroid cancer following scalp irradiation: A reanalysis accounting for uncertainty in dosimetry. *Biometrics* **57,** 689–697.

Simon, S. L., Till, J. E., Lloyd, R. D., Kerber, R. L., Thomas, D. C., Preston-Martin, S., Lyon, J. L., and Stevens, W. (1995). The Utah Leukemia case–control study: Dosimetry methodology and results. *Health Physics* **68,** 460–471.

Stevens, W., Till, J. E., Thomas, D. C., et al. (1992). *Assessment of leukemia and thyroid disease in relation to fallout in Utah: Report of a cohort study of thyroid disease and radioactive fallout from the Nevada test site.* Technical Report. University of Utah, Salt Lake City.

Walker, S. and Mallick, B. K. (1999). Semiparametric accelerated life time models. *Biometrics* **55,** 477–483.

Received April 2001. *Revised August* 2001.
Accepted August 2001.

Appendix

MCMC Details

The prior for the β's were independent normals with mean zero and variance 1000. The prior for θ was a truncated normal with mean 40 and variance 100. The prior for p was uniform$[0.2, 0.8]$. The prior for the ω's was Dirichlet$(\gamma = 1)$ (Berger, 1985, p. 561). The priors for the state-level parameters $(\alpha_{s0}, \alpha_{s1})$ was independent normals with mean zero and variance 100. The Pólya tree prior is as specified in Section 4.

The Metropolis proposals were as follows. The subscript "old" means the current values of the parameters. For the β's, the Metropolis proposals were normal$(\beta_{old}$, 1.0, and similarly for the θ's. For the ω's, the following considerations apply. To accommodate the constraint to the simplex, it is rescaled to $r - 1$ dimensional Euclidean space using a logit transformation. Supressing subscripts, let $z_\ell = \log(\omega_\ell)$. The Jacobian from the ω-space to the z-space is $\Pi_{\ell=1}^{r}\omega_\ell$, whence the complete conditional distribution for z, up to a constant of proportionality, is readily obtained. A normal proposal density with mean the current value and standard deviation 0.5 is used for z, and starting values $\omega_\ell = 1/r$ worked well. The proposals for the state-level variables α_{s0}, α_{s1} were independent normals with the current values as the mean and standard deviation 0.5. The prior is used as the proposal for p.

American Journal of Epidemiology
Copyright © 2003 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 158, No. 1
Printed in U.S.A.
DOI: 10.1093/aje/kwg091

Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study

Victor Kipnis[1], Amy F. Subar[2], Douglas Midthune[1], Laurence S. Freedman[3,4], Rachel Ballard-Barbash[2], Richard P. Troiano[2], Sheila Bingham[5], Dale A. Schoeller[6], Arthur Schatzkin[7], and Raymond J. Carroll[8]

[1] Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD.
[2] Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD.
[3] Department of Mathematics, Statistics and Computer Science, Bar Ilan University, Ramat Gan, Israel.
[4] Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel.
[5] Medical Research Council, Dunn Human Nutrition Unit, Cambridge, United Kingdom.
[6] Department of Nutritional Sciences, University of Wisconsin, Madison, WI.
[7] Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.
[8] Department of Statistics, Texas A&M University, College Station, TX.

Received for publication December 26, 2001; accepted for publication December 3, 2002.

Multiple-day food records or 24-hour dietary recalls (24HRs) are commonly used as "reference" instruments to calibrate food frequency questionnaires (FFQs) and to adjust findings from nutritional epidemiologic studies for measurement error. Correct adjustment requires that the errors in the adopted reference instrument be independent of those in the FFQ and of true intake. The authors report data from the Observing Protein and Energy Nutrition (OPEN) Study, conducted from September 1999 to March 2000, in which valid reference biomarkers for energy (doubly labeled water) and protein (urinary nitrogen), together with a FFQ and 24HR, were observed in 484 healthy volunteers from Montgomery County, Maryland. Accounting for the reference biomarkers, the data suggest that the FFQ leads to severe attenuation in estimated disease relative risks for absolute protein or energy intake (a true relative risk of 2 would appear as 1.1 or smaller). For protein adjusted for energy intake by using either nutrient density or nutrient residuals, the attenuation is less severe (a relative risk of 2 would appear as approximately 1.3), lending weight to the use of energy adjustment. Using the 24HR as a reference instrument can seriously underestimate true attenuation (up to 60% for energy-adjusted protein). Results suggest that the interpretation of findings from FFQ-based epidemiologic studies of diet-disease associations needs to be reevaluated.

bias (epidemiology); biological markers; diet; energy intake; epidemiologic methods; nutrition assessment; questionnaires; reference values

Abbreviations: DLW, doubly labeled water; FFQ, food frequency questionnaire; OPEN, Observing Protein and Energy Nutrition; 24HR, 24-hour dietary recall.

Much of the recent literature on the relation between diet and cancer has been based on analytic epidemiologic studies using food frequency questionnaires (FFQs). A number of large prospective studies of this kind have failed to find a consistent relation between dietary components (such as fat, fiber, and fruits and vegetables) and cancers of the breast, colon, or rectum (1–3), which may be explained by a true lack of diet-cancer associations or, alternatively, by serious methodological limitations of the studies themselves, especially due to FFQ measurement error.

Over the years, investigators have recognized that the reported values from FFQs are subject to substantial error, both systematic and random, that can profoundly affect the design, analysis, and interpretation of nutritional epidemiologic studies (4–6). Dietary measurement error often attenuates (biases toward one) the estimates of disease relative

Reprint requests to Dr. Victor Kipnis, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Executive Plaza North, Room 3124, 6130 Executive Boulevard, MSC 7354, Bethesda, MD 20892-7354 (e-mail: victor_kipnis@nih.gov).

risks and reduces statistical power to detect their significance. Therefore, an important relation between diet and disease may be obscured.

This problem has prompted researchers involved in large epidemiologic investigations to integrate calibration substudies that include a more intensive, but presumably more accurate, reference method, typically multiple-day food records (7) or multiple 24-hour dietary recalls (24HRs) (8). Comparing reference measurements with those from the FFQ enables adjustment for attenuation by using the regression calibration approach (7). However, the correct application of this approach requires that the adopted reference instrument satisfy two critical conditions. Although it may be imperfect and contain measurement error, this error should be independent of 1) true intake and 2) error in the FFQ (9). Throughout this paper, we take these two conditions as requirements for a valid reference instrument.

A great deal of accumulated evidence suggests that common dietary report reference instruments are unlikely to meet these requirements. Studies with the few biomarkers of dietary intake that do qualify as valid reference measurements ("reference" biomarkers), such as doubly labeled water (DLW) for total energy expenditure and urinary nitrogen for protein intake, demonstrate serious systematic biases in all dietary report instruments that may be potentially related (10–16). This has led to proposals for new models of dietary measurement error that might explain why the large prospective studies fail to find a relation between diet and cancer, even were an important relation to exist (9, 17, 18).

For example, Kipnis et al. (9) considered two potential systematic components of dietary measurement error. The first component reflects correlation between error and true intake ("intake-related" bias). The second component ("person-specific" bias) is independent of true intake and represents the difference between total within-person bias and its intake-related component. The existence of person-specific biases was proposed in all dietary report instruments, and a sensitivity analysis demonstrated that correlation between person-specific biases in the FFQ and the reference instrument, if ignored, could lead to serious underestimation of the degree of attenuation in a conventional calibration study. In a subsequent paper, Kipnis et al. (18) provided empirical evidence directly supporting their hypothesis, based on the results from a validation study that included the urinary nitrogen reference biomarker for protein intake. Moreover, based on the urinary nitrogen data, the measurement error model was extended to also include intake-related bias in dietary report reference instruments and was shown to fit the data statistically significantly better than other proposed models.

In this paper, we take this further by analyzing data from the Observing Protein and Energy Nutrition (OPEN) Study that included reference biomarkers for protein (urinary nitrogen) and energy (DLW) intakes, together with a FFQ and a 24HR. This study enabled us to evaluate not only absolute protein intake but also total energy and energy-adjusted protein intakes (19). We were therefore able to investigate the conjecture that energy adjustment substantially reduces measurement error in reported intake and that remaining error can be reliably corrected for by the common approach (20).

MATERIALS AND METHODS

Effect of measurement error

The effects of dietary measurement error on the estimation of disease risks are well known (9). The most important concept is that of attenuation. Consider the disease model

$$R(D|T) = \alpha_0 + \alpha_1 T, \qquad (1)$$

where $R(D|T)$ denotes the risk of disease D on an appropriate scale (e.g., logistic) and T is true habitual intake of a given nutrient, also measured on an appropriate scale. The slope α_1 represents an association between the nutrient intake and disease (e.g., log relative risk). In practice, FFQ-reported intake Q is used instead of unknown true intake T. We assume throughout that dietary measurement error is nondifferential with respect to disease D; that is, reported intake contributes no additional information about disease risk beyond that provided by true intake. To an excellent approximation, fitting model 1 to reported intake leads to estimating not the true risk parameter α_1 but the product $\tilde{\alpha}_1 = \lambda_1\alpha_1$ of α_1 and the slope λ_1 in the linear regression calibration model, $T = \lambda_0 + \lambda_1 Q + \xi$, where ξ denotes random error.

In nutritional studies, the value of λ_1 is usually between 0 and 1 (21), so dietary measurement error leads to underestimation of the true risk parameter. This underestimation is called attenuation, and λ_1 is called the attenuation factor. Values of λ_1 closer to zero lead to more serious underestimation of risk. For example, a true relative risk of 2 would appear as $2^{0.4} = 1.32$ if the attenuation factor were 0.4 and as $2^{0.2} = 1.15$ if the attenuation factor were 0.2.

Measurement error also leads to loss of statistical power for testing disease-exposure associations. Approximately, the sample size required to reach the desired statistical power to detect a given risk is proportional to $N \propto 1/\{\rho^2(Q,T)\sigma_T^2\} = 1/\{\lambda_1^2\sigma_Q^2\}$, where $\rho(Q,T)$ is the correlation between the reported and true intakes and σ_Q^2 and σ_T^2 are the between-person variances of the reported and true intakes, respectively (21). In particular, for a given FFQ, the required sample size is inversely proportional to the squared attenuation factor, λ_1^2. For example, if the true attenuation factor were 0.2, the sample size, calculated by assuming that $\lambda_1 = 0.4$, should be multiplied by $0.4^2/0.2^2 = 4$ to achieve the nominal power.

Estimation of the attenuation factor

Estimation of the attenuation factor λ_1 requires collecting additional reference measurements to compare with the FFQ in the calibration substudy (9). The common approach in nutritional epidemiology uses a more intensive dietary report method as the reference instrument, assuming that it is unbiased at the individual level and that its errors are independent of those in the FFQ (7). In this paper, we contrast this model with the measurement error model of Kipnis et al. (18) that specifies the same general error structure in the dietary report reference instrument (F) as the one for the FFQ (Q). To be fully identifiable, the model requires data from a reference biomarker. The model is specified as

$$\begin{aligned} Q_{ij} &= \mu_{Qj} + \beta_{Q0} + \beta_{Q1}T_i + r_i + \varepsilon_{ij} \\ F_{ij} &= \mu_{Fj} + \beta_{F0} + \beta_{F1}T_i + s_i + u_{ij} \\ M_{ij} &= \mu_{Mj} + T_i + \upsilon_{ij}, \end{aligned} \qquad (2)$$

where μ_{Qj}, μ_{Fj}, and μ_{Mj} are time-specific group intercepts for the FFQ, 24HR, and biomarker, respectively, which sum to zero over j; β_{Q0} and β_{F0} are the overall group intercepts for the FFQ and 24HR; β_{Q1} and β_{F1} are the slopes reflecting intake-related bias for the FFQ and 24HR; r_i and s_i are person-specific biases for the FFQ and 24HR that are independent of true intake T_i, have means zero, variances σ_r^2 and σ_s^2, respectively, and are correlated with the correlation coefficient ρ_{rs}; and ε_{ij}, u_{ij}, and υ_{ij} are within-person random errors for the FFQ, 24HR, and biomarker, with means zero and variances σ_ε^2, σ_u^2, and σ_υ^2, respectively, that are assumed to be independent of each other and of other terms in the model, except that "within-pair" errors (ε_{ij}, u_{ij}), (ε_{ij}, υ_{ij}), and (u_{ij}, υ_{ij}) are allowed to be correlated, if the corresponding measurements are taken contemporaneously.

In the presence of the reference biomarker, model 2 does not require an instrument F to estimate the error components in the FFQ. However, its inclusion enables us to additionally analyze the error structure of the dietary report reference instrument and its relation to that in the FFQ.

The common model may be obtained from model 2 by ignoring information from the reference biomarker and assuming that the dietary report instrument F contains no intake-related bias ($\beta_{F1} = 1$) or person-specific bias ($\sigma_s^2 = 0$). We use the following general form of this model:

$$\begin{aligned} Q_{ij} &= \mu_{Qj} + \beta_{Q0} + \beta_{Q1}T_i + r_i + \varepsilon_{ij}, \\ F_{ij} &= \mu_{Fj} + T_i + u_{ij}. \end{aligned} \qquad (3)$$

When the model parameters are used, the attenuation factor is expressed as

$$\lambda_1 = \frac{\mathrm{cov}(T, Q)}{\mathrm{var}(Q)} = \frac{\beta_{Q1}}{\beta_{Q1}^2 + \sigma_r^2/\sigma_T^2 + \sigma_\varepsilon^2/\sigma_T^2}, \qquad (4)$$

and the correlation of the FFQ and true intake is given by

$$\rho_{Q,T} = \frac{\mathrm{cov}(T, Q)}{\sqrt{\mathrm{var}(T)\mathrm{var}(Q)}} = \frac{\beta_{Q1}}{\sqrt{\beta_{Q1}^2 + \sigma_r^2/\sigma_T^2 + \sigma_\varepsilon^2/\sigma_T^2}}. \qquad (5)$$

Both are estimated by replacing the parameters by their estimates based on the corresponding model 2 or 3. Doing so is essentially equivalent to adjusting for random measurement error in the adopted reference instrument.

The OPEN data

The OPEN Study was conducted by the National Cancer Institute from September 1999 to March 2000. The recruitment procedure, subject characteristics, and detailed study conduct are described in the companion paper in this issue of the *Journal* (22). Briefly, 261 male and 223 female participants aged 40–69 years were healthy volunteers from Montgomery County, Maryland. Each participant was asked to complete a FFQ and a 24HR on two occasions. The FFQ was completed within 2 weeks of visit 1 and then approximately 3 months later, within a few weeks of visit 3. The 24HR was completed at visit 1 and then approximately 3 months later at visit 3. Participants received their DLW dose at visit 1 and returned 2 weeks later (visit 2) to complete the DLW assessment. In addition, repeat DLW measurements were collected from 14 male and 11 female volunteers who received their second DLW dose at the end of visit 2 and returned 2 weeks later to complete their DLW assessment. Participants provided two 24-hour urine collections during the 2-week period between visit 1 and visit 2, verified for completeness by using the PABAcheck method (23). Since approximately 81 percent of nitrogen intake is excreted through the urine (18) and nitrogen constitutes 16 percent of protein, the urinary nitrogen values were adjusted, by dividing by 0.81 and multiplying by 6.25, to estimate individual protein intake.

The adopted FFQ was the Diet History Questionnaire, developed and evaluated at the National Cancer Institute (24–28). The 24HR was a highly standardized version using the five-pass method, developed by the US Department of Agriculture for use in national dietary surveillance (29).

Statistical analysis

Throughout, we applied the logarithmic transformation to energy and protein to make measurement error in the DLW and urinary nitrogen biomarkers additive and homoscedastic and to better approximate normality. In addition to total energy and protein, the reference biomarkers in the OPEN Study enabled us to evaluate dietary measurement error for energy-adjusted protein intake. Because modeling relations between disease and multiple covariates measured with error is beyond the scope of this paper, we assumed that model 1 included only energy-adjusted exposure and that energy was not related to disease. We used two energy adjustment methods: nutrient density and nutrient residual (19). Protein density was calculated as the percentage of energy from protein sources and was then log transformed. The protein residual was calculated from the linear regression of protein on energy intake on the log scale. Both protein density and residual were calculated for each instrument by using the protein and energy intakes as measured by this instrument. The convention used for dealing with biomarker-based derived measures is explained in the Appendix.

For all dietary variables, we excluded extreme outlying values that fell outside the interval given by the 25th percentile minus twice the interquartile range to the 75th percentile plus twice the interquartile range. For each variable and each

TABLE 1. Estimated attenuation factor λ_1 and correlation $\rho(Q, T)$ of food frequency questionnaire-reported intake (Q) and true intake (T)* in the Observing Protein and Energy Nutrition Study, Maryland, September 1999–March 2000

Nutrient	Gender	Model	Attenuation factor λ_1	Correlation $\rho(Q, T)$
Energy	Male	Biomarker based†	0.080 (0.025)‡	0.199 (0.061)
		24HR based§	0.230 (0.037)	0.437 (0.065)
	Female	Biomarker based	0.039 (0.028)	0.098 (0.069)
		24HR based	0.128 (0.044)	0.261 (0.088)
Protein	Male	Biomarker based	0.156 (0.034)	0.323 (0.067)
		24HR based	0.177 (0.043)	0.312 (0.074)
	Female	Biomarker based	0.137 (0.041)	0.298 (0.088)
		24HR based	0.158 (0.046)	0.334 (0.098)
Protein density	Male	Biomarker based	0.404 (0.066)	0.431 (0.063)
		24HR based	0.409 (0.066)	0.497 (0.077)
	Female	Biomarker based	0.316 (0.084)	0.346 (0.087)
		24HR based	0.501 (0.060)	0.778 (0.117)

* As estimated by the model accounting for the reference biomarker of intake or the common model accounting only for the 24-hour recall (24HR) as reference measurements.

† Defined as model 2 in the text.

‡ Numbers in parentheses, standard error.

§ Defined as model 3 in the text.

instrument, no more than six outlying values for men and four for women were excluded from the analyses.

The estimates of the model parameters and their standard errors were obtained by using the method of maximum likelihood under the assumption of normality of the random terms in the models. Standard errors were checked for accuracy by using the bootstrap method. Comparisons of correlated parameters (such as attenuation factors estimated by two models) were performed by comparing the ratios of their differences to the standard errors of the differences calculated by the bootstrap method with the standardized normal distribution.

RESULTS

The descriptive statistics for measurements taken by using the different instruments are provided in the companion paper (22). For energy-adjusted protein, the results for only nutrient density are shown since the results for nutrient residual were similar.

Attenuation and correlation with true intake

Table I displays the estimates of the attenuation factor λ_1 and correlation $\rho(Q, T)$ between the FFQ and true usual intake resulting from applying models 2 and 3 to energy, protein, and energy-adjusted protein. The table contrasts the estimated values when the common approach versus the biomarker-based model was used.

Absolute intakes. The biomarker-based attenuation factors were distressingly close to zero. For example, for women, the attenuation factors for energy and protein were 0.039 and 0.137, respectively. The attenuation factors estimated by using the common approach were substantially higher (underestimating the corresponding attenuation) for energy at 0.128 (p = 0.05 when compared with the biomarker-based attenuation) and somewhat higher for protein at 0.158 (p = 0.73). Results for men showed a similar pattern, with the attenuation factor being statistically significantly overestimated (p < 0.001) when the common approach for energy was used.

The correlations between the FFQ and true intake were also very low. The biomarker-based correlations for energy and protein intakes for women were 0.098 and 0.298, respectively, while the common approach overestimated correlations at 0.261 (p = 0.10) and 0.334 (p = 0.81). For men, the correlation estimated by using the common approach was statistically significantly biased upward (p < 0.001) for energy.

Energy-adjusted intakes. For energy-adjusted intakes, the attenuation factors were somewhat higher (attenuation was lower) than for absolute intakes. For example, for women, the biomarker-based estimate for protein density was 0.316 compared with 0.137 for protein (p = 0.10). Results for men showed a similar pattern, with the highly statistically significant difference in attenuation between absolute and energy-adjusted protein intakes (p < 0.001).

The attenuation factor estimated by using the common approach for women again appeared substantially more optimistic than the biomarker-based estimate at 0.501 versus 0.316 (p = 0.10) for protein density. For men, however, no marked difference was found between the attenuation factors estimated by using the two models. Correlations between FFQ and true intake for energy-adjusted protein displayed the same pattern as those for attenuation factors.

TABLE 2. Variance of true intake and parameters of dietary measurement error in the food frequency questionnaire and 24-hour dietary recall,* the Observing Protein and Energy Nutrition Study, Maryland, September 1999–March 2000

Nutrient	Gender	Model	Variance of true intake $\sigma^2_T \times 10^2$	Slope in regression of FFQ†-reported on true intake β_{Q1}	Slope in regression of 24HR-reported on true intake β_{F1}	Variance of person-specific bias in FFQ $\sigma^2_r \times 10^2$	Variance of person-specific bias in 24HR $\sigma^2_s \times 10^2$	Correlation of person-specific biases in FFQ and 24HR $\rho_{r,\,s}$	Variance of within-person error in FFQ $\sigma^2_\varepsilon \times 10^2$	Variance of within-person error in 24HR $\sigma^2_u \times 10^2$
Energy	Male	Biomarker based‡	2.6 (0.27)§	0.49 (0.15)	0.66 (0.10)	12.2 (1.2)	3.2 (0.61)	0.45 (0.08)	3.2 (0.28)	5.3 (0.48)
		24HR based¶	4.4 0.68)	0.83 (0.15)	1	9.7 (1.2)	0		3.2 (0.28)	5.3 (0.47)
	Female	Biomarker based	2.4 (0.29)	0.24 (0.17)	0.46 (0.13)	11.2 (1.3)	3.2 (0.78)	0.28 (0.11)	3.9 (0.37)	7.9 (0.75)
		24HR based	3.7 (0.81)	0.53 (0.20)	1	10.3 (1.3)	0		3.9 (0.37)	7.9 (0.75)
Protein	Male	Biomarker based	4.4 (0.57)	0.67 (0.15)	0.70 (0.11)	13.3 (1.4)	3.9 (0.94)	0.18 (0.10)	3.7 (0.33)	9.3 (0.82)
		24HR based	6.1 (1.0)	0.55 (0.14)	1	13.5 (1.5)	0		3.7 (0.33)	9.3 (0.82)
	Female	Biomarker based	3.7 (0.71)	0.65 (0.21)	0.60 (0.16)	11.0 (1.5)	2.6 (1.1)	0.24 (0.15)	4.8 (0.46)	12.0 (1.2)
		24HR based	3.9 (1.2)	0.70 (0.25)	1	10.7 (1.6)	0		4.8 (0.46)	12.0 (1.2)
Protein density	Male	Biomarker based	3.1 (0.47)	0.46 (0.08)	0.62 (0.11)	1.6 (0.25)	1.2 (0.50)	0.40 (0.15)	1.2 (0.11)	5.8 (0.51)
		24HR based	2.4 (0.53)	0.60 (0.13)	1	1.4 (0.30)	0		1.2 (0.11)	5.8 (0.51)
	Female	Biomarker based	3.5 (0.72)	0.38 (0.11)	0.39 (0.13)	2.3 (0.36)	1.2 (0.60)	0.94 (0.19)	1.4 (0.13)	6.8 (0.65)
		24HR based	1.7 (0.59)	1.2 (0.36)	1	0.30 (0.76)	0		1.4 (0.13)	6.8 (0.65)

* As estimated by the model accounting for the reference biomarker of intake or the common model accounting only for the 24-hour dietary recall (24HR) as reference measurements

† FFQ, food frequency questionnaire.

‡ Defined as model 2 in the text.

§ Numbers in parentheses, standard error.

¶ Defined as model 3 in the text.

Error structure of the FFQ and 24HR

Intake-related bias. Table 2 demonstrates across-the-board intake-related bias in both FFQ and 24HR measurements. All biomarker-based estimates of slopes β_{Q1} and β_{F1} were substantially smaller than the desired value of 1.0, leading to the flattened slope phenomenon. If anything, energy adjustment seemed to make this phenomenon even more pronounced. The flattened slope in the FFQ estimated by using the common approach is often not seen as clearly. For example, for males, the DLW-based estimate of β_{Q1} for energy intake was 0.49, but the common estimate was 0.83.

Person-specific bias. Table 2 also demonstrates the existence and importance of person-specific biases in reported intakes from the FFQ and 24HR. Compared with the true between-person variance σ_T^2, the person-specific biases σ_s^2 and σ_r^2 were quite dominant for absolute intakes. For example, for females reporting protein intake, the FFQ person-specific bias variance was 0.110 and the 24HR person-specific bias variance was 0.026, quite large compared with the variance of true intake (0.037). Energy adjustment considerably reduced person-specific biases. Continuing with the example above, for protein density, this variance was reduced from 0.110 to 0.023 for the FFQ and from 0.026 to 0.012 for the 24HR, while the variance of true intake remained practically the same (0.035). However, even for energy-adjusted intakes, person-specific biases were still substantial and highly significantly different from zero.

Table 2 also demonstrates substantial positive correlation $\rho_{r,s}$ between person-specific biases in the FFQ and 24HR. The correlation increased after energy adjustment, especially for women.

Within-person random error. For absolute intakes, within-person random variation σ_ε^2 in the FFQ was of the same magnitude as between-person variation σ_T^2 of true intake. Similar to person-specific bias, it was considerably reduced by energy adjustment. As expected because of day-to-day variation in intake, within-person random variation σ_u^2 in the 24HR was substantially greater. Interestingly, relative to variation of true intake, it was only moderately reduced by energy adjustment. In all cases considered, within-person random errors were not statistically significantly correlated across instruments.

"Nonprotein" intake. Using the measurements for protein and energy on each instrument, we also evaluated dietary measurement error for nonprotein-energy-contributed nutrients ("nonprotein" for short), for both absolute nonprotein and energy-adjusted nonprotein intakes. The results for absolute nonprotein intake were similar to the results for energy, and the results for energy-adjusted nonprotein were similar to the results for energy-adjusted protein.

DISCUSSION

In this paper, we focused mostly on the attenuation factor because it directly affects the observed relative risks and the sample size necessary to detect diet-disease associations in epidemiologic studies. The critical requirement for our results that the adopted biomarkers represent valid reference instruments, that is, their errors are unrelated to true intakes and errors in dietary report instruments, is supported by accumulated evidence for both the adjusted urinary nitrogen (18) and DLW (30). The OPEN Study yielded the following conclusions.

First, the impact of FFQ measurement error on total energy and absolute protein intakes was severe and in agreement with the findings of Kipnis et al. (18) for protein intake. Attenuation factors were vexingly close to zero, as were the correlations with true intake.

Second, the impact of measurement error seemed less severe after energy adjustment. As follows from expression 4, the attenuation factor is inversely proportional to the variances of both person-specific bias and within-person random error relative to between-person variation of true intake. Since these relative variances decreased substantially after energy adjustment (table 2) because of correlated errors in reporting protein and energy, energy-adjusted protein was less affected by measurement error compared with absolute protein intake. However, the estimated attenuation factors for energy-adjusted intakes were in the range 0.32–0.41 (table 1), indicating that measurement error still remained an important problem.

Third, the 24HR was seriously flawed, suffering from intake-related bias and from person-specific bias that was correlated with person-specific bias in the FFQ. As a result, it violated both requirements for a valid reference instrument and in most cases substantially misrepresented the impact of measurement error in the FFQ. As follows from formula A1 in the Appendix, bias in the attenuation factor λ_F calculated by using the common approach depends on the sum of the values for slope β_{F1} and expression

$$\rho_{r,s}\sqrt{(\sigma_r^2/\sigma_T^2)(\sigma_s^2/\sigma_T^2)}/\beta_{Q1}\,.$$

Table 2 reveals that, for absolute intakes, the relative variances of person-specific biases in the FFQ and 24HR and the correlation between them were sufficiently large to override the small values of β_{F1} and to raise λ_F above the true attenuation factor λ_1. The same remained true for energy-adjusted protein in women, where the effect of reduced person-specific biases was compensated for by the increased correlation between them. As a result, the 24HR underestimated true attenuation. On the other hand, for energy-adjusted protein in men, the two effects essentially cancelled each other, demonstrating that a flawed reference instrument may sometimes produce a good estimate.

Our results are in line with previous data presented on protein intake. For women in the British Medical Research Council study (18), the urinary-nitrogen-based attenuation factor for protein was 0.187, while the common approach based on a 4-day weighed food record produced an overly optimistic estimate of 0.282. The former is slightly larger than the 0.137 obtained in the OPEN Study, while the latter is noticeably more optimistic than our 24HR-based estimate of 0.158 ($p = 0.08$). The correlations of FFQ with true intake were 0.284 (urinary nitrogen based) and 0.432 (record based) compared with our values of 0.298 (urinary nitrogen based) and 0.334 (24HR based), respectively. Neither difference approaches statistical significance.

An important consideration is whether our results could be affected by the fact that biomarkers in the OPEN Study were collected mostly over one season. We analyzed 24HRs taken in different seasons in cross-sectional national survey data (Continuing Survey of Food Intakes by Individuals 1994–1996) by region and gender, and we found no seasonal fluctuations in energy or protein intakes. However, if seasonality were to exist, it would affect only the estimated mean usual intake and would not change the higher-order parameters presented in tables 1 and 2.

Since DLW measures total energy expenditure, it would be important to adjust the data for long-term weight change to enable DLW to truly represent usual energy intake. Doing so over the 2-week DLW period may introduce only more random error, however, since only a small amount of within-person week-to-week fluctuations in energy balance can be explained by contemporary changes in weight (31). Even using the 3-month OPEN Study period may not adequately represent long-term weight changes, especially given protocol differences in fasting conditions between the first and last visits (22). Nevertheless, when we adjusted individual DLW measurements for the weight change over either the 2-week or 3-month period, the results did not change materially for either absolute or energy-adjusted nutrients.

Recently, Willett (20) suggested that any evaluation of a FFQ would be invalid unless heterogeneity in the study population due to gender, age, and body size was adjusted for. To address this issue, we performed further analyses that included age in 5-year groups and the logarithm of body mass index as covariates in the models. The results did not change substantially except for energy in women; the attenuation factor and correlation of the FFQ with true intake became even closer to zero.

Our results have important implications for nutritional epidemiology. First, they question the ability of FFQs to detect diet-disease associations for absolute nutrient intakes. While some journals have recently required that energy adjustment be used in the analysis of nutrient-disease associations, the practice has been controversial (32, 33). Our data clearly document failure of the FFQ to provide a sufficiently accurate report of absolute protein, nonprotein, and energy intakes to enable detection of their moderate associations with disease. For example, with the attenuation factors of 0.08 for energy intake for males and 0.04 for females, a true relative risk of 2.0 would appear as 1.06 and 1.03, respectively, using the FFQ data. Needless to say, such small relative risks are not detectable in epidemiologic studies since their signal is smaller than the noise caused by confounders. It is plausible that similarly small attenuation factors would be found for many other nutrients, although it would require a suitable reference biomarker for each nutrient to confirm this possibility.

Second, it appears that FFQ-based energy-adjusted nutrient intakes may just be sufficiently accurate to use in large cohort studies to detect moderate diet-disease associations; a relative risk of 2.0 would appear close to 1.3, which could be at the limits of detection. The benefits of adjusting for energy intake have been discussed previously at the general level (19, 32). Our conclusion is necessarily a qualified one, since our study was restricted to energy-adjusted protein and nonprotein intakes. There is no guarantee that the results will be as favorable for nonprotein components such as energy-adjusted fat intake. Even less could be speculated about the effect of energy adjustment for non-energy-contributing nutrients. Nevertheless, until further evidence becomes available on other nutrients, use of energy-adjusted intakes seems the best working approach for nutritional epidemiology, at least under the assumption that energy is not related to disease. Note, however, that biomarker-based attenuation factors for energy-adjusted protein intake are between 0.32 and 0.41, indicating that measurement error has a substantial negative impact on the statistical power of observational epidemiologic studies.

Third, our results throw into question use of the 24HR as a reference instrument for validation/calibration studies. In the OPEN Study, such use substantially overestimated performance of the FFQ for absolute intakes of energy and nonprotein. The results also cast some doubt on the performance of the 24HR as a reference for energy-adjusted intakes. For example, for protein density in women (table 1), the biomarker-based attenuation factor was estimated as 0.3 compared with the 24HR-based estimate of 0.5. Use of the latter would lead to underestimation of the required sample size by a factor of $2.8 = 0.5^2/0.3^2$, with profound effects on the power to detect diet-disease associations.

The OPEN Study provides solid evidence of measurement errors in a FFQ as they pertain to energy intake and both absolute and energy-adjusted protein and nonprotein intakes. Further studies of a similar design are needed to confirm our results, especially to clarify whether 24HRs or multiple-day food records can be used reliably as reference instruments in validation/calibration studies, at least for energy-adjusted intakes. Unfortunately, few dietary biomarkers qualify as valid reference instruments; that is, they have errors unrelated to true intakes and errors in dietary report instruments. Most other biomarkers, such as vitamin C or beta-carotene, measure concentrations of related constituents for which the quantitative relation to dietary intake is unknown and depends on individual characteristics (e.g., concomitant intake of other nutrients, obesity, or smoking habits) (34). Therefore, such concentration-based biomarkers cannot provide valid reference measurements and at best can serve only as correlates of intake. Further work should explore whether a combination of data from dietary report and biomarker measurements for energy or protein can be used to assess dietary exposure variables for which no reference biomarkers exist.

ACKNOWLEDGMENTS

Dr. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-E509106).

REFERENCES

1. Hunter DJ, Spiegelman D, Adami HO, et al. Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. N Engl J Med 1996;334:356–61.
2. Fuchs CS, Giovannucci EL, Colditz GA, et al. Dietary fiber and the risk of colorectal cancer and adenoma in women. N Engl J Med 1999;340:169–76.
3. Michels KB, Giovannucci E, Joshipura KJ, et al. Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers. J Natl Cancer Inst 2000;92:1740–52.
4. Beaton GH, Milner J, Corey P, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Am J Clin Nutr 1979;32:2546–59.
5. Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. Nutr Cancer 1988;11:243–50.
6. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning a sample size required in a cohort study. Am J Epidemiol 1990;132:1185–95.
7. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051–69.
8. Kaaks R, Riboli E. Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. Int J Epidemiol 1997;26(suppl):S15–25.
9. Kipnis V, Carroll RJ, Freedman LS, et al. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. Am J Epidemiol 1999;150:642–51.
10. Bandini LG, Schoeller DA, Cyr HN, et al. Validity of reported energy intake in obese and nonobese adolescents. Am J Clin Nutr 1990;52:421–5.
11. Livingstone MBE, Prentice AM, Strain JJ, et al. Accuracy of weighed dietary records in studies of diet and health. BMJ 1990;300:708–12.
12. Heitmann BL. The influence of fatness, weight change, slimming history and other lifestyle variables on diet reporting in Danish men and women aged 35–65 years. Int J Obes 1993;17:329–36.
13. Heitmann BL, Lissner L. Dietary underreporting by obese individuals—is it specific or non-specific? BMJ 1995;311:986–9.
14. Martin LJ, Su W, Jones PJ, et al. Comparison of energy intakes determined by food records and doubly labeled water in women participating in a dietary-intervention trial. Am J Clin Nutr 1996;63:483–90.
15. Sawaya AL, Tucker K, Tsay R, et al. Evaluation of four methods for determining energy intake in young and older women: comparison with doubly labeled water measurements of total energy expenditure. Am J Clin Nutr 1996;63:491–9.
16. Black AE, Bingham SA, Johansson G, et al. Validation of dietary intakes of protein and energy against 24 urinary N and DLW energy expenditure in middle-aged women, retired men and post-obese subjects: comparisons with validation against presumed energy requirements. Eur J Clin Nutr 1997;51:405–13.
17. Prentice R. Measurement error and results from analytic epidemiology: dietary fat and breast cancer. J Natl Cancer Inst 1996;88:1738–47.
18. Kipnis V, Midthune D, Freedman LS, et al. Empirical evidence of correlated biases in dietary assessment instruments and its implications. Am J Epidemiol 2001;153:394–403.
19. Willett WC. Nutritional epidemiology. Chapter 5. New York, NY: Oxford University Press, 1990.
20. Willett W. Commentary: dietary diaries versus food frequency questionnaires—a case of undigestible data. Int J Epidemiol 2001;30:317–19.
21. Kaaks R, Riboli E, van Staveren W. Calibration of dietary intake measurements in prospective cohort studies. Am J Epidemiol 1995;142:548–56.
22. Subar AF, Kipnis V, Troiano RP, et al. Using intake biomarkers

to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN Study. Am J Epidemiol 2003;158:1–13.

23. Bingham SA, Cummings JH. Urine nitrogen as an independent validatory measure of dietary intake: a study of nitrogen balance in individuals consuming their normal diet. Am J Clin Nutr 1985;42:1276–89.
24. Subar AF, Thompson FE, Smith AF, et al. Improving food frequency questionnaires: a qualitative approach using cognitive interviewing. J Am Diet Assoc 1995;95:781–8.
25. Subar AF, Midthune D, Kulldorff M, et al. Evaluation of alternative approaches to assign nutrient values to food groups in food frequency questionnaires. Am J Epidemiol 2000;152:279–86.
26. Subar AF, Ziegler RG, Thompson FE, et al. Is shorter always better? Relative importance of questionnaire length and cognitive ease on response rates and data quality for two dietary questionnaires. Am J Epidemiol 2001;153:404–9.
27. Subar AF, Thompson FE, Kipnis V, et al. Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: the Eating at America's Table Study. Am J Epidemiol 2001;154:1089–99.
28. Thompson FE, Subar AF, Brown CC, et al. Cognitive research enhances accuracy of food frequency questionnaire reports: results of an experimental validation study. J Am Diet Assoc 2002;102:212–25.
29. Moshfegh AJ, Raper N, Ingwersen I, et al. An improved approach to 24-hour dietary recall methodology. Ann Nutr Metab 2001;45(suppl 1):156.
30. Schoeller DA. Measurement error of energy expenditure in free-living humans by using doubly labeled water. J Nutr 1988; 118:1278–89.
31. Edholm OG, Healy MJR, Wolfe HS, et al. Food intake and energy expenditure in army recruits. Br J Nutr 1970;24:1091–107.
32. Willett W, Howe GR, Kushi L. Adjustment for total energy intake in epidemiological studies. Am J Clin Nutr 1997; 65(suppl):1220S–8S.
33. Freedman LS, Kipnis V, Brown CC, et al. Comments on "Adjustment for total energy intake in epidemiological studies." Am J Clin Nutr 1997;65(suppl):1229S–31S.
34. Kaaks RJ. Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: conceptual issues. Am J Clin Nutr 1997;65(suppl):1232S–9S.

APPENDIX

Derived Reference Measures Based on the Observed Biomarkers

In the OPEN Study, replications of the DLW measurement were available for only a small sample of 25 persons (14 men and 11 women). This fact did not affect the results for total energy intake since the DLW measurements were remarkably consistent across replications. The coefficient of variation in the DLW measurements was only 5.1 percent, in effect indicating that energy expenditure was measured with very little error.

However, a technical difficulty arose in the analysis of nonprotein and energy-adjusted nutrients. The error in the biomarker-based derived reference measures was almost entirely influenced by the error in the urinary nitrogen measurements, where the coefficient of variation was 17.6 percent. As a result, attempting to estimate the within-person variance of the derived reference measurements as a parameter in the model led to relatively large standard errors in the main analysis and to instability in the procedure for bootstrap calculations.

On the basis of these facts, in dealing with the derived reference measurements for nonprotein and energy-adjusted protein and nonprotein intakes, we used the following convention. When defining biomarker-based reference measures for nonprotein as well as nutrient density and nutrient residual, we used the first DLW observation with both the first and second repeat urinary nitrogen observations. In theory, doing so induced some correlation between repeat biomarker-based reference observations, but the DLW measurement error was so small that this correlation could be ignored in practice.

Bias in the Attenuation Factor Based on the Dietary Report Reference Instrument

For a valid reference biomarker M, the attenuation factor is expressed as $\lambda_M = \mathrm{cov}(M, Q)/\mathrm{var}(Q) = \mathrm{cov}(T,Q)/\mathrm{var}(Q)$ (18).

Thus, the biomarker-based attenuation factor λ_M is equal to the true attenuation factor λ_1. However, the attenuation factor λ_F based on the common approach with a dietary report reference instrument is given by $\lambda_F = \mathrm{cov}(F,Q)/\mathrm{var}(Q) = (\beta_{F1}\beta_{Q1}\sigma_T^2 + \mathrm{cov}(r,s))/\mathrm{var}(Q)$.

Taking into account expression 4 for the true attenuation factor λ_1, we can rewrite this expression as

$$\lambda_F = \lambda_1\left(\beta_{F1} + \frac{1}{\beta_{Q1}}\rho_{r,s}\sqrt{\frac{\sigma_r^2\sigma_s^2}{\sigma_T^2\sigma_T^2}}\right). \qquad \text{(A1)}$$

Thus, the attenuation factor λ_F is generally biased. The relative bias, defined by the expression in parentheses, depends on intake-related biases in the FFQ and dietary report instrument F, reflected by slopes β_{Q1} and β_{F1}, respectively; the variances of their person-specific biases relative to variation in true intake, σ_r^2/σ_T^2 and σ_s^2/σ_T^2, respectively; and the correlation $\rho_{r,s}$ between person-specific biases. Values of slope β_{F1} less than one decrease λ_F relative to true attenuation factor λ_1, whereas positive values of

$$\rho_{r,s}\sqrt{(\sigma_r^2/\sigma_T^2)(\sigma_s^2/\sigma_T^2)},$$

as well as values of slope β_{Q1} less than one, increase λ_F.

Biometrika (2005), **92**, 2, *pp.* 399–418
© 2005 Biometrika Trust
Printed in Great Britain

Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies

By NILANJAN CHATTERJEE

Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland 20852, U.S.A.

chattern@mail.nih.gov

and RAYMOND J. CARROLL

Texas A&M University, College Station, Texas 77843-3143, U.S.A.

carroll@stat.tamu.edu

Summary

We consider the problem of maximum-likelihood estimation in case-control studies of gene-environment associations with disease when genetic and environmental exposures can be assumed to be independent in the underlying population. Traditional logistic regression analysis may not be efficient in this setting. We study the semiparametric maximum likelihood estimates of logistic regression parameters that exploit the gene-environment independence assumption and leave the distribution of the environmental exposures to be nonparametric. We use a profile-likelihood technique to derive a simple algorithm for obtaining the estimator and we study the asymptotic theory. The results are extended to situations where genetic and environmental factors are independent conditional on some other factors. Simulation studies investigate small-sample properties. The method is illustrated using data from a case-control study designed to investigate the interplay of BRCA1/2 mutations and oral contraceptive use in the aetiology of ovarian cancer.

Some key words: Case-control study; Gene-environment interaction; Genetic epidemiology; Logistic regression; Population stratification; Profile likelihood; Retrospective study; Semiparametric method.

1. Introduction

The case-control study design gives an efficient way of collecting covariate information for epidemiological studies of rare diseases. Cornfield (1956) showed that the prospective odds ratio of a disease given a covariate is equivalent to the retrospective odds ratio of the covariate given the disease and thus prospective odds ratios are estimable from case-control designs. For discrete covariates, Andersen (1970) and then more generally Prentice & Pyke (1979) showed that fitting a standard prospective logistic regression that ignores the retrospective sampling nature of the design yields the maximum likelihood estimates of the regression parameters under a 'semiparametric' model that allows the covariate distribution to be nonparametric. More recently, Rabinowitz (1997) and Breslow et al.

(2000) used modern semiparametric theory to show that the prospective logistic regression analysis of case-control data is efficient in the sense that it achieves the variance lower bound of the underlying semiparametric model.

It is now believed that the risks of many complex diseases are determined by the combined effects of genetic susceptibility G and environmental or non-genetic exposures E, and, since studies of interactions, especially for rare exposures, typically require a large sample size, efficient designs and analytical methods for gene-environment interaction are vital.

A special feature of the gene-environment interaction problem is that it may often be reasonable to assume that a subject's genetic susceptibility, a factor which is determined from birth, is independent of his/her subsequent environmental exposure. Standard logistic regression analysis, being the semiparametric maximum likelihood solution for the problem that allows an arbitrary covariate distribution, clearly remains a valid option for analysing case-control data. However, the method may not be efficient because it fails to exploit the gene-environment independence assumption. In general, under the case-control design, the variance lower bound for estimators of the regression parameters under particular constraints or models for the covariate distribution will be lower than that of the more general model that allows a completely nonparametric covariate distribution.

In the past, several researchers have presented analytical methods that exploit the gene-environment independence assumption. Piegorsch et al. (1994) noted that, under gene-environment independence and the rare disease assumption, the multiplicative interaction parameter in the logistic regression model can be estimated as the odds ratio between G and E among cases alone. Moreover, they observed that the corresponding case-only estimator of interaction is more precise than the estimator of the interaction parameter from traditional logistic regression analysis involving both cases and controls. When data on both cases and controls are available, assuming rare disease and categorical exposures, Umbach & Weinberg (1997) showed that maximum-likelihood estimators of all the parameters of a logistic regression model can be obtained in a fairly general setting by fitting a suitably constrained log-linear model to the data. They showed that, for simple scenarios that involve dichotomous G, dichotomous E and no confounder, the log-linear model and case-only analysis approach yields the same estimator of the multiplicative interaction parameter in the logistic regression model. Modan et al. (2001), in a specific application, noted that, under gene-environment independence and the rare disease assumption, $\text{pr}(E|G, D=0) = \text{pr}(E|D=0)$, where $D=0$ corresponds to disease-free, i.e. control, subjects. Based on this, they argued that the disease odds ratio associated with E among subjects with genotype $G=g$ can be estimated by a logistic regression analysis that compares the distribution of E among all controls, $\text{pr}(E|D=0)$, with the exposure distribution among cases with $G=g$, $\text{pr}(E|D=1, G=g)$.

The methods have some limitations. First, they all require the risk of the disease to be small for all levels of both genetic and environmental exposures. This assumption can lead to substantial bias in the estimation of the odds ratio parameters even for diseases like cancer, for which the marginal probability of the disease may be small in the population but the disease risk may be high for certain combinations of genetic and environmental exposures (Schmidt & Schaid, 1999). Secondly, the methods of Piegorsch et al. (1994) and Modan et al. (2001) allow estimation of some, but not all, of the parameters of interest in the general logistic regression model. Thirdly, some of the above methods have been described in very simple settings involving only two factors G and E, and it is often not clear how to exploit the gene-environment independence assumption in the most general setting that will, for example, allow for potential confounders or account for factors that could induce association between G and E. The log-linear model framework described by Umbach & Weinberg (1997) for categorical co-factors gives the most general method to date for exploiting the gene-environment independence assumption and can handle some of these issues. For a rich model with many covariates, however, the log-linear modelling approach can easily become cumbersome and intricate. Moreover, in a rich model, the log-linear specification would typically involve a large number of 'nuisance parameters' that characterise the covariate distribution among the controls. When continuous covariates are involved, the number of such nuisance parameters would even increase with the sample size. The asymptotic theory for the lower-dimensional regression parameters of interest in the presence of the high-dimensional nuisance parameters is nonstandard and has not been studied rigorously under the underlying semiparametric setting.

In this paper, we develop a general framework for maximum-likelihood estimation under the gene-environment independence assumption. The proposed method has several unique aspects. First, it is exact in not requiring any rare-disease assumption. Secondly, we develop the methodology in a very general setting so that it retains all the flexibility of traditional logistic regression analysis, such as adjustment for confounders, incorporation of continuous exposures and/or confounders and complex modelling of the regression effects of the risk factors. Thirdly, we show how to incorporate external information about the marginal probability of the disease in the population and hence improve efficiency of parameter estimation. Fourthly, we show how to adjust for bias that may arise when G and E may be related because of their dependence on other common measured factors. Finally, we develop the methodology in a semiparametric framework that allows the distribution of the environmental factors $F(e)$ to be completely nonparametric. Given that in a typical application E might include many factors, both discrete and continuous variables, nonparametric treatment of $F(e)$ is attractive both for avoiding complex modelling and for robustness.

2. Estimation theory and methodology

2·1. *Model and identification*

Let D be the binary indicator of presence, $D=1$, or absence, $D=0$, of a disease. Suppose the prospective risk model for the disease given a subject's genetic factors, G, and environmental risk factors, E, is given by the logistic regression model $\text{pr}(D=1|G,E)=H\{\beta_0+m(G,E;\beta_1)\}$, where $H(x)=\{1+\exp(-x)\}^{-1}$ is the logistic distribution function and $m(.)$ is a known but arbitrary function. Typically, in the standard logistic regression model, one has $m(G,E,\beta_1)=(G,E,G*E)\beta_1$ with the exponents of the parameters in β_1 having the standard exposure odds ratio interpretation. However, more general forms of $m(.)$ could be of interest, especially for interaction studies where different forms of $m(.)$ can be chosen to assess interaction at different scales; see Khouri et al. (1993, § 5.5.3). We assume that the joint distribution of G and E is given by the product form $\mathscr{H}(e,g)=Q(g)F(e)$, where Q and F are the marginal distribution functions of G and E, respectively. Suppose that N_0 controls and N_1 cases are sampled from the conditional distributions $\text{pr}(G,E|D=1)$ and $\text{pr}(G,E|D=0)$, respectively, and let $(G_i,E_i)_{i=1}^{N_0+N_1}$ denote the corresponding covariate data of the N_0+N_1 study subjects.

Before we describe estimation, it is useful to study the identifiability of the parameters. In a nonparametric setting where no assumption is made about the form of the covariate distribution $\mathscr{H}$, it is well known that neither $\mathscr{H}$ nor the intercept parameter β_0 is identifiable from case-control data (Prentice & Pyke, 1979). Under the assumption of gene-environment independence, however, these results may not necessarily be true. Let $\mathscr{B}$ denote the parameter space for β_1 and let $\mathscr{B}^0 \subset \mathscr{B}$ denote the values of β_1 so that $m(G, E, \beta_1)$ depends only on G or only on E, but not both. For example, suppose that $m(G, E, \beta_1)$ corresponds to a standard logistic regression model with $\beta_1 = (\beta_G, \beta_E, \beta_{GE})$, where β_G, β_E and β_{GE} denote the main effect of G, the main effect of E and the interaction between G and E, respectively. In this case, the set $\mathscr{B}^0$ would consist of parameter values of the form $\beta_1 = (\beta_G, 0, 0)$ or $\beta_1 = (0, \beta_E, 0)$, which correspond to either only the main effect of G or only the main effect of E, respectively. Since β_1 is well known to be identifiable from case-control data under general nonparametric $\mathscr{H}$, it follows trivially that β_1 remains identifiable when H is assumed to be of the form $\mathscr{H} = Q \times F$. The identifiability result for the remaining parameters can be stated as follows.

LEMMA 1. *For all* $\beta_1 \notin \mathscr{B}^0$,

$$\text{pr}(E = e, G = g|D = d, \beta_0, \beta_1, Q, F) = \text{pr}(E = e, G = g|D = d, \beta_0^*, \beta_1, Q^*, F^*)$$

if and only if $\beta_0 = \beta_0^*$, $Q = Q^*$ *and* $F = F^*$.

The proof of Lemma 1 is given in the Appendix. Thus, we note that a somewhat surprising consequence of the gene-environment independence assumption is that, except for some boundary situations, the intercept parameter of the logistic regression model β_0 is theoretically identifiable from the retrospective likelihood of case-control data. Although this may seem counter-intuitive, it is easy to see from the proof of Lemma 1 that in general the identifiability of β_0 is intrinsically related to the class of $\mathscr{H}$ that is under consideration.

2·2. *Profile likelihood estimation*

We begin with the following parameterisation of the exposure distributions Q and F. We assume that the genetic factor G for a subject can take values in a fixed set $\{g_0, \ldots, g_J\}$. Thus the distribution Q can be parameterised by the corresponding probability masses $\{q_0, \ldots, q_J\}$. Moreover, using population genetics theory, in many situations the probabilities q_j $(j = 1, \ldots, J)$ can be further modelled as $q_j = q_j(\theta)$, for some known function q_j and some parameter vector θ. For example, if G represents one of the three possible genotypes a subject can have corresponding to a bi-allelic locus, the population frequencies of the three genotypes could be specified in terms of the allele frequency of one of the alleles under the Hardy–Weinberg equilibrium assumption for the underlying population. If no population genetics model assumption is made to specify the q_j's, we will assume in the above notation that θ represents the vector of q_j's themselves.

For parameterisation of the environmental covariate E, we first assume that the nonparametric maximum likelihood estimator of F can allow positive masses only within the set $\mathscr{E} = \{e_1, \ldots, e_K\}$ that represents the unique values of E that are observed in the case-control sample of $N_0 + N_1$ study subjects. Thus, for obtaining the maximum likelihood estimator it is sufficient to consider the class of discrete F that have support points within the set $\mathscr{E}$. Any F in this class can be parameterised with respect to the probability masses $\{\delta_1, \ldots, \delta_K\}$ that it assigns to the points $\{e_1, \ldots, e_K\}$. Let n_{ijk} denote the number of subjects with $D = i$, $G = g_j$ and $E = e_k$. Let the corresponding marginal frequencies for the ith category of disease be $n_{i++} = N_i$, for the jth category of G be n_{+j+} and for the kth category of E be n_{++k}. The loglikelihood for the case-control data is then

$$\begin{aligned} L = \log\{\ell_{\mathrm{CC}}(\beta_0, \beta_1, \theta, \delta)\} &= \sum_{u=1}^{N_0+N_1} \log\{\mathrm{pr}(D_u|G_u, E_u)\,\mathrm{pr}(G_u)\,\mathrm{pr}(E_u)/\mathrm{pr}(D_u)\} \\ &= \sum_{ijk} n_{ijk} \log\{P_{ij}(e_k, \beta_0, \beta_1)\} + \sum_j n_{+j+} \log\{q_j(\theta)\} + \sum_k n_{+k+} \log(\delta_k) \\ &\quad - \sum_i n_{i++} \log\left\{\sum_{lm} P_{il}(e_m, \beta_0, \beta_1) q_l(\theta) \delta_m\right\}, \qquad (1) \end{aligned}$$

where $P_{ij}(e_k, \beta_0, \beta_1) = \mathrm{pr}(D = i|G = j, E = e_k)$.

When the dimension of δ is large, as could be expected when E consists of multiple covariates and/or some of its components are continuous, direct maximisation of the loglikelihood with respect to $(\beta_0, \beta_1, \theta, \delta)$ may be numerically challenging or even infeasible. An alternative approach is first to derive the profile likelihood of the data that is obtained by maximising the likelihood with respect to δ for fixed values of $\gamma = (\beta_0, \beta_1, \theta)$ and then to maximise the profile likelihood with respect to γ. If $\hat{\delta}(\gamma)$ denotes the value of δ that maximises the likelihood for fixed γ, the profile loglikelihood is then $L(\gamma, \hat{\delta}(\gamma)\}$. In Lemma 2, we state an equivalent representation of $L\{\gamma, \hat{\delta}(\gamma)\}$ that is computationally useful.

LEMMA 2. *Define the parameters* $\mu_i = n_{i++}/\{N\,\mathrm{pr}(D = i)\}$ *for* $i = 0, 1$ *and let*

$$P^*_{ij}\{e_k; \gamma, \mu = (\mu_0, \mu_1)\} = \frac{P_{ij}(e_k; \beta_0, \beta_1)\mu_i q_j(\theta)}{\sum_i \sum_j P_{ij}(e_k; \beta_0, \beta_1)\mu_i q_j(\theta)}. \qquad (2)$$

The profile loglikelihood $L\{\gamma, \hat{\delta}(\gamma)\}$ *can be computed as* $L^*\{\gamma, \hat{\mu}(\gamma)\}$, *where*

$$L^*(\gamma, \mu) = \sum_{ijk} n_{ijk} \log P^*_{ij}(e_k; \gamma, \mu), \qquad (3)$$

and $\hat{\mu}(\gamma) = \{\hat{\mu}_0(\gamma), \hat{\mu}_1(\gamma)\}$ *is defined by the solution of the equations*

$$n_{i++} = \sum_k \sum_j n_{++k} P^*_{ij}(e_k; \gamma, \mu) \quad (i = 0, 1). \qquad (4)$$

The proof of the lemma is given in the Appendix and is developed following techniques in Scott & Wild (1997). The main consequence of Lemma 2 is that $L\{\gamma, \hat{\delta}(\gamma)\}$ can be computed without having to maximise the likelihood $L(\gamma, \delta)$ numerically with respect to the potentially high-dimensional nuisance parameter δ. Instead, $L\{\gamma, \hat{\delta}(\gamma)\}$ can be obtained in closed form up to only two additional parameters $\mu = (\mu_0, \mu_1)$, which in turn are defined as the solution of two equations given in (4). The result of this lemma can also be compared to the classical result of Prentice & Pyke (1979) that, when the exposure distribution is unspecified, maximisation of the retrospective likelihood can be achieved by simply fitting the prospective logistic model to the data ignoring the retrospective design. Lemma 2 gives the corresponding simplification for maximum-likelihood estimation under the gene-environment independence assumption and unspecified distribution of E. Prentice & Pyke (1979) also essentially showed that the maximum likelihood estimator of the logistic regression parameters can be obtained by maximising the prospective likelihood of the form $\mathrm{pr}(D|X, \delta = 1)$, where δ is the indicator of whether or not a subject has been selected in the case-control sample and $\mathrm{pr}(\delta = 1|D)$, the probability of selection of a subject given his/her disease status, is fixed at its asymptotic value, which in turn is proportional to μ_D. Similarly, Lemma 2 shows that the maximum likelihood estimator of the regression parameter under the gene-environment independence assumption can be obtained by solving score equations corresponding to the prospective likelihood $P^*_{DG}(E) = \mathrm{pr}(D, G|E, \delta = 1)$. For derivation of the asymptotic distribution theory, however, we will show later that $P^*_{DG}(E)$ cannot be treated as an ordinary likelihood.

For computational convenience we consider further reparameterisation of the problem. Let $G = g_0$ define a reference category for the genetic exposure G. We can now write

$$P^*_{ij}(e_k, \gamma, \mu) = \frac{\exp\{\theta_{ij}(e_k; \gamma, \mu)\}}{1 + \sum_{ij:(ij)\neq(0,0)} \exp\{\theta_{ij}(e_k; \gamma, \mu)\}}, \tag{5}$$

where

$$\begin{aligned}\theta_{ij}(e_k; \gamma, \mu) &= \log\left\{\frac{P^*_{ij}(e_k; \gamma, \mu)}{P^*_{00}(e_k; \gamma, \mu)}\right\}\\ &= i\{\beta_0 + \log(\mu_1/\mu_0)\} + im(g_j, e_k; \beta_1) + \log\{q_j(\theta)/q_0(\theta)\}\\ &\quad + \log\left[\frac{1 + \exp\{\beta_0 + m(g_0, e_k; \beta_1)\}}{1 + \exp\{\beta_0 + m(g_j, e_k; \beta_1)\}}\right].\end{aligned} \tag{6}$$

Thus, $L^*(\gamma, \mu) = \sum_{ijk} n_{ijk} \log P^*_{ij}(e_k; \gamma, \mu)$ depends on μ_0 and μ_1 only through the parameter $\kappa = \beta_0 + \log(\mu_1/\mu_0)$. Moreover, since

$$\frac{\partial}{\partial\kappa} L^*(\gamma, \kappa) = n_{1++} - \sum_k \sum_j n_{++k} P^*_{1j}(e_k, \gamma, \mu),$$

it follows that $\hat{\kappa}(\gamma) = \beta_0 + \log\{\hat{\mu}_1(\gamma)/\hat{\mu}_0(\gamma)\}$, where $\hat{\mu}_1(\gamma)$ and $\hat{\mu}_0(\gamma)$ are defined in equation (4), will satisfy the equation $(\partial/\partial\kappa)L^*(\gamma, \kappa) = 0$. Thus, the semiparametric maximum likelihood estimate of γ can be obtained by solving the equation $\partial L^*(\gamma, \kappa)/\partial(\gamma, \kappa) = 0$ jointly with respect to γ and κ.

Estimation of β_0 in the above approach requires special attention. From the expression for $\theta_{ij}(e_k; \gamma, \mu)$ given in (6), it can be seen that the intercept parameter β_0 is involved in $L^*(\gamma, \kappa)$ not only through κ but also through the term

$$\tau(e_k, g_j, \beta_0, \beta_1) = \log\left[\frac{1 + \exp\{\beta_0 + m(g_0, e_k; \beta_1)\}}{1 + \exp\{\beta_0 + m(g_j, e_k; \beta_1)\}}\right].$$

Thus, in principle, β_0 is identifiable from $L^*(\gamma, \kappa)$ independently of the parameter κ. However, for diseases that are rare for all combinations of g_j and e_k, that is $\tau(e_k, g_j, \beta_0, \beta_1) \simeq 0$ for all j and k, there would be little information about β_0 from $L^*(\gamma, \kappa)$ that is not absorbed in κ. Since the corresponding information matrix is nearly singular, direct optimisation of $L^*(\gamma, \kappa)$ with respect to β_0 using standard methods such as Newton–Raphson can be numerically unstable. To overcome this problem, one strategy that we have found useful is to consider the profile likelihood of β_0 obtained as $L^*\{\beta_0, \hat{\beta}_1(\beta_0), \hat{\theta}(\beta_0), \hat{\kappa}(\beta_0)\}$, where $\{\hat{\beta}_1(\beta_0), \hat{\theta}(\beta_0), \hat{\kappa}(\beta_0)\}$ denotes the solution of the equation $\partial L^*(\beta_0, \beta_1, \theta, \kappa)/\partial(\beta_1, \theta, \kappa) = 0$ for fixed β_0. One can then perform a one-dimensional grid search for the optimal value of β_0 that maximises $L^*\{\beta_0, \hat{\beta}_1(\beta_0), \hat{\theta}(\beta_0), \hat{\kappa}(\beta_0)\}$, possibly on a fixed interval of values.

In the above approach, for rare diseases, the estimate of the parameter β_0 itself can be expected to be imprecise because of intrinsic noninformativeness of the retrospective likelihood. Much more precise estimation of β_0 is possible when the marginal probability of the disease, $\mathrm{pr}(D = 1)$, in the underlying population is known. We can then fix the parameters μ_i for $i = 0, 1$ in $L^*(\gamma, \mu_0, \mu_1)$ at their true values $n_{i++}/\{N\,\mathrm{pr}(D = i)\}$ for

$i = 0, 1$, respectively. In the corresponding expression for $\theta^*_{ij}(e_k, \gamma, \mu)$ given in (6), $\log(\mu_1/\mu_0)$ will be fixed and β_0 will be identifiable from the first term of (6) itself. In this case, the parameterisation $\eta = \{\beta_0, \beta_1, \theta, \kappa = \beta_0 + \log(\mu_1/\mu_0)\}$ is unnecessary and instead the original parameterisation $\eta = (\beta_0, \beta_1, \theta)$ should be used. Hereafter, we will use the generic notation η so that our results are valid for both the cases of $\mathrm{pr}(D = 1)$ being known and $\mathrm{pr}(D = 1)$ being unknown.

2·3. *Asymptotic theory*

In this section, we study the asymptotic properties of the semiparametric maximum likelihood estimator of η. Since we have shown that the estimator can be obtained by solving the equation $\partial L^*(\eta)/\partial\eta = 0$, the asymptotic properties can be studied by estimating-equation theory. Since $L^*(\eta) = \sum_{ijk} n_{ijk} \log P^*_{ij}(e_k, \eta)$, where $P^*_{ij}(e_k, \eta)$ is defined in (5), the estimating function $\partial L^*(\eta)/\partial\eta$ can be expressed as

$$\begin{aligned}\frac{\partial L^*}{\partial\eta} &= \sum_{ijk} n_{ijk}\left[\frac{\partial\theta_{ij}(e_k;\eta)}{\partial\eta} - \sum_{i'j'} \frac{\exp\{\theta_{i'j'}(e_k;\eta)\}}{\sum_{i''j''}\exp\{\theta_{i''j''}(e_k;\eta)\}}\frac{\partial\theta_{i'j'}(e_k;\eta)}{\partial\eta}\right]\\ &= \sum_{u=1}^{N}\left[\frac{\partial\theta_{D_uG_u}(E_u;\eta)}{\partial\eta} - E^*_{DG}\left\{\frac{\partial\theta_{DG}(E;\eta)}{\partial\eta}\middle| E = E_u\right\}\right],\end{aligned} \quad (7)$$

where $E^*_{DG}(.|E)$ denotes expectation with respect to the joint probability distribution for D and G given E that was defined by P^* in (2). Define $\Psi(D_u, G_u, E_u; \eta)$ to be the summand in the second expression of formula (7). We will develop the asymptotic theory in a scenario in which the total sample size $N = N_0 + N_1$ goes to infinity, but the sampling proportions for the cases and controls, namely N_0/N and N_1/N, remain fixed at π_1 and $\pi_0 = 1 - \pi_1$, respectively. We first state a lemma, proved in the Appendix, that will be used repeatedly in the development of the asymptotic theory, because in various places we will need to compute expectations and limits of functions in the case-control sampling scheme.

LEMMA 3. *Under the case-control sampling design described above and for any measurable function $Q(D, G, E)$ of data (D, G, E),*

$$N^{-1}\sum_{u=1}^{N} Q(D_u, G_u, E_u) \rightarrow \int E^*_{DG}\{Q(D, G, E)|E = e\}h(e)dF(e),$$

where the convergence is in probability and $h(e) = \sum_{ij} P_{ij}(e; \beta_0, \beta_1)\mu_i q_j(\theta)$, if we assume that the integral in the above equation exists.

At this point, we note an important subtlety of studying asymptotic theory under the case-control sampling design when the assumption of gene-environment independence is made. If no assumption is made about the joint distribution of (G, E), that is the form of $\mathscr{H}(g, e)$ is left completely unspecified, then from standard case-control sampling theory it follows that $N^{-1}\sum_{u=1}^{N} Q(D_u, G_u, E_u) \rightarrow \tilde{E}_{D,G,E}\{Q(D, G, E)\}$, where the convergence is in probability and where $\tilde{E}_{D,G,E}$ corresponds to expectation with respect to a joint distribution function $\tilde{\mathrm{pr}}(D, G, E)$, so that $\tilde{\mathrm{pr}}(G, E|D) = \mathrm{pr}(G, E|D)$ and $\tilde{\mathrm{pr}}(D) = N_D/N$. This follows because, when the form of $\mathscr{H}$ is left unspecified, one can vary the parameters β_0 and $\mathscr{H}$ without changing the value of the retrospective likelihood $\mathrm{pr}(G, E|D)$ (Roeder et al., 1996, Lemma 1). In particular, one can choose $\tilde{\beta}_0$ and $\tilde{\mathscr{H}}$ so that $\mathrm{pr}_{\beta_0,\beta_1,\mathscr{H}}(G, E|D) = \mathrm{pr}_{\tilde{\beta}_0,\beta_1,\tilde{\mathscr{H}}}(G, E|D)$ and $\mathrm{pr}_{\tilde{\beta}_0,\beta_1,\tilde{\mathscr{H}}}(D) = N_D/N$. However, these results do not hold when one assumes $\mathscr{H}$ to be of the form $Q \times F$ as in this case we have shown that α and $\mathscr{H}$ are uniquely identifiable from the retrospective likelihood, except for some boundary parameter values. Similarly, other standard theories for case-control sampling may not be applicable under the gene-environment independence model.

In Lemma 4, proved in the Appendix, we state the limiting form of the second derivatives of $L^*(\eta)$.

Lemma 4. *We have that*

$$\frac{1}{N}\frac{\partial^2 L^*}{\partial\eta\,\partial\eta^{\mathrm{T}}} \to \int V^*_{DG}\left\{\frac{\partial\theta_{DG}(E;\eta)}{\partial\eta}\middle| E=e\right\} h(e)dF(e) \equiv \mathscr{I},$$

in probability, where $V^*_{DG}(.|E)$ denotes variance with respect to the joint probability distribution for D and G given E that is defined by P^*.

Finally, we state the main asymptotic limiting results, proved in the Appendix.

Proposition 1. *Under suitable regularity conditions, the following results hold*:

(i) *the estimating equations $\partial L^*/\partial\eta \equiv \sum_{i=1}^N \Psi(D_i, G_i, E_i; \eta) = 0$ have a unique, consistent sequence of solutions, $\{\hat\eta^N\}_{N\geqslant 1}$*;

(ii) *if $\Omega = \sum_{d=0}^1 \mu_d[E\{\Psi(D, G, E)|D=d\}]^{\otimes 2}$, then $N^{\frac{1}{2}}(\hat\eta^N - \eta_0) \to N(0, \Sigma)$ in distribution, with*

$$\Sigma = \mathscr{I}^{-1} - \mathscr{I}^{-1}\Omega\mathscr{I}^{-1}. \qquad (8)$$

3. Extensions

3·1. *Population stratification*

Although genetic susceptibility and environmental exposures are unlikely to be causally related at an individual level, these factors may be correlated at a population level because of their dependence on other factors, such as ethnicity. In this section, we briefly describe how to generalise our methods to handle ‘population stratification’. Most of the details and proofs of the theoretical results follow from straightforward generalisation of the results derived in § 2.

We will assume that G and E are independent conditional on a set of variables S so that the joint distribution of G, E and S is given by the product form $H(g, e, s) = Q_s(g) \times F(e, s)$, where $Q_s(g)$ corresponds to the distribution of G given $S = s$ and $F(e, s)$ denotes the joint distribution of E and S. The distribution function $F(e, s)$ will be treated nonparametrically. Let $\mathrm{pr}(G = g_j|S = s)$ be denoted by $q_j(s; \theta)$ with θ being a fixed set of parameters characterising the conditional distribution. If S involves only discrete variables that define a relatively small number of strata, then no modelling of $\mathrm{pr}(G = g_j|S = s)$ is necessary and θ may denote the vector of conditional probabilities themselves. If S involves a relatively large number of variables, possibly including continuous ones, parametric modelling of the distribution $\mathrm{pr}(G|S)$ will be necessary. When G is a binary variable indicating the presence or absence of a certain genetic variation, for example, $\mathrm{pr}(G|S)$ can be parametrically specified through a logistic regression model. We further assume that the disease-risk model is given by $\mathrm{pr}(D = 1|G, E, S) = H\{\beta_0 + m(G, E, S; \beta_1)\}$. Thus, we allow the stratum variables S to be covariates of interest in the disease model. Let (e_k, s_k), for $k = 1, \ldots, K$, be the unique observed values for (E, S) and let n_{ijk} denote the number of subjects in the data with $D = i$, $G = j$ and $(E, S) = (e_k, s_k)$. As before, let $\mu_i = n_{i++}/\{N\,\mathrm{pr}(D = i)\}$.

With this notation, the results of Lemma 4 can be generalised to show that the semiparametric maximum likelihood estimator of $\gamma=(\beta_0, \beta_1, \theta)$ can be obtained by solving the equation $\partial L^*(\gamma, \mu)/\partial(\gamma, \mu)=0$ jointly with respect to (γ, μ), where

$$L^*(\gamma, \mu)=\sum_{ijk} n_{ijk} \log P^*_{ij}(e_k, s_k; \gamma, \mu)$$

and $P^*_{ij}(e_k, s_k; \gamma, \mu)$ is defined by formula (2) with $P_{ij}(e_k, \beta_0, \beta_1)$ and $q_j(\theta)$ replaced by $P_{ij}(e_k, s_k, \beta_0, \beta_1)$ and $q_j(s; \theta)$, respectively. Moreover, $P^*_{ij}(e_k, s_k; \gamma, \mu)$ can be written in the form of expression (5) with $\theta_{ij}(e_k; \gamma, \mu)$ replaced by $\theta_{ij}(e_k, s_k; \gamma, \mu)$, which in turn is defined by equation (6) with $q_j(\theta)/q_0(\theta)$ and $m(g, e_k; \beta_1)$ replaced by $q_j(s; \theta)/q_0(s; \theta)$ and $m(g, e_k, s_k; \beta_1)$, respectively. All the theory that we developed in § 2·3 can now be generalised by replacing E with $E'=(E, S)$ and $q_j(\theta)$ by $q_j(s; \theta)$ everywhere.

3·2. *Frequency-matched case-control studies*

In this section, we comment briefly on the modifications needed for the proposed methodology while dealing with frequency-matched case-control studies in which controls are selected in numbers proportional to the number of cases within strata defined by some matching variables W. The problem of individually-matched case-control studies is addressed in a separate article (Chatterjee et al., 2005). Let $W=w_m$ $(m=1, \ldots, M)$ denote M strata used for matching. To allow for factors, such as race, which may be candidates for both matching and population stratification, we write $W=(W^S, W^{\bar{S}})$, so that W^S represents the elements of W that are included in S, the factors for population stratification. Similarly, we write $S=(S^W, S^{\bar{W}})$, so that S^W denotes elements of S that are included in W. We will assume that G is independent of $(E, W^{\bar{S}})$ conditional on S. We further assume that the regression model is given by

$$\mathrm{pr}(D=1|G, E, S^{\bar{W}}, W)=H\{\beta_{0W}+m(G, E, S^{\bar{W}}, W; \beta_1)\},$$

so that it corresponds to the standard practice of allowing for an independent intercept term for each level of the matching variable $W=w$. Let $\beta_0=(\beta_{01}, \ldots, \beta_{0M})$ be the vector of intercept parameters corresponding to the M different values of W.

With the above notation and definitions, the retrospective likelihood for the matched case-control design can be written as

$$\ell_{\mathrm{MCC}}=\prod_{i=1}^{N_0+N_1} \mathrm{pr}(G_i, E_i, S_i^{\bar{W}}|D_i, W_i),$$

where the conditioning on (D, W) represents the fact that in a matched case-control design subjects are selected into the study based on both the disease status D and the matching variable W. The semiparametric maximum likelihood estimator of $\gamma=(\beta_0, \beta_1, \theta)$ that leaves the joint distribution of (E, S, W) completely unspecified can be derived by following techniques of §§ 2·2, 2·3 and 3·1, with μ_i replaced throughout by μ_{wi}, where $\mu_{wi}=n_{wi++}/\{N\, \mathrm{pr}(D=i|W)\}$, in which n_{wi++} is the number of subjects with $D=i$ and $W=w$ in the sample. In particular, we can show that the semiparametric maximum likelihood estimate of γ can be obtained by jointly solving a set of equations of the form $\partial L^*(\gamma, \kappa)/\partial(\gamma, \kappa)=0$, where $\kappa=(\kappa_1, \ldots, \kappa_M)$ with $\kappa_m=\beta_{0m}+\log(\mu_{1m}/\mu_{0m})$ and $L^*(\gamma, \kappa)=\sum_{ijwk} n_{wijk} \log P^*_{wij}(e_k, s_k; \gamma, \mu)$, in which $P^*_{wij}(e_k, s_k; \gamma, \mu)$ is defined by formula (5) with $\theta_{ij}(e_k; \gamma, \mu)$ replaced by

$$\theta_{wij}(e_k, s_k; \gamma, \mu) = \log\left\{\frac{P^*_{wij}(e_k, s_k; \gamma, \mu)}{P^*_{w00}(e_k, s_k; \gamma, \mu)}\right\}$$

$$= ik_w + im(g_j, e_k, s_k^{\bar{W}}, w; \beta_1) + \log\{q_j(s_k; \theta)/q_0(s_k; \theta)\}$$

$$+ \log\left[\frac{1+\exp\{\beta_{0w} + m(g_0, e_k, s_k^{\bar{W}}, w; \beta_1)\}}{1+\exp\{\beta_{0w} + m(g_j, e_k, s_k^{\bar{W}}, w; \beta_1)\}}\right]. \quad (9)$$

Using the structure of $\theta_{wij}(.)$ we observe that β_{0w} is involved in $L^*(\gamma, \kappa)$ not only through κ_w but also through the last term of expression (9), which we will denote by $\tau(g_j, e_k, s_k^{\bar{W}}, w; \beta_{0w}, \beta_1)$. For rare diseases, however, for which $\tau(g_j, e_k, s_k^{\bar{W}}, w; \beta_{0w}, \beta_1) \simeq 0$ for all values of j and k there would be little information about β_{0w} from $L^*(\gamma, \kappa)$ that is not absorbed in κ_w. In most case-control studies the matching factor W consists of basic demographic factors such as race, sex and age-groups, for which $\text{pr}(D=1|W)$ is available externally, for example from a population registry. In this case, μ_{wi} can be treated as a fixed parameter in the definition of κ_w and hence β_{0w} can be identified through κ_w itself. Use of external information about $\text{pr}(D=1|W)$ is recommended as it would not only resolve any numerical problems that may arise with estimation of the barely identifiable parameters, but would also improve efficiency of estimation of the other regression parameters of interest. An alternative solution for diseases that are extremely rare, such as the example of ovarian cancer we consider in § 5, is to ignore the term $\tau(g_j, e_k, s_k^{\bar{W}}, w; \beta_{0w}, \beta_1)$ in the calculations. Under the rare-disease assumption, we note that the functional form of L^* becomes exactly the same under frequency-matched and traditional unmatched case-control sampling designs, and thus the estimates under the matched design can be obtained by the method described in § 2 for the traditional case-control design with a disease risk model that allows for an independent intercept term for each level of the matching variable W. The estimator for the main effects for W would yield an unbiased estimator not of β_{0w} but of κ_w.

4. Simulation study

4·1. *The factors G and E are independent*

In the first experiment, we study the relative performance of the standard logistic regression analysis and the proposed semiparametric maximum-likelihood estimator under the gene-environment independence model. We assumed that the genetic covariate G is a binary variable, where for example $G=1$ or $G=0$ corresponds to presence or absence of a genetic mutation, respectively. We considered two scenarios: (a) $\text{pr}(G=1)=0{\cdot}065$ and (b) $\text{pr}(G=1)=0{\cdot}26$, corresponding to a rare and a common genetic mutation, respectively. We generated the environmental covariate as $E=\min(10, X)$ where X follows the log-normal distribution for which the mean and variance of the underlying normal distribution are 0 and 1. Given the values of (G, E), we generated a binary disease outcome D from the logistic regression model $\text{logit}\{\text{pr}(D|G, E)\} = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \times E$, with $(\beta_G, \beta_E, \beta_{GE}) = (0{\cdot}26, 0{\cdot}10, 0{\cdot}3)$. We choose the intercept parameter β_0 to be, respectively, $-3{\cdot}2$ and $-3{\cdot}45$ for scenarios (a) and (b) so that in both cases the marginal probability of the disease in the population is 0·05. The parameter values were chosen to reflect modest main effects for both G and E, but strong interaction between G and E. For example, the odds ratio associated with the lower versus the upper quartile of the distribution of E was 1·3 for $G=0$ and 3·1 for $G=1$. The marginal odds ratios for G and E were 2·6 and 2·5, respectively. In each replication of our simulation experiment, we generated data for 500 cases and 500 controls from the above model by sampling the cases and controls from a larger random sample of subjects. We analyse each such case-control dataset using three procedures: standard logistic regression; SPMLE$_1$, which denotes the proposed semiparametric maximum likelihood method under the gene-environment independence model when $\text{pr}(D=1)$ is known; and SPMLE$_2$, which denotes the same procedure but with $\text{pr}(D=1)$ unknown.

Table 1 summarises the simulation results for scenarios (a) and (b). Based on these simulation results we make the following key observations. First, as expected from theory, both the logistic regression and the semiparametric maximum likelihood estimators under the correct conditional independence assumption provide essentially unbiased estimators of all regression parameters. Secondly, the variance ratios of the semiparametric maximum likelihood and logistic regression estimator show that when the gene-environment independence assumption is exploited there is a major efficiency gain for the estimation of β_G and β_{GE}; the gain is quite dramatic for estimation of the interaction parameter β_{GE} and is larger for the study of the rare mutation than for the common mutation. Thirdly, under the gene-environment independence model, incorporating the known $\text{pr}(D=1)$ in the estimation leads to major efficiency gains in the estimation of the regression parameters, the gain being particularly striking for β_{GE}. This observation is particularly interesting given that it is well known that in the standard logistic regression setting, when no assumption is made about the exposure distribution, use of the known marginal probability of the disease in the population only identifies the intercept parameter of the logistic regression model, but does not have any effect on the efficiency of the estimators of the other regression parameters of interest. Fourthly, comparison of the empirical standard errors and the means of the estimated standard errors of the semiparametric maximum likelihood estimator shows that the proposed sandwich variance estimator performs well for realistic parameter values and modest sample sizes.

Table 1. *Simulation study for studying bias and efficiency of semiparametric maximum-likelihood estimators when G and E are independent*: SPMLE$_1$, *the proposed method when the marginal probability* $\text{pr}(D=1)$ *is known*; SPMLE$_2$, *the proposed method when* $\text{pr}(D=1)$ *is unknown*

	Bias			Var ratio		Empirical SE		Estimated SE	
	Logistic regres.	SPMLE$_1$	SPMLE$_2$	$\frac{\text{SPMLE}_1}{\text{Logistic}}$	$\frac{\text{SPMLE}_2}{\text{Logistic}}$	SPMLE$_1$	SPMLE$_2$	SPMLE$_1$	SPMLE$_2$
				Scenario (a): $\text{pr}(G=1)=0·05$					
β_G	0·033	0·021	0·034	0·629	0·818	0·282	0·322	0·275	0·327
β_E	−0·002	0·000	0·002	0·900	0·991	0·035	0·037	0·034	0·035
β_{GE}	−0·032	−0·009	−0·023	0·264	0·535	0·090	0·128	0·087	0·126
θ	·	0·020	0·021	·	·	0·164	0·187	0·167	0·194
				Scenario (b): $\text{pr}(G=1)=0·2$					
β_G	0·016	0·004	0·015	0·709	0·905	0·175	0·198	0·171	0·195
β_E	0·001	0·002	0·001	0·769	0·987	0·038	0·043	0·037	0·041
β_{GE}	−0·013	−0·006	−0·011	0·360	0·717	0·053	0·075	0·052	0·074
θ	·	0·003	−0·001	·	·	0·092	0·105	0·095	0·108

regres., regression; Var, variance; SE, standard error.

The above simulation set-up also allows us to study bias in parameter estimation in existing approximate methods that rely on the rare-disease assumption. Schmidt & Schaid (1999) noted that, even for rare diseases like breast cancer, the 'case-only' analysis approach to interaction that is based on the rare disease assumption can seriously underestimate the logistic regression interaction parameters for studying major susceptibility genes such as the BRCA1 and BRCA2 genes, which are known to confer a very high risk of breast and ovarian cancer. Our simulation gives an alternative relevant scenario involving a continuous environmental exposure variable where the gene or the environmental exposures themselves do not pose a very high risk of the disease, but among the mutation carriers there is a strong dose-response relationship between the risk of the disease and the continuous exposure. We examined the bias in estimation of the interaction parameter β_{GE} in two approximate methods, the case-only analysis (Piegorsch et al., 1994) and the combined control group approach of Modan et al. (2001). We did not implement the log-linear model approach for categorical covariates (Umbach & Weinberg, 1997) as in our simulation the environmental covariate E was continuous. We found that on average the case-only estimates of β_{GX} were 0·189 and 0·212 in scenarios (a) and (b), respectively. The corresponding average estimates obtained from the approach of Modan et al. are 0·194 and 0·229, respectively. Given that the true value of the interaction parameter was 0·30, in each of the scenarios we considered, the approximate methods seriously underestimate the odds-ratio interaction parameter.

4·2. *Factors G and E are independent conditional on S*

We considered a second simulation experiment in which the independence assumption between G and E holds only within subpopulations defined by a stratum variable S. As before, we considered two scenarios, one for a rare mutation and one for a common mutation, but in each situation we now assume that the gene frequency differs across strata defined by S: we took $\theta_1 = \text{pr}(G = 1|S = 1) = 0{\cdot}05$ and $\theta_2 = \text{pr}(G = 1|S = 2) = 0{\cdot}1$ in scenario (a) and $\theta_1 = \text{pr}(G = 1|S = 1) = 0{\cdot}2$ and $\theta_2 = \text{pr}(G = 1|S = 2) = 0{\cdot}4$ in scenario (b). We assumed that $\text{pr}(S = 2) = 0{\cdot}3$. Also, as before, we generated the environmental covariate as $E = \min(10, X)$, where X follows a log-normal distribution, but we allowed the mean parameter for the underlying normal distribution to be different across strata defined by S. In particular, we used the values of $\mu_1 = 0$, $\mu_2 = 0{\cdot}67$ and $\sigma_1 = \sigma_2 = 1$ so that the 75th percentile of the distribution of $X|S = 1$ corresponds to only the 50th percentile of the distribution of $X|S = 2$. We also assumed that the stratification variable S is a risk factor for the disease and hence is part of the risk model. We allowed both a main effect, β_S, and an interaction of S with G, β_{GS}, in the disease risk model with the true parameter values being log(2) and log(3), respectively. As before, we assumed $(\beta_G, \beta_X, \beta_{GX}) = (0{\cdot}26, 0{\cdot}1, 0{\cdot}3)$. We generated 500 simulated datasets, each dataset consisting of observations on (G, X, S) for 500 cases and 500 controls. We analysed each such case-control dataset using three procedures: standard logistic regression; SPMLE(CS), which denotes the proposed method under the correctly specified independence model that assumes G is independent of E given S; and SPMLE(MS), based on a misspecified independence model that assumes of G is independent of (E, S). For both of the latter two procedures, we assumed $\text{pr}(D)$ was known.

The results in Table 2 stimulate the following key observations. First, when the correct model is that G and E are independent given S, but we assume the misspecified model in which G is independent of both E and S, estimators of β_G, β_S and β_{GS} can be seriously

Table 2. *Simulation study for studying bias and efficiency of semiparametric maximum-likelihood estimators when G and E are independent conditional on a stratification variable S*: SPMLE(CS), *our method when the probability model for G and E given S is correctly specified*; SPMLE(MS), *our method when this model is misspecified*

	Bias			MSE ratio		Empirical SE		Estimated SE	
	Logistic	SPMLE	SPMLE	SPMLE(CS)	SPMLE(MS)	SPMLE	SPMLE	SPMLE	SPMLE
	regres.	(CS)	(MS)	Logistic	Logistic	(CS)	(MS)	(CS)	(MS)
			$\mathrm{pr}(G=1\mid S=1)=0{\cdot}05$ and $\mathrm{pr}(G=1\mid S=2)=0{\cdot}1$						
β_G	0·036	0·016	0·518	0·680	1·605	0·397	0·323	0·393	0·338
β_E	−0·002	−0·001	0·008	0·827	0·847	0·030	0·029	0·029	0·029
β_S	−0·008	−0·009	0·078	0·981	1·195	0·153	0·150	0·154	0·150
β_{GE}	−0·044	−0·021	−0·026	0·273	0·242	0·088	0·081	0·088	0·085
β_{GS}	−0·005	0·018	−0·940	0·879	3·386	0·508	0·338	0·481	0·343
			$\mathrm{pr}(G=1\mid S=1)=0{\cdot}2$ and $\mathrm{pr}(G=1\mid S=2)=0{\cdot}4$						
β_G	0·014	0·017	0·473	0·874	3·292	0·278	0·263	0·269	0·252
β_E	−0·002	0·001	0·016	0·736	0·865	0·037	0·036	0·039	0·038
β_S	0·003	0·003	0·342	0·976	3·235	0·221	0·213	0·222	0·211
β_{GE}	−0·007	−0·007	−0·025	0·472	0·569	0·054	0·054	0·054	0·056
β_{GS}	−0·025	−0·025	−1·101	0·953	11·468	0·326	0·273	0·337	0·269

regres., regression; MSE, mean squared error; SE, standard error.

biased, with the bias of the interaction parameter being the most striking. Secondly, the ratio of the mean squared error for SPMLE(CS) and for the logistic regression analysis shows that when the correct conditional independence model was exploited there was a major efficiency gain in estimating β_G, β_E and β_{GE}, the gain being most dramatic for estimation of β_{GE}. The corresponding ratio of the mean squared errors for SPMLE(MS) shows that for those parameters, where SPMLE(MS) produces large bias, the mean squared error for SPMLE(MS) tends to be much larger than that for the logistic regression analysis. For the parameters β_E and β_{GE}, however, where there is very little bias in SPMLE(MS), both SPMLE(MS) and SPMLE(CS) have similar mean squared errors. Thus, if we adjust properly for the stratification variable S in the independence model, we can correct for bias in estimating β_G, β_S and β_{GS} and yet can retain the efficiency advantage resulting from the gene-environment independence assumption. Thirdly, comparison of the empirical standard errors and the means of the estimated standard errors of the semiparametric maximum likelihood estimators shows that the proposed variance estimator performs well under the population-stratification model.

5. Israeli ovarian cancer study

In this section, we apply the proposed methodology to data from a population-based case-control study based on all ovarian cancer patients identified in Israel between 1 March 1994 and 30 June 1999 (Modan et al., 2001). For each case, two controls were selected from the central population registry matched by age within two years, area of birth and place and length of residence. Blood samples were then collected from the cases and the controls in order to test for the presence of mutation in the two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. In addition, the subjects were interviewed to collect data on reproductive/gynaecological history such as parity, number of years of oral contraceptive use and gynaecological surgery. The main goal of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynaecological risk factors of ovarian cancer.

Modan et al. (2001) studied the interaction between BRCA1/2 mutations and two known reproductive risk factors for ovarian cancer, namely oral contraceptive use and parity. They pointed out that, since BRCA1/2 mutations were very rare among ovarian cancer controls, traditional logistic regression analysis would yield very imprecise estimates of the various regression parameters of interest. Thus they considered alternative efficient methods of analysis that exploit the likely scenario that the status of BRCA1/2 mutations is independent of the reproductive risk factors. In particular, they estimated the odds ratio of ovarian cancer associated with the reproductive risk factors separately for carriers and non-carriers by using the combined common control group approach that we described in § 1. In addition, to test if the effects of the reproductive risk factors are different for BRCA1/2 mutation carriers and non-carriers, the authors performed the 'case-only' analysis of interaction (Piegorsch et al., 1994).

We reanalysed the data using the proposed maximum likelihood method under the gene-environment independence assumption. Our analysis included 832 cases and 747 controls who did not have bilateral oophorectomy, were interviewed for risk factor information and successfully tested for BRCA1/2 mutations. There were 240 carriers, but only 12 among the controls. Similarly to Modan et al., we coded reported parity values greater than 10 to be 10. In addition, we deleted three women with extreme oral contraceptive use, of at least 250 months, as they became highly 'influential' for the estimation of regression parameters. We considered the following logistic regression model for risk of ovarian cancer:

$$\begin{aligned}\text{logit}\{\text{pr}(D=1)\} &= \beta_0 + \beta_{\text{BRCA1/2}} I(\text{BRCA1/2}) + \beta_{\text{OC}}\text{OC} + \beta_{\text{Par}}\text{Parity} \\ &\quad + \beta_{\text{BRCA1/2*OC}} I(\text{BRCA1/2})*\text{OC} \\ &\quad + \beta_{\text{BRCA1/2*Par}} I(\text{BRCA1/2})*\text{Parity} + \gamma^{\text{T}} Z,\end{aligned}$$

where $I(\text{BRCA1/2})$ denotes the 0–1 indicator of carrying at least one BRCA1/2 mutation, OC denotes years of oral contraceptive use, Parity denotes the number of children and Z denotes the set of all co-factors that Modan et al. used to adjust their regression analysis; Z included the main effects of age, as a categorical variable defined by decades, ethnic background, being Ashkenazi or non-Ashkenazi, the presence of personal history of breast cancer, PHB, history of gynaecological surgery, and family history of breast or ovarian cancer, FHBO, where 0 corresponds to no history in the family, 1 to one breast cancer case in the family and 2 to ovarian cancer or two or more breast cancer cases in the family.

Next we considered an appropriate model for gene-environment independence. Clearly, a personal history of breast cancer and family history of breast/ovarian cancer cannot be assumed to be independent of BRCA1/2 status as mutations in these genes are known to increase dramatically the risk of these familial cancers. Moreover, BRCA1/2 mutation frequency has been reported in the past to vary by age and ethnicity. Given that some of these factors can also be related to oral contraceptive use and parity, we make the assumption of independence between mutation and reproductive risk factors only conditional on $S = (\text{Age}, \text{Ethnicity}, \text{PHB}, \text{FHBO})$. Given that the total number of strata defined by S is large, estimation of the genotype frequencies individually for each stratum would be imprecise. Thus, we considered the following parametric model for specification of the carrier frequencies:

$$\text{logit}\{\text{pr}(G=1|S)\} = \theta_0 + \theta_{\text{Age}} I(\text{Age} \geqslant 50) + \theta_{\text{Eth}} I(\text{Non-Ashkenazi}) \\ + \theta_{\text{PH}} I(\text{PHB}=1) + \theta_{1\text{FH}} I(\text{FHBO}=1) + \theta_{2\text{FH}} I(\text{FHBO}=2). \quad (10)$$

Modan et al. had reported a total of 1326 cases of peritoneal or epithelial ovarian cancer during the five-year study period, in a baseline population of approximately 1·5 million. Thus, the marginal probability for the disease for the underlying population is small, at about $\text{pr}(D=1) = 8{\cdot}7 \times 10^{-4}$. Therefore, based on the discussion in § 3·2, we note that we can analyse data from this age-matched case-control study using methods developed for ordinary case-control studies as long as we allow for an independent intercept term for each of the age-strata that were used for matching. Although the cases and controls were matched to within two years, to avoid problems with sparse cells we allowed an independent intercept term only for every 10-year interval. This approximation, the validity of which requires assumptions similar to those required for unconditional logistic regression analysis of matched data, is reasonable for this study.

Table 3 shows the estimates and 95% confidence intervals corresponding to the regression parameters associated with the main covariates of interest: BRCA1/2, oral contraceptive use and parity. Two sets of estimates and confidence intervals are shown, one corresponding to an ordinary logistic regression analysis of the case-control data and the other corresponding to our method estimated under the conditional gene-environment independence model. Based on the ordinary logistic regression estimates of the main effect parameters, we first observe that, among childless women, for whom Parity = 0, and who never used oral contraceptives, BRCA1/2 mutation is associated with a dramatic increase in risk of ovarian cancer, with odds ratio exp(3·58) = 35·87. Among BRCA1/2 non-carriers, both higher parity and longer use of oral contraceptives are associated with decreased risk of ovarian cancer, with the associated odds ratio parameters estimated to be respectively 0·95 and 0·94 for Parity; both of these results are borderline statistically significant at the 5% level. The estimates of the interaction parameters from the logistic regression analysis suggest that, among BRCA1/2 carriers, the risk of ovarian cancer decreases even more strongly with increasing parity, with odds ratio $\exp(-0{\cdot}058) \times \exp(-0{\cdot}199) = 0{\cdot}77$, but increases slightly with longer oral contraceptive use, with odds ratio $\exp(-0{\cdot}047) \times \exp(0{\cdot}056) = 1{\cdot}01$. However, the confidence intervals for the interaction parameters are very wide, suggesting that the point estimates are imprecise and hence hard to interpret.

Table 3. *Parameter estimates and confidence intervals for the risk model in the Israeli ovarian cancer study*

	Ordinary logistic regression		MLE with G-E independence given S	
	Estimate	95% CI	Estimate	95% CI
BRCA1/2	3·58	(2·27, 4·89)	3·15	(2·51, 3·79)
OC use	−0·047	(−0·098, 0·003)	−0·051	(−0·102, −0·001)
Parity	−0·058	(−0·121, 0·004)	−0·061	(−0·125, 0·002)
OC*BRCA1/2	0·056	(−0·149, 0·260)	0·089	(0·021, 0·150)
Parity*BRCA1/2	−0·199	(−0·626, 0·229)	−0·036	(−0·141, 0·068)

MLE, maximum likelihood estimate; G-E, gene-environment; CI, confidence interval; S = (Age, Ethnicity, Personal history of breast cancer, Family history of breast/ovarian cancer); OC, oral contraceptive.

Inspection of the parameter estimates from the semiparametric maximum likelihood method with the gene-environment independence model suggests similar types of association to those from the logistic-regression analysis. However, the precisions of the estimates are greater for all the terms involving BRCA1/2, the gain being particularly striking for the interaction terms. In particular, under the gene-environment independence model, the interaction between BRCA1/2 mutation and oral contraceptive use is statistically significant, suggesting that, unlike for non-carriers, the risk of breast cancer for carriers did not decrease with increasing oral contraceptive use. For carriers, the association between oral contraceptive use and risk of ovarian cancer, if any, is positive, with odds ratio $\exp(-0{\cdot}051 + 0{\cdot}089) = 1{\cdot}034$, and 95% confidence interval (0·977, 1·095). The interaction estimate between Parity and BRCA1/2 suggests that the decrease in risk of ovarian cancer associated with increased parity is modestly larger for carriers than for non-carriers, but this difference is not statistically significant.

Table 4 shows the maximum likelihood estimates corresponding to the model for carrier frequency $\mathrm{pr}(G = 1|S)$. Although these parameters do not have any causal interpretation and are not generally of biological interest, they can be useful for descriptive purposes. For example, as expected, prevalence of a BRCA1/2 mutation is significantly higher among women with either a personal history of breast cancer or family history of breast/ovarian cancer. Moreover, we observe that BRCA1/2 mutation frequency is significantly lower among non-Ashkenazi Jewish women compared to Ashkenazi women. There is also some evidence, with p-value = 0·05, that carrier frequency was smaller among women older than 50 than among younger women.

Table 4. *Parameter estimates and confidence intervals for the logistic regression model for* $\mathrm{pr}(G = 1|S)$ *in the Israeli ovarian cancer study, with risk factors ethnicity, age, personal history of breast cancer, family history of breast cancer and family history of breast/ovarian cancer*

	θ_0	θ_{Eth}	θ_{Age}	θ_{PH}	θ_{1FH}	θ_{2FH}
Estimate	−3·78	−1·31	−0·28	1·59	0·71	1·32
95% ci	(−4·40, −3·16)	(−1·74, −0·890)	(−0·638, 0·071)	(1·01, 2·18)	(0·20, 1·21)	(0·74, 1·90)

ci, confidence interval.

A version of the dataset is available together with the software from the website http://dceg.cancer.gov/people/ChatterjeeNilanjan.html under the software link. This dataset consists of the real data on disease status, D, and non-genetic co-factors, X. For reasons of privacy, however, the real genetic data are not publicly available. Instead, the data consist of simulated genetic data, G, generated using the conditional distribution of $[G|D, X]$ as specified by the parameter estimates obtained from the real data.

6. Discussion

Case-control studies with modest sample sizes often have very little power for studying interaction and other hypotheses of interest using the standard logistic regression analysis. In such situations, epidemiological researchers currently have been prone to exploit the efficiency advantage from the gene-environment independence assumption through the case-only approach that yields estimate of the multiplicative interaction parameter in the logistic regression model under the rare disease assumption (Piegorsch et al., 1994). This analysis, however, is limited. It discards all the information from controls and hence loses the ability to estimate the main effect parameters of the logistic regression model which are required for deriving the various alternative scientific parameters of interest. In this paper, we have considered estimation of regression parameters under the gene-environment independence assumption in a very general logistic regression model that uses data from both cases and controls and hence can estimate all of the parameters of interest.

However, we recommend cautious use of the gene-environment independence assumption. Simulation studies reported in § 4·2, as well as those in Albert et al. (2001), show that methods that use the gene-environment independence assumption when the assumption is not true may produce severe bias in parameter estimation. We have proposed a possible remedy for minimising such bias by explicitly accounting for observable factors, denoted by S, that can potentially be related to both G and E.

Methods for exploiting the gene-environment independence assumption could be practically useful without concerns about bias in many important situations. For 'randomised exposure' such as the treatment assigned in a randomised trial, the gene-environment independence assumption would be satisfied by the definition of randomisation. The assumption of gene-environment independence is also very likely to be satisfied for external environmental agents, e.g. carcinogens from a nearby chemical factory, exposure to which is not directly controlled by an individual's own behaviour. When an exposure depends on subject's individual behaviour, on the other hand, the independence assumption should be used more cautiously. There could be spurious association between G and E for established risk factors such as smoking because family history of lung cancer, which is associated with G, may also influence a subject to change his/her smoking behaviour. There could also be direct association. For example, genetic polymorphisms in the smoking metabolism pathway may not only modify a subject's risk from smoking, but may also influence a subject's degree of addiction to smoking.

When violation of the gene-environment independence seems plausible, because of direct or indirect association, effort should be made to validate the assumption empirically. However, tests for independence within a given study may have very little power, and empirical evidence from external data sources should be investigated. When substantial uncertainty remains about the validity of the assumption because of lack of empirical data or for other reasons, positive findings based on proposed methodology should be considered as preliminary screen which should be pursued with high 'priority' in future epidemiological studies.

In practice, genetic and/or environmental exposure data can be also missing on certain study subjects, by design or by change. Umbach & Weinberg (1997) described a number of alternative designs in which genetic and/or environmental exposure data are collected only on a subset of controls. They showed how different parameters of interest can be estimated under different designs using the approximate log-linear model approach for categorical variables. Further research is warranted to extend the proposed maximum-likelihood methodology to handle missing data in genetic as well as environmental exposures. Such extensions will also be useful to haplotype-based associated studies where genetic effects are modelled in terms of 'haplotypes', the combinaton of alleles at multiple loci in a single chromosome, but the exact haplotype configuration in two chromosomes of some subjects cannot be derived with certainty from available locus-specific genotype data.

Acknowledgement

Carroll's research was supported by a grant from the National Cancer Institute, and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences. We would like to thank Dr Sholom Wacholder for sharing his valuable insights about gene-environment studies.

Appendix

Proofs

Proof of Lemma 1. By Lemma 1 of Roeder et al. (1996), it follows that the probability equality of our Lemma 1 holds only if

$$d\mathscr{H}^*(G, E) = \frac{[1+\exp\{\beta_0^* + m(G, E; \beta_1)\}]/[1+\exp\{\beta_0 + m(G, E; \beta_1)\}]d\mathscr{H}(G, E)}{\sum_g \int_e [1+\exp\{\beta_0^* + m(g, e; \beta_1)\}]/[1+\exp\{\beta_0 + m(g, e; \beta_1)\}]d\mathscr{H}(g, e)}.$$

If $\mathscr{H}$ is of the product form $Q \times F$, $\mathscr{H}^* \neq \mathscr{H}$ could be of the product form only if $\beta_1 \in \mathscr{B}^0$. Thus, if $\beta_1 \notin \mathscr{B}^0$, then $F = F^*$ and $Q = Q^*$. Moreover, since $\mathscr{H}^* = \mathscr{H}$, it also follows that $\beta_0 = \beta^*$. □

Proof of Lemma 2. By equating the partial derivatives of the loglikelihood given in equation (1) with respect to $\delta_1, \ldots, \delta_K$, we can easily show that $\hat{\delta}_k(\gamma)$ will satisfy the equation

$$\delta_k = \frac{n_{++k}}{\sum_{ij} P_{ij}(e_k, \beta)\hat{\mu}_i q_j(\theta)}, \tag{A1}$$

where

$$\hat{\mu}_i = \frac{n_{i++}}{\text{pr}(D=i)} = \frac{n_{i++}}{\sum_{j'k'} P_{ij'}(e_{k'}, \beta) q_{j'}(\theta)\delta_{k'}}. \tag{A2}$$

If we now substitute the left-hand side of (A1) for δ_k into the loglikelihood of the data defined in (1), we obtain

$$\begin{aligned} L\{\gamma, \hat{\delta}(\gamma)\} &= \sum_{ijk} n_{ijk} \log P_{ij}(e_k, \beta_0, \beta_1) + \sum_j n_{+j+} \log q_j(\theta) \\ &\quad + \sum_k n_{+k+} \log \frac{n_{++k}}{\sum_{ij'} P_{ij'}(e_k; \beta_0, \beta_1)\hat{\mu}_i(q_{j'}(\theta)} - \sum_i n_{i++} \log \frac{n_{i++}}{\hat{\mu}_i(\gamma)}, \end{aligned}$$

which is equivalent to $L^*\{\gamma, \hat{\mu}(\gamma)\}$ up to constant terms. Moreover, if we substitute (A1) into (A2) it can be seen that $\hat{\mu}_i(\gamma)$, for $i = 0, 1$, are given by solutions of the equations

$$n_{i++} = \sum_{k'} n_{++k'} \frac{\sum_{j'} P_{ij'}(e_{k'}; \beta) q_j(\theta)\mu_i}{\sum_{ij} P_{ij}(e_{k'}; \beta) q_j(\theta)\mu_i} \quad (i = 0, 1),$$

which are in turn equivalent to the equations given in (4). Thus Lemma 2 is proved. □

Proof of Lemma 3. First, we note that, by the law of large numbers,

$$\begin{aligned} \frac{1}{N}\sum_{u=1}^{N} Q(D_u, G_u, E_u) &\equiv \frac{N_0}{N}\frac{1}{N_0}\sum_{u=1}^{N_0} Q(D_u, G_u, E_u) + \frac{N_1}{N}\frac{1}{N_1}\sum_{u=1}^{N_1} Q(D_u, G_u, E_u) \\ &= \mu_0 \,\text{pr}(D=0) E\{Q(D, G, E)|D=0\} \\ &\quad + \mu_1 \,\text{pr}(D=1) E\{Q(D, G, E)|D=1\} + o_p(1). \end{aligned} \tag{A3}$$

Using Bayes' rule we can write

$$\text{pr}(D=d) E\{Q(D, G, E)|D=d\} = \int \left[\sum_j \{Q(d, g_j, e) q_j(\theta)\} \right] dF(e).$$

Thus, we can write the limiting expression in (A3) as

$$\int_e \left\{ \sum_{ij} \frac{Q(i, g_j, e) P_{ij}(e; \beta_0, \beta_1) \mu_i q_j(\theta)}{h(e)} \right\} h(e) dF(e).$$

The proof of Lemma 3 follows if we note that $P^*_{ij}(e; \gamma, \mu) = \mu_i P_{ij}(e; \beta_0, \beta_1) q_j(\theta)/h(e)$. □

Proof of Lemma 4. By applying the chain rule of derivatives to formula (7) we have

$$\frac{1}{N} \frac{\partial^2 L^*}{\partial \eta\, \partial \eta'} = \sum_{u=1}^{N} \left[\frac{\partial^2 \theta_{D_u G_u}(E_u, \eta)}{\partial \eta\, \partial \eta'} - E^*_{DG} \left\{ \frac{\partial^2 \theta_{DG}(E, \eta)}{\partial \eta\, \partial \eta'} \middle| E = E_u \right\} \right]$$

$$- \sum_{u=1}^{N} \sum_{ij} \frac{\partial \theta_{ij}(E_u, \eta)}{\partial \eta'} \frac{\partial}{\partial \eta} P^*_{ij}(E_u; \eta).$$

Using Lemma 3, we can now show that the first term in the above expression goes to zero in probability. Furthermore, with some algebra it can be shown that

$$\sum_{ij} \frac{\partial \theta_{ij}(E_u, \eta)}{\partial \eta} \frac{\partial}{\partial \eta} P^*_{ij}(E_u; \eta) = V^* \left\{ \frac{\partial \theta_{DG}(E_u, \eta)}{\partial \eta} \middle| E = E_u \right\}.$$

The proof of Lemma 4 now easily follows from the result of Lemma 3. □

Proof of Proposition 1. (i) The main condition for consistency, that is the asymptotic unbiasedness of the score equation $\sum_{i=1}^{N} \Psi(D_i, G_i, E_i; \eta) = 0$, follows from direct application of Lemma 3. In Lemma 4, we have further shown that $-\partial/\partial\eta \{N^{-1} \sum_{i=1}^{N} \Psi(D_i, G_i, E_i; \eta)\} \to \mathscr{I}$ in probability, where $\mathscr{I}$ is a positive definite matrix. Moreover, from (6) it is easy to see that the first and second derivatives of $\theta_{ij}(E; \eta)$ with respect to η can be uniformly bounded in an open neighbourhood of η_0. This can be used to show that the convergence in Lemma 4 holds uniformly in an open neighbourhood of η_0. The proof now follows using results of Foutz (1977).

(ii) The asymptotic normality of the estimator follows from standard application of the central limit theorem. To derive the form of the asymptotic variance, we need to prove that

$$\Gamma \equiv \operatorname{cov} N^{-1/2} \sum_{u=1}^{N} \Psi(D_i, G_i, E_i; \eta) = \mathscr{I} - \Omega. \tag{A4}$$

Let $\Phi(D; \eta) = E\{\Psi(D, G, E; \eta) | D\}$ and $\tilde{\Psi}(D, G, E; \eta) = \Psi(D, G, E; \eta) - \Phi(D; \eta)$. We can now write

$$\Gamma = \sum_d \frac{N_d}{N} \operatorname{cov}\{\Psi(D, G, E; \eta) | D = d\} = \sum_d \frac{N_d}{N} E\{\tilde{\Psi}^{\otimes 2}(D, G, E; \eta) | D = d\}$$

$$= \sum_d \mu_d \sum_j \int \tilde{\Psi}^{\otimes 2}(D, G, E; \eta)\, \mathrm{pr}(D = d | G = g_j, E = e) q_j dF(e).$$

By reordering the sums and the integral in the last expression we can easily show that

$$\Gamma = \int E^*\{\tilde{\Psi}^{\otimes 2}(D, G, E; \eta) | E = e\} h(e) dF(e).$$

Since

$$\tilde{\Psi}^{\otimes 2}(D, G, E; \eta) = \Psi^{\otimes 2}(D, G, E; \eta) + \Phi^{\otimes 2}(D; \eta) - 2\Psi(D, G, E; \eta)\Phi(d; \eta)^{\mathrm{T}}$$

and $E^*\{\Psi(D, G, E; \eta)^{\otimes 2} | E = e\} = V^*\{\Psi(D, G, E; \eta) | E = e\}$, the proof of formula (A4) will follow if we can show that

$$\sum_d \mu_d \Phi(d; \eta)^{\otimes 2} = \int E^*\{\Psi(D, G, E; \eta)\Phi(D; \eta)^{\mathrm{T}} | E = e\} h(e) dF(e), \tag{A5}$$

$$\sum_d \mu_d \Phi(d; \eta)^{\otimes 2} = \int E^*\{\Phi(D; \eta)^{\otimes 2} | E = e\} h(e) dF(e). \tag{A6}$$

To prove (A5), we first define

$$W(D, E; \eta) = E\{\Psi(D, G, E; \eta) | D, E\} = E^*\{\Psi(D, G, E; \eta) | D, E\}$$

and note that $E\{W(D, E; \eta) | D\} = \Phi(D; \eta)$. It is easily seen that the right-hand side of (A5) can be written as

$$\int E_D^*\{\Phi(D; \eta) W(D, e; \eta) | E = e\} h(e) dF(e). \tag{A7}$$

Now we observe that $\mathrm{pr}^*(D|E) = \mathrm{pr}(D|E)\mu_D / h(E)$ and write (A7) as

$$\begin{aligned}\int \sum_D \mu_D\, \mathrm{pr}(D|E = e)\Phi(D; \eta) W(D, e; \eta) dF(e) &= \sum_D \mu_D\, \mathrm{pr}(D)\Phi(D; \eta) \int \frac{W(D, e; \eta)\, \mathrm{pr}(D|E = e) dF(e)}{\mathrm{pr}(D)} \\ &= \sum_D \mu_D \Phi^{\otimes 2}(D; \eta).\end{aligned}$$

This proves (A5). The proof of (A6) follows from similar steps. □

References

Albert, P. S., Ratnasinghe, D., Tangrea, J. & Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interaction. *Am. J. Epidemiol.* **154**, 687–93.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. Statist. Soc.* B **32**, 283–301.

Breslow, N. E., Robins, J. M. & Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**, 447–55.

Chatterjee, N., Kalaylioglu, Z. & Carroll, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, interactions and joint effects. *Genet. Epidemiol.* **28**, 138–56.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, **4**, Ed. J. Neyman, pp. 135–48. Berkeley, CA: University of California Press.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *J. Am. Statist. Assoc.* **72**, 147–9.

Khouri, M. J., Beaty, T. H. & Cohen, B. H. (1993). *Fundamentals of Genetic Epidemiology.* Oxford University Press.

Modan, M. D., Hartge, P. et al. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New Engl. J. Med.* **345**, 235–40.

Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statist. Med.* **13**, 153–62.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Rabinowitz, D. (1997). A note on efficient estimation from case-control data. *Biometrika* **84**, 486–8.

Roeder, K., Carroll, R. J. & Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *J. Am. Statist. Assoc.* **91**, 722–32.

Schmidt, S. & Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *Am. J. Epidemiol.* **150**, 878–85.

Scott, A. J. & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.

Umbach, D. M. & Weinberg, C. M. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statist. Med.* **16**, 1731–43.

[*Received September* 2003. *Revised October* 2004]

Genetic Epidemiology 28: 138–156 (2005)

Exploiting Gene-Environment Independence in Family-Based Case-Control Studies: Increased Power for Detecting Associations, Interactions and Joint Effects

Nilanjan Chatterjee,[1]* Zeynep Kalaylioglu,[2] and Raymond J. Carroll[3]

[1]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Rockville, Maryland*
[2]*Information Management System, Rockville, Maryland*
[3]*Texas A&M University, College Station, Texas*

Family-based case-control studies are popularly used to study the effect of genes and gene-environment interactions in the etiology of rare complex diseases. We consider methods for the analysis of such studies under the assumption that genetic susceptibility (*G*) and environmental exposures (*E*) are independently distributed of each other within families in the source population. Conditional logistic regression, the traditional method of analysis of the data, fails to exploit the independence assumption and hence can be inefficient. Alternatively, one can estimate the multiplicative interaction between *G* and *E* more efficiently using cases only, but the required population-based *G-E* independence assumption is very stringent. In this article, we propose a novel conditional likelihood framework for exploiting the within-family *G-E* independence assumption. This approach leads to a simple and yet highly efficient method of estimating interaction and various other risk parameters of scientific interest. Moreover, we show that the same paradigm also leads to a number of alternative and even more efficient methods for analysis of family-based case-control studies when parental genotype information is available on the case-control study participants. Based on these methods, we evaluate different family-based study designs by examining their relative efficiencies to each other and their efficiencies compared to a population-based case-control design of unrelated subjects. These comparisons reveal important design implications. Extensions of the methodologies for dealing with complex family studies are also discussed. *Genet. Epidemiol.* 28:138–156, 2005. © 2004 Wiley-Liss, Inc.

Key words: case-control studies; case-only designs; conditional likelihood; conditional logistic regression; family-based studies; gene-environment independence; gene-environment interactions; genetic epidemiology; matched case-control studies; population stratification.

Grant sponsor: National Cancer Institute; Grant number: CA-57030; Grant sponsor: National Institute of Environmental Health Sciences; Grant number: P30-ES09106.

*Correspondence to: N. Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, 6120 Executive Blvd, EPS Room 8038, Rockville, MD 20852. E-mail: chattern@mail.nih.gov

Received 11 June 2004; Accepted 7 July 2004
Published online 8 December 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20049

INTRODUCTION

The risks of many complex diseases are determined by a combination of the effects of genetic susceptibilities and environmental exposures. Advances in human genome research have thus led to epidemiologic investigations not only of the effects of genes alone, but also of their effects in combination with environmental exposures. The risk of a disease in relation to two exposures can be studied at various scales including statistical interaction (multiplicative or additive), joint effects, and the subgroup effect of one exposure within strata defined by the other exposures. Although the utility of a specific risk parameter depends on the scientific context and is sometimes a matter of debate, collectively various risk parameters of interest involving genetic and environmental exposures can be important for understanding biological and public health effects of the exposures, for targeting high-risk subjects for intervention, for individual risk prediction, and for enhancing the power to detect the association of the disease with one exposure (e.g., a gene) by selecting subjects to study based on the other exposure (e.g., the environment). Studies of genetic and environmental exposures together, whether designed to evaluate statistical interactions or other parameters of scientific interest, can be cost-prohibitive due to the typical requirement of large sample sizes to achieve reasonable statistical power. Thus, efficient designs as well as efficient analytic methods that can reduce required sample size have become an important area of genetic-epidemiologic research.

Case-control study designs, in which diseased (cases) and non-diseased subjects (controls) are compared with respect to their exposure history, are now increasingly being used to study the role of genes and gene-environment interactions in the etiology of rare diseases. In population-based case-control designs, cases and controls are randomly selected from the diseased and non-diseased subjects that arise in an underlying population. Typically, the cases and controls in such designs are unrelated. In contrast, in family-based case-control designs, controls are selected from families of the cases. Both population- and family-based designs have advantages and disadvantages [Witte et al., 1999; Gauderman et al., 1999; Weinberg and Umbach, 2000]. While selection of population-based controls may be logistically more convenient, family-based designs can offer protection against spurious association induced by population stratification or admixture. Even when bias due to population stratification or admixture is not a concern, for efficiency reasons family-based designs may be preferred in studies of gene-environment interaction involving rare genetic variants [Witte et al., 1999; Gauderman, 2002].

The focus of this report is family-based designs where data on both genotype and environmental exposures are available on cases and related matching controls within families. Such designs may include, for example, designs where healthy siblings [Curtis, 1997; Spielman and Ewens, 1998] or cousins [Witte et al., 1999] of diseased subjects are selected as controls. Traditional analysis of such family-based studies involves conditional-logistic regression that restricts comparison of cases and controls to be within the family. The underlying theory of this approach relies on the conditional likelihood of the observed disease configuration data within matched-sets (family) given the risk factor information of the individual subjects. This method does not require any assumptions on the distribution of the risk factors in the underlying population.

In this article, we develop a new paradigm of conditional likelihoods for efficient analysis of family-based case-control studies when genetic susceptibility and environmental exposures can be assumed to be independently distributed of each other within families in the source population.

The efficiency advantage of methods that can exploit the gene-environment independence assumption was first observed in the context of population-based case-control studies. If genetic (*G*) and environmental (*E*) risk-factors can be assumed to be independently distributed in the underlying population, then the multiplicative interaction (also known as statistical interaction) between *G* and *E* for a rare disease can be estimated as the odds ratio between *G* and *E* among cases alone: the corresponding case-only estimate of interaction can be much more precise than the corresponding estimate of the interaction parameter from standard logistic regression analysis that involves both cases and controls but does not exploit the independence assumption [Piegorsch et al., 1994]. For discrete covariates, Umbach and Weinberg [1997] and then more generally Chatterjee and Carroll [2005] have developed efficient methods for estimating all of the parameters in a logistic regression model using data from both cases and controls and utilizing the *G-E* independence assumption.

The *G-E* independence assumption has been exploited also for the case-parent-trio design, an alternative family-based design that is known to be powerful for studying the effects of genes alone. The assumption was first used implicitly to show that the multiplicative interaction between *G* and *E* can be estimated from a case-parent-trio design that genotypes cases and their parents and determines environmental exposures of the cases [Schaid, 1999]. In particular, the likelihood based methods for analysis of such data rely on the assumption that conditional on parental genotypes, an individual's exposure status is independent of his/her genotype, a relatively weak independence assumption that is not affected by spurious association between genotype and exposure status in the general population that may be created due to hidden substructure [Umbach and Weinberg, 2000; Thomas, 2000].

Within the context of family-based studies, the choice of the case-control or the case-parents design for studies of gene-environment interaction depends on a number of different considerations [Weinberg and Umbach, 2000]. Besides various practical issues such as the availability of parents in the case-parent-trio design and availability of sibling/cousin for case-control designs, an important consideration is the relative efficiency of these designs for estimation of various parameters of interest. Efficiency comparisons for estimation of the multiplicative interaction parameter have revealed that either the case-control design or the case-parent design can be superior depending on whether the effect of the gene under study is dominant or recessive, respectively [Witte et al. 1999; Gauderman, 2002]. It is worth noting that in all of these previous efficiency comparisons, the conditional likelihood method that is used for analysis of case-parent designs relies on the *G-E* independence assumption, while the traditional conditional logistic regression method used for analysis of case-control design does not exploit any such assumption.

In this article, we develop an alternative conditional likelihood framework for analyzing family-based case-control studies. Our approach relies on the assumption that genetic and environmental exposures are independently distributed within families. This family-level independence assumption, similar to the *G-E* independence assumption required for the case-parent design, is relatively weak in the sense that it is less likely to be affected by spurious association between *G* and *E* in the population. We show that when the underlying independence assumption is valid, the method can lead to major efficiency gains over traditional conditional logistic regression analysis of family-based case-control studies.

We also show that our conditional likelihood framework can be used to analyze data from a novel hybrid design that obtains genotype data of parents in addition to collecting genotype and environmental exposure data on cases and family-based controls. In particular, we show that a sibling-case-control design with parental genotype information, when analyzed using our new conditional likelihood approach, can be far superior to the sibling-case-control, case-parent, or population-based case-control design for estimation of different parameters of interest, and in a wide variety of situations. Various extensions of the methods for general family-based studies are also described.

BACKGROUND OF CONDITIONAL LOGISTIC REGRESSION (CLR)

MODEL AND NOTATION

For a major part of this article, we consider designs with 1:1 case-control matching. Later, we discuss how the methodology can be extended to more general types of family studies that may involve more than one case and/or control per family. We first describe a set of model assumptions that are required for traditional CLR analysis. Let (D_1, D_2), (G_1, G_2), and (E_1, E_2) denote the 0-1 disease indicators, genotypes, and environmental exposures for a pair of relatives. We assume that within a given family *F*, the risk of disease for two relatives is conditionally independent given their covariate information, and that the prospective risk model for the disease for the j^{th} $(j = 1$ or $2)$ relative is given by the logistic regression model

$$\text{pr}(D_j = 1|G_j, E_j, F) = H\{\alpha_F + m(G_j, E_j; \beta)\}, \quad (1)$$

where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(\bullet)$ is a known but arbitrary function. Let $R_F(G_1, G_2, E_1, E_2)$ denote the joint distribution of (G_1, G_2) and (E_1, E_2) within family *F*. Standard CLR analysis allows $R_F(G_1, G_2, E_1, E_2)$ to be completely arbitrary.

There are two important features of the above model that need special attention. First, model (1) allows the family-specific intercept parameter α_F to account for potential heterogeneity in disease risk between different families. Such heterogeneity may arise, for example, if there are other sources of familial aggregation that cannot be explained by the genetic and environmental exposures under study.

Second, the model (1) allows the joint log-odds-ratio (log-relative-risk assuming rare disease) function $m(G, E, \beta)$ to be of very general form, so that it can include many different kinds of interaction models. The standard logistic model corresponds to $m(G, E, \beta) = \beta_G G + \beta_E E + \beta_{GE} G * E$, where $\exp(\beta_G)$ is the relative-risk (assuming rare disease) associated with the gene variant in the absence of the environmental exposure (the main effect of *G*), $\exp(\beta_E)$ is the relative-risk associated with the exposure in the absence of the gene-variant (the main effect of *E*), and $\exp(\beta_{GE})$ is the multiplicative interaction between *G* and *E*, which measures how the relative-risk associated with exposure changes with genotypes, or, equivalently, how the relative-risk associated with the gene variant changes with the exposure, with the changes being measured in the ratio scale. Many studies of gene and environment focus on estimation of the multiplicative interaction parameter β_{GE}, but estimation only of such statistical interaction may not necessarily contribute to an understanding of the biological or public health effects of the two exposures [Thompson, 1991; Clayton and McKeigue, 2001]. Thus, when genetic and environmental factors are being studied together, it is important to consider modelling and estimation approaches that allow flexibility of estimation of various different parameters of interest. Examples of such parameters include the joint effect of the exposures G and E, the effect of E in sub-groups of subjects with different genetic exposures, the effect of G in sub-groups of subjects with different environmental exposures, and interaction between G and E at various different scales [Khoury, et al., 1993]. We will describe all of our methodologies in the general setting of model (1), which allows testing and estimation of all the risk-parameters of interest.

We assume M relative pairs are sampled into the study so that each pair has one diseased (case) and one non-diseased (control) subject. For the i^{th} such matched set, let D_{i0}, G_{i0}, and E_{i0} denote the disease status, genotype, and environmental exposure for the control and D_{i1}, G_{i1} and E_{i1} denote the corresponding values for the cases.

TRADITIONAL CONDITIONAL LIKELIHOOD

We now describe the conditional likelihood that forms the basis for traditional CLR analysis of family-based or other types of individually matched case-control studies. In the context of our model and notation, this conditional likelihood for the i^{th} matched set is given by

$$\begin{aligned} L_{i,\text{CLR}} =& \Pr(D_{i1} = 1,\, D_{i0} = 0 | D_{i1} + D_{i0} = 1, \\ & G_{i1}, G_{i0}, E_{i1}, E_{i0}) \\ =& \frac{\exp\{m(G_{i1}, E_{i1}; \beta)\}}{\exp\{m(G_{i1}, E_{i1}; \beta)\} + \exp\{m(G_{i0}, E_{i0}; \beta)\}}. \end{aligned} \quad (2)$$

In (2), conditioning on the *set* event $D_{i1} + D_{i0} = 1$ reflects the constraint that by design the total number of cases in each matched set is exactly equal to one. For each matched set i, the conditional likelihood is formed based on the probability of the observed disease configuration for the members of the matched set, conditional on their risk factor information G and E and the ascertainment event $D_{i1} + D_{i0} = 1$. We observe that for studying the effect of genes alone using the sibling-case-control design, Spielman and Ewens [1998] previously proposed a Monte Carlo test-procedure, known as Sib-TDT (SDT), based on within-family permutation of genotypes. The CLR method can be viewed as an alternative likelihood-based analysis approach that is efficient as well as flexible in the sense that it allows both testing and estimation of risk parameters, can adjust for co-factors, and can be used to study gene-environment interaction.

There are several features of the above conditional likelihood that make the CLR analysis very flexible. First, computation of the CLR likelihood, as shown in the second line of formula (2), is free of the family-specific intercept parameters α_{F_i} and hence does not require any modelling assumptions about possible mechanisms of heterogeneity in disease risk between different families that cannot be explained by the genetic and environmental risk factors under study. Moreover, the likelihood in formula (2) is constructed based on probabilities that condition on all the risk factor information in a matched set and, hence, is free of any assumption about $R_F(G_1, G_2, E_1, E_2)$, the joint distribution of the risk factors in pairs of relatives. At the same time, however, the method, being distribution free, cannot exploit the gene-environment independence assumption when it is reasonable to do so.

PROPOSED METHODOLOGY

A NOVEL CONDITIONING PRINCIPLE

We propose to exploit an appropriate G-E independence assumption based on a conditional likelihood that does not condition on all of the genotype information of the individual subjects in a matched set. The straightforward choice for such a conditional likelihood is $\Pr(D_{i1} = 1, D_{i0} = 0, G_{i1}, G_{i0} | D_{i1} + D_{i0} = 1, E_{i1}, E_{i0})$. However, a complication with such a conditional likelihood is that its computation depends on the joint genotype frequencies for pairs of relatives. In the presence of population substructure, genotype frequencies may vary across families and estimation of the regression parameters in the presence of family-specific genotype/allele frequency parameters is not possible without a strong modelling assumption about the distribution of the gene in different families. Since a major motivation of family-based designs is to avoid making this kind of assumption, we propose an alternative conditional likelihood that does not involve the genotype-frequency parameters and yet can exploit the G-E independence assumption. The most general form of such a conditional likelihood is given by

$$L_{i,\text{general}} = \Pr(D_{i1} = 1, D_{i0} = 0, G_{i1}, G_{i0} | D_{i1} + D_{i0} = 1, \mathcal{G}_i, E_{i1}, E_{i0}), \quad (3)$$

where the conditioning event $\mathcal{G}_i$ denotes a variable that contains partial, but not all, information about the genotypes of the subjects in the matched set i, chosen in such a way that $L_{i,\text{general}}$ remains free of genotype-frequency parameters. Similar ideas for conditioning on partial genotype information to remove distributional assumption on genotypes have been previously used in the context of case-parent-trio studies of genetic association [Clayton, 1999; Cordell and Clayton, 2002; Rabinowitz and Laird, 2000]. In what follows, we will show how such an idea can be utilized for efficient analysis of gene-environment interaction in the context of case-control studies.

THE NEW CONDITIONING PARADIGM IN THE STANDARD CASE-CONTROL DESIGN

We now describe how the conditioning principle outlined above can be applied to analyze data from standard family-based case-control designs where genotype and exposure data are available only for selected cases and related matching controls. The required *G-E* independence assumption for this design is that the joint genotype and exposure status are independent for pairs of relatives within each family in the source population. Formally speaking, this assumption for a given family F implies that the joint genotype and exposure distribution $R_F(G_1, G_2, E_1, E_2)$ can be expressed as

$$R_F(G_1, G_2, E_1, E_2) = Q_F(G_1, G_2) \times V_F(E_1, E_2), \quad (4)$$

where Q_F and V_F are the family specific distributions of (G_1, G_2) and (E_1, E_2), respectively. We also assume that the joint genotype distribution for any pair of relatives is symmetric, that is $Q_F(g_1, g_2) = Q_F(g_2, g_1)$, an assumption that automatically holds under the Mendelian law of inheritance within families. Hereafter, we will describe the assumption stated in (4) as "Type-I Independence" to distinguish it from an alternative independence assumption that we will use later for case-control designs with parental genotype information.

Observe that independence assumption (4) is much weaker than the population-based *G-E* independence assumption that is required for the case-only design. The population-based independence assumption, for example, may be violated due to the effects of a hidden population substructure [Umbach and Weinberg, 2000] or the influence of family history, a factor that is clearly related to susceptibility genes, on lifestyle factors such as smoking [Thomas, 2000]. Since related subjects share both ethnic and family history background, the within-family independence assumption (4) is much less likely to be affected by spurious association due to these factors. In particular, the assumption is the weakest for the sibling-case-control design, because siblings share ethnic and family history background. Cousins, on the other hand, only partially share these factors and thus the required assumption for the cousin case-control design is stronger. Possible ways for further relaxing this assumption based on a conditional independence model will be discussed later.

In the setting of the standard case-control design, we propose the conditioning event $\mathcal{G}_i$ in the likelihood (3) to be $\mathcal{G}_i^S$, the *set* of genotypes that is observed in the i^{th} matched pair. For matched pairs where observed genotypes are discordant, $\mathcal{G}_i$ contains the information on the two different types of genotypes that are observed for that pair, but does not specify the individual genotype of the case (G_{i1}) and the control (G_{i0}). In the Appendix, we show that under a rare disease assumption, with this definition of $\mathcal{G}_i$ the proposed conditional likelihood (3) can be computed as

$$L_{i,\text{CC}} = \frac{\exp\{m(G_{i1}, E_{i1}; \beta)\}}{\sum_{j=0}^{1} \left[\exp\{m(G_{ij}, E_{i1}; \beta)\} + \exp\{m(G_{ij}, E_{i0}; \beta)\}\right]}. \quad (5)$$

The sum in the denominator of $L_{i,\text{CC}}$ essentially constitutes four subjects corresponding to the four genotype-exposure configurations: (G_{i0}, E_{i0}), (G_{i1}, E_{i1}), (G_{i0}, E_{i1}), and (G_{i1}, E_{i0}). Thus, $L_{i,\text{CC}}$ is the same as the standard conditional likelihood for a 1:3-matched design, where the two additional subjects with genotype-exposure configuration (G_{i0}, E_{i1}) and (G_{i1}, E_{i0}) can be viewed as "pseudo" family members obtained by exchanging the genotypes of the observed family members: under the *G-E* independence assumption, such "pseudo" subjects are as equally likely to appear in a family as the observed subjects in that family. In this spirit, the proposed methodology has an intriguing similarity with the conditional logistic regression analysis of case-parent trio designs [Self et al., 1991], which is also based on pseudo-sibs for the observed case that could have been observed given the genotype of the parents.

Several other observations can be made from the final form of the conditional likelihood (5). First, similar to the standard conditional likelihood, $L_{i,\mathrm{CLR}}$ in (2), $L_{i,\mathrm{CC}}$ is free of the family-specific intercept parameters α_{F_i}. Second, although $L_{i,\mathrm{CC}}$ relies on the assumption of independence between G and E within each family, it is otherwise quite flexible in the sense that it does not depend on any assumption about $V_F(E_1, E_2)$, the family-specific distributions of joint exposure status for two relatives. Finally, because of the standard CLR form, estimates of the regression parameters β that maximize $L_{i,\mathrm{CC}}$, as well as corresponding asymptotic variance estimates, can be obtained by using standard and widely available CLR software.

In the Appendix, we derive standard errors for the estimates of β for a general class of designs that include all the designs described in this report. The relevant formula is (A.5). Also, in the Appendix we show that the proposed method is asymptotically at least as efficient as standard CLR. Indeed, what we show is that the asymptotic covariance matrix of estimates of β that maximize $L_{i,\mathrm{CC}}$ can be written as

$$V = I^{-1}(\beta) = \{I_1(\beta) + I_2(\beta)\}^{-1},$$

where $I_1(\beta)$ is the information matrix corresponding to the standard conditional likelihood (2) and $I_2(\beta)$ is a non-negative definite matrix. This automatically proves that $\{I_1(\beta) + I_2(\beta)\}^{-1} \leq \{I_1(\beta)\}^{-1}$ in the sense $\{I_1(\beta) + I_2(\beta)\}^{-1} - \{I_1(\beta)\}^{-1}$ is always a non-positive-definite matrix, thus implying that the proposed method is asymptotically at least as efficient as standard CLR.

Finally, we observe that the within family G-E independence assumption can be also exploited to construct powerful permutation-based methods for testing a global null hypothesis of no association. The pivot statistics could be given by any measure of distance, such as the sum of square differences, between the joint genotype and exposure frequencies of the cases and those of the controls. The permutation distribution of the statistics can be then generated by randomly switching G and E status within matched pairs of cases and controls. Unlike standard permutation tests, where covariates are permuted together, under the G-E independence assumption the two type of exposures should be permuted independently of one another.

THE NEW CONDITIONING PARADIGM IN CASE-CONTROL DESIGN WITH PARENTAL GENOTYPE DATA

For simplicity of notation, we will describe the proposed method in the context of a case-sibling-control design with parental genotype information. The method easily extends to alternative types of matched case-control designs as long as genotype data are available for parents of both cases and controls.

We assume M matched case-control sibling-pairs are sampled into a study. We use the same data structure and notation that we have introduced earlier for the standard family-based case-control designs. In addition, we define $\mathcal{G}_i^P = (G_{iM}, G_{iF})$ to be the parental (mother and father) genotype data for the i^{th} matched pair. The key assumption we exploit in this design setting is that genotype and exposure status for pairs of relatives in the source population are independently distributed conditional on their parental genotype information. Thus, if (G_1, G_2) and (E_1, E_2) denote the joint genotype and exposure status for a pair of siblings and $\mathcal{G}^P$ denotes their parental genotype information, the required independence assumption can be stated formally as

$$\mathrm{pr}(G_1, G_2, E_1, E_2|\mathcal{G}^P) = \mathrm{pr}(G_1, G_2|\mathcal{G}^P) \times \mathrm{pr}(E_1, E_2|\mathcal{G}^P). \tag{6}$$

Moreover, assuming a Mendelian mode of inheritance given parental genotypes, we can write $\mathrm{pr}(G_1, G_2|\mathcal{G}^P) = \mathrm{pr}(G_1|\mathcal{G}^P) \times \mathrm{pr}(G_2|\mathcal{G}^P)$. The family-based independence assumption stated in formula (6) is very weak in the sense that it is robust to the effects of various factors such as the presence of hidden population sub-structure or the influence of family history on lifestyle-related exposures. We will describe the assumption stated in (6) as "Type-II Independence".

In the setting described above, we propose to use the conditioning event ($\mathcal{G}_i$) in the general conditional likelihood (3) to be the *parental* genotype information ($\mathcal{G}_i^P$). We observe that $\mathcal{G}_i^P$ is a larger event than $\mathcal{G}_i^S$, the *set* genotype information in a case-control pair, in the sense that $\mathcal{G}_i^P$ contain the information of all possible values for $\mathcal{G}_i^S$. Thus, it is expected that when parental genotype data are available, conditioning on $\mathcal{G}_i^P$ will be more efficient than the conditioning on $\mathcal{G}_i^S$. The general conditional likelihood (3) with $\mathcal{G}_i = \mathcal{G}_i^P$ can be computed as

$$L_{i,\text{CCGP}} = \frac{\exp\{m(G_{i1},E_{i1};\beta)\}\text{pr}\left(G_{i1}|\mathcal{G}_i^P\right)\text{pr}\left(G_{i0}|\mathcal{G}_i^P\right)}{\sum_{G'_{i1}\in H_{\mathcal{G}_i^P}}\exp\{m(G'_{i1},E_{i1};\beta)\}\text{pr}\left(G'_{i1}|\mathcal{G}_i^P\right)+\sum_{G'_{i0}\in H_{\mathcal{G}_i^P}}\exp\{m(G'_{i0},E_{i0};\beta)\}\text{pr}\left(G'_{i0}|\mathcal{G}_i^P\right)}, \quad (7)$$

where $H_{\mathcal{G}_i^P}$ denotes all possible offspring genotypes associated with the parental genotypes G_i^P. The quantities $\text{pr}\left(G_{ij}|\mathcal{G}_i^P\right)$, the genotype probability of an offspring given the genotype of the parents, can be computed as fixed constants assuming a standard Mendelian mode of inheritance within family. Derivation of formula (7) follows assuming rare disease and calculations similar to those for the derivation of formula (5), and is given in the Appendix. In terms of computation, estimates of β maximizing $L_{i,\text{CCGP}}$ and associated standard errors can be obtained using standard CLR software that allows incorporation of offset terms. The general form of the covariance matrix of the parameter estimates is given in equation (A.5) in the Appendix.

There are connections between L_{CCGP} and the traditional conditional-likelihood for case-parents-trio (CPT) data [Self et al., 1991; Schaid, 1999], which is given by

$$L_{i,\text{CPT}} = \frac{\exp\{m(G_{i1},E_{i1},\beta)\}\text{pr}\left(G_{i1}|\mathcal{G}_i^P\right)}{\sum_{G'_{i1}\in H_{\mathcal{G}_i^P}}\exp\{m(G'_{i1},E_{i1},\beta)\}\text{pr}\left(G'_{i1}|\mathcal{G}_i^P\right)}. \quad (8)$$

The numerators of both $L_{i,\text{CCGP}}$ and $L_{i,\text{CPT}}$ correspond to the i^{th} case, with the expressions being the same up to a constant term. However, there are differences in the denominator. While the denominator of $L_{i,\text{CPT}}$ consists of all possible offspring the i^{th} pair of parents could have had with the offsprings' environmental exposure the same as that of the observed case (E_{i1}), the denominator of $L_{i,\text{CCGP}}$ consists of those of $L_{i,\text{CPT}}$ plus all possible offspring the i^{th} pair of parents could have had with the offsprings' environmental exposure the same as that of the observed control (E_{i0}).

In recent years, various extensions of $L_{i,\text{CPT}}$ have been developed for utilizing data on multiple offspring in nuclear families [Clayton, 1999; Cordell and Clayton, 2002; Cordell et al., 2004; Kraft et al., 2004]. All of these methods, however, condition on the exact phenotype (disease status) of the individual offsprings. As a result, in these methods, unaffected subjects (controls) are only indirectly informative, in the sense that they can be utilized to infer missing parental genotype data, but once the parental genotype information is available/inferred, the unaffected subjects are ignored in the respective conditional likelihoods. In contrast, $L_{i,\text{CCGP}}$ conditions only on the *set* of phenotypes defined by the event $D_1 + D_0 = 1$, instead of the individual phenotypes D_1 and D_0 themselves. In this approach, the unaffected offspring remain informative even when complete parental genotype information is available. The environmental exposure status (E_0) of the unaffected subject allows estimation of β_E, the main effect parameter associated with E, which cannot be estimated from $L_{i,\text{CPT}}$. Moreover, incorporation of the unaffected subjects leads to major increase in efficiency for estimation of the multiplicative interaction parameter (β_{GE}) and other related quantities (see Tables I and II). It is, however, important to note that $L_{i,\text{CCGP}}$, similar to $L_{i,\text{CC}}$, requires the assumption that the selection of a case-control pair of relatives does not depend on the individual G and E status of the relatives [Hsu et al., 2000].

SIMULATION STUDIES INVOLVING DIFFERENT DESIGNS AND ANALYTIC METHODS FOR FAMILY-BASED CASE-CONTROL STUDIES

In this section, we report simulation studies of the relative efficiency of different study designs and analytic methods for estimation of various risk-parameters of interest using data from nuclear families. In particular, we considered three designs: (A) the Sibling-Case-Control (SCC) design with G and E available on the matched cases and controls; (B) the Case-parent-trio (CPT) design with G available on cases and their parents and E available on the cases; and (C) the Sibling-Case-Control design with genotyped parents (SCCGP). The analytic methods we compared are: (1) Traditional conditional likelihood ($L_{i,\text{CLR}}$) for design (A); (2) the proposed conditional likelihood

TABLE I. Dominant Gene: Bias and efficiencies of alternative family-based designs[a] and analytic methods[m] for evaluation of different risk parameters

				SCC^{d2}					
		CPT^{d1}: L^{m1}_{CPT}		L^{m2}_{CLR}		L^{m3}_{CC}		SCCGP^{d3}: L^{m4}_{CCGP}	
$\{p_G, p_E\}$[b] $\{\beta_G, \beta_E\}$[c]	Risk-parameters[d]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]
0.01, 0.2 log(7), log(1.3)	OR(G I E=1)	0.09	1.13	−0.01	1.62	0.03	2.30	0.08	3.49
	OR(E I G=1)	NA	NA	−0.02	2.89	0.09	5.52	0.10	5.67
	MI_{GE}	0.18	0.81	−0.02	2.90	0.11	5.52	0.13	5.67
	AI_{GE}	NA	NA	−0.02	2.90	0.09	5.56	0.11	5.64
0.2, 0.2 log(1.3), log(1.3)	OR(G I E=1)	0.05	1.04	0.00	0.82	0.02	1.04	0.04	1.52
	OR(E I G=1)	NA	NA	0.00	0.93	0.02	1.19	0.03	1.40
	MI_{GE}	0.06	0.94	−0.00	1.05	0.02	1.34	0.04	1.64
	AI_{GE}	NA	NA	0.00	1.05	0.02	1.40	0.04	1.72
0.01, 0.5 log(7), log(1.12)	OR(G I E=1)	0.09	0.82	−0.01	0.66	0.02	0.78	0.08	1.38
	OR(E I G=1)	NA	NA	−0.01	2.49	0.09	3.89	0.10	4.18
	MI_{GE}	0.19	0.70	−0.01	2.39	0.10	3.58	0.12	3.83
	AI_{GE}	NA	NA	−0.01	2.48	0.09	3.91	0.11	4.18
0.2, 0.5 log(1.3), log(1.12)	OR(G I E=1)	0.04	0.79	−0.00	0.56	0.01	0.63	0.03	1.05
	OR(E I G=1)	NA	NA	0.00	0.94	0.02	1.14	0.02	1.34
	MI_{GE}	0.05	0.79	0.00	0.96	0.02	1.16	0.03	1.37
	AI_{GE}	NA	NA	0.00	0.94	0.02	1.17	0.03	1.38

[a] Designs: ^{d1}Case-parents trio, ^{d2}Sibling case-control, and ^{d3}Sibling case-control w. genotyped parents. Methods^{m}: Conditional-likelihoods described in formulae ^{m1}(8), ^{m2}(2), ^{m3}(5), and ^{m4}(7).

[b] Genotype (G=Aa/aa) and exposure (E=1) frequencies.

[c] True values for main effects of G and E.

[d] OR(G I E=1): OR for G among subjects with E=1; OR(E I G=1): OR for E among subjects with G=1; MI_{GE}; multiplicative-interaction; AI_{GE}: additive interaction.

[e] Relative bias evaluated as (true value − mean estimated value)/true value.

[f] Relative efficiencies compared to a population-based case-control design with the same number of cases and 1:1 case-control ratio.

($L_{i,\mathrm{CC}}$) for design (A); (3) the efficient conditional likelihood ($L_{i,\mathrm{CPT}}$) method for analysis of designs (B) [Self et al., 1991; Schaid, 1999]; and (4) the proposed conditional likelihood ($L_{i,\mathrm{CCGP}}$) for design (C).

THE SIMULATION DESIGN

We assumed that the gene variant of interest is a bi-allelic locus with the wild and variant-type alleles being denoted by A and a, respectively. We considered two distinct settings of interest, one for rare variants and the other for common variants. Within each setting, we considered dominant and recessive models for the effect of the gene-variant. We also assumed a binary environmental exposure and considered two scenarios, one involving a common exposure and the other involving a rare exposure.

We simulated data for nuclear families consisting of two siblings and their parents using the following setup. We simulated a family-specific allele frequency parameter to allow for population-substructure. For each family (F), we simulated an allele frequency parameter (θ_F) by first generating a random variable u_F from the normal distribution with mean parameter μ and variance σ^2 and then transforming u_F to the 0-1 scale as $\theta_F = \exp(u_F)/\{1 + \exp(u_F)\}$. We chose the variance parameter σ^2 to be 0.5 so that the $\pm 2\sigma$ limit of this distribution corresponds to approximately 15-fold variation in allele frequency across different families. We chose the mean parameter μ in such a way that the marginal probability of the genotype variant of interest (Aa or aa for the dominant model and aa for the recessive model) in the underlying population is fixed at 0.01 for the setting of a rare variant and 0.2 for the setting of a common variant. Given the allele frequency parameter θ_F for a family, we generated the genotype data for the parents assuming Hardy-Weinberg-Equilibrium and that the parents are

TABLE II. Recessive gene: bias and efficiencies of alternative family-based designs[a] and analytic methods[m] for evaluation of different risk parameters

				SCC^{d2}					
		CPT^{d1}: L^{m1}_{CPT}		L^{m2}_{CLR}		L^{m3}_{CC}		SCCGP^{d3}: L^{m4}_{CCGP}	
$\{p_G, p_E\}$[b] $\{\beta_G, \beta_E\}$[c]	Risk-parameters[d]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]	Bias[e]	RE[f]
0.01, 0.2 log(7), log(1.3)	OR(G∣E=1)	0.11	2.81	−0.02	1.69	0.04	2.74	0.10	5.43
	OR(E∣G=1)	NA	NA	−0.03	2.49	0.12	4.62	0.14	5.74
	MI$_{GE}$	0.21	2.04	−0.04	2.49	0.14	4.09	0.18	5.12
	AI$_{GE}$	NA	NA	−0.03	2.51	0.12	4.67	0.15	5.78
0.2, 0.2 log(1.3), log(1.3)	OR(G∣E=1)	0.05	1.28	0.00	0.69	0.02	0.95	0.04	1.73
	OR(E∣G=1)	NA	NA	0.00	0.92	0.02	1.26	0.03	1.45
	MI$_{GE}$	0.06	1.13	−0.00	0.96	0.03	1.40	0.04	1.80
	AI$_{GE}$	NA	NA	0.00	0.93	0.02	1.39	0.04	1.86
0.01, 0.5 log(7), log(1.12)	OR(G∣E=1)	0.10	2.06	−0.01	0.90	0.02	1.14	0.09	2.99
	OR(E∣G=1)	NA	NA	−0.01	2.26	0.11	4.33	0.13	5.26
	MI$_{GE}$	0.20	1.77	−0.01	2.26	0.13	3.93	0.15	4.83
	AI$_{GE}$	NA	NA	−0.01	2.26	0.11	4.33	0.14	5.28
0.2, 0.5 log(1.3), log(1.12)	OR(G∣E=1)	0.04	1.14	−0.00	0.59	0.01	0.70	0.04	1.29
	OR(E∣G=1)	NA	NA	0.00	0.79	0.02	1.05	0.02	1.13
	MI$_{GE}$	0.05	0.80	0.00	0.73	0.02	0.95	0.04	1.13
	AI$_{GE}$	NA	NA	0.00	0.72	0.02	0.98	0.03	1.17

[a] Designs: ^{d1}Case-parents trio, ^{d2}Sibling case-control, and ^{d3}Sibling case-control w. genotyped parents. Methods[m]: Conditional-likelihoods described in formulae ^{m1}(8), ^{m2}(2), ^{m3}(5), and ^{m4}(7).

[b] Genotype (G=aa) and exposure (E=1) frequencies.

[c] True values for main effects of G and E.

[d] OR(G∣E=1): OR for G among subjects with E=1; OR(E∣G=1): OR for E among subjects with G=1; MI$_{GE}$; multiplicative-interaction; AI$_{GE}$: Additive interaction.

[e] Relative bias evaluated as (true value−mean estimated value)/true value.

[f] Relative efficiencies compared to a population-based case-control design with the same number of cases and 1:1 case-control ratio.

independent. Given the genotype of the parents, we generated the genotypes for a pair of siblings based on a standard Mendelian mode of inheritance. We generated the environmental exposures for a pair of siblings by first generating a pair of correlated random variables (E_1^*, E_2^*) from a bivariate normal distribution with marginal means zero, marginal variances one and a correlation parameter ρ. We then dichotomized E_1^* and E_2^* into two binary 0-1 exposure variables E_1 and E_2 so that the marginal probability of exposure ($E = 1$) for the underlying population is 0.2 for the setting of a rare exposure and 0.5 for the setting of a common exposure. In our basic simulation setting (Tables I and II), we fixed the correlation parameter $\rho = 0.3$ so that it represents only a modest correlation between the environmental exposures for a pair of siblings. Later (Figs. 1 and 2), we explore the effect of varying ρ on the efficiencies of different designs and analytic methods.

We simulated the family-specific intercept term (α_F) to allow for heterogeneity in disease-risk between families that cannot be accounted for by G and E. For a given family F, we generated α_F from the normal distribution with mean α and variance τ^2. We chose $\tau^2 = 1$ so that the $\pm 2\sigma$ limit of this distribution corresponds to an approximately 50-fold variation in disease risk between families due to unknown factors. We fixed the mean parameter α at different values for different settings so that the marginal probability of the disease in the population, $\mathrm{pr}(D = 1)$, is always fixed at 0.01. Given α_F, we generated the disease outcome for each sibling, independent of the other, using the logistic regression model

$$\mathrm{pr}(D_j = 1|G_j, E_j, \alpha_F) = \frac{\exp\{\alpha_F + \beta_G f(G) + \beta_E E + \beta_{GE} f(G) * E\}}{1 + \exp\{\alpha_F + \beta_G f(G) + \beta_E E + \beta_{GE} f(G) * E\}}, \tag{9}$$

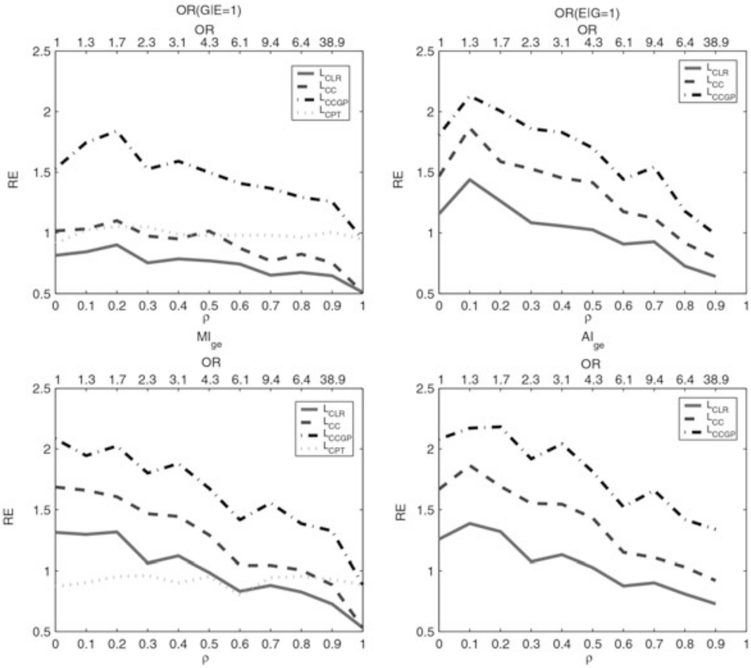


Fig. 1. Dominant gene: Relative efficiency (RE) of alternative family-based designs and analytic methods as a function of sibling-correlation (ρ) in exposure (E). The top axis shows the correlation in the binary OR scale. The methods compared are $\mathcal{L}_{\text{CLR}}$ (solid line), $\mathcal{L}_{\text{CC}}$ (dashed line), $\mathcal{L}_{\text{CPT}}$ (dotted line), and $\mathcal{L}_{\text{CCGP}}$ (dashed/dotted line). MI$_{GE}$ and AI$_{GE}$ indicate multiplicative and additive interaction parameters, respectively. Values of the relative efficiency are evaluated based on 500 simulations. In each simulation, data for different designs are generated based on 5,000 cases.

where $f(G)$ is a binary 0-1 function reflecting the mode of effect of the gene: $f(Aa/aa) = 1$ for dominant and $f(aa) = 1$ for recessive. In the basic simulation setting (Tables I and II), we chose the main effect parameters β_G to be $\log(7)$ when $\text{pr}\{f(G) = 1\} = 0.01$ and $\log(1.3)$ when $\text{pr}\{f(G) = 1\} = 0.2$, so that the two settings corresponds to a high-penetrance rare variant and a low-penetrance common variant, respectively. Similarly, we chose β_E to be $\log(1.3)$ when $\text{pr}(E = 1) = 0.2$ and $\log(1.12)$ when $\text{pr}(E = 1) = 0.5$, so that the main effect of E is stronger when the exposure is rare. We fixed β_{GE} to be $\log(3)$, which corresponds to a strong multiplicative interaction between G and E.

Following this scheme, we first generated data for a large number of randomly sampled nuclear families. Treating these randomly selected families as the underlying population, we then selected 5,000 families with one diseased and one non-diseased sibling. During analysis of data from each design, we only retained the appropriate genotype and environmental exposure information for that design and discarded the rest of the information.

In the above simulation setting, we used two types of random effects, namely u_F and α_F, to generate variations in genotype frequencies and disease-risk, respectively, across families. We chose these random effects to be uncorrelated with each other so that there is no population-level association between genetic exposure and disease risk that cannot be explained by the direct effect of the gene on risk of the disease other than

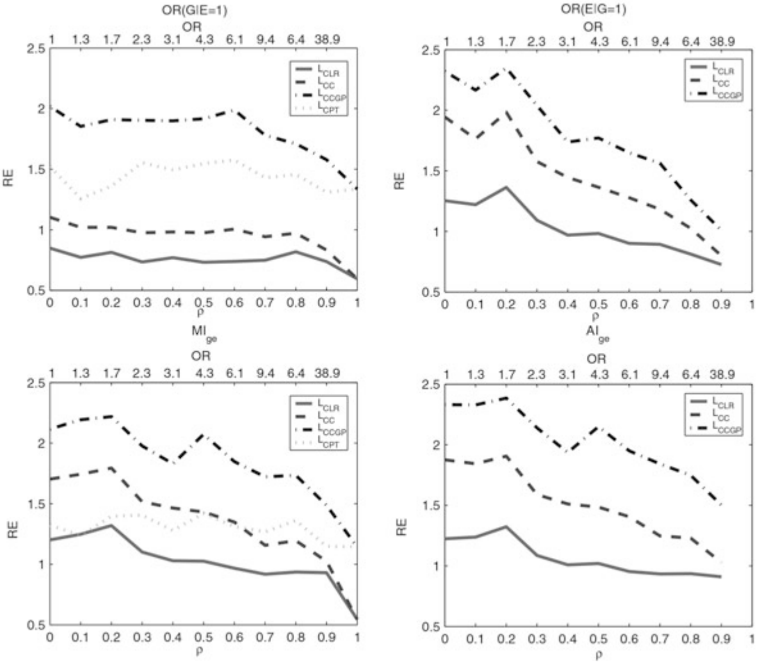


Fig. 2. Recessive gene: Relative efficiency (RE) of alternative family-based designs and analytic methods as a function of sibling-correlation (ρ) in exposure (E). The top axis shows the correlation in the binary OR scale. The methods compared are $\mathcal{L}_{\text{CLR}}$ (solid line), $\mathcal{L}_{\text{CC}}$ (dashed line), $\mathcal{L}_{\text{CPT}}$ (dotted line), and $\mathcal{L}_{\text{CCGP}}$ (dashed/dotted line). MI$_{GE}$ and AI$_{GE}$ indicate multiplicative and additive interaction parameters, respectively. Values of the relative efficiency are evaluated based on 500 simulations. In each simulation, data for different designs are generated based on 5,000 cases.

through the parameters β_G and β_{GE} in the risk model (9). Thus, in this setting, bias due to population-stratification not being a concern, a case-control study based on unrelated subjects is an alternative valid design. We chose the population-based case-control design (PCC), analyzed with standard logistic regression, to be the common reference point for evaluating the relative efficiencies of various family-based designs and analytic methods. To simulate data for this design, we first generated data for a large number of randomly sampled nuclear families and then selected 5,000 cases and 5,000 controls from 10,000 independent families.

For evaluating the efficiencies of different designs and methods, we considered four different parameters of epidemiologic interest: (1) $\text{OR}(G|E=1)=\exp(\beta_G+\beta_{GE})$: the odds ratio associated with the gene variant among subjects with environmental exposure ($E=1$); (2) $\text{OR}(E|G=1)=\exp(\beta_E+\beta_{GE})$: the odds ratio associated with the environmental exposure among subjects with a variant genotype ($G=1$); (3) $\text{MI}_{GE}=\exp(\beta_{GE})$: the multiplicative interaction between G and E; and (4) $\text{AI}_{GE}=\exp(\beta_G+\beta_E+\beta_{GE})-\exp(\beta_G)-\exp(\beta_E)+1$: the additive interaction between G and E [Khoury et al., 1993]. All of the designs except the case-parent-trio design yield an estimate of all of the four types of association/interaction parameters; the case-parent trio design cannot estimate the main effect of E (β_E) and hence also cannot estimate $\text{OR}(E|G=1)$ and AI_{GE}.

Strictly speaking, estimates for a particular type of association/interaction parameter from different designs are not directly comparable due to differences in scale of measurement. That is, while the model for analyzing the CPT design is defined in terms of within family relative risk parameters, those for family-based and population-based case-control designs are defined in terms of within- and between-family odds-ratio parameters, respectively. In spite of such differences, one common goal of all of the designs is to test for hypotheses about different types of association/interaction parameters. Thus, we evaluated different designs based on their relative powers for rejecting the null hypothesis about different association/interaction parameters in the respective underlying scales. For each type of association/interaction parameter of interest (θ), we evaluated the quantity $\tau = \overline{\log(\hat{\theta})}/\text{sd}\{\log(\hat{\theta})\}$, where $\overline{\log(\hat{\theta})}$ and $\text{sd}\{\log(\hat{\theta})\}$ are the empirical mean and the standard error of the estimate of θ from a given design over different simulated data sets. The ratio of τ^2 for two designs estimates the asymptotic relative efficiency of the two designs, i.e., the inverse-ratio of the sample sizes required by the two designs to reject the null hypothesis of no association/interaction, i.e., $\log(\theta) = 0$. For rare diseases, such as the one considered in our simulation setting, the differences between the mean estimates of parameters $\overline{\log(\hat{\theta})}$ from different designs are small and thus the relative efficiencies of different designs are mostly determined by the precision of parameter estimates $1/\text{var}\{\overline{\log(\hat{\theta})}\}$.

RESULTS: BIAS AND EFFICIENCY

Table I (dominant gene) and Table II (recessive gene) show the results of simulation experiments with fixed sets of parameter values. We make several key observations from Tables I and II, as follows. For the scenario of "low penetrance common variant" ($\beta_G = \log(1.3), \text{pr}\{f(G) = 1\} = 0.2$), all of the proposed analytic methods had negligible percentage bias in estimating the "true" parameters of the underlying disease-risk model. For the scenario of "high-penetrance rare variant," the novel methods produced noticeable, but modest ($\leq 15\%$), bias in parameter estimates. The bias likely arises due to the rare disease approximation, because in this setting the risk of disease for subjects with both the genetic and the environmental exposure was as high as 30%. Comparison of the traditional (L_{CLR}) and the proposed method (L_{CC}) of analyzing the SCC design shows the major efficiency advantages of the latter approach for all of the four different parameters. Comparison of the CPT design and the SCC design shows that the latter design, when analyzed using our new method (L_{CC}), was superior for estimation of the multiplicative interaction term. For estimation of $\text{OR}(G|E = 1)$, however, the CPT design was generally more efficient. The SCCGP design, when analyzed using our approach (L_{CCGP}), had the highest efficiency among all of the designs for all of the four different association/interaction parameters. The efficiencies of all of the family-based designs relative to the PCC design decreased as either the genetic or the environmental exposure becomes more common.

Figures 1 (for dominant gene) and 2 (for recessive gene) show the effect of varying the parameter ρ that determines the correlation between the exposure variables in a pair of siblings. For these graphs, we chose other parameter values of the simulation in such a way so that they reflect an intermediate situation between the "high-penetrant rare gene" and "the low-penetrant common gene" scenarios we considered for Table I and Table II. In particular, we fixed $\Pr(G = 1) = 0.1$, $\Pr(E = 1) = 0.3$, $\beta_G = \log(1.6)$, $\beta_E = \log(1.12)$, and $\beta_{GE} = \log(3)$.

Overall, for all of the four types of parameters, the efficiencies of the SCC and SCCGP designs, relative to the PCC design, decreased as ρ increased. The efficiency of the CPT design did not depend on ρ as this design does not involve the sibling controls. Comparison of the traditional (L_{CLR}) and novel conditional likelihood method (L_{CC}) for analyzing the SCC design shows that the efficiency advantage of the latter method remained fairly constant over a wide range of value of ρ: only at very high values of ρ did the difference between the two method starts diminishing. When $\rho = 1$, which corresponds to two siblings having identical exposures, the two methods are identical and hence had the same efficiency. The efficiency of the SCCGP design also remained substantially higher than all the other designs except for extremely high values of ρ. At $\rho = 1$, the SCCGP and CPT design are identical. At $\rho = 1$, none of the family-based designs can estimate the association parameter $OR(E|G = 1)$ and the additive interaction parameter AI_{GE}.

Inspection of the efficiencies of the family-based designs relative to the PCC design suggests that in our simulation setting, where the gene-variant is moderately common, i.e., $\Pr(G = 1) = 0.1$, the

SCC design, when analyzed using the traditional method (L_{CLR}), had a lower efficiency than the PCC design for high values of ρ ($\rho > 0.5$). In contrast, when analyzed with the novel method (L_{CC}), the efficiency of the SCC design remained higher in a wider range of ρ. The efficiency of the SCCGP design remained higher than that for the PCC design for almost all values of ρ.

RESULTS: TYPE-I ERROR RATE

Under a global null hypothesis of no association of the disease with either G or E, the conditional probabilities $\Pr(D_{i1} = 1, D_{i0} = 0, G_{i1}, G_{i0} | D_{i1} + D_{i0} = 1, G_i, E_{i1}, E_{i0})$ become free of the intercept parameters irrespective of whether the disease is rare or not. Thus, the tests based on our proposed likelihood will be valid under this global null hypothesis whether the disease is rare or not. We evaluated the empirical type-I error rates of the proposed methods for testing weaker null hypotheses that specify only certain parameters of interest to be null, but leave the other parameters unspecified.

We consider the simulation setting of a high-penetrant dominant rare gene, a scenario where we had observed modest bias in parameter estimation due to violation of the rare disease assumption. We simulated data under four types of null hypotheses corresponding to (1) $\mathrm{MI}_{GE} = 1$, (2) $\mathrm{OR}(G|E = 1) = 1$, (3) $\mathrm{OR}(E|G = 1)$, and (4) $\mathrm{AI}_{GE} = 0$. For generating data under (1), we chose β_G to be $\log(7)$ and β_E to be $\log(1.3)$ as before, but set $\beta_{GE} = 0$. For generating data under (2), we chose $\beta_{GE} = -\beta_G = -\log(7)$ and $\beta_E = \log(1.3)$. Similarly, for generating data under (3), we chose $\beta_{GE} = -\beta_E = -\log(1.3)$ and $\beta_G = \log(7)$. Finally, for generating data under (4), we chose $\beta_G = \log(7)$ and $\beta_E = \log(1.3)$ and then solved for that value of β_{GE} so that $\exp(\beta_G + \beta_E + \beta_{GE}) - \exp(\beta_G) - \exp(\beta_E) + 1 = 0$. Table III shows the empirical type-I error rates of the Wald tests (5% significance level) associated with the likelihoods L_{CC} and L_{CCGP}. We observe that for each type of null hypotheses, the test procedures maintained the nominal α -level very well. We also found the tests to be unbiased in extensive simulation studies in the other settings of Tables I and II (data not shown). These results are also consistent with the fact that in all of the scenarios in Tables I and II where we had observed modest bias in parameter estimation, the direction of bias was always towards the null value of the parameters.

TABLE III. High-penetrant rare dominant gene: empirical type-I error rates of wald-tests with 5% significance levels

Methods	H_0: MI_{GE}=1	H_0: OR(G∣E=1)=1	H_0: OR(E∣G=1)=1	H_0: AI_{GE}=0
L_{CC}	0.042	0.048	0.040	0.036
L_{CCGP}	0.040	0.052	0.056	0.030

GENERAL FAMILY DATA WITH *r* AFFECTED AND *s* UNAFFECTED SUBJECTS

In this section, we briefly outline how the proposed methods for analyzing 1:1 family-matched case-control studies can be utilized for more general family studies that collect genotype and environmental exposure data for more than one affected and/or unaffected family members. Suppose there are M families sampled into a study and each family defines a matched set consisting of comparable cases and controls in the family, all of whom have data on both G and E. Suppose the i^{th} family defines a matched set consisting of r_i cases and s_i controls with a total of $n_i = r_i + s_i$ subjects. Let D_{i0} and D_{i1} denote the set of controls and cases, respectively, for the i^{th} family.

Liang [1987] proposed a pairwise pseudo-likelihood approach for analysis of matched case-control studies in which the contribution of a matched set of r_i cases and s_i controls is given by the product of the usual conditional likelihoods ($L_{i,\mathrm{CLR}}$) for 1:1 matched studies for all possible $r_i \times s_i$ case-control pairs within that matched set. Under the gene-environment independence assumption, we propose to use a similar pseudo-likelihood approach based on case-control pairs, except that for each pair we use an appropriate efficient conditional likelihood instead of the traditional conditional likelihood. More explicitly, the pseudo-likelihood for the data can be written in the general form

$$L = \prod_{i=1}^{M} \prod_{j \in D_{i0}, k \in D_{i1}} L_{(jk)_i,*}, \tag{10}$$

where $L_{(jk)_i,*}$ denotes an appropriate conditional- or pseudo-conditional likelihood for the $(j,k)^{\mathrm{th}}$ case-control pair within the i^{th} matched set: the likelihood should be chosen efficiently according to whether and what kind of parental genotype data are available for that pair. In particular, when parental genotype data (both parents) are available for both subjects in the pair, $L_{(jk)_i,*}$ can be defined to be the conditional likelihood $L_{(jk)_i,\mathrm{CCGP}}$

(formula 8). When parental genotype data (both parents) are available only for the case in the pair, $L_{(jk)_i,*}$ can be defined to be $L_{(jk)_i,\mathrm{CC-CPT}} = L_{(jk)_i,\mathrm{CC}} \times L_{(jk)_i,\mathrm{CPT}}$, a pseudo-likelihood that efficiently combines information from case-control and case-parents-trio data. In all other cases, we propose to use $L_{(jk)_i,\mathrm{CC}}$ (formula 5). Finally, we observe that if for some families only cases and their parents, but no matching controls, are available, the contribution of these families in the above pseudo-likelihood can be defined by $L_{i,\mathrm{CPT}}$, the usual conditional-likelihood for case-parents-trio data.

In the pseudo-likelihood (10), the contribution of a matched set is obtained by taking the product of the contributions of all different case-control pairs within the set pretending that the different pairs from the same family are independent. From the theory of estimating-equations [Godambe, 1991], it is well known that such pseudo-likelihood methods produce consistent estimates of regression parameters even if in truth there is correlation among paired units within the same family. For variance estimation, however, the correlation within a family needs to be accounted for. A sandwich covariance matrix estimator that can account for such correlation is given in the Appendix: the formula is (A.9). In addition, bootstrap sampling with matched-sets as the sampling units can be used to obtain covariance matrix estimates that can account for within-family correlation.

DISCUSSION

We have proposed a new paradigm of conditional likelihood for analysis of family-based case-control studies. This approach, with a rare disease approximation, leads to a variety of simple but highly efficient methods of estimating statistical interaction and other risk parameters of interest involving genetic and environmental exposures. These methods exploit within family *G-E* independence assumptions that are much less stringent than the *G-E* independence assumption that has been previously utilized for case-only and population-based case-control studies. To the best of our knowledge, the proposed method involving the likelihood $L_{i,\mathrm{CC}}$ represents the first successful effort for exploiting the within-family "Type-I independence" assumption in the context of family-based case-control studies. Moreover, the likelihood $L_{i,\mathrm{CC}}$ can be used for efficient analysis of any other type of matched case-control study, the required assumption being that G and E are independent within "matched subjects" in the population. The proposed method involving the likelihood $L_{i,\mathrm{CCGP}}$, exploiting the "Type-II independence" assumption, represents the first approach to a unified analysis of data from family-based case-control studies that include parental genotype information.

An important aspect of the proposed general conditional likelihood ($L_{i,\mathrm{general}}$) is that it simultaneously conditions on a *set* phenotype event ($D_1 + D_0$) and a *set* genotype event (G). Historically, the idea of conditioning on a *set* phenotype event has been used in the setting of traditional conditional logistic regression ($L_{i,\mathrm{CLR}}$) analysis of matched case-control studies. In this approach, however, one conditions on individual covariate information of the case-control subjects. The idea of conditioning on a *set* genotype event has also existed for a while in the literature of family studies. The well known likelihood ($L_{i,\mathrm{CPT}}$) of case-parent-trio design is formed by conditioning on the parental genotypes ($\mathcal{G}^P$) of the cases, which yield the *set* of all possible genotypes for the offspring. Recent extensions of $L_{i,\mathrm{CPT}}$ for dealing with missing parental genotype information has also been based on conditioning on various types of *set* genotype events [Clayton, 1999; Cordell and Clayton 2002; Rabinowitz and Laird, 2000]. All of these methods, however, condition on individual phenotype information of the family members. In this article, we show how the two approaches of conditioning on a *set* genotype event and conditioning on a *set* phenotype event can be unified through the general conditional likelihood ($L_{i,\mathrm{general}}$), resulting in novel and efficient methods for analysis of matched case-control studies with or without parental genotype information.

Our simulation studies clearly demonstrate the efficiency advantage of our methods ($L_{i,\mathrm{CC}}$ and $L_{i,\mathrm{CCGP}}$) over the traditional conditional logistic regression method for analysis of family-based case-control and case-parents studies. These results also reveal some intriguing design implications. Several previous studies have compared the relative efficiencies of sibling-case-control (SCC) and case-parent trio (CPT) designs for estimation of the multiplicative interaction parameter: they generally concluded that while the former design tends to be superior for dominant genes, the latter design is more efficient for recessive genes [Witte et al., 1999; Gauderman, 2002]. However, in these studies the method employed for analysis of the CPT design implicitly assumes *G-E* independence, but that for the SCC design does not exploit any such assumption. In our study, when we analyzed both designs using similar independence assumptions, we found not only that the efficiency advantage of the SCC design over CPT design for dominant genes is even greater than reported before, but also that the SCC design can be more efficient than the CPT design even for recessive genes. In terms of other parameters of interest, a weakness of the CPT design is that it cannot be used to estimate either the additive interaction or the association parameter for the environmental exposures. This design, however, is quite efficient for estimation of the genetic association parameter. The SCC design, on the other hand, although it produces an estimate of all different parameters of interest, was inefficient for estimation of the genetic association parameter $OR(G|E=1)$.

Our simulation studies also demonstrate the optimality of the family-based case-control design that includes parental genotype information. Although the potential promise of such a hybrid design has been discussed previously [Weinberg and Umbach, 2000], the actual efficiency of such a design has not been evaluated. It is worth noting that we compare efficiencies of different family-based designs with a fixed number of families, but different designs require different amounts of data collection within a family. The sibling-case-control design with parental genotype information, for example, requires one additional genotyping compared to an ordinary sibling-case-control design: note that the sibling control need not be genotyped if parental genotype data are available. It also requires one additional exposure assessment compared to an ordinary case-parents design. Given that family members of cases are usually well motivated, the effort required for such additional data collection may be worthwhile, considering the potential for large efficiency gain.

Comparison of the efficiency of the sibling-case-control designs (SCC and SCCGP) with that of the population-based case-control (PCC) design shows that our methodology increases the utility of the former design for a wider set of situations. Previous studies, as well as our simulations, demonstrate that the SCC design, when analyzed with traditional methods, generally tends to be less efficient than the PCC design, except when the genetic variant under study is rare and/or sibling correlation in the environment exposure is modest. When analyzed with our methods, the SCC design retains the efficiency advantage over the PCC design for estimation of interaction and some of the other parameters of interest in a wider range of values for the gene-frequency and the sibling-correlation parameters. The SCCGP design retained the efficiency advantage for an even wider range of parameter values. Of course, one can also increase efficiency of the PCC design, based on methods that can exploit G-E independence. The required population level independence assumption, however, is much stronger and more likely to be violated due to spurious association as described below.

Although exploiting *G-E* independence leads to major efficiency gains, some caution is needed for use of this assumption. The case-only estimate of interaction, which relies on a population-based *G-E* independence assumption, has been shown to be severely biased when the assumption is violated [Albert et al., 2001]. Even if no direct association exists, association between *G* and *E* in the population may arise, for example, due to hidden population sub-structure across which genotype and exposure frequency may vary or by influence of family history on an individual's behavior regarding established risk factors such as smoking.

Although the family-based independence assumptions we exploit are much less stringent in general, it is important to realize that the Type-I independence assumption we exploit for case-control studies is robust to spurious association only for the sibling-case-control design. If cousins are selected as controls, they may partially but not completely share ethnic and family history background with the cases; thus, the possibility of spurious association remains. One possible remedy for minimizing such bias is to consider a conditional *G-E* independence model that can adjust for co-factors *S*, such as the ethnic origins of the unrelated parents, by specifying the difference in genotype-frequencies between a pair of relatives as a parametric function of the differences in *S* between the relatives. We have found that under this relaxed independence assumption, the general conditional likelihood in formula (3) can be used with $\mathcal{G} = \mathcal{G}^S$, the unordered genotype information in a matched pair, to jointly estimate the regression parameters of interest and the additional parameters of the conditional independence model.

The *G-E* independence assumption can be violated also due to direct association between *G* and *E*. Genetic polymorphisms in the smoking metabolism pathway, for example, may not only modify a subject's risk from smoking, but also can influence his/her level of addiction to smoking. When the plausibility of such direct association exists, the advantage of the case-control design is that it has the option of being analyzed by the standard conditional logistic regression method that does not require the independence assumption. The case-parents design, however, intrinsically relies on the independence assumption and thus can lead to biased parameter estimates.

Our novel conditional likelihood framework opens several areas of further research. We have assumed rare disease for the simplification of conditional likelihood calculations. From the derivation shown in the Appendix it can be seen that the precise assumption required here is that the probability of at least one disease occurrence in a pair of individuals is small for all combinations of risk factors, a slightly stronger assumption than that is typically made for an individual subject in population-based case-control studies. In our simulation study, where overall disease prevalence was 1% in the population, we observed no bias in testing and only modest bias in parameter estimation even in situations where the disease-risk was high for certain combinations of *G* and *E*. Future work, however, is needed to study the impact of the rare disease assumption for more common diseases. The general conditional likelihood of the form (3) is valid whether the disease is rare or not: for common diseases, however, these likelihoods would involve the family-specific intercept parameters (α_F). Thus, one way of relaxing the rare disease assumption would be to assume a parametric random effect model for the distribution of α_F across families and estimate the corresponding parameters from the conditional likelihoods themselves [Pfeiffer et al., 2002].

Similar to the traditional conditional-logistic-regression (CLR) analysis, the proposed conditional likelihood procedures assume disease status is conditionally independent within families. However, if the locus under study is in linkage disequilibrium with another disease susceptibility locus, such a conditional independence model may not capture the correlation within a family that contributes more than two members in the study [Weinberg and Umbach, 2000]. In such a situation, the proposed pair-wise pseudo-likelihood (10) together with the robust variance estimator (A.9) still remains a valid method for analysis of the data. Alternative methods for dealing with residual correlation that have been previously developed in the context of traditional CLR [Siegmund et al., 2000; Rieger et al., 2001], also could be adopted in the setting of the novel conditional likelihoods.

When studying the effect of a gene through haplotypes is of interest, the proposed methodologies can be applied for studying haplotype-environment interaction, but some extensions are required to deal with phase ambiguity. We have demonstrated how to exploit parental genotype information on case-control subjects when both parents of a subject are available to be recruited. In practice, however, genotype information may be available only for one parent for some case-control subjects. One way of efficiently utilizing incomplete parental genotype information would be to choose the conditioning event $\mathcal{G}$ in $L_{i,general}$ (formula 3) to be the minimal sufficient statistics for the gene-frequency parameters [Rabinowitz and Laird, 2000]. Various alternative strategies, such as modelling mating-type parameters [Kraft et al., 2004], that have been proposed in the past for dealing with missing parental genotypes in case-parents studies, could be also useful for developing extensions of the proposed methodologies. These and other extensions of the methodologies will be studied in future publications.

ACKNOWLEDGEMENTS

Research was supported by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

REFERENCES

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interaction. Am J Epidemiol 154:687–693.

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika (in press).

Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. Am J Hum Genet 65:1170–1177.

Clayton D, McKeigue PM. 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1356–1360.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data:

application to *HLA* in type 1 diabetes. Am J Hum Genet 70: 124–141.

Cordell HJ, Barratt BJ, Clayton DG. 2004. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol 26:167–185.

Curtis D. 1997. Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333.

Gauderman WJ. 2002. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21:35–50.

Gauderman WJ, Witte JS, Thomas DC. 1999. Family-based association studies. J N Cancer Inst 26:31–37.

Godambe VP. 1991. Estimating functions. Oxford: Oxford University Press.

Hsu L, Zhao LP, Aragaki C. 2000. A note on a conditional-likelihood approach for family-based association studies of candidate genes. Human Hered 50:194–200.

Khoury MJ, Beaty TH, Cohen BH. 1993. Fundamentals of genetic epidemiology. Oxford: Oxford University Press.

Kraft P, Palmer C, Woodward J, Turunen J, Minassian S, Paunio T, Lonnqvist J, Peltonen L, Sinsheimer J. 2004. RHD maternal-fetal genotype incompatibility and schizophrenia: Extending the MFG test to include multiple siblings and birth order. Eur J Hum Genet 12:192–198.

Liang KY. 1987. Extended Mantel-Haenzel estimating procedure for multivariate logistic regression models. Biometrics 43:289–299.

Pfeiffer R, Gail MH, Pee D. 2002. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. Biometrika 88:933–948.

Piegorsch WW, Weinberg CR, Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. Stat in Med 13:153–162.

Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Hered 50:211–223.

Rieger RH, Kaplan NL, Weinberg CR. 2001. Efficient use of siblings in testing for linkage and association. Genet Epidemiol 20:175–191.

Schaid DJ. 1999. Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273.

Self SG, Longton G, Kopecky KJ, Liang KY. 1991. On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61.

Siegmund KD, Langholz B, Kraft P, Thomas DC. 2000. Testing linkage disequilibrium in sibships. Am J Hum Genet 67:244–288.

Spielman RS, Ewens JE. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458.

Thomas DC. 2000. Case-parents design for gene-environment interaction by Schaid. Genet Epidemiol 19:461–463.

Thompson WD. 1991. Effect modifications and limits of biological inference from epidemiologic data. J Clin Epidemiol 44:221–232.

Umbach DM, Weinberg CM. 1997. Designing and analyzing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743.

Umbach DM,Weinberg CM. 2000. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 66:251–261.

Weinberg CR, Umbach DM. 2000. Choosing a retrospective design to assess joint genetic and environmental contribution to risk. Am J Epidemiol 152:197–203.

Witte JS, Gauderman WJ, Thomas DC. 1999. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149:693–705.

APPENDIX

DERIVATION OF PROPOSED CONDITIONAL LIKELIHOOD IN THE GENERAL CASE

Recall that (D_{i0}, G_{i0}, E_{i0}) and (D_{i1}, G_{i1}, E_{i1}) are the data for the control and case, respectively. In addition, $\mathcal{G}_i$ is the conditioning event, and $\mathcal{H}_{\mathcal{G}_i}$ is the set of ordered pairs (G_{i1}, G_{i0}) that are consistent with the information in $\mathcal{G}_i$. We also denote by F_i the i^{th} family. Note that for the i^{th} pair,

$$\begin{aligned}\mathcal{L}_{i,\text{general}} =& \operatorname{pr}(D_{i1}=1, D_{i0}=0, G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)\\ =& \operatorname{pr}(D_{i1}=1, D_{i0}=0 | D_{i1}+D_{i0}=1, G_{i1}, G_{i0}, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)\\ &\times \operatorname{pr}(G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)\\ =& \operatorname{pr}(D_{i1}=1, D_{i0}=0 | D_{i1}+D_{i0}=1, G_{i1}, G_{i0}, E_{i1}, E_{i0}, F_i)\\ &\times \operatorname{pr}(G_{i1}, G_{i0} | D_{i1}+D_{i0}=1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i) = \mathcal{L}_{i1} \times \mathcal{L}_{i2}.\end{aligned}$$

The term $\mathcal{L}_{i1}$ is given in equation (2). It is easily seen that

$$\mathcal{L}_{i2} = \frac{\operatorname{pr}\{D_{i1}+D_{i0}=1 | G_{i1}, G_{i0}, E_{i1}, E_{i0}, F_i\} \operatorname{pr}(G_{i1}, G_{i0} | \mathcal{G}_i, E_{i1}, E_{i0}, F_i)}{\sum_{(G'_{i1}, G'_{i0}) \in H_{G_i}} \operatorname{pr}\{D_{i1}+D_{i0}=1 | G'_{i1}, G'_{i0}, E_{i1}, E_{i0}, F_i\} \operatorname{pr}(G'_{i1}, G'_{i0} | \mathcal{G}_i, E_{i1}, E_{i0}, F_i)}.$$

By assumption, G and E are independent given $\mathcal{G}_i$ and F_i, and in addition the conditioning event removes the family effect, so that $\operatorname{pr}(G_{i1}, G_{i0} | \mathcal{G}_i, E_{i1}, E_{i0}, F_i) = \operatorname{pr}(G_{i1}, G_{i0} | \mathcal{G}_i)$. We thus have that

$$\mathcal{L}_{i2} = \frac{\operatorname{pr}\{D_{i1}+D_{i0}=1 | G_{i1}, G_{i0}, E_{i1}, E_{i0}, F_i\} \operatorname{pr}(G_{i1}, G_{i0} | \mathcal{G}_i)}{\sum_{(G'_{i1}, G'_{i0}) \in \mathcal{H}_{\mathcal{G}_i}} \operatorname{pr}\{D_{i1}+D_{i0}=1 | G'_{i1}, G'_{i0}, E_{i1}, E_{i0}, F_i\} \operatorname{pr}(G'_{i1}, G'_{i0} | \mathcal{G}_i)}. \tag{A.1}$$

In addition, using (1), we have that

$$\begin{aligned}
&\mathrm{pr}(D_{i1}+D_{i0}=1|G_{i1},G_{i0},E_{i1},E_{i0},F_i)\\
&=\mathrm{pr}(D_{i1}=1,D_{i0}=0|G_{i1},G_{i0},E_{i1},E_{i0},F_i)+\mathrm{pr}(D_{i1}=0,D_{i0}=1|G_{i1},G_{i0},E_{i1},E_{i0},F_i)\\
&=\frac{\exp(\alpha_{F_i})\times\left[\exp\{m(G_{i1},E_{i1};\beta)\}+\exp\{m(G_{i0},E_{i0};\beta)\}\right]}{\left[1+\exp\{\alpha_{F_i}+m(G_{i1},E_{i1};\beta)\}\right]\times\left[1+\exp\{\alpha_{F_i}+m(G_{i0},E_{i0};\beta)\}\right]} \qquad (A.2)\\
&\approx\exp(\alpha_{F_i})\times\left[\exp\{m(G_{i1},E_{i1};\beta)\}+\exp\{m(G_{i0},E_{i0};\beta)\}\right],
\end{aligned}$$

where the approximation in the last step is based on the assumption of rare disease. Thus, combining (A.1) and (A.2), we see that

$$\mathcal{L}_{i2}=\frac{\left[\sum_{j=0}^{1}\exp\{m(G_{ij},E_{ij},\beta)\}\right]\mathrm{pr}(G_{i1},G_{i0}|\mathcal{G}_i)}{\sum_{(G'_{i1},G'_{i0})\in\mathcal{H}_{\mathcal{G}_i}}\left[\sum_{j=0}^{1}\exp\left\{m(G'_{ij},E_{ij},\beta)\right\}\right]\mathrm{pr}(G'_{i1},G'_{i0}|\mathcal{G}_i)}. \qquad (A.3)$$

Combining (2) and (A.3), we find that

$$\mathcal{L}_{i,\mathrm{general}}=\frac{\exp\{m(G_{i1},E_{i1},\beta)\}\mathrm{pr}(G_{i1},G_{i0}|\mathcal{G}_i)}{\sum_{(G'_{i1},G'_{i0})\in\mathcal{H}_{\mathcal{G}_i}}\left[\sum_{j=0}^{1}\exp\left\{m(G'_{ij},E_{ij},\beta)\right\}\right]\mathrm{pr}(G'_{i1},G'_{i0}|\mathcal{G}_i)}, \qquad (A.4)$$

which is the natural generalization of (7).

For the standard matched family-based case-control design where we set $\mathcal{G}_i$ to be the *set* of genotypes for the i^{th} case-control pair, under the assumption that $\mathrm{pr}(G_{i1}=g_1,G_{i0}=g_0|\mathcal{G}_i)=\mathrm{pr}(G_{i1}=g_0,G_{i0}=g_1|\mathcal{G}_i)$, we have that $\mathrm{pr}(G_{i1},G_{i0}|\mathcal{G}_i)=1/2$, thus verifying (5). For the parental genotype case-control design, we have already verified (7) simply by setting $\mathcal{G}_i=\mathcal{G}_i^{\mathrm{P}}$, the full parental genotype information for the i^{th} case-control pair.

STANDARD ERROR ESTIMATION

In this section, we show that the asymptotic covariance matrix of the estimate, $\widehat{\beta}$, that maximizes (A.4) can be estimated as follows:

$$\mathrm{cov}(\widehat{\beta})=\left[\sum_{i=1}^{M}\left\{\frac{R_i(\widehat{\beta})}{A_i(\widehat{\beta})}-\frac{C_i(\widehat{\beta})C_i^{\mathrm{T}}(\widehat{\beta})}{A_i^2(\widehat{\beta})}\right\}\right]^{-1}, \qquad (A.5)$$

where if $m_\beta(G,E,\beta)$ and $m_{\beta\beta}(G,E,\beta)$ are the vector and matrix of first and second partial derivatives of $m(G,E,\beta)$ with respect to β, then

$$\begin{aligned}
A_i(\beta)&=\sum_{(g'_1,g'_0)\in\mathcal{H}_{G_i}}\left[\sum_{j=0}^{1}\exp\left\{m(g'_j,E_{ij},\beta)\right\}\right]\mathrm{pr}(G_{i1}=g'_1,G_{i0}=g'_0|\mathcal{G}_i);\\
C_i(\beta)&=\sum_{(g'_1,g'_0)\in\mathcal{H}_{G_i}}\left[\sum_{j=0}^{1}m_\beta(g'_j,E_{ij},\beta)\exp\left\{m(g'_j,E_{ij},\beta)\right\}\right]\mathrm{pr}(G_{i1}=g'_1,G_{i0}=g'_0|\mathcal{G}_i);\\
R_i(\beta)&=\sum_{(g'_1,g'_0)\in\mathcal{H}_{G_i}}\left[\sum_{j=0}^{1}m_\beta(g'_j,E_{ij},\beta)m_\beta^{\mathrm{T}}(g'_j,E_{ij},\beta)\exp\left\{m(g'_j,E_{ij},\beta)\right\}\right]\\
&\quad\times\mathrm{pr}(G_{i1}=g'_1,G_{i0}=g'_0|\mathcal{G}_i).
\end{aligned}$$

To show (A.5), an alternative formulation of (A.4) is useful. No longer insisting that D_{i1} is the case, so that $(D_{i1},D_{i0})=(1,0)$ or $(0,1)$, equation (A.4) actually shows that

$$\begin{aligned}&\mathrm{pr}(D_{i1} = d_1, D_{i0} = d_0, G_{i1} = g_1, G_{i0} = g_0 | D_{i1} + D_{i0} = d_1 + d_0 = 1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)\\ &= \frac{\exp\{d_1 m(g_1, E_{i1}, \beta) + d_0 m(g_0, E_{i0}, \beta 1)\}\mathrm{pr}(G_{i1} = g_1, G_{i0} = g_0 | \mathcal{G}_i)}{A_i(\beta)}.\end{aligned} \tag{A.6}$$

This means that the derivative with respect to β of the loglikelihood for the i^{th} observation is

$$\ell_i(\beta) = -C_i(\beta)/A_i(\beta) + D_{i1} m_\beta(G_{i1}, E_{i1}, \beta) + D_{i0} m_\beta(G_{i0}, E_{i0}, \beta). \tag{A.7}$$

It is easily seen that the Hessian is

$$\begin{aligned}\ell_{i,\beta}(\beta) = &-\frac{R_i(\beta)}{A_i(\beta)} + \frac{C_i(\beta)C_i^{\mathrm{T}}(\beta)}{A_i^2(\beta)} + D_{i1} m_{\beta\beta}(G_{i1}, E_{i1}, \beta) + D_{i0} m_{\beta\beta}(G_{i0}, E_{i0}, \beta)\\ &- \frac{\sum_{(g_1', g_0') \in \mathcal{H}_{\mathcal{G}_i}} \left[\sum_{j=0}^{1} m_{\beta\beta}(g_j', E_{ij}, \beta) \exp\left\{m(g_j', E_{ij}, \beta)\right\}\right] \mathrm{pr}(G_{i1} = g_1', G_{i0} = g_0' | \mathcal{G}_i)}{A_i(\beta)}.\end{aligned}$$

Using (A.6), it is easy to see that the expectation of the sum of the last three terms conditional on $(D_{i1} + D_{i0} = 1, \mathcal{G}_i, E_{i1}, E_{i0}, F_i)$ equals zero, and hence that the expected Fisher information for the i^{th} observation is $R_i(\beta)/A_i(\beta) - C_i(\beta)C_i^{\mathrm{T}}(\beta)/A_i^2(\beta)$, thus proving (A.5).

ASYMPTOTIC EFFICIENCY THEORY

Both $\mathcal{L}_{i1}$ and $\mathcal{L}_{i2}$ are conditional distributions depending on the parameter β. Hence the derivatives of their logarithms $\mathcal{L}_{i1,\beta}(\beta)$ and $\mathcal{L}_{i2,\beta}(\beta)$ behave as if they are log-likelihood scores, and when summed over the data each have information matrices $I_1(\beta)$ and $I_2(\beta)$, respectively. The claim that our method is asymptotically more efficient than conditional logistic regression then follows if we can show that $E\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\} = 0$. To see this, note that by their definitions as derivatives of logarithms of conditional probabilities, and making the rare disease assumption, we have that

$$0 = E\{\mathcal{L}_{i1,\beta}(\beta) | D_{i1} + D_{i0} = 1, G_{i1}, G_{i0}, \mathcal{G}_i, E_{i1}, E_{i0}, F_i\}. \tag{A.8}$$

Of course,

$$E\left\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\right\} = E\left[E\left\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta) | D_{i1} + D_{i0} = 1, G_{i1}, G_{i0}, \mathcal{G}_i, E_{i1}, E_{i0}, F_i\right\}\right],$$

where the interior expectation is with respect to the distribution of (D_{i1}, D_{i0}) given $D_{i1} + D_{i0} = 1$. However, $\mathcal{L}_{i2,\beta}(\beta)$ is a function only of $(D_{i1} + D_{i0} = 1, G_{i1}, G_{i0}, G_i, E_{i1}, E_{i0}, F_i)$, and does not depend otherwise on (D_{i1}, D_{i0}), so that by (A.8),

$$E\left\{\mathcal{L}_{i1,\beta}(\beta)\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\right\} = E\left[E\{\mathcal{L}_{i1,\beta}(\beta) | D_{i1} + D_{i0} = 1, G_{i1}, G_{i0}, \mathcal{G}_i, E_{i1}, E_{i0}, F_i\}\mathcal{L}_{i2,\beta}^{\mathrm{T}}(\beta)\right] = 0.$$

This verifies the claim.

SANDWICH VARIANCE ESTIMATOR FOR PAIRWISE PSEUDO-LIKELIHOOD

Let $l_{(jk)_i}(\beta) = \partial \log \mathcal{L}_{(jk)_i,*}/\partial\beta$ and $V_{(jk)_i}(\beta) = \partial \log \mathcal{L}_{(jk)_i,*}/\partial\beta\partial\beta^T$ denote the vector of first-derivatives and matrix of second-derivatives, respectively, of the log-pseudo-likelihood for the $(jk)^{\text{th}}$ pair of the i^{th} family. The covariance matrix of $\widehat{\beta}$ then can be estimated by the sandwich estimator defined as follows:

$$\mathrm{cov}|(\widehat{\beta}) = Q_3^{-1}(\widehat{\beta})Q_4(\widehat{\beta})Q_3^{-1}(\widehat{\beta}); Q_3(\beta) = \sum_{i=1}^{M} \sum_{j \in D_{i0}, k \in D_{i1}} V_{(jk)_i}(\beta);$$

$$Q_4(\beta) = \sum_{i=1}^{M} \left\{\sum_{j \in D_{i0}, k \in D_{i1}} \ell_{(jk)_i}(\beta)\right\}\left\{\sum_{j \in D_{i0}, k \in D_{i1}} \ell_{(jk)_i}(\beta)\right\}^{\mathrm{T}};$$

Chapter 4
Nonparametric and Semiparametric Regression for Independent Data

By Hua Liang

About the Author. Hua Liang received his PhD in Statistics in 2001 from Texas A&M University under the direction of Professor Raymond J. Carroll. He is now Professor of Statistics at George Washington University. American Statistical Association, International Mathematical Statistics, and among his honors, he is an Elected Member of the International Statistical Institute, Fellow of the Royal Statistical Society, and Alexander von Humboldt Fellow. He has published two books and more than 120 journal articles. He has served as Associate Editor of *Journal of the American Statistical Association*, *Biostatistics*, *Electronic Journal of Statistics*, and *Journal of Nonparametric Statistics*. Hua is indebted to Ray for his long-term mentoring, help, and encouragement.

Selected Papers on Nonparametric and Semiparametric Regression for Independent Data

[NSRI-1]-[179] Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Annals of Statistics*, 10, 1224–1233.

[NSRI-2]-[49] Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, 53, 573–585.

[NSRI-3]-[257] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92, 477–489.

[NSRI-4]-[103] Carroll, R. J., Ruppert, D. and Welsh, A. (1998). Local estimating equations. *Journal of the American Statistical Association*, 93, 214–227.

[NSRI-5]-[145] Ruppert, D. and Carroll, R. J. (2000). Spatially adaptive penalties for spline fitting. *Australia and New Zealand Journal of Statistics*, 42, 205–223.

Consider the linear model

$$y_i = \mathbf{x}_i^T\beta + \sigma_i\varepsilon_i, i = 1, \cdots, n,$$

where β is an unknown parameter vector and the $\{\varepsilon_i\}$ are i.i.d. errors. It is well known that ordinary least squares (LS) estimators are unbiased and consistent, but are not efficient when errors are heteroscedastic, and the usual standard error estimators of LS estimators are biased. Hence the usual confidence intervals and test statistics are biased and may lead to incorrect conclusions. Weighted LS estimators with weights that are inversely proportional to the σ_i^2 yield the most efficient estimators of β. Because the σ_i^2 are rarely available in real applications, an immediate question is how to estimate them.

Prior to Carroll (1982 [NSRI-1]), it was generally assumed that $\sigma_i^2 = H(\mathbf{x}_i, \theta)$ or $H(\mathbf{x}_i^T\theta)$, where H is a known parametric function with an unknown parameter

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_4,

vector θ. One therefore needs only to estimate the parameter vector θ. The parametric model $H(\cdot)$ might be misspecified. The parametric variance estimator may be also sensitive to outliers. As shown by Carroll and Ruppert (1988), potential outliers are likely to have a negative impact on estimation of β. It is therefore of substantial interest to develop robust estimation for the variance function.

Carroll (1982 [NSRI-1]) filled this gap by extending the parametric variance function $H(\mathbf{x}_i, \theta)$ to a nonparametric/semiparametric function as $\sigma_i^2 = H(\mathbf{c}_i)$ with $\mathbf{x}_i = (1, \mathbf{c}_i)^T$ and $\mathbf{c}_i$ being a scalar, or $\sigma_i^2 = H(\mathbf{x}_i^T \beta)$, where $H(\cdot)$ is an unknown and smooth function. He used nonparametric regression to estimate $H(\cdot)$ and, most importantly, theoretically proved that one can still get full efficiency for estimating linear regression parameters, that is, the corresponding estimators of β are adaptive (Stein, 1956; Bickel, 1982). Its numerical superiority over its competitors was demonstrated by Matloff, Rose, and Tai (1984). After 30 years, one may feel that this idea is simple and the results are trivial. However, at that time, the associate editor did not believe the results and insisted that the author provided a detailed proof. The paper was written before the Latex era, and hence the proof was written by hand for over 100 pages.

Since this pioneer work, a lot of research has been done in this direction. For example, Müller and Stadtmüller (1987) studied nonparametric regression with heteroscedastic errors; Davidian and Carroll (1987) developed a general theory for variance function estimation with emphasis on estimation of structural parameters. Robinson (1988) investigated the effects of heteroscedasticity in semiparametric models. More recently, Ma, Chiou, and Wang (2006) studied efficient estimation for heteroscedastic semiparametric models.

A second area that Ray has made a pioneer contribution is nonparametric and semiparametric regression with covariates measured with errors. For parametric measurement error modeling, various methods have been developed to correct bias for measurement error, and several consistent estimators of the parameters of interest have been derived when validation data are available. Carroll and Wand (1991 [NSRI-2]) considered semiparametric estimation and inference in a logistic measurement error model: $P(Y = 1|X = x) = G(\beta_0 + \beta_1 x)$ with $G(t) = \{1 + \exp(-t)\}^{-1}$. One does not observe X, but instead observes its surrogate W. The likelihood method or its variations were traditionally applied to estimate β_0 and β_1. Consider the $Y|W$ model

$$P(Y = 1|W) = \int G(\beta_0 + \beta_1 x) f_{x|w}(x|w) dx, \tag{4.1}$$

where $f_{x|w}(x|w)$ is the conditional density of X given W. If this density can be parameterized with a nuisance parameter ξ, one may use validation data to estimate this nuisance parameter, and plug the estimated value in $f_{x|w}(x|w)$. Then $(\beta_0, \beta_1)^T$ can be estimated using (4.1). However, this parametric method is difficult to implement and has poor numerical performance for moderate sample sizes or when the conditional density $f_{x|w}(x|w)$ is misspecified.

Carroll and Wand (1991 [NSRI-2] proposed to estimate the probability $P(Y = 1|W)$ by nonparametric regression using the validation data. Suppose there are n_1

observations of the validation data $\{(Y_i, X_i, W_i); i = 1, \cdots, n_1\}$, and n_2 observations of the primary data $\{(Y_i, W_i); i = 1 + n_1, \cdots, n_1 + n_2\}$. Write

$$\begin{aligned}
B(x,y,\beta) &= G^y(\beta_0 + \beta_1 x)\{1 - G(\beta_0 + \beta_1 x)\}^{1-y}, \\
\hat{f}_W(w) &= \frac{1}{n_1 h}\sum_{i=1}^{n_1} K_h(w - W_i), \\
D_n(y,w,\beta) &= \frac{1}{n_1 h}\sum_{i=1}^{n_1} B(X_i, y, \beta) K_h(w - W_i), \\
C_n(y,w,\beta) &= \frac{1}{n_1 h}\sum_{i=1}^{n_1} B_\beta(X_i, y, \beta) K_h(w - W_i),
\end{aligned}$$

where B_β is the derivative of $B(x,y,\beta)$ with respect to β, $K_h(\cdot) = K(\cdot/h)$ with $K(\cdot)$ being a kernel function and h being a bandwidth. Then they estimated β by solving

$$n^{-1/2}\sum_{j=1}^{n_1} U(Y_j, X_j, \beta) + n^{-1/2}\sum_{i=n_1+1}^{n_1+n_2} \frac{D_n(Y_i, W_i, \beta)}{C_n(Y_i, W_i, \beta)} = 0,$$

where $U(y,x,\beta)$ is the likelihood score for $(Y,X) = (y,x)$. Under mild conditions, the resulting estimators were shown to be asymptotically normal and numerically superior to their alternatives such as the estimators developed by Rosner et al. (1989) and a modification of the Stefanski and Carroll (1985) estimators in most situations.

A third area that Ray has made a major contribution is estimation in semiparametric single index models. Carroll et al. (1997 [NSRI-3]) studied generalized partially linear single-index models (GPLSiM) of the form:

$$g\{\mu(\mathbf{x},\mathbf{z})\} = \eta_0(\alpha_0^T\mathbf{x}) + \beta_0^T\mathbf{z}, \text{ with } \|\alpha_0\| = 1, \tag{4.2}$$

where $\mu(\mathbf{x},\mathbf{z}) = E(Y|\mathbf{x},\mathbf{z})$, $g(\cdot)$ is a known link function, $\eta_0(\cdot)$ is an unknown smooth function, α_0 and β_0 are $p \times 1$ and $q \times 1$ parameter vectors of primary interest. Model (4.2) presents a novel and very general structure for the conditional mean of Y given $\mathbf{x}, \mathbf{z}$, and is flexible to cover various commonly used models. For instance, if $g(\cdot)$ is an identity function, (4.2) reduces to partially linear single index models (Liang et al., 2010; Wang et al., 2010); furthermore, if $\beta_0 \equiv 0$, (4.2) reduces to single index models (Ichimura, 1993). More generally, when $\beta_0 = 0$, (4.2) is simply a generalized linear model with an *unknown* link function (Weisberg and Welsh, 1994). If $\mathbf{x}$ is scalar, (4.2) is a generalized partially linear model (Severini and Staniswalis, 1994), which can be further simplified to popular partially linear models when the link function is identical. For more details, see Wahba (1984), Engle et al. (1986), Speckman (1988), Mammen and van de Geer (1997), Liang, Wang, and Carroll (2007). Härdle, Liang, and Gao (2000) give a fairly comprehensive survey for partially linear models.

Carroll et al. (1997 [NSRI-3]) suggested the use of the semiparametric profile likelihood principle to estimate α_0, β_0 and $\eta_0(v)$ by locally approximating $\eta_0(v)$ by a linear function $\eta_0(v) \approx \eta_0(u) + \eta_0'(u)(v-u) \equiv a + b(v-u)$, for v in a neighborhood of u, where $a = \eta_0(u)$ and $b = \eta_0'(u)$. Let K be a symmetric probability density function, and $K_h(t) = K(t/h)/h$ be a rescaling of K. Give an initial estimator $(\hat{\alpha}_1, \hat{\beta})$ and set $\hat{\alpha} = \hat{\alpha}_1/\|\hat{\alpha}_1\|$. Denote by $Q(w,y)$ the quasilikelihood function (Severini and Staniswalis, 1994). The estimation procedure for estimating α_0, β_0 and $\eta_0(\cdot)$ is described in the following algorithm.

Step 1: Find $\hat{\eta}(u; h, \hat{\alpha}, \hat{\beta}) = \hat{a}$ by maximizing the local quasilikelihood

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{a + b(\hat{\alpha}^T\mathbf{X}_i - u) + \hat{\beta}^T\mathbf{Z}_i\right\}, Y_i\right] K_h(\hat{\alpha}^T\mathbf{X}_i - u), \tag{4.3}$$

with respect to a and b.

Step 2: Update $(\hat{\alpha}, \hat{\beta})$ by maximizing

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{\hat{\eta}(\alpha^T\mathbf{X}_i; h, \hat{\alpha}, \hat{\beta}) + \beta^T\mathbf{Z}_i\right\}, Y_i\right], \tag{4.4}$$

with respect to α and β.

Step 3: Continue Steps 1 and 2 until convergence.

Step 4: Fix (α, β) at its estimated value from Step 3. The final estimate of $\eta_0(\cdot)$ is $\hat{\eta}(u; h, \hat{\alpha}, \hat{\beta}) = \hat{a}$ where $(\hat{a}, \hat{b})$ is obtained by maximizing (4.3).

The resulting estimators are called fully iterated estimators. They also proposed an alternative algorithm for implementation simplicity, which was referred to as the one-step estimator; i.e., given value $\hat{\alpha}$, find $\hat{\eta}(u; h, \hat{\alpha}) = \hat{a}$ by maximizing the local quasilikelihood

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{a + b(\hat{\alpha}^T\mathbf{X}_i - u) + \beta^T\mathbf{Z}_i\right\}, Y_i\right] K_h(\hat{\alpha}^T\mathbf{X}_i - u), \tag{4.5}$$

with respect to a, b, and β. Given $\hat{\alpha}$ and the estimator $\hat{\eta}(u; h, \hat{\alpha})$ one estimates $\hat{\beta}$ by maximizing

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{\hat{\eta}(\hat{\alpha}^T\mathbf{X}_i; h, \hat{\alpha}) + \beta^T\mathbf{Z}_i\right\}, Y_i\right]. \tag{4.6}$$

Under general conditions, the fully iterated estimators of α_0 and β_0 were shown to be semiparametrically efficient, while the one-step estimators are not efficient and need undersmoothing but is easier to implement. This is a substantial addition to the semiparametric profile likelihood literature. Since then, semiparametric profile likelihood methods have been developed in various partial linear model settings, for example, Lin and Carroll (2001, 2006) for longitudinal data, Li and Liang (2008) and Liang et al. (2010) for variable selection in semiparametric models.

Generalized estimating equations (GEE) (Diggle et al., 2002) is a popular technique for modeling longitudinal data because it relies on specification of only the first two moments, and under mild regularity conditions, parameter estimators using GEEs are consistent. Suppose there are independent observations $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots \mathbf{Y}_n)$, with the $\mathbf{Y}$'s possibly vector-valued. Let $\psi(\mathbf{Y}, \theta)$ be a parametric estimating function for θ that has the same dimension as θ and satisfies $E\psi(\mathbf{Y}, \theta) = 0$, where θ is a parametric vector of interest. Then the GEE estimators of θ can be obtained by solving the estimating equation

$$0 = \sum_{i=1}^{n} \psi(\mathbf{Y}_i, \theta). \tag{4.7}$$

Carroll, Ruppert, and Welsh (1998 [NSRI-4]) extended the parametric model to allow θ to depend on a predictor nonparametrically as $\theta(x)$. They used local polynomials to approximate $\theta(x)$; that is, $\theta(x) \approx \sum_{j=0}^{p} \mathbf{b}_j (x - x_0)^j$ in a neighborhood of x_0 for some $p \geq 0$, where $\mathbf{b}_j = \theta^{(j)}(x_0)/j!$. They proposed to solve for $(\mathbf{b}_0, \ldots, \mathbf{b}_p)$ using $q \times (p+1)$ equations

$$0 = \sum_{i=1}^{n} w_h(X_i, x_0)[\mathbf{G}_p(X_i - x_0) \otimes \psi\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_j (X_i - x_0)^j\}], \tag{4.8}$$

where $\mathbf{G}_p^T(v) = \left(1, v, v^2, \ldots, v^p\right)$, $w_h(x, x_0)$ is a local weight, which can be a kernel function, with h being a tuning constant. The final estimates are $\hat{\theta}(x_0) = \hat{\mathbf{b}}_0$. This framework covers most general estimation methods. For example, in ordinary nonparametric regression, the response is $\mathbf{Y} = Y$ and $\psi(\mathbf{Y}, \mathbf{v}) = Y - v$; in generalized linear models with the mean function $\mu(x)$, and variance proportional to $V(x)$, the response is $\mathbf{Y} = Y$, and $\psi(\mathbf{Y}, \theta) = \{Y - \mu(\theta)\}\mu^{(1)}(\theta)/V(\theta)$, where $\mu^{(1)}(\theta) = (\partial/\partial\theta)\mu(\theta)$; and in varying coefficient generalized linear models, $\mathbf{Y} = (x, Y, Z)$, $\theta = (\theta_0, \theta_1^T(x))^T$ and the mean function is $\mu(\theta_0 + \theta_1^T(x)\mathbf{Z})$. The estimating function is $\psi(\mathbf{Y}, \theta) = \left[\{Y - \mu(\theta_0 + \theta_1^T(x)\mathbf{Z})\}\mu^{(1)}(\theta_0 + \theta_1^T(x)\mathbf{Z})/V(\theta_0 + \theta_1^T(x)\mathbf{Z})\right]$ $(1, \mathbf{Z})^T$.

It is worth mentioning that this approach ignores within-subject correlation; i.e., the resulting estimator is equivalent to that given by Lin and Carroll (2000) with the working independence structure. The idea of local estimating equation has been further applied to model longitudinal data with semiparametric models. See, for example, Lin and Carroll (2001, 2006).

The fourth area Ray has made a major contribution is the use of penalized splines (P-splines) for estimation in semiparametric models. Consider the nonparametric model $Y_i = m(X_i) + e_i$ for $i = 1, \cdots, n$, where X_i is univariate. The basic idea of a P-spline (Eilers and Marx, 1996) is to approximate a nonparametric function $m(\cdot)$ by a regression spline with a moderate number of knots and estimate $m(\cdot)$ by incorporating penalties of regression coefficients. Specifically, let $m(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - \zeta_k)_+^p$, where $p \geq 1$ is an integer and $\zeta_1 < \cdots < \zeta_K$ are fixed knots, $a_+ = \max(a, 0)$. Estimators $\hat{\beta}(\lambda)$ of β are defined as the minimizer of

$$\sum_{i=1}^{n} \{Y_i - m(X_i;\beta)\}^2 + \lambda \sum_{k=1}^{K} b_k^2, \tag{4.9}$$

where λ is a smoothing parameter. As a consequence, $m(x)$ is estimated by $m(x;\hat{\beta})$. Ruppert and Carroll (2000 [NSRI-5]) creatively modified the penalty term in (4.9) to $\sum_{k=1}^{K} \lambda(\kappa_k) b_k^2$, where $\lambda(\cdot)$ is a penalty function with knots at $\kappa_1, \cdots, \kappa_K$ rather than a constant (Eilers and Marx, 1996). This generalization makes P-splines perform well for functions that rapidly oscillate in some regions, and smooth in other regions. The resulting P-spline estimators are competitive with regression splines that adaptively select knots especially for regression functions with significant spatial inhomogeneity. An attractive feature of P-splines is that they reduce computational complexity by using a smaller number of knots compared to smoothing splines especially when used to fit additive models, for which the backfitting algorithm is generally used. They are also more flexible than regression splines and are less sensitive to knot allocation.

We expect that this local penalty spline can be implemented through fitting a linear mixed model (Laird and Ware, 1982). As shown by Brumback, Ruppert, and Wand (1999), $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon$, where $\mathbf{X}$ is an $n \times (p+1)$ matrix whose jth column is $(X_1^{j-1}, \cdots, X_n^{j-1})^T$ for $1 \leq j \leq p+1$, and $\mathbf{Z}$ is an $n \times K$ matrix whose jth column is $\{(X_1 - \zeta_j)_+^p, \cdots, (X_n - \zeta_n)_+^p\}^T$ for $1 \leq j \leq K$. Selection of $\lambda(\cdot)$ is equivalent to selecting the covariance structure of the random efforts $\mathbf{b}$, and calculations can be easily implemented through use of linear mixed model functions available in Splus/R and SAS. Ruppert, Wand, and Carroll (2003) provide a comprehensive survey on P-splines for semiparametric models.

References

Other publications by Ray Carroll cited in this chapter.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1091.

Liang, H., Wang, S,. and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika 94*, 185–198.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, 95. 520–534.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96, 1045–1056.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, 68, 69–88.

Ruppert, D., Wand, M. P., and Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Publications by other authors cited in this chapter.

Bickel, P. J. (1982). On adaptive estimation. *Annals of Statistics*, 10, 647–671.

Brumback, B. A., Ruppert, D., and Wand, M. (1999). Comments on 'variable selection and function estimation in additive nonparametric regression using a data-based prior' by Shively, Kohn, and Wood. *Journal of the American Statistical Association*, 94, 794–797.

Diggle, P., Heagerty, P., Liang, K., and Zeger, S. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford: Oxford University Press.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties (with comments). *Statistical Science*, 11, 89–121.

Engle, R., Granger, C., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81, 310–320.

Härdle, W., Liang, H., and Gao, J. T. (2000). *Partially Linear Models*. Heidelberg: Springer Physica.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of singleindex models. *Journal of Econometrics*, 58, 71–120.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Annals of Statistics*, 36, 261–286.

Liang, H., Liu, X., Li, R., and Tsai, C. L. (2010). Estimation and testing for partially linear single index models. *Annals of Statistics*, 38, 3811–3836.

Ma, Y., Chiou, J.-M., and Wang, N. (2006). Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika* 93, 75–84.

Mammen, E. and van de Geer, S. (1997). Penalized estimation in partial linear models. *Annals of Statistics*, 25, 1014–1035.

Matloff, N., Rose, R., and Tai, R. (1984). A comparison of two methods for estimating optimal weights in regression analysis. *Journal of Statistical Computation and Simulation*, 19, 265–274.

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics*, 15, 610–625.

Robinson, P. M. (1988). Root n-consistent semiparametric regression. *Econometrica* 56, 931–954.

Rosner, B., Willett, W. C. and Spiegelmann, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8, 1051–1069.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89, 501–511.

Speckman, P. E. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, 50, 413–436.

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, 13, 1335–1351.

Stein, C. (1956). Efficient nonparametric testing and estimation. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I.* Berkeley and Los Angeles: University of California Press, pp. 187–195.

Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Statistical Analyses for Time Series*. Tokyo: Institute of Statistical Mathematics, pp. 319–329.

Wang, J.-L., Xue, L., Zhu, L., and Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *Annals of Statistics*, 38, 246–274.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link. *Annals of Statistics*, 22, 1674–1700.

The Annals of Statistics
1982, Vol. 10, No. 4, 1224-1233

ADAPTING FOR HETEROSCEDASTICITY IN LINEAR MODELS[1]

BY RAYMOND J. CARROLL

University of North Carolina at Chapel Hill

In a heteroscedastic linear model, it is known that if the variances are a parametric function of the design, then one can construct an estimate of the regression parameter which is asymptotically equivalent to the weighted least squares estimate with known variances. We show that the same is true when the only thing known about the variances is that they are determined by an unknown but smooth function of the design or the mean response.

1. Introduction. We are interested in efficient regression parameter estimation in a heteroscedastic linear model given by

$$Y_{ij} = x_i'\beta + \sigma_i\varepsilon_{ij},\ i = 1, \cdots, n, j = 1, \cdots, m_i, \Sigma m_i = N. \tag{1.1}$$

Here Y_{ij} is the response of the jth replicate at the design point x_i (a p-vector), β is the unknown regression parameter of interest, $\{\sigma_i\}$ express the heteroscedasticity in the model and $\{\varepsilon_{ij}\}$ are i.i.d. with variance one and distribution function F assumed symmetric about zero but otherwise unknown. Theoretical analysis of the model (1.1) has traditionally fallen into one of the two areas we describe below.

The parametric approach generally assumes

$$\sigma_i^2 = H(x_i, \theta) \quad \text{or} \quad H(x_i'\beta, \theta), \quad H \text{ known}. \tag{1.2}$$

See Hildreth and Houck (1968), Froehlich (1973), Dent and Hildreth (1977), Box and Hill (1974), Jobson and Fuller (1980) and Carroll and Ruppert (1982). Once a parametric assumption such as (1.2) is made, one computes estimates of the $r \times 1$ unknown parameter θ, next estimates

$$\hat{\sigma}_i^2 = H(x_i, \hat{\theta}) \quad \text{or} \quad H(x_i'\hat{\beta}, \hat{\theta}),$$

as appropriate, and then constructs a weighted estimate of β. If we denote the weighted least squares estimate based on the true weights by $\hat{\beta}_T$ and the weighted estimate based on the estimates $\{\hat{\sigma}_i\}$ by $\hat{\beta}_E$, we get a well-known result:

RESULT 1. For the parametric approach, in large samples there is no cost due to estimating $\{\sigma_i\}$, i.e., $\hat{\beta}_T$ and $\hat{\beta}_E$ have the same limiting normal distribution.

This result is proved rigorously and extended to robust estimation by Carroll and Ruppert (1982); Carroll (1982) shows it holds even if the dimension p of β increases with N, e.g., $p^2/N \to 0$ generally suffices. See also Williams (1975).

The nonparametric approach differs quite radically. Here,

$$\sigma_i^2 = H(x_i) \quad \text{or} \quad H(x_i'\beta), \quad H \text{ unknown}. \tag{1.3}$$

Since $H(\cdot)$ is assumed completely unknown, the standard method is to get information about $H(\cdot)$ by replication ($m_i > 1$). Fuller and Rao (1978) consider the situation often seen

Received May 1981; revised April 1982.

[1] University of North Carolina at Chapel Hill, Department of Statistics. This work was supported by the Air Force Office of Scientific Research under Contract AFOSR-80-0080.

AMS 1970 *subject classification.* Primary 62J05; secondary 62G35.

Key words and phrases. Linear models, regression, heteroscedasticity, nonparametric regression.

in practice that the number of design points $n \to \infty$, but each m_i stays bounded. Their method is to fit least squares estimates ($\hat{\beta}_L$) to the data, compute predicted values ($t_i = x_i'\hat{\beta}_L$) and residuals ($r_{ij} = Y_{ij} - t_i$) and estimate

$$\hat{\sigma}_i^2 = \frac{1}{m_i} \textstyle\sum_j r_{ij}^2. \tag{1.4}$$

With these estimates one then performs weighted least squares, obtaining what we shall denote by $\hat{\beta}_{FR}$. By delicate and very interesting calculations, they obtain an important result which had not been previously known or appreciated:

RESULT 2. In the nonparametric approach, there is a cost due to not knowing $\{\sigma_i\}$, i.e., $\hat{\beta}_T$ and $\hat{\beta}_{FR}$ have different limiting distributions.

We explore here the possibility of closing the wide gap between Results 1 and 2, at least in an asymptotic sense. Specifically, we explore methods for which the nonparametric approach (1.3) is used but for which Result 1 obtains. In other words, we will show that situations exist in which nothing specific is known about the variance function, but estimation of β can be done asymptotically as well as if the variance function were completely known.

A key feature of many—but as Fuller and Rao (1978) note, not all—heteroscedastic regression problems is that the variances appear to be smooth functions of the design or mean response; we use the term "smooth" loosely here, but generally will mean that the variance function $H(\cdot)$ has a continuous first derivative. This smoothness suggests that if x_1 and x_2 are very close, so too should be $H(x_1)$ and $H(x_2)$. This suggests that information about $H(x_1)$ can be obtained from data at x_2. Hence, we will study the nonparametric models

$$\sigma_i^2 = H(x_i) \quad \text{or} \quad H(x_i'\beta), \quad H \text{ unknown but smooth.} \tag{1.5}$$

This approach of sharing information contrasts with that of the nonparametric method (1.3)—(1.4), which only uses data at x_1 to estimate $H(x_1)$. By sharing information we should now get good consistent estimates of $H(\cdot)$, which enables us in certain circumstances to get better estimates of β for which Result 1 holds.

We specifically consider only two cases. In Section 2 we discuss simple linear regression, while in Section 3 we assume that the variance is a smooth function of the mean. The technical details are not trivial and the notation is rather messy, but the basic idea is simple and can be described as follows. Under the second part of (1.5) for example, we have

$$E(Y_{ij} - x_i'\beta)^2 = H(x_i'\beta).$$

Thus, for the residuals r_{ij} we have

$$Er_{ij}^2 = E(Y_{ij} - x_i'\hat{\beta}_L)^2 \approx H(x_i'\beta). \tag{1.6}$$

Equation (1.6) puts us in the realm of nonparametric regression of squared residuals on a function $H(\cdot)$. Even if one goes no further, there is already a huge literature which can be exploited to define nonparametric regression estimates (Watson, 1964; Rosenblatt, 1969; Stone, 1977; Mack and Silverman, 1980; Johnston, 1982); this we do. If one goes further and makes the often reasonable assumption that $H(\cdot)$ is monotone, isotonic regression could be used (Wright, 1978). Such isotonic estimates should work well in practice but we have been unable to develop a theory for them.

Throughout this paper, x and β refer to p-vectors, while α and c are scalars. For example, the heteroscedastic simple linear regression model is, with $x_i' = (1, c_i)$ and $\beta' = (\alpha_0, \alpha_1)$,

$$Y_{ij} = x_i'\beta + \sigma_i\varepsilon_{ij} = \alpha_0 + \alpha_1 c_i + \sigma_i\varepsilon_{ij}. \tag{1.7}$$

NOTE ADDED IN PROOF. After this paper was accepted for publication, I was informed by Professor N. Matloff (Department of Statistics and Electrical and Computer Engineering, University of California at Davis) that in 1978 his student Dr. Robin Lawrence Rose, in an unpublished dissertation, proposed methods of estimation similar to those investigated here, and performed Monte-Carlo experiments for these methods.

2. Simple linear regression. We first consider simple linear regression. This is the only case for which we have been able to obtain results in which the variance is a function of the design alone, as would be the case in the random coefficient model of Hildreth and Houck (1978). In the next section, we discuss the situation in which the variance is a function of the mean response.

Thus, in this section, the model is given by (1.7), where

$$\sigma_i^2 = H(c_i), \quad H(\cdot) \text{ unknown.}$$

Much of the literature for the nonparametric regression problem assumes that the independent or predictor variables are themselves random. In order to make the most efficient presentation, we will follow this lead, making the assumption for model (1.7) that $\{\varepsilon_{ij}\}$ and $\{c_i\}$ are sets of i.i.d. random variables independent of one another. After the statement of Theorem 1, we will discuss the case that $\{c_i\}$ is a set of fixed constants. We will first present and discuss the assumptions, and then state the first result.

First, from Watson (1964) and Johnston (1982), a plausible kernel-type estimate of H is

$$\hat{H}_N(c) = \sum_{i=1}^n \sum_{j=1}^{m_i} r_{ij}^2 K\left(\frac{c_i - c}{b(N)}\right) \left\{\sum_{i=1}^n \sum_{j=1}^{m_i} K\left(\frac{c_i - c}{b(N)}\right)\right\}^{-1}. \tag{2.1}$$

The weighted estimate $\hat{\beta}_w$ is formed by setting

$$\hat{\sigma}_i^2 = \hat{H}_N(c_i)$$

and then performing weighted least squares.

In order that information about the scalar function $H(\cdot)$ can be shared and in order to avoid being subject to the Fuller and Rao Result 2, we need the design to be eventually dense in a set such as an interval. This will enable us to estimate $H(\cdot)$ uniformly well. One can do this under the following assumption.

ASSUMPTION 1. $\{x_i\}$ have density function f positive on its compact support $\mathscr{I}$. Further, on $\mathscr{I}$, f has two continuous derivatives.

Note that Assumption 1 is really designed for regression problems and not for factorial designs. Naturally, we also require that H be smooth:

ASSUMPTION 2. H and its first derivative are continuous on $\mathscr{I}$.

In order to make sure that no infinite weights occur in our weighted regression, we need

ASSUMPTION 3. H has a positive infimum on $\mathscr{I}$.

We also need some assumptions on the kernel $K(\cdot)$ and bandwidth $b(N)$ in (2.1).

ASSUMPTION 4. $K(\cdot)$ is a symmetric density function. It has compact support, three continuous derivatives, and its support includes an open set containing $\mathscr{I}$.

ASSUMPTION 5. The bandwidth $b(N)$ satisties $Nb(N)^4 \to 0$.

ASSUMPTION 6. The bandwidth $b(N)$ satisfies $N^{5/4}b(N)^4 \to \infty$.

Finally, we make assumptions relating to the uniqueness of the design; these are reasonably standard assumptions even in the parametric approach. Recall $x' = (1, c)$.

Assumption 7. $E(xx')$ and $E\{xx'H^{-1}(c)\}$ are positive definite.

Theorem 1. *Under Assumptions* 1–7 *and the condition that* $E\varepsilon_1^6 < \infty$, $\hat\beta_W$ *and* $\hat\beta_T$ *have the same normal limit distribution, with mean* β *and covariance* $N^{-1}Exx'H^{-1}(c)$.

The proof is in Section 5. Assumptions 5 and 6 probably can be weakened.

Remarks. The problem discussed in this section is a special case of one in which the variance is a function of the design and $p \geq 2$; when $p \geq 3$, H is a function of a vector argument. For larger values of p, the rate of convergence to H of the estimator (2.1) will be slower and the proof given in the appendix will break down, since Theorem 5.1 will not be true. We believe brute force (Taylor series) can be used to extend Theorem 1 to the case $p \geq 3$, but an alternative approach would be preferable.

Theorem 1 can be extended to the case where the predictor variables $\{c_i\}$ are fixed constants by assuming that they "act i.i.d." in all essential aspects; this is rather untidy. Alternatively, one could replace (2.1) by the Priestley-Chao estimator studied by Benedetti (1977). For this estimator, certain technical difficulties can be avoided because there is no random denominator term as in (2.1); the assumptions, however, remain basically unchanged with the exception of Assumption 1.

3. Variance a function of the mean. Here we consider the model (1.1) with

$$\sigma_i^2 = H(x_i'\beta) = H(\tau_i), \quad H \text{ unknown.} \tag{3.1}$$

The variance is often considered a function of the mean as in (3.1) because residual plots fall in a fan-shaped pattern; see Box and Hill (1974), Bickel (1978), Jobson and Fuller (1980) and Carroll and Ruppert (1982).

Note that

$$\tau_i = \text{true mean response} = x_i'\beta$$

and define the predicted values as

$$t_i = x_i'\hat\beta_L,$$

where $\hat\beta_L$ is least squares estimator. Following the same reasoning as in the previous section, the estimator of H becomes

$$\hat H_N(s) = \{Nb(N)\}^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m_i} r_{ij}^2 K\left(\dfrac{t_i - s}{b(N)}\right)\left[\{Nb(N)\}^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m_i} K\left(\dfrac{t_i - s}{b(N)}\right)\right]^{-1},$$

and the estimated variances are $\{\hat H_N(t_i)\}$.

Theorem 2. *Under the assumptions of Theorem* 1, *but in Assumption* 7 *replacing* $H(c)$ *by* $H(x'\beta)$, $\hat\beta_W$ *and* $\hat\beta_T$ *have the same normal limit distribution with mean* β *and covariance* $N^{-1}E\{xx'H^{-1}(x'\beta)\}$.

The proof is in Section 5.

4. A Monte-Carlo study. We performed a small Monte-Carlo experiment to see if the previous results make any sense even in an ideal situation. We took the model to be simple linear regression.

$$Y_i = \alpha_0 + \alpha_1 c_i + \sigma_i\varepsilon_i, \qquad i = 1, \cdots, N = 60. \tag{4.1}$$

Here $\{\varepsilon_i\}$ are standard normal random variables, $(\alpha_0, \alpha_1) = (50, 60)$, and $\{c_i\}$ are i.i.d.

uniform on the interval (−½, ½). The normal random numbers were generated by the IMSL routine GGNPM, while the uniform numbers used GGUBS. The number of Monte-Carlo simulations for each situation was 500.

We estimated the function H by $\hat{H}_N$ of (2.1), with

$$K(v) = \begin{cases} 3(1 - |v|)^2/2 & |v| \leq 1, \\ 0 & |v| \geq 1, \end{cases} \tag{4.2}$$

$$b(N) = 0.13. \tag{4.3}$$

The particular choice for $b(N)$ was arbitrary, although on average approximately 8 observations are used in constructing $\hat{H}_N$ at each design point. While $K(\cdot)$ does not strictly satisfy Assumption 5, it does have a continuous first derivative which should suffice. In Table 1, the weighted least squares estimate with weights generated by $\hat{H}_N$ is denoted NONPAR. The least squares estimate is LSE.

Three models for the variances were considered. The first, given by Jobson and Fuller (1980), is

$$\sigma_t^2 = a_1 + a_2\tau_t^2, \qquad \tau_t = \alpha_0 + \alpha_1 c_t. \tag{4.4}$$

For our simulations we chose $a_1 = 100$, $a_2 = 0.25$. Our second model is one of more severe heteroscedasticity,

$$\sigma_t = a_1 \exp(a_2 |\tau_t|), \tag{4.5}$$

where $a_1 = 0.25$ and $a_2 = 0.04$. This type of model is mentioned by Bickel (1978). The third model is one of severe heteroscedasticity

$$\sigma_t = a_1 \exp(a_2\tau_t^2), \tag{4.6}$$

where

$$(a_1, a_2) = (¼, 1/3200).$$

We also constructed a third estimator PARM based on the parametric model (4.4). Our intentions in doing this were (a) to see if the nonparametric estimate is at all reasonable when compared to an estimate based on a correct parametric model (4.4) for the variances, and (b) to see if the nonparametric estimate is more robust than the parametric estimate if the variance model is badly misspecified, i.e., (4.6) holds but estimation is done as if (4.4) holds.

TABLE 1

Results of the Monte-Carlo Study of Section 4 for the model $\tau_t = EY_t = 50 + 60\, c_t$, c_t *Uniform* (−½, ½). *The models for* $\mathrm{Var}(Y_t) = \sigma_t^2$ *are*: *Model 1* $\sigma_t^2 = a_1 + a_2\tau_t^2$, $(a_1, a_2) = (100, 0.25)$; *Model 2* $\sigma_t = a_1\exp(a_2|\tau_t|)$, $(a_1, a_2) = (0.25, 0.04)$; *Model 3* $\sigma_t = a_1\exp(a_2\tau_t^2)$, $(a_1, a_2) = (0.25, 1/3200)$.

Estimator	Variance Model	$\alpha_0 = 50$		$\alpha_1 = 60$	
		Bias	MSE*	Bias	MSE*
LSE	1	.052	13.27	.925	172.07
PARM	1	.040	13.27	.711	138.58
NONPAR	1	.066	14.80	.727	144.46
LSE	2	.016	12.87	.106	231.20
PARM	2	.011	11.45	.049	100.00
NONPAR	2	.007	10.18	.037	80.34
LSE	3	.004	10.98	.032	200.41
PARM	3	.003	9.85	.016	96.09
NONPAR	3	.002	9.19	.014	88.88

* *The actual* MSE *for Model* 2 *is the figure given divided by* 10^2, *while the figure for Model* 3 *should be divided by* 10^3.

The estimate PARM is constructed as follows:

(i) Define P as in Jobson and Fuller (1980).
(ii) Let $\hat{\beta}_L$ = LSE, with $\hat{\beta}'_L = (\hat{\alpha}_0, \hat{\alpha}_1)$.
(iii) Let r^2 be the vector of squared residuals, i.e. squares of $Y_i - x_i'\hat{\beta}_L$.
(iv) Minimize $(r^2 - P\hat{\mathbf{a}})'(r^2 - P\hat{\mathbf{a}})$ for $\hat{\mathbf{a}} \geq 0$, where $\hat{\mathbf{a}}' = (\hat{a}_1, \hat{a}_2)$
(v) Define $\hat{\sigma}_i^2 = \hat{a}_1 + \hat{a}_2(\hat{\alpha}_0 + \hat{\alpha}_1 c_i)^2$
(vi) Compute a weighted estimate $\hat{\beta}_p$ and residuals $r_{iw} = Y_i - x_i'\hat{\beta}_p = Y_i - \hat{\alpha}_{0p} - \hat{\alpha}_{1p}c_i$.
(vii) Repeat steps (iv) and (v), replacing $(\hat{\alpha}_0, \hat{\alpha}_1)$ by $(\hat{\alpha}_{0p}, \hat{\alpha}_{1p})$ in (v).
(viii) Recompute a weighted estimate, call it PARM.

The outcomes of the simulations are given in Table 1. The results are quite encouraging and suggest that there are instances where our nonparametric estimation of the variances can work well, particularly for larger sample sizes.

5. Proof. Because the details are lengthy, we sketch the proofs only for the case $m_i \equiv 1$. As a shorthand notation, identify Assumptions 1-7 as A1, A2, $\cdots$, A7. Consider first simple linear regression in Section 2. We have the following.

THEOREM 5.1. *If the supremum is taken over the support of the design $\mathscr{I}$ (assumed compact), then*

$$N^{1/4}\sup|\hat{H}_N(c) - H(c)| \to_p 0.$$

PROOF OF THEOREM 5.1. Rewrite (2.1) as $\hat{H}_N = \hat{G}_N/\hat{f}_N$ and

$$\begin{aligned}\hat{G}_N(c) &= \hat{G}_{N1}(c) - 2\hat{G}_{N2}(c) + \hat{G}_{n3}(c) \\ &= \{Nb(N)\}^{-1}\textstyle\sum_{i=1}^N [\sigma_i^2\varepsilon_i^2 - 2\sigma_i\varepsilon_i x_i'(\hat{\beta}_L - \beta) + \{x_i'(\hat{\beta}_L - \beta)\}^2]K\left(\dfrac{c_i - c}{b(N)}\right).\end{aligned}$$

Because both $K(\cdot)$ and the support of the design are bounded and $\hat{\beta}_L = \beta + O_p(N^{-1/2})$, we have

$$N^{1/4}\sup|\hat{G}_{N3}(c)| \to_p 0.$$

Routine but detailed weak convergence arguments using Theorem 12.3 of Billingsley (1968), A4-A7 and $E\varepsilon^6 < \infty$ can be used to show that

$$\begin{array}{ll} N^{1/4}\sup|\hat{G}_{N2}(c)| \to_p 0, & N^{1/4}\sup|\hat{f}_N(c) - E\hat{f}_N(c)| \to_p 0, \\ N^{1/4}\sup|\hat{G}_{N1}(c) - E\hat{G}_{N1}(c)| \to_p 0, & N^{1/4}\sup|E\hat{G}_{N1}(c)/E\hat{f}_N(c) - H(c)| \to 0. \end{array} \tag{5.1}$$

The first part of (5.1) is simple enough. It follows from direct weak convergence arguments after one shows that, from the central limit theorem, one can replace $\hat{\beta}_L - \beta$ by

$$\{E(xx')\}^{-1}N^{-1}\textstyle\sum_{i=1}^N x_i\sigma_i\varepsilon_i.$$

The second and third parts of (5.1) can be shown directly by weak convergence arguments, but they are also essentially known from the nonparametric regression literature. The fourth part of (5.1) is merely tedious algebra; one has to be quite careful with end points.

Now recall once again that the model for Theorem 1 is

$$Y_i = x_i'\beta + \sigma_i\varepsilon_i = \alpha_0 + \alpha_1 c_i + \sigma_i\varepsilon_i, \qquad \mathrm{Var}(\varepsilon_i) = 1.$$

PROOF OF THEOREM 1. First note because $E\varepsilon^2 < \infty$ and A7 holds, $N^{1/2}(\hat{\beta}_T - \beta)$ has the normal limit distribution claimed in Theorem 1. It thus suffices to show that

$$N^{1/2}(\hat{\beta}_w - \hat{\beta}_T) \to_p 0.$$

Recall, $\hat{\beta}_w$ is the weighted estimator based on the adaptive weights (2.1). Because $\hat{\beta}_T$ is asymptotically normal, the design is bounded, $H(c) > 0$ and $E\varepsilon^6 < \infty$, one can use Theorem 5.1 to see that it suffices to show that

$$N^{-1/2} \Sigma\, x_i \varepsilon_i \{\hat{H}_N(c_i) - H(c_i)\}/\sigma_i^3 \to_p 0, \tag{5.2}$$

where $x_i' = (1, c_i)$ as before. By the proof of Theorem 5.1 it suffices to show

$$N^{-1/2} \sum x_i \varepsilon_i [\hat{G}_N(c_i) - E\hat{G}_{N1}(c_i)]/\sigma_i^3 E\hat{f}_N(c_i) \to_p 0, \tag{5.3}$$

$$N^{-1/2} \sum x_i \varepsilon_i E\hat{G}_{N1}(c_i)[\hat{f}_N(c_i) - E\hat{f}_N(c_i)]/\sigma_i^3 (E\hat{f}_N(c_i))^2 \to_p 0, \tag{5.4}$$

$$N^{-1/2} \sum x_i \varepsilon_i \{E\hat{G}_{N1}(c_i)/E\hat{f}_N(c_i) - H(c_i)\}/\sigma_i^3 \to_p 0. \tag{5.5}$$

In the expressions above, $E\hat{G}_{N1}(c_i)$ refers to $E\hat{G}_{N1}(c)$ evaluated at $c = c_i$, and similarly for $E\hat{f}_N(c_i)$. We will only sketch (5.3) as (5.4) and (5.5) are much easier. Rewrite (5.3) as

$$\begin{aligned} &\{N^{3/2} b(N)\}^{-1} \textstyle\sum_i \sum_j \dfrac{x_i \varepsilon_i}{\sigma_i^3 E\hat{f}_N(c_i)} \left\{ \sigma_j^2 K\left(\dfrac{c_j - c_i}{b(N)}\right) - E\hat{G}_{N1}(c_i) \right\} \\ &\qquad + \{N^{3/2} b(N)\}^{-1} \textstyle\sum_i \sum_j \dfrac{x_i \varepsilon_i}{\sigma_i^3 E\hat{f}_N(c_i)} \\ &\qquad \cdot [\sigma_j^2(\varepsilon_j^2 - 1) - 2\sigma_j \varepsilon_j x_j'(\hat{\beta}_L - \beta) + \{x_j'(\hat{\beta}_L - \beta)\}^2] K\left(\dfrac{c_j - c_i}{b(N)}\right). \end{aligned} \tag{5.6}$$

Each term in (5.6) converges in probability to zero. The first term and the first part of the second term only require computing second moments, remembering that $\{x_i\}$ and $\{\sigma_i\}$ are uniformly bounded and noting that $\{E\hat{f}_N(c_i)\}$ are bounded away from zero. The third part of the second term is easy. For the second part of the second term, it suffices to prove the result when we replace $\hat{\beta}_L - \beta$ by

$$\{E(xx')\}^{-1} N^{-1} \textstyle\sum_{i=1}^N x_i \varepsilon_i \sigma_i. \tag{5.7}$$

Having done this, one then computes second moments. In these steps the full strength of the assumption $E\varepsilon^6 < \infty$ is used.

We next sketch the proof for Theorem 2. The first step is a version of Theorem 5.1. Recall that $\tau_i = x_i'\beta = EY_i$, $t_i = x_i'\hat{\beta}_L$. The definitions of $\hat{H}_N$ and $\hat{\beta}_w$ are given in Section 3, while $\hat{f}_N$ is the inverted term in the definition of $\hat{H}_N$.

THEOREM 5.2.

$$N^{1/4} \sup |\hat{H}_N(s) - H(s)| \to_p 0.$$

PROOF OF THEOREM 5.2. It is first of all possible to show by weak convergence techniques that

$$N^{1/4} \sup |\hat{f}_N(s) - E_* \hat{f}_N(s)| \to_p 0, \tag{5.8}$$

where $f(\cdot)$ is the density of $\{x_i'\beta\}$,

$$\hat{f}_N(s) = \{Nb(N)\}^{-1} \textstyle\sum_{i=1}^N K\left(\dfrac{t_i - s}{b(N)}\right),$$

$$E_* \hat{f}_N(s) = E\{Nb(N)\}^{-1} \textstyle\sum_{i=1}^N K\left(\dfrac{\tau_i - s}{b(N)}\right);$$

i.e., E_* means we replace t_i by $\tau_i = x_i'\beta$ and then take expectations. To show (5.8), first recall that the support of the design is bounded, so that $|t_i - \tau_i| = O_p(N^{-1/2})$ uniformly in i. This means that, uniformly in i,

$$\{(t_i - \tau_i)/b(N)\}^3 \to_p 0.$$

Using this and compactness, one expands to get

$$\begin{aligned} &\hat{f}_N(s) = \{Nb(N)\}^{-1} \textstyle\sum_{i=1}^N \\ &\left\{ K\left(\dfrac{\tau_i - s}{b(N)}\right) + \left(\dfrac{t_i - \tau_i}{b(N)}\right) K'\left(\dfrac{\tau_i - s}{b(N)}\right) + \dfrac{1}{2}\left(\dfrac{t_i - \tau_i}{b(N)}\right)^2 K''\left(\dfrac{\tau_i - s}{b(N)}\right) \right\} + o_p(1). \end{aligned} \tag{5.9}$$

That the third term on the r.h.s. of (5.9) convergences in probability to zero at rate $N^{1/4}$ follows directly from A6. Denote the first term by $V_{N1}(s)$ and the second by $V_{N2}(s)$. The same weak convergence argument used in Theorem 5.1 shows $N^{1/4}\{V_{N1}(s) - E_* \hat{f}_N(s)\}$ converges in probability to zero uniformly on compacts. Dealing with $V_{N2}(s)$ is quite tricky. One first shows that it suffices to make the substitution (5.7) for $\hat{\beta}_L - \beta$. Then, a simple second moment computation shows that the finite dimensional distributions of the resulting modified process

$$V_{N2}^*(s) = \{N^2 b^2(N)\}^{-1} \textstyle\sum_i \sum_j x_i'\{E(xx')\}^{-1} x_j \sigma_j \varepsilon_j K'\left(\dfrac{\tau_i - s}{b(N)}\right)$$

converge in probability to zero; here, as in the tightness argument to follow, we use the fact that the support of K strictly includes the support of $\{x_i'\beta\}$ and, since K is a symmetric density, $\int K'(y)\,dy = 0$. Finally, tightness can be proven by using Theorem 12.3 of Billingsley (1968) (use his equation (12.51) with $\gamma = 2$ and $\alpha = 1 + a$, a very small); in doing this calculation, one must separate the cases $|t_2 - t_1| \geq db(N)$ and $< db(N)$ for a large constant d (t_1 and t_2 refer to Billingsley's notation). Because of (5.8) and Theorem 5.1, we now only need to prove Theorem 5.2 for

$$(5.10)\qquad H_N^*(s) = \{Nb(N)\}^{-1} \textstyle\sum_{i=1}^N r_i^2 K\left(\dfrac{t_i - s}{b(N)}\right) - \{Nb(N)\}^{-1} E \textstyle\sum_{i=1}^N \varepsilon_i^2 \sigma_i^2 K\left(\dfrac{x_i'\beta - s}{b(n)}\right).$$

One first makes the expansion of (5.10), as in (5.9), about $K((\tau_i - s)/b(N))$, and then argues as above and in the proof of Theorem 5.1; the assumption $E\varepsilon^6 < \infty$ is again vital here.

Proof of Theorem 2. As in the proof of Theorem 1 we must show

$$(5.11)\qquad N^{-1/2} \textstyle\sum x_i \varepsilon_i \{\hat{H}_N(t_i) - H(\tau_i)\}/\sigma_i^3 \to_p 0.$$

The proof parallels that of Theorem 1. Here, the difficult case is to show

$$(5.12)\qquad N^{-1/2} \textstyle\sum_{i=1}^N x_i \varepsilon_i \{\hat{G}_N(t_i) - Q_N(\tau_i)\}/\{\sigma_i^3 E\hat{f}_N(\tau_i)\} \to_p 0,$$

where

$$\hat{H}_N = \hat{G}_N/\hat{f}_N, \qquad Q_n(x) = \{Nb(N)\}^{-1} E \textstyle\sum \varepsilon_i^2 \sigma_i^2 K\left(\dfrac{\tau_i - x}{b(N)}\right).$$

Rewrite (5.12) as

$$\begin{aligned} &\{N^{3/2}b(N)\}^{-1} \textstyle\sum_i \sum_j \dfrac{x_i \varepsilon_i}{\sigma_i^3 E\hat{f}_N(\tau_i)} r_j^2 \left\{K\left(\dfrac{t_j - t_i}{b(N)}\right) - K\left(\dfrac{\tau_j - \tau_i}{b(N)}\right)\right\} \\ (5.13)\qquad & \\ &+ \{N^{3/2}b(N)\}^{-1} \textstyle\sum_i \sum_j \dfrac{x_i \varepsilon_i}{\sigma_i^3 E\hat{f}_N(\tau_i)} \left\{r_j^2 K\left(\dfrac{\tau_j - \tau_i}{b(N)}\right) - b(N) Q_N(\tau_n)\right\}. \end{aligned}$$

By a messy argument similar to that of (5.10), the second term in (5.13) can be shown to converge in probability to zero. For the first term, it suffices to show that for every $M > 0$,

$$(5.14)\qquad \sup_{|\Delta| \leq M} |V_N(\Delta)| \to_p 0,$$

where

$$V_N(\Delta) = \{N^{3/2}b(N)\}^{-1} \textstyle\sum_i \sum_j \dfrac{x_i \varepsilon_i r_j^2}{\sigma_i^3 f(\tau_i)} \left\{K\left(\dfrac{\tau_j - \tau_i}{b(N)} + \dfrac{(x_i' - x_j')\Delta}{N^{1/2} b(N)}\right) - K\left(\dfrac{\tau_j - \tau_i}{b(N)}\right)\right\}.$$

Because, uniformly in i,

$$\{x_i'(\hat{\beta}_L - \beta)\}^2 = O_p(N^{-1}),$$

by A6 we must merely show (5.14) for the process $V_{N^*}(\Delta)$ which, in $V_N(\Delta)$, replaces r_j^2 by

$$\sigma_j^2 \varepsilon_j^2 - 2\sigma_j \varepsilon_j x_j'(\hat{\beta}_L - \beta).$$

Divide V_{N^*} into the two processes

$$V_{N^*}(\Delta) = V_{N^*}^{(1)}(\Delta) + V_{N^*}^{(2)}(\Delta).$$

We now invoke the results of Bickel and Wichura (1971) on multiparameter stochastic processes, changing their equation (3) to

$$E\,|X(B)|^2 \leq \mu(B)^{1+\gamma},$$

for some $\gamma > 0$. This shows (in order) that it suffices to show the results when we replace

$$N^{-1} \sum x_i x_i'$$

in the definition of $\hat{\beta}_L - \beta$ by $E(xx')$, and then that $V_{N^*}^{(j)}$ is tight with finite dimensional distributions converging in probability to zero. This proves (5.14) and completes the proof of Theorem 2.

NOTE. Handwritten detailed proofs are available from the author.

Acknowledgment. I would like to express my thanks to David Ruppert and Kent Bailey for stimulating discussions which led to this work. I would also like to thank the referees and associate editor for reading a difficult first draft, and to the editor for insisting on a readable presentation.

REFERENCES

BENEDETTI, J. K (1977). On the nonparametric estimation of regression functions. *J. Roy. Statist. Soc. B* **39** 248-253.

BICKEL, P. J. (1978). Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *Ann. Statist.* **6** 266-291.

BICKEL, P. J. and WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656-1670.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BOX, G. E. P. and HILL, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics* **16** 385-389.

CARROLL, R. J. (1982). Estimation in heteroscedastic models when there are many parameters. *J. Statist. Plann. Infer.* To appear.

CARROLL, R. J. and RUPPERT, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10** 429-441.

DENT, W. T. and HILDRETH, C. (1977). Maximum likelihood estimation in random coefficient models. *J. Amer. Statist. Assoc.* **72** 69-72.

FROEHLICH, B. R . (1973). Some estimators for a random coefficient regression model. *J. Amer. Statist. Assoc.* **68** 329-335.

FULLER, W. A and RAO, J. N. K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Ann. Statist.* **6** 1149-1158.

HILDRETH, C. and HOUCK, J. P. (1968). Some estimators for a linear model with random coefficients. *J. Amer. Statist. Assoc.* **63** 584-595.

JOBSON, J. D. and FULLER, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *J. Amer. Statist. Assoc.* **75** 176-181.

JOHNSTON, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 404-414.

LENTH, R. V. (1977). Robust splines. *Commun. Statist. A* **6** 847-854.

MACK, Y. P. and SILVERMEN, B. W. (1980). Weak and strong uniform consistency of kernel regression estimates. Preprint.

ROSENBLATT, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis II.* Academic, New York, 25-31.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595-645.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā A* **26** 359-372.

WILLIAMS, J. S. (1975). Lower bounds on convergence rates of weighted least squares to best linear unbiased estimators. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.) 555–569. Academic, New York.

WRIGHT, I. W. and WEGMAN, E. J. (1980). Isotonic, convex and related splines. *Ann. Statist.* **8** 1023–1035.

WRIGHT, F. T. (1978). Estimating strictly increasing regression functions. *J. Amer. Statist. Assoc.* **73** 636–639.

9810 PARKWOOD DRIVE
BETHESDA, MARYLAND 20814

J. R. Statist. Soc. B (1991)
53, *No.* 3, *pp.* 573–585

Semiparametric Estimation in Logistic Measurement Error Models

By R. J. CARROLL† and M. P. WAND

Texas A&M University, College Station, USA

[Received July 1989. Final revision February 1990]

SUMMARY

We describe semiparametric estimation and inference in a logistic regression model with measurement error in the predictors. The particular measurement error model consists of a primary data set in which only the response Y and a fallible surrogate W of the true predictor X are observed, plus a smaller validation data set for which (Y, X, W) are observed. Except for the underlying assumption of a logistic model in the true predictor, no parametric distributional assumption is made about the true predictor or its surrogate. We develop a semiparametric parameter estimate of the logistic regression parameter which is asymptotically normally distributed and computationally feasible. The estimate relies on kernel regression techniques. For scalar predictors, by a detailed analysis of the mean-squared error of the parameter estimate, we obtain a representation for an optimal bandwidth.

Keywords: BANDWIDTH SELECTION; DENSITY ESTIMATION; ERRORS IN VARIABLES; GENERALIZED LINEAR MODELS; KERNEL REGRESSION; LOGISTIC REGRESSION; MAXIMUM LIKELIHOOD; MEASUREMENT ERRORS MODELS; NONPARAMETRIC REGRESSION; PROBIT REGRESSION

1. INTRODUCTION

1.1. *Motivation and Literature*

This paper describes semiparametric estimation and inference in a logistic regression model with measurement error in the predictors. The primary concern is the univariate predictor case, although our main asymptotic normality result can be extended to higher dimensions. We first describe the example which motivated this work, and then the general model.

In the Nurses Health Study described by Rosner *et al.* (1989), the relationship between breast cancer (Y, a binary variable) and long-term dietary saturated fat (X) was examined prospectively. The primary data set consisted of a cohort of 89538 women, but instead of observing X, a surrogate W was observed, namely a self-administered quantitative food frequency questionnaire. To understand the relationship between X and W, 173 nurses became part of a validation study, in which Y, W and X were observed. X was not observed exactly but diet was measured sufficiently often in the validation data set, for one week at four different points in the year, that we may assume that X is known. The question to be addressed is how to use the validation data to obtain good estimates of logistic regression parameters.

†*Address for correspondence*: Department of Statistics, Texas A&M University, College Station, TX 77843–3143, USA.

0035–9246/91/53573 $2.00

The logistic model is $\mathrm{pr}(Y=1|X=x)=F(\beta_0+\beta_1 x)$, where $F(v)=\{1+\exp(-v)\}^{-1}$. We assume that Y and W are independent given X. If $f_{X|W}$ is the conditional density of X given W, then in the primary study the observed data follow the model

$$\mathrm{pr}(Y=1\,|\,W=w)=\int F(\beta_0+\beta_1 x)\,f_{X|W}(x\,|\,w)\,\mathrm{d}x. \tag{1.1}$$

It is well known that a logistic regression of Y on W leads to inconsistent estimates of β_0 and β_1 (Stefanski and Carroll, 1985). In the Nurses Health Study, the measurement error is large and the asymptotic bias considerable; see Rosner *et al.* (1989).

The topic of binary regression when predictors are measured with error has been the subject of several recent papers. In this literature, Y has been unobserved in the validation data although extensions to our case are straightforward. The methods can be categorized as

(a) fully parametric,
(b) efficient semiparametric and
(c) approximate corrections for attenuation.

Carroll *et al.* (1984) and Schafer (1987) parameterize $f_{X|W}$ with a nuisance vector ξ. They compute a pseudo-maximum-likelihood estimate, using the following algorithm:

(a) estimate ξ from the validation data;
(b) pretend that in model (1.1) ξ is known and equal to its estimated value, thus yielding a pseudolikelihood function;
(c) estimate $\beta=(\beta_0,\beta_1)^{\mathrm{T}}$ by maximizing the pseudolikelihood over the primary data.

This method in the logistic case is difficult to compute, performs poorly in moderate sample sizes and is non-robust to misspecification of the conditional density $f_{X|W}$.

Stefanski and Carroll (1987) take a semiparametric approach. Specifically, W given X is assumed to be normally distributed with constant variance and mean linear in X; the mean and variance parameters are denoted by ξ. The marginal density of X, f_X, is assumed unknown. A sufficiency argument gives rise to a set of estimating equations depending on ξ and β. Again, ξ is estimated from the validation study and β is then estimated from the primary data assuming that ξ is known. This method makes fewer assumptions than the previous method, and works well for even moderate sample sizes, but the robustness against non-normal measurement error has not been explored.

A third approach is based on small measurement error asymptotics; see Stefanski and Carroll (1985), Stefanski (1985), Whittemore and Keller (1988) and Rosner *et al.* (1989). These methods pretend that the difference between X and W is 'small', which gives rise to estimates of β which partially correct for measurement error. There are again nuisance parameters ξ, which are estimated from the validation study. These methods work very well in practice, even though they are only partial corrections for attenuation. An asymptotic theory is given in Carroll and Stefanski (1990).

In this paper, we consider a fourth approach, namely using nonparametric kernel regression methods in the validation data to estimate the probability function (1.1). A semiparametric estimate of β results from this method, with an estimated bandwidth and an asymptotically normal limit distribution. The next subsection outlines this approach. Independently, Pepe and Fleming (1991) consider a problem similar to ours with W a discrete random variable.

1.2. *Description of Method*

We assume throughout the paper that the joint, marginal and conditional densities of X and W have two bounded and continuous derivatives.

Define

$$B(x, y, \beta) = F(\beta_0 + \beta_1 x)^y \{1 - F(\beta_0 + \beta_1 x)\}^{1-y}. \tag{1.2}$$

The likelihood function for $Y = y$ given $W = w$ is given by

$$L(y, w, \beta) = E\{B(X, y, \beta) \mid W = w\}. \tag{1.3}$$

Equation (1.3) describes a nonparametric regression problem: regressing $B(X, y, \beta)$ on W. No numerical integration is required to obtain an estimate of the likelihood function. We propose to estimate equation (1.3) by a kernel regression.

In what follows, n_1 is the size of the validation data set, n_2 the size of the primary data set and $n_2/n_1 = \lambda$. Derivatives with respect to β are denoted by subscripts.

Let K be a symmetric density function, and let h be a bandwidth or window width. Define

$$\hat{f}_W(w) = (n_1 h)^{-1} \sum_1^{n_1} K\{(w - W_i)/h\},$$

$$D_n(y, w, \beta) = (n_1 h)^{-1} \sum_1^{n_1} B(X_i, y, \beta) K\{(w - W_i)/h\},$$

$$C_n(y, w, \beta) = (n_1 h)^{-1} \sum_1^{n_1} B_\beta(X_i, y, \beta) K\{(w - W_i)/h\}.$$

The estimated likelihood function for $Y = y$ given $W = w$ is $L_n(y, w, \beta) = D_n(y, w, \beta)/\hat{f}_W(w)$, while the estimated likelihood score is $H_n(y, w, \beta) = C_n(y, w, \beta)/D_n(y, w, \beta)$. Let $l(y, x, \beta) = (1, x)^T \{y - F(\beta_0 + \beta_1 x)\}$ be the likelihood score for $(Y, X) = (y, x)$.

In principle, there is information about β in both validation and primary data sets. The validation data contribute terms to the likelihood based on (Y, X), while the primary data are values of (Y, W). We use both types of information in our estimate.

Recall that $n = n_1 + n_2$. We propose that we estimate β by $\hat{\beta}$, the solution to

$$n^{-1/2} \sum_{j=1}^{n_1} l(Y_j, X_j, \beta) + n^{-1/2} \sum_{i=n_1+1}^{n} H_n(Y_i, W_i, \beta) = 0. \tag{1.4}$$

If $f_{X|W}$ were known, then we would replace H_n by H in equation (1.4) to obtain the likelihood equations. We shall quantify how much would be gained by making this replacement, in effect considering the cost due to estimating H by nonparametric regression.

In solving equation (1.4) we can use a scoring method with the starting value of β taken to be $\hat{\beta}^{(0)}$, the maximum likelihood estimate based on the validation data. An alternative estimator can be found for β by performing just one iteration of Newton's method with $\hat{\beta}^{(0)}$ as the starting value. Let $\hat{\beta}^{(1)}$ be this estimator. It may be shown that $\hat{\beta}$ and $\hat{\beta}^{(1)}$ have the same limit distribution.

An interesting feature of this problem is that when $\beta_1 = 0$, for $i \geqslant n_1 + 1$, it can be shown that the estimated score is unbiased, i.e. $EH_n(Y_i, W_i, \beta) = 0$. Thus, the classic bias–variance trade-off that we associate with nonparametric regression disappears when $\beta_1 = 0$.

1.3. *Arrangement of Paper*

In the next section, we summarize the asymptotic behaviour of $\hat{\beta}$ and $\hat{\beta}^{(1)}$ when $h \to 0$, where h is deterministic. In Section 3, we indicate methods for selecting h from the validation data, one of which is easy to implement. Section 4 describes the results of a simulation study.

2. ASYMPTOTICS FOR DETERMINISTIC BANDWIDTHS

2.1. *Introduction*

A practical problem occurs with this method as a result of edge effects. The estimated likelihood $H_n(y, w, \beta)$ will be unreliable near the boundaries of the validation data, where there are few observations and the weighted averaging of kernel regression becomes asymmetric. In addition, the primary data set, being larger, is expected to have observations W_i outside the range of the primary data. Used blindly, this would mean extrapolating the kernel fit $H_n(y, w, \beta)$ outside the range of the data used in its construction. Such extrapolation is dangerous, and robustness considerations dictate that it be avoided.

One method for overcoming this is to evaluate $H_n(y, w, \beta)$ only for those w in a fixed set interior to the support of W. Such a restriction is similar in spirit to the so-called Mallows method of robust regression, which downweights observations on the basis of leverage.

What follows are formal calculations, rather than detailed proofs. These calculations can be justified if the summations for $i \geqslant n_1 + 1$ in equation (1.4) are taken over those W_i in a fixed set interior to the support of W.

2.2. *Main Result*

Make the following definitions:

$$C(y, w, \beta) = f_W(w) E\{B_\beta(X, y, \beta) \mid W = w\};$$

$$D(y, w, \beta) = f_W(w) E\{B(X, y, \beta) \mid W = w\};$$

$$H(y, w, \beta) = C(y, w, \beta)/D(y, w, \beta);$$

$$M_\beta(\beta) = (1 + \lambda)^{-1} E\{l_\beta(Y, X, \beta) + \lambda H_\beta(Y, W, \beta)\}.$$

Also, let $\hat{\beta}^*$ stand for either $\hat{\beta}$ or $\hat{\beta}^{(1)}$. Since $H_n \to H$, by a Taylor series argument

$$n^{1/2}(\hat{\beta}^* - \beta) \approx -\{G_1(n, \beta) + G_2(n, \beta)\}^{-1} n^{1/2}\{G_3(n, \beta) + G_4(n, \beta)\}, \qquad (2.1)$$

for random variables $G_i(n, \beta)$, $i = 1, \ldots, 4$, specified in equation (A.2) of Appendix A. If we assume that $h^4 n \rightarrow 0$ and $nh^2 \rightarrow \infty$, then calculations outlined in Appendix A indicate that $n^{1/2}(\hat{\beta}^* - \beta)$ is asymptotically normally distributed with mean zero and covariance matrix given by

$$n \operatorname{cov}(\hat{\beta}^* - \beta) \rightarrow M_\beta(\beta)^{-1} \Gamma(\beta) M_\beta(\beta)^{-1}, \tag{2.2}$$

where

$$\Gamma(\beta) = (1 + \lambda)^{-1}[E\{l(Y, X, \beta)\, l(Y, X, \beta)^{\mathrm{T}}\} + \lambda E\{H(Y, W, \beta)\, H(Y, W, \beta)^{\mathrm{T}}\} + \lambda^2 \zeta(\beta)] \tag{2.3}$$

$$\zeta(\beta) = \sum_{y=0}^{1} \sum_{z=0}^{1} E\{L(z, W, \beta)\, L(y, W, \beta)\, Q(X, z, W, \beta)\, Q(X, y, W, \beta)^{\mathrm{T}} f_W^2(W)\}$$

$$Q(x, y, w, \beta) = \frac{B_\beta(x, y, \beta)\, D(y, w, \beta) - B(x, y, \beta)\, C(y, w, \beta)}{D^2(y, w, \beta)} \tag{2.4}$$

We indicate in Section 5 and Appendix A that this result can be extended to arbitrary dimensions for X and W.

Remark 1. Each term in equation (2.3) has a distinct source. The first is the contribution of the validation data set: the Fisher information for β from validation. The second term is the Fisher information from the primary data set if $f_{X|W}$ were known. The third term represents the cost due to not knowing $f_{X|W}$. For the Nurses Health Study, if we assume that (X, W) are jointly normally distributed, then from information in Rosner *et al.* (1989) we conclude that X and W have standard deviations 4.6 and 5.9 respectively, and that, given W, X has a mean linear in W with slope 0.47 and standard deviation 3.7. Rosner *et al.* (1989) conclude for this example that the slope estimate obtained by regressing Y on W approximates $0.5\beta_1$, not β_1 itself; this is the effect of the measurement error. We also assume that X and W have the same means, which we take to be zero by centring. If we choose $\beta_0 = -5.0$ and, following Rosner *et al.* (1989), $\beta_1 = -0.018$, then the contribution due to estimating H is less than 1% of the total standard error of $\hat{\beta}_1$. If, however, $\beta_1 = -0.3$, then nearly 70% of the standard error is due to estimating H. This latter value of β_1 is used merely as an illustration, as it is much larger than would be expected in this study. □

Remark 2. In most applications, the primary data set is large relative to the validation data, i.e. λ is large. In the Nurses Health Study, $\lambda = 517.6$. In such cases, there is little information about β in the validation data set, and there will be little difference between solving equation (1.4) and solving

$$n^{-1/2} \sum_{i=n_1+1}^{n} H_n(Y_i, W_i, \beta) = 0. \tag{2.5}$$

The changes in the asymptotics when solving equation (2.5) is that $M_\beta(\beta) = E\{H_\beta(Y, W, \beta)\}$, $\Gamma(\beta) = E\{H(Y, W, \beta)\, H(Y, W, \beta)^{\mathrm{T}}\} + \lambda \zeta(\beta)$ and $n = n_2$ in expression (2.2). □

Remark 3. When λ is extremely large, the main component in covariance (2.2) is $\zeta(\beta)$, which comes from the uncertainty in nonparametric estimation in the validation data. This gives one motivation to make the validation data set sufficiently large that the randomness incurred by the nonparametric regression does not dominate. □

Remark 4. Equations (2.2) and (2.3) give some insight into the design of a study, in particular the choice of size of the validation study. If we assume that (X, W) are jointly normally distributed, then we can compute covariance (2.2) for various values of (β, λ) to obtain a sense of an acceptable λ. For the Nurses Health Study, following the assumptions of remark 1, with $\beta_0 = -5$ and $\beta_1 = -0.018$, we estimate that the effect of using a validation sample size $n_1 = 3750$ instead of the actual size $n_1 = 173$ is to decrease the standard error for estimating β_1 by less than 5%. The standard error can be reduced by approximately 40% by observing X for all study participants. However, if $\beta_1 = -0.3$, these figures are 73% and 88% respectively. □

Remark 5. If we model $f_{X|W}$ parametrically, a result similar to equations (2.2) and (2.3) occurs: an extra component in covariance due to estimating parameters via the validation study. See Section 4 for details. □

Remark 6. The covariance matrix (2.2) can be estimated by replacing $M_\beta(\beta)$, $\Gamma(\beta)$ and $\zeta(\beta)$ by their method-of-moments estimators. □

3. BANDWIDTH SELECTION

The proof of asymptotic normality with covariance (2.2) is sketched in Appendix A, under the condition that $nh^4 + (nh^2)^{-1} \to 0$. Unfortunately, this tells us nothing about selection of the bandwidth h. We might take the view that h should be varied over a wide range to see whether the estimates and inference are sensitive to h. We have sympathy with this data analytic viewpoint, but there is also value in letting h be determined by the data. In this section, we discuss automatic bandwidth selection. As a first step we derive a higher order expansion of the covariance of an asymptotically equivalent form of $n^{1/2}(\hat{\beta} - \beta)$.

Define

$$a_1 = \frac{\lambda}{2(1+\lambda)} \int z^2 K(z)\, \mathrm{d}z \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W, \beta)\, f_W(W)\, f_2(X, W)\} \tag{3.1}$$

$$a_2 = \lambda \int K^2(z)\, \mathrm{d}z \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W, \beta)\, f_W(W)\, B(X, y, \beta)/D(y, W, \beta)\} \tag{3.2}$$

$$f_2(x, w) = \{(\partial^2/\partial w^2)\, f_{X,W}(x, w)\}/f_{X,W}(x, w) \tag{3.3}$$

$$a_3(n, h) = M_\beta(\beta)^{-1}\{(nh^4)^{1/2} a_1 - (nh^2)^{-1/2} a_2\}.$$

In Appendix B, we outline a result showing that, for some random variable Z_n^* and matrix A, $n^{1/2}(\hat{\beta}^* - \beta) = Z_n^* + o_p\{(nh^4)^{1/2} + (nh^2)^{-1/2}\}$, where

$$E Z_n^* (Z_n^*)^{\mathrm{T}} = M_\beta(\beta)^{-1}\Gamma(\beta)\, M_\beta(\beta)^{-1} + a_3(n, h)\, a_3(n, h)^{\mathrm{T}} + (nh)^{-1} A. \tag{3.4}$$

Equation (3.4) shows that the bandwidth does not affect the covariance except to smaller order terms. However, our second-order expansion does suggest a plug-in method for bandwidth selection; see below.

At the end of Section 1.2, we discussed the fact that $EH_n(Y, W, \beta) = 0$ when $\beta_1 = 0$. We can show in this case that $a_1 = a_2 = a_3(n, h) = (0, 0)^{\mathrm{T}}$. Hence, equation (3.4) leads to choosing $h = \infty$, which suggests one reason why reliable automatic bandwidth selection based on plugging into equation (3.4) will be difficult, in general.

Remark 7. In terms of rates of convergence for estimating the linear combination $\gamma^{\mathrm{T}}\beta$, the 'optimal' h minimizes $\{\gamma^{\mathrm{T}} a_3(n, h)\}^2$. Except when $\beta_1 = 0$, this h is of the order $n^{-1/3}$, much smaller than the usual order of $n^{-1/5}$ common in nonparametric regression. In fact, the usual rate is prohibited in our calculations, if we are to estimate β at the rate $n^{1/2}$. □

The matrix A in equation (3.4) is complicated. For example, the contribution to A from the first-order linear expansion of H_n about H is $M_\beta(\beta)^{-1} J_2 M_\beta(\beta)^{-1}$, where

$$J_2 = \lambda \int K^2(z)\,\mathrm{d}z \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W, \beta)\, Q^{\mathrm{T}}(X, y, W, \beta)\, f_W(W)\}. \tag{3.5}$$

The higher order terms are even more complex, perhaps too much so to be useful in plug-in bandwidth selection.

In the simulations to follow, we used a simple *ad hoc* bandwidth selection method. We took $h = \hat{\sigma}_W n^{-1/3}$., where $\hat{\sigma}_W$ is the estimated standard deviation of W in the validation data set; a robust scale could be used as well. This method, while *ad hoc*, does have the correct rate of convergence and is easily programmed. Unlike plug-in rules based on equation (3.4), it has the virtue of being stable even when $\beta_1 \approx 0$.

4. PARAMETRIC PROBLEMS

In parametric problems, the form of the limit distribution of $\hat{\beta}$ is the same as equation (2.2). Writing the densities as $f_{X|W}(x|w, \eta)$ and $f_W(w|\eta)$, the likelihood of an observed $(Y, X, W) = (y, x, w)$ in the validation data is $B(x, y, \beta)\, f_{X|W}(x|w, \eta)\, f_W(w|\eta)$. If

$$L(y, w, \beta, \eta) = \int B(x, y, \beta)\, f_{X|W}(x|w, \eta)\,\mathrm{d}x,$$

then $L_*(y, w, \beta, \eta) = f_W(w|\eta)\, L(y, w, \beta, \eta)$ is the likelihood of an observed $(Y, W) = (y, w)$ in the primary data set. The full maximum likelihood estimate of (β, η) follows standard lines.

Let $H(y, w, \beta, \eta) = (\partial/\partial\beta) \log L(y, w, \beta, \eta)$. Let $\hat{\eta}$ be the maximum likelihood estimate of η in the validation data and define

$$\psi(x, w, \eta) = (\partial/\partial\eta) \log\{f_{X|W}(x|w, \eta)\, f_W(w|\eta)\}.$$

If $\hat{\beta}$ is the pseudo-maximum-likelihood estimate obtained by replacing η by $\hat{\eta}$ and maximizing the pseudolikelihood in β, then $\hat{\beta}$ has the limit distribution (2.2), with the exception that in equation (2.3) we replace $\zeta(\beta)$ by

$$\zeta_{\mathrm{parm}}(\beta) = EH_\eta (E\psi_\eta)^{-1} E\psi\psi^{\mathrm{T}} (E\psi_\eta)^{-1} EH_\eta^{\mathrm{T}}.$$

5. MONTE CARLO STUDY

To understand the performance of the method when applied to data, we undertook a small Monte Carlo study.

The performance of our method was assessed in comparison with some other standard methods. The first was the usual logistic coefficient replacing X by W. Second was the estimate of Rosner *et al.* (1989), which divides the usual estimate by the slope of the regression of X on W in the validation data. The third was a modification of the Stefanski and Carroll (1985) estimate; see Stefanski (1989). We modified this slightly by applying the method not to W but to $W_* = (W - v_0)/v_1$, where v_0 and v_1 are the least squares intercept and slope from the regression of W on X. The final estimate was that of Whittemore and Keller (1988), p. 1060. Their parameters A and Ω were estimated from the validation data by assuming a linear regression of X on W.

The method proposed here started with the usual logistic regression estimates and used five iterations of the scoring method. The functions $\hat{f}_W$, C_n and D_n were assessed on a grid of 41 points covering the range of the validation data, with grid points being the equally spaced percentiles of W, i.e. the minimum, 0.025 percentile, 0.05 percentile, etc. Between grid points, linear interpolation was used. The sums over i in the formulae for these functions were assessed only for those W_i in the primary data which were in the range of the validation data.

The competing estimates are all parametric in nature and are based on specific parametric assumptions. If these assumptions hold, then we expect that these methods will outperform our method. Our simulations were based on the idea of calibrating our method in situations ideal for the parametric methods, i.e. linear additive normal measurement error. We also tested the methods against two moderate model departures, and against a severe model departure.

We took the primary data set to be of size $n_2 = 2000$ with the validation study of size $n_1 = 250$. We took $\beta_0 = -1.10$, -2.20, -3.66 and $\beta_1 = 0.80$. The three choices of β_0 represent cases where the expected numbers of times $Y = 1$ are 500, 200 and 50 respectively.

There were three sampling situations. In the first, the random variables (X, W) were normally distributed according to the model $W = X + U$, where (X, U) are independent with zero means and standard deviations 0.50.

The second sampling situation consisted of $W = XU$, where X is as in the previous example and U is the negative exponential distribution, so that the variance of W is twice that of X, as in the previous case. This is only a moderately heteroscedastic situation. However, a plot of X against W for a single data set, given in Fig. 1, suggests that the semiparametric method might do poorly in this case, because of the rapid change of the regression near $W = 0$.

The third sampling situation is of a moderate model breakdown for the parametric methods. We took X to be uniform on the interval (0, 5), and $W = X^2/15$. The values of β_0 were -2.7, -3.8 and -5.26. We say that this is a moderate model breakdown because the plot of X against W is reasonably linear for much of the range of W.

The fourth model situation represents severe model breakdown. We took W to be uniform on the interval $(-\pi/2, 5\pi/2)$, and $X = \cos W$.

Clearly this last choice can be criticized on the grounds that the parametric

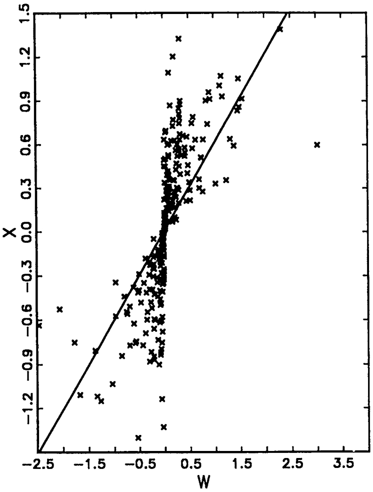


Fig. 1. Plot of X against W for a randomly generated sample of size 250: here, X is normally distributed with mean zero and standard deviation 0.5, while $W=XU$ and U follows a negative exponential distribution with unit mean

methods are inappropriately applied. However, if we only simulated situations where the parametric assumptions were appropriate, then we would find the unsurprising result that parametric is better than semiparametric in practice. What is of interest here is to see whether the semiparametric method does reasonably well in straightforward cases, and better in cases of model breakdown.

There were 100 iterations of the experiment. The results are given in Table 1, where we report median values for $\hat{\beta}_1$, the median absolute error (MAE) and the 90th percentile of the absolute errors. The semiparametric method does remarkably well in the normal model and is the clear winner in the quadratic and cosine models. As expected, its performance for rare events in the heteroscedastic model is relatively poorer compared with the method of Rosner *et al.* (1989), but it is still not unacceptable.

6. GENERALIZATIONS AND DISCUSSION

Most applications of logistic regression involve non-scalar predictors. One important example consists of the case where the predictors are (Z, X). Here Z is observable, X is scalar and unobservable, and W is the scalar surrogate for X. If the distribution of X given W is independent of Z, then our results apply with only the following minor modifications. In equation (1.2), define $B(x, z, y, \beta) = F(\beta_0^{\mathrm{T}}z+\beta_1 x)^y\{1-F(\beta_0^{\mathrm{T}}z+\beta_1 x)\}^{1-y}$, while in equation (1.3) define $L(y, z, w, \beta) = E\{B(X, z, y, \beta) \mid W=w\}$. These changes lead to redefined functions such as $D_n(y, z, w, \beta)$, $D(y, z, w, \beta)$ and $l(Y, X, Z, \beta)$. Otherwise, the results are unchanged.

TABLE 1
Results of a simulation study†

Estimate	*Results for the following values of β:*								
	$\beta_0 = -1.09$			$\beta_0 = -2.20$			$\beta_0 = -3.66$		
	Median	*MAE*	*90th*	*Median*	*MAE*	*90th*	*Median*	*MAE*	*90th*
Additive normal									
Usual	0.39	0.413	0.496	0.40	0.404	0.532	0.36	0.436	0.680
W-K	0.78	0.100	0.259	0.80	0.135	0.362	0.72	0.314	0.688
R-W-S	0.78	0.096	0.244	0.78	0.126	0.340	0.72	0.312	0.726
S-C	0.77	0.164	0.376	0.82	0.153	0.467	0.78	0.333	0.757
Semi-p	0.79	0.099	0.309	0.80	0.135	0.380	0.72	0.314	0.688
Heteroscedastic normal									
Usual	0.41	0.395	0.508	0.40	0.404	0.524	0.40	0.401	0.685
W-K	1.36	0.563	1.052	1.25	0.453	0.805	1.09	0.421	0.871
R-W-S	0.79	0.137	0.311	0.78	0.152	0.370	0.77	0.255	0.606
S-C	0.55	0.289	0.462	0.53	0.290	0.483	0.48	0.331	0.653
Semi-p	0.88	0.109	0.245	0.90	0.133	0.405	0.91	0.329	0.776
Cosine model									
Usual	0.00	0.800	0.803	0.00	0.800	0.806	0.00	0.800	0.807
W-K	−0.01	0.806	0.856	−0.01	0.812	0.879	0.00	0.801	0.928
R-W-S	−0.29	1.975	8.926	0.31	2.183	14.949	0.30	4.765	20.277
S-C	0.00	0.800	0.803	0.00	0.800	0.806	0.00	0.800	0.808
Semi-p	0.90	0.101	0.214	0.90	0.133	0.307	0.92	0.217	0.513
	$\beta_0 = -2.70$			$\beta_0 = -3.80$			$\beta_0 = -5.26$		
Quadratic model									
Usual	0.67	0.131	0.191	0.63	0.173	0.235	0.64	0.158	0.288
W-K	1.92	1.116	1.615	0.86	0.154	0.423	0.53	0.300	0.536
R-W-S	0.74	0.064	0.126	0.69	0.109	0.177	0.70	0.101	0.238
S-C	0.67	0.127	0.188	0.63	0.170	0.232	0.65	0.155	0.286
Semi-p	0.80	0.035	0.095	0.80	0.045	0.116	0.82	0.091	0.213

†Median is the median of the estimates, MAE is the median absolute error and 90th is the 90th percentile of the absolute error. Usual is the ordinary logistic regression estimate, W-K is the Whittemore and Keller (1988) estimate, R-W-S is the estimate of Rossner *et al.* (1989), S-C is the modified Stefanski and Carroll (1985) method discussed in the text and Semi-p is the semiparametric method. The models are as discussed in the text.

If W is non-scalar or if X given W is not independent of Z, then nonparametric regression must be performed in more than one dimension. We show in Appendix A that if W has dimension d, if all densities have p bounded and continuous derivatives and if K is a pth-order kernel function, then result (2.2) holds as long as $nh^{2p} \rightarrow 0$ and $nh^{2d} \rightarrow \infty$. The case discussed in Section 2 is $d = 1$ and $p = 2$. Note that the dimension of W need not be the same as the dimension of X. However, such a method is hardly practical if $d = 10$. The problem of suitable methods for higher dimensional surrogates remains open.

ACKNOWLEDGEMENTS

We are grateful to the referees for their helpful comments which led to significant improvements in the presentation of our results. Our research was supported by the National Institutes of Health and Sonderforschungsbereich 303, University of Bonn.

APPENDIX A

This section discusses asymptotic normality of the estimates of β. We shall assume that W is of dimension d, that the kernel K is a pth-order kernel, that $nh^{2d} \rightarrow \infty$ and that $nh^{2p} \rightarrow 0$. The case discussed in Section 2 is $p=2$ and $d=1$.

Since the size of the validation data set is $O(n)$, $n^{1/2}$-consistent estimates $\hat{\beta}^{(0)}$ of β are already available from the validation data. Thus, the analysis of the estimate $\hat{\beta}^{(1)}$ will follow standard lines of argument, as will that of $\hat{\beta}$. Rather than to use considerable space in tedious but standard arguments, we have chosen to take approximation (2.1) as our starting point.

Dropping the dependence on β, define $Q(x, y, w)$ by equation (2.4) and

$$Q_*(x, y, w) = Q(x, y, w)/D(y, w, \beta).$$

Where appropriate, we shall suppress arguments to individual terms, e.g. $H_{n,i}$ for $H_n(Y_i, W_i, \beta)$. In the subsequent calculations, we shall also use the notation

$$\tilde{Q}_{n,i} = (n_1 h^d)^{-1} \sum_{j=1}^{n_1} K\{(W_j - W_i)/h\}\, Q(X_j, Y_i, W_i). \tag{A.1}$$

An analogous definition is ascribed to $\tilde{Q}_{*,n,i}$. Let $\eta_n = nh^{2p} + (nh^{2d})^{-1}$. To order $o_p(\eta_n^{1/2})$, we can show that

$$n^{1/2}(\hat{\beta}^{(1)} - \beta) \approx -\left(n^{-1} \sum_{1}^{n_1} l_{i,\beta} + n^{-1} \sum_{n_1+1}^{n} H_{n,i,\beta}\right)^{-1} n^{1/2}\left(n^{-1} \sum_{1}^{n_1} l_i + n^{-1} \sum_{n_1+1}^{n} H_{n,i}\right). \tag{A.2}$$

Note that $G_i(n, \beta)$, $i = 1, \ldots, 4$, from approximation (2.1) are the four normalized sums in this expression. Except for terms of order $o_p(\eta_n^{1/2})$, we can write $H_{n,i} = H_i + \tilde{Q}_{n,i} - \tilde{Q}_{*,n,i}(D_{n,i} - D_i)$. Making this substitution and then taking derivatives with respect to β, we can show by computing first and second moments that

$$n^{-1} \sum_{1}^{n_1} l_{i,\beta} + n^{-1} \sum_{n_1+1}^{n} H_{n,i,\beta} - M_\beta(\beta) = o_p(\eta_n^{1/2}). \tag{A.3}$$

We shall use equations (A.2) and (A.3) in Appendix B.

Since $H_{n,i} - H_i - \tilde{Q}_{n,i} = o_p(n^{-1/2})$ under our conditions on the bandwidth h, for result (2.2) we need to show that Z_n is asymptotically normal with mean zero and covariance $\Gamma(\beta)$, where

$$Z_n = n^{-1/2}\left\{\sum_{j=1}^{n_1} l_j + \sum_{i=n_1+1}^{n} (H_i + \tilde{Q}_{n,i})\right\}.$$

Let $\mathcal{G}_{n_1}$ denote the observations in the validation data, i.e. $(X_i, W_i, Y_i)_1^{n_1}$. Define $\alpha_n = E(\tilde{Q}_{n,n} | \mathcal{G}_{n_1})$. A standard bias calculation shows that, except for terms of order $o_p(1)$,

$$\alpha_n = (\lambda/n_2) \sum_{j=1}^{n_1} S_j,$$

where

$$S_j = \sum_{y=0}^{1} L(y, W_j, \beta)\, Q(X_j, y, W_j)\, f_W(W_j).$$

It is easy to show by a covariance calculation that

$$n^{-1/2} \sum_{n_1+1}^{n} (\tilde{Q}_{n,i} - \alpha_n) = o_p(1),$$

and hence that

$$Z_n = n^{-1/2} \left\{ \sum_{j=1}^{n_1} (l_j + \lambda S_j) + \sum_{i=n_1+1}^{n} H_i \right\} + o_p(1). \tag{A.4}$$

The right-hand side of equation (A.4) has covariance $\Gamma(\beta)$, and it is clearly asymptotically normally distributed.

APPENDIX B

The purpose of this section is to verify equation (3.4). We are taking the dimension of W and X to be $d = 1$, and the order of the kernel to be $p = 2$; see equation (A.1). On the basis of the introductory comments in Appendix A, specifically equations (A.2) and (A.3), it suffices to compute the mean and covariance of

$$Z_n^* = n^{-1/2} \left[\sum_{1}^{n_1} l_i + \sum_{n_1+1}^{n} \{H_i + \tilde{Q}_{n,i} - \tilde{Q}_{*,n,i}(D_{n,i} - D_i)\} \right] = U_{1n} + U_{2n} + U_{3n} - U_{4n}.$$

Obviously, since they are scores, $E(l_i) = E(H_i) = 0 = E(U_{1n}) = E(U_{2n})$.

Define $f_2(x, w)$ as in equation (3.3). Standard bias calculations show that to order $o(\eta_n^{1/2})$

$$\begin{aligned} E(U_{3n}) &= n^{1/2} h^2 a_1 \\ E(U_{4n}) &= (n^{1/2} h)^{-1} a_2, \end{aligned} \tag{B.1}$$

where a_1 and a_2 are given by equations (3.1) and (3.2) respectively.

The covariance terms take some effort to compute. Of course, $E(U_{1n} U_{1n}^{\mathrm{T}}) = (1 + \lambda)^{-1} E(ll^{\mathrm{T}})$, $E(U_{2n} U_{2n}^{\mathrm{T}}) = \lambda E(HH^{\mathrm{T}})/(1 + \lambda)$ and $E(U_{1n} U_{2n}) = 0$. By direct and fairly easy calculation, to order $o(\eta_n)$, $E(U_{1n} U_{3n}) = E(U_{1n} U_{4n}) = 0$.

Define $c_1(K) = \int K^2(z)\, \mathrm{d}z$. It is also relatively easy to show that $E(U_{2n} U_{3n}^{\mathrm{T}}) = O(h^2)$, and that to order $o(\eta_n)$

$$E(U_{2n} U_{4n}^{\mathrm{T}}) = \frac{\lambda}{nh} c_1(K) \sum_{y=0}^{1} E\{L(y, W, \beta)\, H(y, W, \beta)\, Q_*^{\mathrm{T}}(X, y, W)\, B(X, y, W)\, f_W(W)\}. \tag{B.2}$$

A somewhat lengthier calculation yields that

$$\begin{aligned} \operatorname{cov}(U_{3n}) = {} & \lambda^2 \zeta(\beta)/(1 + \lambda) + \lambda (nh)^{-1} c_1(K) \\ & \times \sum_{y=0}^{1} E\{L(y, W, \beta)\, Q(X, y, W)\, Q^{\mathrm{T}}(X, y, W)\, f_W(W)\}. \end{aligned} \tag{B.3}$$

We can show that $E(U_{2n} U_{3n}^{\mathrm{T}}) = O(h^2)$, and that $E(U_{3n} U_{4n}^{\mathrm{T}})$ can be written as

$$E(U_{3n}U_{4n}^{\mathrm{T}}) = \lambda^2 (nh)^{-1} c_1(K) \sum_{y,\,z=0}^{1} E\{L(y, W, \beta)\, L(z, W, \beta)\, f_W^2(W)$$

$$\times\, B(X, z, \beta)\, Q(X, y, W) Q_*(X, z, W)^{\mathrm{T}}\} + o(\eta_n). \tag{B.4}$$

Finally, we note that $E(U_{4n}U_{4n}^{\mathrm{T}})$ is, to order $o(\eta_n)$, the sum of the following two terms:

$$\lambda^2 (nh)^{-1} \alpha(K) \sum_{y,\,z=0}^{1} \int L(z, w, \beta)\, L(y, w, \beta)\, Q_*(x_2, z, w)\, Q_*^{\mathrm{T}}(x_2, y, w)\, B(x_1, z, \beta)$$

$$\times\, B(x_1, y, \beta)\, f_{X,W}(x_1, w)\, f_{X,W}(x_2, w)\, f_W^2(w)\, \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}w; \tag{B.5}$$

$$\lambda^2 (nh)^{-1} \alpha(K) \sum_{y,\,z=0}^{1} \int L(z, w, \beta)\, L(y, w, \beta)\, Q_*(x_1, z, w)\, Q_*^{\mathrm{T}}(x_2, y, w)\, B(x_1, z, \beta)$$

$$\times\, B(x_2, y, \beta)\, f_{X,W}(x_1, w)\, f_{X,W}(x_2, w)\, f_W^2(w)\, \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}w, \tag{B.6}$$

where

$$\alpha(K) = \int K(z_1)\, K(z_2)\, K(z_1 + z_3)\, K(z_2 - z_3)\, \mathrm{d}z_1\, \mathrm{d}z_2\, \mathrm{d}z_3.$$

We can collect terms to compute equation (3.4). The matrix J_2 in equation (3.5) comes from the second part of equation (B.3). The terms (B.2), (B.4), (B.5) and (B.6) arise from U_{4n}, which is the error in linearizing H_n about H.

REFERENCES

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. (1984) On errors-in-variables for binary regression models. *Biometrika*, **71**, 19–26.

Carroll, R. J. and Stefanski, L. A. (1990) Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Am. Statist. Ass.*, **85**, 652–663.

Pepe, M. S. and Fleming, T. R. (1991) A general nonparametric method for dealing with errors in missing or surrogate data. *J. Am. Statist. Ass.*, to be published.

Rosner, B., Willett, W. C. and Spiegelman, D. (1989) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist. Med.*, **8**, 1075–1093.

Schafer, D. (1987) Covariate measurement error in generalized linear models. *Biometrika*, **74**, 385–389.

Stefanski, L. A. (1985) The effects of measurement error on parameter estimation. *Biometrika*, **72**, 583–592.

——— (1989) Correcting data for measurement error in generalized linear models. *Communs Statist.* A, **18**, 1715–1734.

Stefanski, L. A. and Carroll, R. J. (1985) Covariate measurement error in logistic regression. *Ann. Statist.*, **13**, 1335–1351.

——— (1987) Conditional scores and optimal scores for generalized linear measurement error models. *Biometrika*, **74**, 703–716.

Whittemore, A. S. and Keller, J. B. (1988) Approximations for errors in variables regression. *J. Am. Statist. Ass.*, **83**, 1057–1066.

Generalized Partially Linear Single-Index Models

R. J. CARROLL, Jianqing FAN, Irène GIJBELS, and M. P. WAND

The typical generalized linear model for a regression of a response Y on predictors $(\mathbf{X}, \mathbf{Z})$ has conditional mean function based on a linear combination of $(\mathbf{X}, \mathbf{Z})$. We generalize these models to have a nonparametric component, replacing the linear combination $\boldsymbol{\alpha}_0^T\mathbf{X} + \boldsymbol{\beta}_0^T\mathbf{Z}$ by $\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}$, where $\eta_0(\cdot)$ is an unknown function. We call these *generalized partially linear single-index models* (GPLSIM). The models include the "single-index" models, which have $\beta_0 = 0$. Using local linear methods, we propose estimates of the unknown parameters (α_0, β_0) and the unknown function $\eta_0(\cdot)$ and obtain their asymptotic distributions. Examples illustrate the models and the proposed estimation methodology.

KEY WORDS: Asymptotic theory; Generalized linear models; Kernel regression; Local estimation; Local polynomial regression; Nonparametric regression; Quasi-likelihood.

1. INTRODUCTION

1.1 Motivation

The Framingham Heart Study (Kannel et al. 1986) comprises a series of exams taken 2 years apart. For the purpose of illustration, we use Exam #3 as the baseline. The dataset includes 1,615 men age 31–65, with the outcome indicating the occurrence of coronary heart disease (CHD) within an 8-year period following Exam #3; there were 128 such cases of CHD. Predictors used in this example are patient's age, smoking status, and serum cholesterol level, in addition to systolic blood pressure (SBP) at Exam #3, the latter being the average of two measurements taken by different examiners during the same visit.

For these data, let the response Y be the incidence of CHD and let Z be the indicator of smoking status. The other covariates used are a vector, denoted by $\mathbf{X}$, consisting of the three variables X_1 (age of patient), X_2 ($= \log(\text{SPB} - 25)$), and X_3 ($= \log(\text{cholesterol level})$). An ordinary logistic regression model says that the logit of CHD probabilities satisfies

$$\operatorname{logit}\{P(\text{CHD}|\mathbf{X}, Z)\} = \gamma_0 + \alpha_0^T\mathbf{X} + \beta_0 Z. \tag{1}$$

The advantage of the linear-logistic model lies not only in its computational convenience, but also (and more importantly) in the ease of interpretation of the model parameters and our ability to make inference about them.

As we discuss in Section 3.2, some curvature is not captured by this linear-logistic model. This article is concerned with simple semiparametric alternatives to the fully parametric model (1) that allow for such curvature but yet retain the ease of interpretation of parameters such as α_0 and β_0. In this particular example, our generalization consists of two parts: (a) the linear combination $\alpha_0^T\mathbf{X}$ enters the model via a nonparametric link function, and (b) smoking status $\beta_0 Z$ enters the model as a logistic offset. Combining (a) and (b) suggests the simple model

$$\operatorname{logit}\{P(\text{CHD}|\mathbf{X}, Z)\} = \eta_0(\alpha_0^T\mathbf{X}) + \beta_0 Z \tag{2}$$

for some completely unknown function η_0. Model (2) retains much of the ease of interpretation of model (1), in the sense that nonzero components of α_0 or β_0 indicate a "significant" predictor of CHD, but model (2) allows for curvature in the logit.

The purpose of this article is to introduce versions of (2) for generalized linear and quasi-likelihood models, describe a way to fit such models, and derive an asymptotic theory that allows inference about the parameters (α_0, β_0). In the rest of this section, we describe the general class of models of interest to us here, which we call *generalized partially linear single-index models* (GPLSIM). We show that these models are a natural combination and generalization of simpler models already in the literature, namely single-index models and partially linear models. Further sections deal with fitting and making inference about GPLSIM. In particular, we present a class of asymptotically optimal estimators of the unknown parameters.

1.2 The Models

We consider semiparametric versions of generalized linear models where a response Y is to be predicted by covariates $(\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ and $\mathbf{Z}$ are possibly vector-valued predictors of lengths p and q. Generalized linear models are derived as follows. The conditional density of Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ belongs to a canonical exponential family

$$f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x}, \mathbf{z}) = \exp[y\theta(\mathbf{x}, \mathbf{z}) - \mathcal{B}\{\theta(\mathbf{x}, \mathbf{z})\} + C(y)] \tag{3}$$

for known functions $\mathcal{B}$ and C. In parametric generalized linear models, the unknown regression function $\mu(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \mathcal{B}'\{\theta(\mathbf{x}, \mathbf{z})\}$ is modeled linearly via a link function g by

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \gamma_0 + \boldsymbol{\alpha}_0^T\mathbf{x} + \boldsymbol{\beta}_0^T\mathbf{z}. \tag{4}$$

R. J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843. Jianqing Fan is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. Irène Gijbels is Associate Professor, Institut de Statistique, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. M. P. Wand is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, Sydney 2052, Australia. Carroll's research was supported by National Cancer Institute grant CA-57030. Fan's research was supported by National Science Foundation (NSF) grant DMS-9203135 and an NSF postdoctoral fellowship. Gijbels' research was supported by the Programme d'Action de Recherche Concertée, No. 93-98-164. The authors thank the referee and associate editor for many helpful comments that significantly improved the article.

Journal of the American Statistical Association
June 1997, Vol. 92, No. 438, Theory and Methods

If $g = (\mathcal{B}')^{-1}$ (the inverse function of $\mathcal{B}'$), then g is the canonical link function (see McCullagh and Nelder 1989 for more details).

In many practical situations, however, the linear model (4) is not complex enough to capture the underlying relationship between the response variable and its associated covariates. Indeed, some components can be highly nonlinear. A natural generalization of (4) is to allow only some of the predictors to be modeled linearly, with others being modeled nonlinearly. This leads us to consider the class of GPLSIM,

$$g\{\mu(\mathbf{x},\mathbf{z})\} = \eta_0(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}, \quad \text{with} \quad \|\boldsymbol{\alpha}_0\| = 1. \quad (5)$$

The restriction $\|\boldsymbol{\alpha}_0\| = 1$ is required for identifiability.

Model (5) is flexible enough to cover a variety of situations. When $\boldsymbol{\beta}_0 = 0$ or, equivalently, there are no predictors $\mathbf{Z}$, (5) is simply a generalized linear model with an *unknown* link function. The problem of the "missing link" function in generalized linear models has been considered previously by Weisberg and Welsh (1994). In other contexts, when only the mean function is specified, this problem is known as the nonparametric single-index model (Härdle, Hall, and Ichimura 1993). The appeal of these models is that by focusing on an index $\boldsymbol{\alpha}_0^T\mathbf{X}$, the so-called "curse of dimensionality" in fitting multivariate nonparametric regression functions is avoided (albeit at the cost of some loss in flexibility). Other recent work on estimation in the framework of single-index models was done by Bonneu, Delecroix, and Hristache (1995).

The meaning of the single-index parameter $\boldsymbol{\alpha}_0$ deserves a short explanation. Here we basically follow the lead of Li (1991), who noted three points:

a. Clearly, as a practical matter, lowering dimensionality before fitting data is important (Li's remark 1.2 goes even further and suggests that in many cases this is the crucial step), and the appeal of single-index models is that they provide a readily interpretable means of performing this reduction.
b. If $\eta_0(\cdot)$ is monotone, then $\boldsymbol{\alpha}$ takes on the same general meaning as "effect" parameters as would occur in ordinary linear models.
c. Given an estimated "direction" $\boldsymbol{\alpha}_0$, model criticism becomes a more manageable proposition.

Severini and Staniswalis (1994) considered model (5) but with $\eta_0(\boldsymbol{\alpha}_0^T\mathbf{x})$ replaced by $\gamma(\mathbf{x})$, a p-variate function. Hunsberger (1994) considered model (5) but with $\mathbf{X}$ scalar, so that $p = 1$ and $\boldsymbol{\alpha}_0 = 1$. In this case model (5) becomes

$$g\{\mu(\mathbf{x},\mathbf{z})\} = \eta_0(\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}. \quad (6)$$

Model (6) is particularly popular in the spline literature. (See, e.g., Chen 1988, Cuzick 1992, Heckman 1986, Speckman 1988, and Wahba 1984, where it is called the partial spline model or the partially linear model.) Recently, Mammen and van de Geer (1995) studied penalized quasi-likelihood estimation in partially linear models.

A different approach to modeling (and coping with the "curse of dimensionality") is through generalized additive models (GAM's) (see Hastie and Tibshirani 1990). These models replace the nonparametric component of (5) by a sum of nonparametric functions over the components of $\mathbf{X}$. When they adequately fit the data, the GPLSIM (5) have the obvious advantage of being more parsimonious, although they are clearly more difficult to compute given the existence of commercial software for GAM's. We have in our own work combined the two to fit models of the form (5), with an estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}_0$ obtained using our techniques and then GAM applied to $\mathbf{Z}$ and $\hat{\boldsymbol{\alpha}}^T\mathbf{X}$. In this context, one can think of our techniques as providing a preliminary dimension reduction. Clearly, an important issue for future work is to test for model misspecification of the GPLSIM against a richer class of models.

There are also various schools of thought about the need to use parsimonious parametric models (see Royston and Altman 1994, and the discussions therein). GPLSIM fall somewhere between the fully parametric flexible models of Royston and Altman (1994) and the almost fully nonparametric models of Hastie and Tibshirani (1990).

1.3 Aim and Outline

In the context of the unknown link function, the single-index model, or the model with $\boldsymbol{\alpha}_0 = 1$, our method differs from those methods previously cited in that we use local linear rather than simple kernel regression methods. Our aim is to estimate the unknown parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ and the unknown function $\eta_0(\cdot)$ in the full model (5), thus generalizing both the single-index model and the partially linear model. Our work also applies to quasi-likelihood models, where only the relationship between the mean and the variance is specified. In this situation estimation of the mean can be achieved by replacing the conditional log-likelihood $\ln f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x},\mathbf{z})$ by a *quasi-likelihood function* $Q\{\mu(\mathbf{x},\mathbf{z}),y\}$. If the conditional variance is modeled as $\text{var}(Y|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) = \sigma^2 V\{\mu(\mathbf{x},\mathbf{z})\}$ for some known positive function V, then the corresponding quasi-likelihood function $Q(w,y)$ satisfies

$$\frac{\partial}{\partial w} Q(w,y) = (y-w)/V(w) \quad (7)$$

(McCullagh and Nelder 1989, chap. 9). The quasi-score (7) possesses properties similar to those of the usual likelihood score function.

In Section 2 we propose estimation procedures, and in Section 3 we illustrate their performance via simulation and examples. In Sections 4 and 5 we describe distribution theory. In Section 6 we present the result showing asymptotic efficiency of the parametric estimators (in the semiparametric sense). In Section 7 we provide methods for estimating the standard errors of the parametric and nonparametric parts of the model. The usual method for estimating standard errors is to derive a formula for the asymptotic covariance matrix, and then plug into this formula to obtain an estimated covariance matrix. Unfortunately, as a general principle this has the drawback that the formula for the asymptotic covariance matrix requires additional nonparametric regression. We derive consistent covariance matrix estimates that avoid these additional nonparametric regressions. We give some implementation details in Sections 3.2 and 8, and discuss the issue of incorporating interactions in the model in Section 9. Proofs are given in the Appendix.

2. MAXIMUM QUASI-LIKELIHOOD

2.1 The Estimation Method

Under model (5), the primary interest is to estimate $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, and $\eta_0(\cdot)$. Because $\eta_0(\cdot)$ is modeled nonparametrically, it is natural to consider *local* quasi-likelihood. However, efficient estimation of the global parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ requires using all data points and hence should rely on the *global* quasi-likelihood. In local quasi-likelihood, we approximate $\eta_0(\cdot)$ locally by a linear function

$$\eta_0(v) \approx \eta_0(u) + \eta_0'(u)(v-u) \equiv a + b(v-u)$$

for v in a neighborhood of u, where $a = \eta_0(u)$ and $b = \eta_0'(u)$. Let K be a symmetric probability density function and let $K_h(t) = K(t/h)/h$ be a rescaling of K. The function K is usually called a kernel function, and the parameter h is called the bandwidth. For $i = 1, \ldots, n$, a sample $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is observed. The local quasi-likelihood is really a weighted quasi-likelihood, with weights $K_h(\boldsymbol{\alpha}^T\mathbf{X}_i - u)$.

The estimation procedure for estimating $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$ is as follows:

Step 0 (Initialization step). Fit a parametric generalized linear model to obtain initial values $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}})$, and set $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_1/\|\hat{\boldsymbol{\alpha}}_1\|$.

Step 1. Find $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \hat{a}$ by maximizing the local quasi-likelihood

$$\sum_{i=1}^{n} Q[g^{-1}\{a + b(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}, Y_i]K_h(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) \tag{8}$$

with respect to a and b. We take h to be an estimate of the bandwidth that is optimal for estimation of $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$.

Step 2. Update $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ by maximizing

$$\sum_{i=1}^{n} Q[g^{-1}\{\hat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}_i; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\beta}^T\mathbf{Z}_i\}, Y_i] \tag{9}$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Step 3. Continue Steps 1 and 2 until convergence.

Step 4. Fix $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ at its estimated value from Step 3. The final estimate of $\eta_0(\cdot)$ is $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \hat{a}$, where $(\hat{a}, \hat{b})$ is obtained by maximizing (8). At this final step, we take h to be an estimate of the bandwidth that is optimal for estimation of $\eta_0(\cdot)$ when $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are known.

The basic idea behind the foregoing algorithm is simple: estimate $\eta_0(\cdot)$ locally via (8), and then use all of the data and (9) to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, with $\hat{\eta}(\cdot)$ replacing $\eta_0(\cdot)$. We briefly discuss an alternative estimator in Section 4.1. We recommend calculating $\hat{\eta}(\cdot; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ at a *fixed* but fine grid of points and using linear interpolation to calculate the other values of $\hat{\eta}(\cdot; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ when needed.

The estimation procedure involves choosing a smoothing parameter on two quite different levels. In Steps 1 and 2 of the algorithm the aim is estimation of the parametric part $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, and hence here the bandwidth h should be optimal for this task. In Step 4, however, the goal is to estimate the nonparametric part $\eta_0(\cdot)$, and hence the bandwidth h should be optimal in this respect.

Finally, we mention that following work of Severini and Staniswalis (1994), maximizing

$$\sum_{i=1}^{n} Q[g^{-1}\{\hat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}_i; h, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{Z}_i\}, Y_i] \tag{10}$$

instead of (9) leads to estimates that are asymptotically equivalent to those resulting from the foregoing algorithm. We make use of this fact later, but for brevity we do not provide the calculations. The statement is true when working with the function Q as in (7), but it does not hold for completely arbitrary functions Q.

2.2 Alternatives

The algorithm suggested here uses local linear weighted fits based on kernel weights with a fixed global bandwidth. One may replace these by more sophisticated smoothers, such as those using higher-degree polynomials, locally varying bandwidths, nearest neighbor weights, and so on. Other nonkernel smoothers, such as splines, also may be used.

3. NUMERICAL EXAMPLES

3.1 Simulation

We ran a small simulation study with $n = 200$ and data generated according to the "sine-bump" model

$$Y_i = \sin\{\pi(\alpha^T X_i - A)/(B - A)\} + \beta Z_i + \varepsilon_i,$$

where the X_i are trivariate with independent uniform (0, 1) components, $Z_i = 0$ for i odd and $Z_i = 1$ for i even, and the ε_i are normally distributed with mean 0 and variance .01. The parameters were $\alpha = (1, 1, 1)/\sqrt{3}$ and $\beta = .3$. We took $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$ to ensure that the design was relatively thick in the tails. The number of replications was 100.

In this particular simulation, the GPLSIM estimates are far more accurate than the ordinary least squares (OLS) estimates, which are badly biased, and comparable to estimates obtained using nonlinear least squares based on the sinusoidal model. Table 1 displays the results of five randomly selected outcomes of the simulations. Note that although GPLSIM estimates are asymptotically efficient in the semiparametric sense (see Sec. 6), asymptotically they are more variable than fully parametric estimators computed at the correct model (see Theorem 4, Sec. 5.2), and this intrinsic difference between semiparametric and (correctly specified) parametric modeling exhibits itself here in the coefficient for Z. Not only are the GPLSIM estimates better than the OLS estimates, but they also do a reasonably effective job of fitting the data; see Figure 1.

Finally, we evaluated the accuracy of the estimated standard errors (defined in Sec. 7). In this simulation the cov-

Table 1. Results From Five Randomly Chosen Samples From the Sine-Bump Simulation Study

	Ordinary least squares				*Nonlinear least squares*				*GPLSIM*			
	X_1	X_2	X_3	Z	X_1	X_2	X_3	Z	X_1	X_2	X_3	Z
Est.	.564	.498	.659	.403	.595	.571	.565	.286	.595	.568	.569	.274
s.e.	.361	.368	.298	.069	.012	.013	.012	.012	.013	.013	.013	.026
Est.	.218	.142	.966	.216	.568	.568	.595	.277	.563	.574	.595	.281
s.e.	.766	.781	.207	.054	.010	.009	.009	.010	.010	.010	.010	.022
Est.	−.126	−.512	−.85	.263	.579	.580	.572	.310	.581	.580	.571	.310
s.e.	1.137	.97	.596	.059	.010	.010	.009	.010	.011	.011	.010	.023
Est.	.851	−.264	−.453	.351	.567	.590	.575	.300	.565	.599	.568	.307
s.e.	.796	1.364	1.349	.068	.010	.011	.011	.011	.012	.013	.013	.023
Est.	−.881	.396	−.261	.323	.587	.574	.570	.283	.592	.569	.571	.291
s.e.	.697	1.309	1.496	.064	.010	.010	.010	.010	.011	.010	.010	.020
MSE	.67	.79	.76	3.9e−3	1.1e−4	1.2e−4	1.1e−4	1.1e−4	1.4e−4	1.6e−4	1.3e−4	2.7e−4
MAE	.69	.75	.74	5.0e−2	8.5e−3	8.6e−3	8.5e−3	9.0e−3	9.6e−3	9.9e−3	9.0e−3	1.3e−2
mdE	.69	.79	.76	3.8e−2	7.9e−3	7.5e−3	6.9e−3	8.6e−3	8.6e−3	8.6e−3	8.2e−3	1.1e−2

NOTE: Mean squared error (MSE), mean absolute error (MAE), and median absolute error (mdE) values for the whole simulation are also given.

erage probabilities for nominal 95% confidence intervals were 94%, 96%, and 98% for the three components of $\mathbf{X}$, and 94% for Z. At least for this sample size and this model, the standard error estimates seem reasonably accurate.

3.2 Example: Framingham Data

The Framingham data were described in Section 1.1; Y corresponds to incidence of CHD and Z to smoking status. In this discussion we use disease and smoker to denote these variables. For covariates we used X_1, X_2, and X_3 as described in Section 1.1, with each variable scaled to lie between 0 and 1. To avoid problems with sparse data near the boundaries, after some experimentation we used only those data with a single-index value in range [.4, 1.2] for curve estimation. This excluded 45 of the 1,615 observations. We applied our methodology to the model

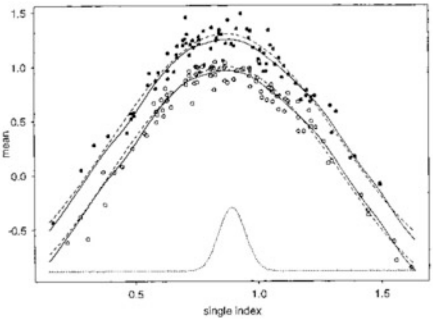


Figure 1. Curve Estimates for a Single Replication of the Sine-Bump Simulation Study. The data are shown by open circles for $Z = 0$ and closed circles for $Z = 1$. The solid curves correspond to the estimates of the underlying mean function when $Z = 0$ and $Z = 1$. The dashed curves are the true mean functions. The dotted curve is the kernel weight used in the local fitting process.

$$\begin{aligned}\operatorname{logit}\{P(\text{disease} &= 1|\text{age, trblood, logchol, smoker})\} \\ &= \eta_0\{\alpha_{01}(\text{age}) + \alpha_{02}(\text{trblood}) \\ &\qquad + \alpha_{03}(\text{logchol})\} + \beta_0(\text{smoker}).\end{aligned}$$

We used the bandwidth h_{opt} defined in (17), obtaining nearly identical results with or without the modification suggested in the discussion centering on (18). Table 2 displays the results of our analysis. For the purpose of illustration, we have compared these results to those obtained by ordinary logistic regression, which in this context is simply another way of estimating the "direction" α_0. We made the ordinary logistic regression coefficients for age, trblood, and logchol comparable to the single-index analysis by making their Euclidean norm equal to 1.0, and adjusted their standard error estimates accordingly.

Figure 2 shows the estimates of (a) η_0 and (b) the conditional probability of heart disease for both smokers and nonsmokers. An interesting feature of this figure is the curvature of $\hat{\eta}$ when the single index becomes greater than .8. We checked this curvature in two ways. First, we used the ordinary logistic regression estimates to define a single index, and then to this index and the smoking indicator fit a partially linear model to the data using the GAM procedure of S-PLUS. The resulting estimate also showed curvature, of the same form as displayed in Figure 2. We also fit an ordinary GAM with nonparametric components in age, trblood, and logchol, and found a nonlinear structure with the "flatness" of Figure 2 for age.

We compared the GPLSIM fit to others as follows. First, we formed the estimated single-index $U = \hat{\alpha}^T\mathbf{X}$, then ran

Table 2. Framingham Heart Study

	age	*trblood*	*logchol*	*smoker*
Ordinary logistic	.43	.57	.69	.57
s.e.	.10	.13	.11	.25
GPLSIM	.37	.65	.66	.59
s.e.	.086	.11	.12	.24

NOTE: "trblood" is transformed systolic blood pressure, "logchol" is the log of serum cholesterol, and "smoker" is smoking status. The ordinary logistic coefficients for age, trblood, and logchol have been normalized to have Euclidean norm equal to 1.0, and the standard errors have been adjusted appropriately.

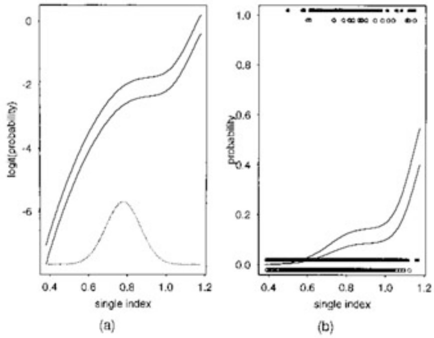


Figure 2. Curve Estimates for the Framingham Heart Study Data. (a) Solid curves correspond to estimates of logit P (heart disease) for smokers (upper curve) and nonsmokers (lower curve) against the estimated single-index described in the text. The dotted curve is the kernel weight used in the local linear fitting process. (b) Estimates of P (heart disease) for smokers (upper curve) and nonsmokers (lower curve) against the single index described in the text. The solid dot denotes smokers, the hollow dot, nonsmokers.

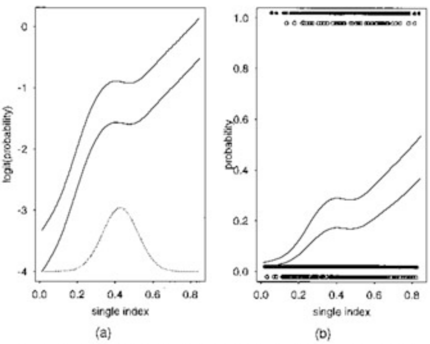


Figure 3. Curve Estimates for the Munich Dust Study Data. (a) Solid curves correspond to estimates of logit pr (Bronchitis) for smokers (upper curve) and nonsmokers (lower curve) against the estimated single-index described in the text. The dotted curve is the kernel weight used in the local linear fitting process. (b) Estimates of pr (Bronchitis) for smokers (upper curve) and nonsmokers (lower curve) against the single index described in the text. The solid dot denotes smokers; the hollow dot, nonsmokers.

a partially linear model in U (nonparametric) and Z (parametric) using the default "GAM" procedure in S-PLUS. We also ran a standard GAM with smoothers for each of X_1, X_2, and X_3, with Z entering as a parametric offset. In this case, surprisingly the GPLSIM had a *smaller* estimated deviance than the GAM, even though it had ≈ 8 more degrees of freedom.

One can also view this example as an informal model diagnostic of the logistic linear regression model via embedding it into the GPLSIM. Our result indicates certain departures from the logistic linear regression model; using the same informal method described in the previous paragraph, the linear logistic and the full GAM are not statistically significantly different, but the linear logistic and the GPLSIM are statistically significantly different.

3.3 Example: Dust Irritation Data

In occupational medicine one important issue is the assessment of the health hazard of specific harmful substances in a working area. We consider here the specific problem of estimating risk of bronchitis in a dust-burdened mechanical engineering plant in Munich.

The regressor variables $\mathbf{X}$ are X_1, the logarithm of 1.0 plus the average dust concentration in the working area over the period of time in question, and X_2, the duration of exposure. Also available was smoking status, Z. The data were described by Ulm (1991) as a possible example of a threshold regression model and were further analyzed by Küchenhoff and Carroll (1997). There were 1,246 observations. Little correlation among the variables was observed.

Table 3. Munich Dust Study

	trdust	*duration*	*smoker*
Ordinary logistic	.403	.915	.68
s.e.	.103	.045	.176
GPLSIM	.222	.975	.668
s.e.	.089	.021	.178

NOTE: "trdust" is transformed dust concentration, "duration" is the duration of exposure, and "smoker" is smoking status. The ordinary logistic coefficients for trdust and duration have been normalized to have Euclidean norm equal to 1.0, and the standard errors have been adjusted appropriately.

Table 3 displays the results of an ordinary logistic and GPLSIM fit to the data, and Figure 3 shows the logit and probability of bronchitis for smokers and nonsmokers. There is an important curvature in these data, which are not well fitted by an ordinary logistic model. As suggested by Küchenhoff and Carroll (1997), this curvature may reflect a threshold effect on concentration. The single-index model provides a slightly worse fit than a full GAM, although not a statistically significant one; we compared these using the deviances from GAM as implemented in S-PLUS, ignoring the effect of estimation of the single index. When compared to the GAM, an ordinary logistic model had an observed level of significance $< .0001$.

4. DISTRIBUTION THEORY: NONPARAMETRIC PART

4.1 Introduction

When $\boldsymbol{\alpha}_0$ is given as is the case in partially linear models or can be estimated at reasonable accuracy (e.g., by the average derivative method or sliced inverse regression), the following simple estimator is attractive from an implementation viewpoint. With the given value of $\hat{\boldsymbol{\alpha}}$, find $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}}) = \hat{a}$ by maximizing the local quasi-likelihood

$$\sum_{i=1}^{n} Q[g^{-1}\{a + b(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \boldsymbol{\beta}^T\mathbf{Z}_i\}, Y_i]K_h(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u), \qquad (11)$$

with respect to a, b, and $\boldsymbol{\beta}$. Because $\hat{\boldsymbol{\beta}}$ here is obtained locally, it can be improved to use all of the data, as follows.

Given $\hat{\boldsymbol{\alpha}}$ and the estimator $\hat{\eta}(u; h, \hat{\boldsymbol{\alpha}})$, one estimates $\hat{\boldsymbol{\beta}}$ by maximizing

$$\sum_{i=1}^{n} Q[g^{-1}\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T \mathbf{X}_i; h, \hat{\boldsymbol{\alpha}}) + \boldsymbol{\beta}^T \mathbf{Z}_i\}, Y_i] \tag{12}$$

with respect to $\boldsymbol{\beta}$. We call this noniterative procedure the *one-step estimator* and the algorithm of Section 2.1 the *fully iterated algorithm*. Based on he distribution theory provided in this and the next section, it is clear that both algorithms have their own merits. The fully iterated algorithm is at least as efficient as the one-step algorithm, but the one-step estimator achieves the same efficiency in some important applications with added computational convenience.

Note that in (11) we are maximizing the local quasi-likelihood with respect to $(a, b, \boldsymbol{\beta})$. This reflects the main difference from the estimation algorithm of Section 2.1, where we maximize with respect to (a, b) only. The foregoing idea can also be expanded to the case where $\boldsymbol{\alpha}$ is unknown by iteratively maximizing (11) and (12); one needs only to replace the first $\hat{\boldsymbol{\alpha}}$ in (12) by $\boldsymbol{\alpha}$ and maximize the modified (12) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. See an earlier version of this article (Carroll, Fan, Gijbels, and Wand 1995) for details.

In this section we investigate properties of the estimators of the nonparametric part $\eta_0(\cdot)$ of (5) when α_0 is either known or estimated to the order $O_P(n^{-1/2})$ (i.e., at the usual parametric rate). The distribution theory depends on two cases: (a) the one-step approach, where $\boldsymbol{\beta}_0$ is estimated locally as in (11); and (b) the fully iterated approach (8), where $\boldsymbol{\beta}_0$ is estimated at parametric rates and thus $\eta_0(\cdot)$ can be estimated asymptotically as well as if $\boldsymbol{\beta}_0$ were known.

4.2 One-Step Estimate of the Nonparametric Part

Let $\rho_l(t) = \{dg^{-1}(t)/dt\}^l/[\sigma^2 V\{g^{-1}(t)\}]$, $l = 1, 2$, and denote the marginal density of $U = \boldsymbol{\alpha}_0^T \mathbf{X}$ by $f(\cdot)$. For the model (3) with the canonical link function $g = (B')^{-1}$, we have $\rho_2\{g(\mu)\} = \sigma^2 V(\mu)$. Define $\kappa_j = \int t^j K(t)\,dt$, $\nu_j = \int t^j K^2(t)\,dt$, and

$$\boldsymbol{\Sigma}(u) = E\left[\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\} \times \begin{pmatrix} 1 & \mathbf{Z}^T \\ \mathbf{Z} & \mathbf{Z}\mathbf{Z}^T \end{pmatrix} \middle| U = u\right], \tag{13}$$

$$q_1(x, y) = \{y - g^{-1}(x)\}\rho_1(x),$$

$$m_i = m_i(U_i) = \eta_0(U_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i,$$

$W_i =$ first element of the vector $q_1(m_i, Y_i)\boldsymbol{\Sigma}^{-1}(u)(1, \mathbf{Z}_i^T)^T$,

and

$$d(u) = \text{first diagonal element of the matrix } \boldsymbol{\Sigma}^{-1}(u).$$

Theorem 1. Consider the maximizer of the local quasi-likelihood (11). Then, as $n \to \infty$, $h \to 0$, and $nh \to \infty$, under Condition 1 in the Appendix,

$$(nh)^{1/2}\left(\begin{bmatrix} \hat{\eta}(u) - \eta_0(u) \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{bmatrix} - \frac{\kappa_2}{2}\,\eta_0''(u)h^2\boldsymbol{\Sigma}^{-1}(u)E \times \left[\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\}\begin{pmatrix} 1 \\ \mathbf{Z} \end{pmatrix}\middle| U = u\right]\right) \xrightarrow{D} \text{normal}\left[0, \frac{\nu_0}{f(u)}\,\boldsymbol{\Sigma}^{-1}(u)\right]. \tag{14}$$

In fact, we have the asymptotic expansion

$$\hat{\eta}(u) - \eta_0(u) = (\kappa_2/2)\eta_0''(u)h^2 + \frac{1}{nf(u)}\sum_{i=1}^{n} W_i K_h(\boldsymbol{\alpha}_0^T \mathbf{X}_i - u) + o_P\{(nh)^{-1/2} + h^2\}, \tag{15}$$

and hence

$$(nh)^{1/2}\left\{\hat{\eta}(u) - \eta_0(u) - \frac{\kappa_2}{2}\,\eta_0''(u)h^2\right\} \xrightarrow{D} \text{normal}\left[0, \frac{\nu_0}{f(u)}\,d(u)\right]. \tag{16}$$

Remark 1. Consider the situation where $\sigma^2 V(\mu) \equiv \sigma^2$ and $E(\mathbf{Z}|\mathbf{X}) = 0$. For this normal model with the identity link, the quasi-likelihood estimates are the OLS estimates. It is easily seen that $d(u) = \sigma^2$. Hence in this particular case, even though $\boldsymbol{\beta}_0$ is estimated locally, the bias and variance of $\hat{\eta}(u)$ are the same as if $\boldsymbol{\beta}_0$ were known.

Remark 2. The rate results in Theorem 1 continue to hold when the variance function is misspecified; that is, $\text{var}(Y|\mathbf{X}, \mathbf{Z}) \neq \sigma^2 V\{\mu(\mathbf{X}, \mathbf{Z})\}$. One must change the matrix $\boldsymbol{\Sigma}(u)$ to reflect the misspecification of the variance function. (See Fan, Heckman, and Wand 1995 for such a modification.)

4.3 Fully Iterated Estimate of the Nonparametric Part

For the fully iterative estimator, the parametric component can be estimated at root-n rate. Thus in Step 4 the local smoothing is carried out as if $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ were known. The results for the nonparametric component are easy: (16) continues to hold, replacing $d(u)$ by $d_*(u) = (E[\rho_2\{\eta_0(u) + \boldsymbol{\beta}_0^T \mathbf{Z}\}|U = u])^{-1}$. This result coincides with the univariate result given by Fan et al. (1995).

4.4 Bandwidth Selection

The results in the previous section suggest bandwidth estimators in the spirit of that of Ruppert, Sheather, and Wand (1995). For example, consider estimation of $\eta_0(\cdot)$ at the final step. For a given function $w(\cdot)$ with compact support, minimizing the asymptotic weighted mean squared error with weight $f(\cdot)w(\cdot)$ yields the optimal global bandwidth

$$h_{\text{opt}} = C(K)n^{-1/5}\left\{\frac{\int d_*(u)w(u)\,du}{\int \eta_0''(u)^2 f(u)w(u)\,du}\right\}^{1/5}, \tag{17}$$

where $C(K) = (\nu_0 \kappa_2^{-2})^{1/5}$.

The Framingham example in Section 3.2 treats the case where both Y and Z are $0-1$ variables, so we briefly describe a rough rule for choosing the bandwidth in this context. Extension to other contexts is straightforward. For the Bernoulli likelihood with logit link,

$$d_*(u)^{-1} = \frac{e^{\eta_0(u)}\{1-\zeta_0(u)\}}{\{1+e^{\eta_0(u)}\}^2} + \frac{e^{\eta_0(u)+\beta_0}\zeta_0(u)}{\{1+e^{\eta_0(u)+\beta_0}\}^2},$$

where $\zeta_0(u) = P(Z=1|U=u)$. Let $\hat{\eta}_Q(\cdot)$ be the quadratic and $\hat{\zeta}_L(\cdot)$ be the linear logistic regression estimates of $\eta_0(\cdot)$ and $\zeta_0(\cdot)$. Let $\hat{\beta}$ be the estimate of β_0 from the previous iteration. Then the integral on the numerator of (17) can be estimated by direct replacement of $\eta_0(\cdot)$, $\zeta_0(\cdot)$, and β_0 by $\hat{\eta}(\cdot)$, $\hat{\zeta}_L(\cdot)$, and $\hat{\beta}$. An estimate for the integral on the denominator is $n^{-1}\sum_{i=1}^{n} \hat{\eta}_Q''(U_i)^2 w(U_i)$. A sensible choice for w is the indicator function on the range of the U_i, with approximately 10% clipped off each end to avoid boundary problems. This results in an estimated bandwidth, $\hat{h}_{\text{opt}}$, for use in Step 4 of the fully iterated algorithm. The rule will give close to optimal answers when the true logit$\{\eta_0(\cdot)\}$ and logit$\{\zeta_0(\cdot)\}$ are approximated reasonably well by a quadratic and a straight line.

A sensible rule for choice of h in Step 1 is more difficult. A relatively ad hoc possibility is

$$\hat{h}_{\text{opt}} \times n^{1/5} \times n^{-1/3} = \hat{h}_{\text{opt}} \times n^{-2/15}, \qquad (18)$$

because this guarantees that the required bandwidth has correct order of magnitude for the conjectured optimal asymptotic performance. (See Remark 3 in Sec. 5.1 for more details.)

5. DISTRIBUTION THEORY: PARAMETRIC PARTS

We now study estimation for the parametric components $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. We treat the one-dimensional case $(p=1)$, for which $\boldsymbol{\alpha}_0 = 1$ and $\boldsymbol{\alpha}_0^T\mathbf{X} = X$, separately. Because in this case the one-step estimator has the advantage of being noniterative, we also provide its distribution theory.

5.1 The Scalar X Case: Partially Linear Models

The following theorem for the one-step estimate shows that one iteration leads already to a root-n consistent estimator.

Theorem 2. Let $\hat{\beta}$ be the one-step estimate that maximizes the quasi-likelihood (12) with $\boldsymbol{\alpha} = 1$. Because $U = \boldsymbol{\alpha}_0^T\mathbf{X} = X$, write $\Sigma(U) = \Sigma(X)$ in Theorem 1. Under Conditions 1 and 2 in the Appendix, as $n \to \infty$, $nh^4 \to 0$, and $nh^2/\log(1/h) \to \infty$,

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \mathbf{B}^{-1}\boldsymbol{\Sigma}_1\mathbf{B}^{-1}), \qquad (19)$$

where $\mathbf{B} = E[\rho_2\{\eta_0(X) + \boldsymbol{\beta}_0^T\mathbf{Z}\}\mathbf{Z}\mathbf{Z}^T]$, $\boldsymbol{\Sigma}_1 = \mathbf{B} + E\{\gamma(X)\gamma^T(X)\mathbf{e}_1^T\boldsymbol{\Sigma}^{-1}(X)\mathbf{e}_1\}$, $\gamma(u) = E[\rho_2\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z}\}\mathbf{Z}|X = u]$, and $\mathbf{e}_1$ is the unit vector with 1 in the first position.

Theorem 3. Under the conditions of Theorem 2, for the fully iterated estimator defined by (9) with $\boldsymbol{\alpha} = 1$, with

$$\rho_2(\cdot) = \rho_2\{\eta_0(X) + \boldsymbol{\beta}_0^T\mathbf{Z}\},$$

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \mathbf{B}_2^{-1}), \qquad (20)$$

provided that $\boldsymbol{\beta}$ is maximized in a consistent neighborhood of $\boldsymbol{\beta}_0$. Here

$$\mathbf{B}_2 = E\{\mathbf{Z}\mathbf{Z}^T\rho_2(\cdot)\} - E\left[\frac{E\{\mathbf{Z}\rho_2(\cdot)|X\}E\{\mathbf{Z}^T\rho_2(\cdot)|X\}}{E\{\rho_2(\cdot)|X\}}\right].$$

The same result holds for the estimator defined by (9) under the weaker condition that $nh^6 \to 0$.

Remark 3. Theorem 2, which concerns the one-step estimator, has an important restriction on the bandwidth h, which precludes the nearly universally familiar optimal bandwidth rates for nonparametric regression, in which h is proportional to $n^{-1/5}$. Basically, our conditions require that to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ at the rate $n^{-1/2}$, one must undersmooth the nonparametric part $\eta_0(\cdot)$. The need to undersmooth to obtain usual rates of convergence is standard in the kernel literature and has analogs in the spline literature (Hastie and Tibshirani 1990, pp. 154–155). This undersmoothing is required for the estimator defined by (9). However, for the estimator defined by (10), in the linear regression single-index model with no $\mathbf{Z}$, ordinary bandwidth rates are permissible, as shown by Härdle et al. (1993), who suggested maximizing (10) simultaneously in the bandwidth and the parameters. Hunsberger (1994) and Severini and Staniswalis (1994) showed the same thing for the partially linear model (see also Severini and Wong 1992). Because ordinary bandwidths "work" for single-index models and also for partially linear models, it is reasonable to suppose that they also work for the combination, namely our GPLSIM's. A brief sketch of an argument was provided in an appendix of an earlier version of this article (Carroll et al. 1995), verifying that ordinary bandwidth rates are possible for full GPLSIM when (10) is maximized.

Remark 4. In the normal model with identity link function, an interesting simplification occurs. We set $E(\mathbf{Z}) = 0$ without loss of generality and define $q(\mathbf{X}) = E(\mathbf{Z}|\mathbf{Z})$. Then $\mathbf{B}_2 = \sigma^{-2}E\{\text{var}(\mathbf{Z}|\mathbf{X})\}$, whereas the asymptotic variance (19) for the one-step estimator is

$$\sigma^2[\{E\mathbf{Z}\mathbf{Z}^T\}^{-1} + Eq(\mathbf{X})q(\mathbf{X})^T \times \{1 - q(\mathbf{X})^T(E\mathbf{Z}\mathbf{Z}^T)^{-1}q(\mathbf{X})\}^{-1}].$$

Because $\mathbf{B}_2^{-1} = \sigma^2\{E\mathbf{Z}\mathbf{Z}^T - q(\mathbf{X})q(\mathbf{X})^T\}^{-1}$, one can easily see that the fully iterated estimator is uniformly as efficient or more efficient than the one-step estimator. However, when $\mathbf{X}$ and $\mathbf{Z}$ are independent, the one-step estimator is as efficient as the fully iterated estimator. Hence the one-step estimator is preferable when $\mathbf{X}$ and $\mathbf{Z}$ are weakly correlated, because it requires no iteration.

5.2 The Multivariate X Case: General Model

For a given $\hat{\eta}$, let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ maximize the global quasi-likelihood (9). We assume that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are in a $\sqrt{n}$ neighborhood of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$; that is, $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_P(n^{-1/2})$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$. Denote a generalized inverse of a square matrix $\mathbf{A}$ by $\mathbf{A}^{-1}$.

Theorem 4. Under Conditions 1 and 2 in the Appendix, the foregoing assumptions, and the restrictions on the bandwidths as stated in Theorem 3, for the estimators defined by (9) and (10),

$$n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\end{pmatrix}\xrightarrow{D}\text{normal}(0,\mathbf{Q}^{-1}), \tag{21}$$

where, if $\rho_2(\cdot)=\rho_2\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X})+\boldsymbol{\beta}_0^T\mathbf{Z}\}$,

$$\mathbf{Q}=E\left[\rho_2(\cdot)\left\{\begin{matrix}\mathbf{X}\eta_0'(U)\\ \mathbf{Z}\end{matrix}\right\}\left\{\begin{matrix}\mathbf{X}\eta_0'(U)\\ \mathbf{Z}\end{matrix}\right\}^T\right]$$
$$-E\left(\rho_2(\cdot)\left\{\begin{matrix}\mathbf{X}\eta_0'(U)\\ \mathbf{Z}\end{matrix}\right\}\right.$$
$$\left.\times\left[\begin{matrix}E\{\mathbf{X}\eta_0'(U)\rho_2(\cdot)|U\}/E\{\rho_2(\cdot)|U\}\\ E\{\mathbf{Z}\rho_2(\cdot)|U\}/E\{\rho_2(\cdot)|U\}\end{matrix}\right]^T\right).$$

Remark 5. When $\sigma^2V(\mu)=\sigma^2$, with identity link and no $\boldsymbol{\beta}$ component, Theorem 4 reduces to the result of Härdle et al. (1993) for the single-index model.

Remark 6. Consult Remark 3 after Theorem 3 for discussion of the bandwidth conditions.

6. ASYMPTOTIC EFFICIENCY IN THE SEMIPARAMETRIC SENSE

In this section we derive the information bound for the semiparametric model (3) and (5). This information bound turns out to be the matrix $\mathbf{Q}$ given in Theorem 4. Thus the estimator from Theorem 4 achieves the information lower bound and is efficient in the semiparametric sense.

To state the information bound, let us define the parameter space. Assume that η_0 is a completely unknown function with a continuous second derivative and that the joint density of $\mathbf{X}$ and $\mathbf{Z}$ with respect to some measure exists and is completely unknown.

Theorem 5. Under the foregoing assumptions, the information matrix for the semiparametric model (3) and (5) is $\mathbf{Q}$ given in Theorem 4.

7. INFERENCE AND STANDARD ERRORS

A consistent estimate of σ^2 is the weighted mean squared error of the residuals Y_i against their predicted mean, with weights $1/V\{\hat{\mu}(\mathbf{X}_i,\mathbf{Z}_i)\}$; one can use $n-l_n-p-q$ df, where l_n is the effective number of parameters used in estimating $\eta_0(\cdot)$. The rest of this section discusses estimating the other variance terms.

7.1 Estimation in Partially Linear Models: Scalar X

When X is scalar, so that $\boldsymbol{\alpha}_0=1$ is known, each of the terms in the limiting covariance matrices (19) and (20) can be estimated by nonparametric regression techniques. We focus on (20), for which this fairly tedious process can be replaced by a simple consistent alternative based on the usual expansions for quasi-likelihood. The derivations are based on the simple form (9), instead of taking derivatives in (10), because these are more complex to compute.

Set $U_i=\boldsymbol{\alpha}_0^T\mathbf{X}_i=X_i$ and $\tilde{\mathbf{Z}}=(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)^T$ and let $\tilde{\mathbf{A}}$ be diagonal with elements ρ_{2i}, where $\rho_{2i}\equiv\rho_2\{\eta(U_i)+\boldsymbol{\beta}^T\mathbf{Z}_i\}$. Further, set $\tilde{\boldsymbol{\eta}}=\{\eta(U_1),\ldots,\eta(U_n)\}^T$ and let $\tilde{\varepsilon}$ be the vector with ith element $\eta(U_i)+\boldsymbol{\beta}^T\mathbf{Z}_i+(Y_i-\mu_i)/(\sigma^2V_i\rho_{1i})$, where $\mu_i=g^{-1}\{\eta(U_i)+\boldsymbol{\beta}^T\mathbf{Z}_i\}$ and $V_i=V(\mu_i)$. The smoothing matrix is the $n\times n$ matrix

$$\tilde{\mathbf{S}}=\left[\begin{matrix}\mathbf{e}_1^T\{\mathbf{U}(U_1)^T\tilde{\mathbf{A}}\mathbf{K}(U_1)\mathbf{U}(U_1)\}^{-1}\mathbf{U}(U_1)^T\tilde{\mathbf{A}}\mathbf{K}(U_1)\\ \vdots\\ \mathbf{e}_1^T\{\mathbf{U}(U_n)^T\tilde{\mathbf{A}}\mathbf{K}(U_n)\mathbf{U}(U_n)\}^{-1}\mathbf{U}(U_n)^T\tilde{\mathbf{A}}\mathbf{K}(U_n)\end{matrix}\right], \tag{22}$$

where $\mathbf{U}(u_0)$ is the $n\times 2$ matrix with the first column all 1's and the second column with the terms $(U_i-u_0)/h$, and $\mathbf{K}(u_0)$ is diagonal with elements $K_h(U_i-u_0)$.

Here is the motivation for $\tilde{\mathbf{S}}$. For fixed $\boldsymbol{\beta}$ and u_0, note that the intercept $a(u_0)$ and h times the slope $b(u_0)$ from the local quasi-likelihood regression are the iterative solutions to the equation

$$\left[\begin{matrix}a(u_0)\\ hb(u_0)\end{matrix}\right]$$
$$=\left\{\sum_{k=1}^{n}\mathbf{U}_k(u_0)\mathbf{U}_k(u_0)^TK_h(U_k-u_0)A_k(u_0)\right\}^{-1}$$
$$\times\sum_{k=1}^{n}\mathbf{U}_k(u_0)K_h(U_k-u_0)A_k(u_0)$$
$$\times\{a(u_0)+b(u_0)(U_i-u_0)+(Y_i-\mu_i)/(\sigma^2V_i\rho_{1i})\}, \tag{23}$$

where $\mathbf{U}_k(u_0)=\{1,(U_k-u_0)/h\}^T$ and $A_k(u_0)=\rho_2\{\eta(u_0)+\boldsymbol{\beta}^T\mathbf{Z}_k\}$. Setting $u_0=U_i$ for $i=1,\ldots,n$ and multiplying both sides of (23) by $\mathbf{e}_1^T$ yields (22).

The following argument has similarities to equation (6.22) of Hastie and Tibshirani (1990, p. 154). Because of the local nature of the fit, the term $b(u_0)(U_i-u_0)$ in the last part of (23) can be ignored asymptotically. This means that the local quasi-likelihood algorithm is asymptotically equivalent to solving in $\boldsymbol{\beta}$ and η the equations

$$\boldsymbol{\beta}=(\tilde{\mathbf{Z}}^T\tilde{\mathbf{A}}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{A}}(\tilde{\varepsilon}-\tilde{\boldsymbol{\eta}})$$

and

$$\tilde{\eta}=\tilde{\mathbf{S}}(\tilde{\varepsilon}-\tilde{\mathbf{Z}}\boldsymbol{\beta}).$$

This means that the estimate of $\boldsymbol{\beta}_0$ is asymptotically equivalent to solving $\boldsymbol{\beta}=\tilde{\mathbf{H}}_1\tilde{\varepsilon}$, where

$$\tilde{\mathbf{H}}_1=\{\tilde{\mathbf{Z}}^T\tilde{\mathbf{A}}(\mathbf{I}-\tilde{\mathbf{S}})\tilde{\mathbf{Z}}\}^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{A}}(\mathbf{I}-\tilde{\mathbf{S}}).$$

Because $\tilde{\varepsilon}$ has covariance matrix $\tilde{\mathbf{A}}^{-1}$, an approximate covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\tilde{\mathbf{H}}_1\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{H}}_1^T$. One can show this estimate yields asymptotically consistent standard errors for $\hat{\boldsymbol{\beta}}$.

7.2 Estimation in General Models: Multivariate X

When $\boldsymbol{\alpha}_0$ is unknown, there are again two strategies: Nonparametric regression techniques can be used to estimate the terms in (21), or we can again develop directly a consistent estimate of (21). We build on the notation in Section 7.1.

Let $\tilde{\mathbf{Q}}$ be the $n \times p$ matrix with the ith row given as $\eta'(U_i)\mathbf{X}_i^T$, and let $\tilde{\mathbf{R}} = (\tilde{\mathbf{Q}}, \tilde{\mathbf{Z}})$. Let

$$\mathbf{P}_{\boldsymbol{\alpha}}^* = \begin{bmatrix} \mathbf{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^T & 0 \\ 0 & \mathbf{I} \end{bmatrix},$$

and let $\tilde{\varepsilon}$ be the vector with ith element $\eta(U_i) + \eta'(U_i)(\boldsymbol{\alpha}^T\mathbf{X}_i) + (\boldsymbol{\beta}^T\mathbf{Z}_i) + (Y_i - \mu_i)/\{\sigma^2 V_i \rho_{1i}\}$. Remembering that we must have $\|\boldsymbol{\alpha}\| = 1$ for identifiability, note that we find $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ by solving

$$0 = \tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\tilde{\varepsilon} - \tilde{\eta}) - \tilde{\mathbf{R}}^T\tilde{\mathbf{A}}\tilde{\mathbf{R}}\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \theta\boldsymbol{\alpha} \\ 0 \end{pmatrix},$$

where θ is a Lagrange multiplier associated with the constraint $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$. Of course, the same argument used in deleting a term explained following (23) is used here. Multiplying both sides by $P_{\boldsymbol{\alpha}}^*$ and solving, we find that $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T = (\mathbf{P}_{\boldsymbol{\alpha}}^*\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}\tilde{\mathbf{R}})^{-}\mathbf{P}_{\boldsymbol{\alpha}}^*\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\tilde{\varepsilon} - \tilde{\eta})$. Remembering that $\tilde{\eta} = \tilde{\mathbf{S}}(\tilde{\varepsilon} - \tilde{\mathbf{Q}}\boldsymbol{\alpha} - \tilde{\mathbf{Z}}\boldsymbol{\beta})$, we find after some algebra that $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T = \tilde{\mathbf{H}}_2\tilde{\varepsilon}$ and

$$\tilde{\mathbf{H}}_2 = \{\mathbf{P}_{\boldsymbol{\alpha}}^*\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\mathbf{I} - \tilde{\mathbf{S}})\tilde{\mathbf{R}}\}^{-}\mathbf{P}_{\boldsymbol{\alpha}}^*\tilde{\mathbf{R}}^T\tilde{\mathbf{A}}(\mathbf{I} - \tilde{\mathbf{S}}).$$

The estimated (and consistent) covariance matrix is $\tilde{\mathbf{H}}_2\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{H}}_2^T$.

8. IMPLEMENTATION

To cut down on the computational labor at the curve estimation stages, we used fast binned approximations (see, e.g., Fan and Marron 1994 and Härdle and Scott 1992). Binning methods can also be used for fast computation of the standard error estimates. Details of such calculations ere given by Turlach and Wand (1995). An S-PLUS/Fortran module for fitting GPLSIM in certain special cases is available from World Wide Web site http://www.agsm.unsw.edu.au/~wand/software.html.

9. DISCUSSION

Model (5) does not explicitly deal with interactions between $\mathbf{X}$ and $\mathbf{Z}$; for example, of the form

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \eta_{z_1}(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}_2, \qquad (24)$$

where $\mathbf{z} = (z_1, \mathbf{z}_2)$ with z_1 binary. However, our methods can be modified to handle (24). The local quasi-likelihood (8) should be replaced by

$$\begin{aligned} &\sum_{i=1}^{n} Q[g^{-1}\{a_0 + b_0(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_{2,i}\}, Y_i] \\ &\quad \times K_{h_0}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u)I(Z_{1,i} = 0) \\ &\quad + \sum_{i=1}^{n} Q[g^{-1}\{a_1 + b_1(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_{2,i}\}, Y_i] \\ &\quad \times K_{h_1}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u)I(Z_{1,i} = 1), \end{aligned}$$

where h_0 and h_1 are bandwidths for η_0 and η_1. The estimators for η_0 and η_1 are $\hat{\eta}_0(u) = \hat{a}_0$ and $\hat{\eta}_1(u) = \hat{a}_1$. One can modify the global quasi-likelihood analogously.

Model (5) also allows modeling interactions of the form

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \eta_0\{\boldsymbol{\alpha}_0^T\mathbf{x} + (\mathbf{x}^T, \mathbf{z}^T)\boldsymbol{\Lambda}(\mathbf{x}^T, \mathbf{z}^T)^T\} + \boldsymbol{\beta}_0^T\mathbf{z},$$

where $\boldsymbol{\Lambda}$ is the parameter matrix for interactions. This model is included in (5) by forming a new and longer X vector. One can also incorporate partial interaction terms in (5), which would reduce the number of effective parameters.

APPENDIX: PROOFS

Here we outline the key ideas for proving Theorems 1, 2, 4, and 5. Details can be found in an earlier draft of this article (Carroll et al. 1995). The methods for proving Theorem 3 are similar.

A.1 Conditions

For simplicity of notation, here we absorb σ^2 into $V(\cdot)$, so that the variance of Y given $(\mathbf{Z}, \mathbf{X})$ is $V\{\mu(\mathbf{Z}, \mathbf{X})\}$. Denote $q_l(x, y) = (\partial^l/\partial x^l)Q\{g^{-1}(x), y\}$, $l = 1, 2, 3$. Then

$$q_1(x, y) = \{y - g^{-1}(x)\}\rho_1(x)$$

$$\text{and} \quad q_2(x, y) = \{y - g^{-1}(x)\}\rho_1'(x) - \rho_2(x), \qquad (\text{A.1})$$

where $\rho_l(t) = \{dg^{-1}(t)/dt\}^l/V\{g^{-1}(t)\}$ is introduced in Section 4.2. In Condition 1, u is a generic argument for Theorem 1, and the condition must hold uniformly in u for Theorems 2–4.

Condition 1.

a. The function $q_2(x, y) < 0$ for $x \in \mathbb{R}$ and y in the range of the response variable.
b. The marginal density of $\boldsymbol{\alpha}_0^T\mathbf{X}$ is positive and continuous at the point u.
c. The function $\eta_0''(\cdot)$ is continuous at the point u.
d. $g''(\cdot)$ and $V(\cdot)$ are continuous functions.
e. With $R = \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}$, $E\{q_1^2(R, Y)|U = t\}$, $E\{q_1^2(R, Y)\mathbf{Z}|U = t\}$ and $E\{q_1^2(R, Y)\mathbf{Z}\mathbf{Z}^T|U = t\}$ are continuous in t at the point u. Moreover, $E[q_2^2\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\}] < \infty$ and $E[q_1^{2+\delta}\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\}] < \infty$, for some $\delta > 2$.
f. The kernel K is a symmetric density function with bounded support.
g. The random vector $\mathbf{Z}$ is assumed to have a bounded support.

Condition 2.

a. The marginal density of $\boldsymbol{\alpha}^T\mathbf{X}$ is positive and uniformly continuous for $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$. Further, $\boldsymbol{\alpha}_0^T\mathbf{X}$ has a positive density on its support D.
b. The function $\eta_0''(\cdot)$ is continuous in $u \in D$.
c. The density function of $\mathbf{X}$ has a continuous second derivative.
d. The function $V''(\cdot)$ and $g'''(\cdot)$ are continuous.
e. With $R = \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}$, $E\{q_1^2(R, Y)|\mathbf{X} = u\}$, $E\{q_1^2(\eta_0(R, Y)\mathbf{Z}|\mathbf{X} = u\}$ and $E\{q_1^2(R, Y)\mathbf{Z}\mathbf{Z}^T|\mathbf{X} = u\}$ are twice differentiable in $u \in D$,

Proof of Theorem 1

Let $c_n = (nh)^{-1/2}$, $U_i = \boldsymbol{\alpha}_0^T\mathbf{X}_i$,

$$\mathbf{X}_i^* = \begin{pmatrix} 1 \\ (U_i - u)/h \\ \mathbf{Z}_i \end{pmatrix},$$

and

$$\hat{\boldsymbol{\beta}}^* = \begin{pmatrix} c_n^{-1}\{\hat{a} - \eta_0(U)\} \\ c_n^{-1} h\{\hat{b} - \eta_0'(u)\} \\ c_n^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \end{pmatrix},$$

and let $f(\cdot)$ denote the density function of $U_i = \boldsymbol{\alpha}_0^T \mathbf{X}_i$. Denote further $\bar{\eta}_i = \bar{\eta}_i(u) = \eta_0(u) + \boldsymbol{\beta}_0^T \mathbf{Z}_i + \eta_0'(u)(U_i - u)$. If $(\hat{a}, \hat{b}, \hat{\boldsymbol{\beta}})^T$ maximizes (11), then $\hat{\boldsymbol{\beta}}^*$ maximizes

$$l_n(\boldsymbol{\beta}^*) = h \sum_{i=1}^{n} [Q\{g^{-1}(c_n \boldsymbol{\beta}^{*T} \mathbf{X}_i^* + \bar{\eta}_i), Y_i\} - Q\{g^{-1}(\bar{\eta}_i), Y_i\}] K_h(U_i - u)$$

with respect to $\boldsymbol{\beta}^*$. The concavity of the function $l_n(\boldsymbol{\beta}^*)$ is ensured by Condition 1a. By a Taylor expansion of the function $Q(g^{-1}(\cdot), Y_i)$ we obtain that

$$l_n(\boldsymbol{\beta}^*) = \mathbf{W}_n^T \boldsymbol{\beta}^* + \frac{1}{2}\, \boldsymbol{\beta}^{*T} \mathbf{A}_n \boldsymbol{\beta}^* \{1 + o_P(1)\}, \qquad (A.2)$$

$$\mathbf{W}_n = h c_n \sum_{i=1}^{n} q_1(\bar{\eta}_i, Y_i) \mathbf{X}_i^* K_h(U_i - u),$$

and

$$\mathbf{A}_n = h c_n^2 \sum_{i=1}^{n} q_2(\bar{\eta}_i, Y_i) \mathbf{X}_i^* \mathbf{X}_i^{*T} K_h(U_i - u).$$

Define

$$\mathbf{A}(\mathbf{Z}) = \begin{pmatrix} 1 & 0 & \mathbf{Z}^T \\ 0 & \kappa_2 & 0 \\ \mathbf{Z} & 0 & \mathbf{Z}\mathbf{Z}^T \end{pmatrix}$$

and

$$\mathbf{B}(\mathbf{Z}) = \begin{pmatrix} \nu_0 & 0 & \nu_0 \mathbf{Z}^T \\ 0 & \nu_2 & 0 \\ \nu_0 \mathbf{Z} & 0 & \nu_0 \mathbf{Z}\mathbf{Z}^T \end{pmatrix}.$$

It can be shown that $\mathbf{A}_n = -f(u) E[\rho_2(\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}) \mathbf{A}(\mathbf{Z}) | U = u] + o_P(1) \equiv -A + o_P(1)$. Therefore, by (A.1),

$$l_n(\boldsymbol{\beta}^*) = \mathbf{W}_n^T \boldsymbol{\beta}^* - \frac{1}{2}\, \boldsymbol{\beta}^{*T} \mathbf{A} \boldsymbol{\beta}^* + o_P(1). \qquad (A.3)$$

By applying the convexity lemma (see Pollard 1991), we obtain that $\hat{\boldsymbol{\beta}}^* = \mathbf{A}^{-1} \mathbf{W}_n + o_P(1)$. Hence the asymptotic normality of $\hat{\boldsymbol{\beta}}^*$ will follow from that of $\mathbf{W}_n$, which we establish next. By the definition of $\mathbf{W}_n$, it can be shown that

$$\begin{aligned} E\mathbf{W}_n = {} & c_n^{-1} \frac{1}{2}\, \eta_0''(u) h^2 f(u) E \\ & \times [\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\}(\kappa_2, 0, \kappa_2 \mathbf{Z}^T)^T | U = u] \\ & + o(c_n^{-1} h^2) \end{aligned} \qquad (A.4)$$

and that $\operatorname{var}(\mathbf{W}_n) = f(u) E[\rho_2\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\} \mathbf{B}(\mathbf{Z}) | U = u] + o(1) \equiv \mathbf{B} + o(1)$. Using Condition 1e, it can be shown that Liapounov's condition is satisfied and hence $\hat{\boldsymbol{\beta}}^*$ is asymptotically normal. This establishes Theorem 1.

Proof of Theorem 2

Lemma A.1. Let C and D be compact sets in $\mathbb{R}^d$ and $\mathbb{R}^p$ and let $f(\mathbf{x}, \boldsymbol{\theta})$ be a continuous function in $\boldsymbol{\theta} \in C$ and $\mathbf{x} \in D$. Assume that $\hat{\boldsymbol{\theta}}(\mathbf{x}) \in C$ is continuous in $\mathbf{x} \in D$ and is the unique maximizer of $f(\mathbf{x}, \boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_n(\mathbf{x}) \in C$ be a maximizer of $f_n(\mathbf{x}, \boldsymbol{\theta})$. If

$$\sup_{\boldsymbol{\theta} \in C, \mathbf{x} \in D} |f_n(\mathbf{x}, \boldsymbol{\theta}) - f(\mathbf{x}, \boldsymbol{\theta})| \to 0,$$

then

$$\sup_{\mathbf{x} \in D} |\hat{\boldsymbol{\theta}}_n(\mathbf{x}) - \hat{\boldsymbol{\theta}}(\mathbf{x})| \to 0, \quad \text{as} \quad n \to \infty.$$

Proof of Theorem 2

First, we note that under Condition 2, by a result of Mack and Silverman (1982), (A.3) holds uniformly in $u \in D$. By the convexity lemma, it also holds uniformly in $\boldsymbol{\beta}^* \in C$ and $u \in D$ for any compact set C. Lemma A.1 then yields

$$\sup_{u \in D} |\hat{\boldsymbol{\beta}}^*(u) - \mathbf{A}^{-1} \mathbf{W}_n(u)| \xrightarrow{P} 0, \qquad (A.5)$$

where $\hat{\boldsymbol{\beta}}^*(u)$ and $\mathbf{W}_n(u)$ are defined in the proof of Theorem 1, except that here we stress the dependence on u. So, by considering the first element of the vectors in (A.5), we have

$$\sup_{u \in D} \left| \hat{\eta}(u) - \eta_0(u) - \frac{1}{n f(u)} \sum_{i=1}^{n} W_i K_h(X_i - u) \right| = o_P(c_n),$$

where $f(u)$ is the density of X_i and W_i is the first element of the vector $q_1(\bar{\eta}_i, Y_i) \boldsymbol{\Sigma}^{-1}(u)(1, \mathbf{Z}_i^T)^T$, with $\bar{\eta}_i = \bar{\eta}_i(u) = \eta_0(u) + \boldsymbol{\beta}_0^T \mathbf{Z}_i + \eta_0'(u)(U_i - u)$. Moreover, the following stronger result holds:

$$\sup_{u \in D} \left| \hat{\eta}(u) - \eta_0(u) - \frac{1}{n f(u)} \sum_{i=1}^{n} W_i K_h(X_i - u) \right| = O_P\{h^2 c_n + c_n^2 \log^{1/2}(1/h)\}. \qquad (A.6)$$

Let $\hat{\theta} = n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, $\hat{m}_i = \hat{\eta}(X_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i$, and $m_i = \eta_0(X_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i$. Then $\hat{\theta}$ maximizes

$$l_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} [Q\{g^{-1}(\hat{m}_i + n^{-1/2} \boldsymbol{\theta}^T \mathbf{Z}_i), Y_i\} - Q\{g^{-1}(\hat{m}_i), Y_i\}]. \qquad (A.7)$$

By Taylor's expansion, we have

$$l_n(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} q_1(\hat{m}_i, Y_i) \boldsymbol{\theta}^T \mathbf{Z}_i + \frac{1}{2}\, \boldsymbol{\theta}^T \mathbf{B}_n \boldsymbol{\theta} \qquad (A.8)$$

and

$$\mathbf{B}_n = \frac{1}{n} \sum_{i=1}^{n} [Y_i \rho_1'\{g^{-1}(\hat{m}_i + \xi_{ni})\} - \rho_3\{g^{-1}(\hat{m}_i + \xi_{ni}')\}] \mathbf{Z}_i \mathbf{Z}_i^T,$$

with ξ_{ni} and ξ_{ni}' between 0 and $n^{-1/2} \boldsymbol{\theta}^T \mathbf{Z}_i$, independent of Y_i, and with $\rho_3(x) = -g^{-1}(x) \rho_1'(x) - \rho_2(x)$. It can be shown that

$$\begin{aligned} \mathbf{B}_n &= -E\rho_2\{\eta_0(X) + \boldsymbol{\beta}_0^T \mathbf{X}\} \mathbf{Z}\mathbf{Z}^T + o_P(1) \\ &\equiv -\mathbf{B} + o_P(1). \end{aligned} \qquad (A.9)$$

Using similar arguments as for obtaining (A.9), we get

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^{n} q_1(\hat{m}_i, Y_i) \mathbf{Z}_i \\ & \quad = n^{-1/2} \sum_{i=1}^{n} q_1(m_i, Y_i) \mathbf{Z}_i \\ & \qquad + n^{-1/2} \sum_{i=1}^{n} q_2(m_i, Y_i) \{\hat{\eta}(X_i) - \eta_0(X_i)\} \mathbf{Z}_i \\ & \qquad + O_P(n^{1/2} \|\hat{\eta} - \eta_0\|_\infty^2). \end{aligned}$$

By (A.6), the second term in the foregoing expression can be expressed as

$$
\begin{aligned}
& n^{-3/2}\sum_{i=1}^{n} q_2(m_i, Y_i) f(X_i)^{-1} \sum_{j=1}^{n} W_j K_h(X_j - X_i)\mathbf{Z}_i \\
&\qquad + O_P\{n^{1/2} c_n^2 \log^{1/2}(1/h)\} \\
&\quad \equiv T_{n1} + O_P\{n^{1/2} c_n^2 \log^{1/2}(1/h)\}.
\end{aligned}
$$

Now define $v_j = v(X_j, Y_j, \mathbf{Z}_j)$ as the first element of $q_1(m_j, Y_j)\mathbf{\Sigma}^{-1}(1, \mathbf{Z}_j^T)^T$. Using the definition of $\bar{\eta}_j(X_i)$, we obtain $\bar{\eta}_j(X_i) - m_j = O((X_j - X_i)^2)$, and thus

$$
\begin{aligned}
T_{n1} &= n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n} q_2(m_i, Y_i) f(X_i)^{-1} v_j K_h(X_j - X_i)\mathbf{Z}_i \\
&\qquad + O_P(n^{1/2}h^2) \\
&\equiv T_{n2} + O_P(n^{1/2}h^2).
\end{aligned}
$$

It can be shown via calculating the second moment that

$$
T_{n2} - T_{n3} \xrightarrow{P} 0, \tag{A.10}
$$

where $T_{n3} = -n^{-1/2}\sum_{j=1}^{n}\gamma(X_j)v_j$ with $\gamma(u) = E[\rho_2\{\eta_0(u) + \beta_0^T\mathbf{Z}\}\mathbf{Z}|X = u]$. Combining (A.7)–(A.10), we obtain that $l_n(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\Omega(X_i, Y_i, \mathbf{Z}_i) - \frac{1}{2}\boldsymbol{\theta}^T\mathbf{B}\boldsymbol{\theta} + o_P(1)$, where $\Omega(X_i, Y_i, \mathbf{Z}_i) = q_1(m_i, Y_i)\mathbf{Z}_i - \gamma(X_i)v_i$. By the convexity lemma, we find that $\hat{\boldsymbol{\theta}} = \mathbf{B}^{-1}n^{-1/2}\sum_{i=1}^{n}\Omega(X_i, Y_i, \mathbf{Z}_i) + o_P(1)$, from which it follows that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \mathbf{B}^{-1}\mathbf{\Sigma}_1\mathbf{B}^{-1})$, as claimed.

Proof of Theorem 4

We use the notation $U = \boldsymbol{\alpha}_0^T\mathbf{X}$, $\hat{U} = \hat{\boldsymbol{\alpha}}^T\mathbf{X}$ and $f(\cdot)$ for the density function of U. The proof relies on two steps, which we state first and prove afterward. The first step consists of an expansion for $\hat{\eta}$ (at an argument u_0). We show that

$$
\begin{aligned}
&\hat{\eta}(u_0; h, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \eta_0(u_0) \\
&\quad = n^{-1}\sum_{i=1}^{n} K_h(U_i - u_0) \\
&\qquad \times \frac{\varepsilon_i}{f(u_0)E\{\rho_2(\cdot)|U = u_0\}} \\
&\qquad - (\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}_0^T)\,\frac{E\{\mathbf{Z}\rho_2(\cdot)|U = u_0\}}{E\{\rho_2(\cdot)|U = u_0\}} \\
&\qquad - (\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)\,\frac{E\{\mathbf{X}\rho_2(\cdot)\eta_0'(\cdot)|U = u_0\}}{E\{\rho_2(\cdot)|U = u_0\}} + o_P(n^{-1/2}),
\end{aligned} \tag{A.11}
$$

where "·" denotes the argument $\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}$ and $\varepsilon_i = \{Y_i - \mu(\cdot)\}\rho_1(\cdot)$ with a similar convention.

The second step is as follows. Introduce the shorthand notations

$$
\mathbf{\Lambda}_i = \begin{bmatrix} \mathbf{X}_i\eta_0'(U_i) \\ \mathbf{Z}_i \end{bmatrix}
$$

and

$$
\mathbf{P}\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{I} - \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} + o_P(1).
$$

We show that

$$
\begin{aligned}
&\mathbf{P}\boldsymbol{\alpha}\mathbf{Q}n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \\
&\quad = n^{-1/2}\textstyle\sum_{i=1}^{n}\varepsilon_i\mathbf{P}\boldsymbol{\alpha}\left[\mathbf{\Lambda}_i - \frac{E\{\mathbf{\Lambda}\rho_2(\cdot)|U_i\}}{E\{\rho_2(\cdot)|U_i\}}\right] + o_P(1).
\end{aligned} \tag{A.12}
$$

Because ε_i has variance ρ_{2i}, the right side of (36) has the covariance matrix $\mathbf{P}\boldsymbol{\alpha}\mathbf{Q}\mathbf{P}\boldsymbol{\alpha}$, verifying the statement of Theorem 4.

Proof of (A.11)

Let $a = \eta_0(u_0)$ and $b = h\eta'(u_0)$. The local linear estimates solve

$$
0 = n^{-1}\sum_{i=1}^{n} K_h(\hat{U}_i - u_0)\begin{bmatrix} 1 \\ (\hat{U}_i - u_0)/h \end{bmatrix}\{Y_i - \hat{\mu}(\cdot)\}\hat{\rho}_1(\cdot),
$$

where $\hat{\mu}(\cdot) = \mu\{\hat{a} + \hat{b}(\hat{U}_i - u_0)/h + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}$, and similarly for $\hat{\rho}_1(\cdot)$. Via Taylor series and using the conditions on h, we obtain

$$
\begin{aligned}
0 &= n^{-1}\sum_{i=1}^{n} K_h(U_i - u_0)\begin{bmatrix} 1 \\ (U_i - u_0)/h \end{bmatrix}\{Y_i - \mu_*(\cdot)\}\rho_{1*}(\cdot) \\
&\qquad - B_{n1}\begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix} - (\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}_0^T)B_{n2} - (\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)B_{n3} \\
&\qquad + o_P(n^{-1/2}) + O_P(h^2),
\end{aligned}
$$

where $\mu_*(\cdot) = \mu\{a + b(U_i - u_0)/h + \boldsymbol{\beta}_0^T\mathbf{Z}_i\}$ and $\rho_{1*}(\cdot)$ is defined similarly. Here $B_{n,j}$ $(j = 1, 2, 3)$ are the resulting sample matrices of kernel form. Solving the foregoing linearized equation and substituting $B_{n,j}$ with their asymptotic counterparts, we obtain (A.11).

Proof of (A.12). Recall that (9) and (10) lead to asymptotically equivalent estimates. Consider (9) and use the expansion

$$
\begin{aligned}
&\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) \\
&\quad = \hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\
&\qquad + \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) \\
&\quad = \hat{\eta}'(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)\mathbf{X}_i + \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\
&\qquad - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) + o_P(n^{-1/2}) \\
&\quad = \eta_0'(\boldsymbol{\alpha}_0^T\mathbf{X}_i)(\hat{\boldsymbol{\alpha}}^T - \boldsymbol{\alpha}_0^T)\mathbf{X}_i + \hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\
&\qquad - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i) + o_P(n^{-1/2}),
\end{aligned} \tag{A.13}
$$

where we dropped the dependence on h for notational simplicity. The second term is handled by (A.13). With θ as the Lagrange multiplier, we know that $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is the solution to

$$
\begin{aligned}
0 &= \theta\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ 0 \end{pmatrix} \\
&\qquad + n^{-1/2}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{X}_i\hat{\eta}'(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \\ \mathbf{Z}_i \end{bmatrix} \\
&\qquad \times [Y_i - \mu\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}] \\
&\qquad \times \rho_1\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}_i\}.
\end{aligned}
$$

We can expand $\hat{\eta}(\cdot)$ about $\eta_0(\cdot)$ using (A.11). Write $\mu_i = \mu\{\eta_0(U_i) + \boldsymbol{\beta}_0^T\mathbf{Z}_i)\}$, and similarly for ρ_{ji}. Make the further definition

$$A_{\boldsymbol{\alpha},\boldsymbol{\beta}} = E\left[\rho_2(\cdot)\left\{\begin{matrix}\mathbf{X}\eta_0'(\cdot)\\ \mathbf{Z}\end{matrix}\right\}\left\{\begin{matrix}\mathbf{X}\eta_0'(\cdot)\\ \mathbf{Z}\end{matrix}\right\}^T\right].$$

By the Taylor series, and using (A.13), we have that (using that $nh^4 \to 0$)

$$\begin{aligned} 0 &= \theta\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ 0\end{pmatrix} + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\varepsilon_i - n^{-1/2} \\ &\quad\times \sum_{i=1}^n \boldsymbol{\Lambda}_i(\hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}_0^T)\mathbf{Z}_i\rho_{2i}(\cdot) \\ &\quad - n^{-1/2}\sum_{i=1}^n \rho_{2i}\boldsymbol{\Lambda}_i\{\hat{\eta}(\hat{\boldsymbol{\alpha}}^T\mathbf{X}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i)\} + o_P(1) \\ &= \theta\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ 0\end{pmatrix} + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\varepsilon_i - A_{\boldsymbol{\alpha},\boldsymbol{\beta}}n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\end{pmatrix} \\ &\quad - n^{-1/2}\sum_{i=1}^n \rho_{2i}\boldsymbol{\Lambda}_i\{\hat{\eta}(\boldsymbol{\alpha}_0^T\mathbf{X}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}_i)\} + o_P(1). \end{aligned}$$

We now invoke (A.11), which implies that

$$\begin{aligned} 0 &= \theta\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ 0\end{pmatrix} + n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\varepsilon_i - Qn^{1/2}\begin{pmatrix}\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\end{pmatrix} \\ &\quad - n^{-1/2}\sum_{i=1}^n \boldsymbol{\Lambda}_i\rho_{2i}n^{-1}\sum_{j=1}^n K_h(U_j - U_i) \\ &\quad \times \frac{Y_j - \mu\{\eta_0(U_i) + \boldsymbol{\beta}_0^T\mathbf{Z}_j\}}{f(U_i)E\{\rho_2(\cdot)|U_i\}}\,\rho_1\{\eta_0(U_i) + \boldsymbol{\beta}_0^T\mathbf{Z}_j\}. \qquad \text{(A.14)} \end{aligned}$$

Only the last term is of interest, and hence we focus on it. Interchanging the summations, we get

$$\begin{aligned} n^{-1/2}\sum_{i=1}^n &\left[n^{-1}\sum_{j=1}^n \boldsymbol{\Lambda}_j\rho_{2j}K_h(U_j - U_i)\right. \\ &\quad\left.\times\ \frac{Y_i - \mu\{\eta_0(U_j) + \boldsymbol{\beta}_0^T\mathbf{Z}_i\}}{f(U_j)E\{\rho_2(\cdot)|U_j\}}\,\rho_1\{\eta_0(U_j) + \boldsymbol{\beta}_0^T\mathbf{Z}_i\}\right]. \end{aligned}$$

The term in the square brackets, being a nonparametric regression, is essentially the same as

$$n^{-1/2}\sum_{i=1}^n \varepsilon_i\,\frac{E\{\boldsymbol{\Lambda}\rho_2(\cdot)|U_i\}}{E\{\rho_2(\cdot)|U_i\}}, \qquad \text{(A.15)}$$

for a symmetric kernel. Combining (A.14) and (A.15), and multiplying by $\mathbf{P}\boldsymbol{\alpha}$, we obtain (A.11).

A.5 Proof of Theorem 5

Let $h(\mathbf{x},\mathbf{z})$ be the joint density of $(\mathbf{X},\mathbf{Z})$. Then, under the semiparametric model (3) and (5), the joint density of $(\mathbf{X},Y,\mathbf{Z})$ is given by

$$f(\mathbf{x},y,\mathbf{z}) = \exp[y\theta(\mathbf{x},\mathbf{z}) - \mathcal{B}\{\theta(\mathbf{x},\mathbf{z})\} + \mathcal{C}(y)]h(\mathbf{x},\mathbf{z}), \qquad \text{(A.16)}$$

where $\theta(\mathbf{x},\mathbf{z}) = g_0 \circ g^{-1}\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}\}$ with $\|\boldsymbol{\alpha}_0\| = 1$ and g_0 as the canonical link function. Define

$$\begin{aligned} P_1 &= \{\text{Model (A.16) with given } \eta_0(\cdot), \text{and } h\}, \\ P_2 &= \{\text{Model (A.16) with given } \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \text{and } h(\cdot)\}, \end{aligned}$$

and

$$P_3 = \{\text{Model (A.16) with given } \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0 \text{ and } \eta_0(\cdot)\}.$$

Then the score function for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ under the parametric model P_1 is given by

$$\dot{l} = \{Y - \mu(\mathbf{X},\mathbf{Z})\}g_1'\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \beta_0^T\mathbf{Z}\}\begin{pmatrix}\eta_0'(\boldsymbol{\alpha}_0^T\mathbf{X})\mathbf{X}\\ \mathbf{Z}\end{pmatrix}.$$

where $g_1 = g_0 \circ g^{-1}$. The tangent space (Bickel et al. 1993, p. 50) of the nonparametric model P_2 can be shown to be $\dot{P}_2 = [\{Y - \mu(\mathbf{X},\mathbf{Z})\}g_1'(\cdot)a(\boldsymbol{\alpha}_0^T\mathbf{X}),$ for all $a \in L_2]$, and the tangent space of the nonparametric model P_3 is given by $\dot{P}_3 = [b(\mathbf{X},\mathbf{Z}) \in L_2 : Eb(\mathbf{X},\mathbf{Z}) = 0]$. Then, by theorem 3.4.1 of Bickel et al. (1993), the efficient score function of $(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)$ under model (A.16) is the projection of $\dot{l}$ into the orthogonal complement of the linear space $\dot{P}_2 + \dot{P}_3$—namely, $\dot{l}^* = \dot{l} - \prod(\dot{l}|\dot{P}_2 + \dot{P}_3)$. The information matrix for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ is just $E(\dot{l}^*)(\dot{l}^*)^T$, where $\prod(\dot{l}|\dot{P}_2 + \dot{P}_3)$ is the projection of $\dot{l}$ into $\dot{P}_2 + \dot{P}_3$. Because $\dot{P}_2 \perp \dot{P}_3$ and $\dot{l} \perp \dot{P}_3$, the projection $\prod(\dot{l}|\dot{P}_2 + \dot{P}_3) = \prod(\dot{l}|\dot{P}_2)$ is to find a vector function of form $(Y - \mu)g_1'(\cdot)a(\boldsymbol{\alpha}_0^T\mathbf{X})$ such that $E\|\dot{l} - (Y - \mu)g_1'(\cdot)a(\boldsymbol{\alpha}_0^T\mathbf{X})\|^2$ is minimized. By conditioning on $\boldsymbol{\alpha}_0^T\mathbf{X}$, one can easily find that

$$\prod(\dot{l}|\dot{P}_2) = (Y - \mu)g_1'(\cdot)\left[\begin{matrix}E\{\mathbf{X}\eta_0'(U)\rho_2(\cdot)|U\}/E\{\rho_2(\cdot)|U\}\\ E\{\mathbf{Z}\rho_2(\cdot)|U\}/E\{\rho_2(\cdot)|U\}\end{matrix}\right],.$$

where $U = \boldsymbol{\alpha}_0^T\mathbf{X}$. Using this, it is now easy to verify that $Q = E(\dot{l}^*)(\dot{l}^*)^T$.

[Received April 1995. Revised August 1996.]

REFERENCES

Bickel, P. J., Klaassen, A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Bonneu, M., Delecroix, M., and Hristache, M. (1995), "Semiparametric Estimation of Generalized Linear Models and Related Models," unpublished manuscript.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1995), "Generalized Partially Linear Single-Index Models," Discussion Paper 9506, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

Chen, H. (1988), "Convergence Rates for Parametric Components in a Partly Linear Model," *The Annals of Statistics*, 16, 136–146.

Cuzick, J. (1992), "Semiparametric Additive Regression," *Journal of the Royal Statistical Society*, Ser. B, 54, 831–843.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150.

Fan, J., and Marron, J. S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21, 157–178.

Härdle, W., and Scott, D. W. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics*, 7, 97–128.

Hastie, T. J., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Heckman, N. (1986), "Spline Smoothing in a Partly Linear Model," *Journal of the Royal Statistical Society*, Ser. A, 48, 244–248.

Hunsberger, S. (1994), "Semiparametric Regression in Likelihood-Based Models," *Journal of the American Statistical Association*, 89, 1354–1365.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B., and Kjelsberg, M. O. (1986), "Overall and Coronary Heart

Disease Mortality Rates in Relation to Major Risk Factors in 325,348 Men Screened for MRFIT," *American Heart Journal*, 112, 825–836.

Küchenhoff, H., and Carroll, R. J. (1997), "Segmented Regression With Errors in Predictors," *Statistics in Medicine*, to appear.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.

Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 61, 405–415.

Mammen, E., and van de Geer, S. (1995), "Penalized Estimation in Partial Linear Models," Technical Report 95-05, University of Leiden, The Netherlands.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

Royston, P., and Altman, D. G. (1994), "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Applied Statistics*, 43, 429–467.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

Turlach, B. A., and Wand, M. P. (1995), "Fast Computation of Auxiliary Quantities in Local Polynomial Regression," unpublished manuscript.

Ulm, K. (1991), "A Statistical Method for Assessing a Threshold in Epidemiological Studies," *Statistics in Medicine*, 10, 341–349.

Wahba, G. (1984), "Partial Spline Models for Semiparametric Estimation of Functions of Several Variables," in *Statistical Analysis of Time Series*, Proceedings of the Japan–U.S. Joint Seminar, Tokyo, pp. 319–329.

Weisberg, S., and Welsh, A. H. (1994), "Estimating the Missing Link Function," *The Annals of Statistics*, 22, 1674–1700.

Local Estimating Equations

Raymond J. CARROLL, David RUPPERT, and Alan H. WELSH

Estimating equations have found wide popularity recently in parametric problems, yielding consistent estimators with asymptotically valid inferences obtained via the sandwich formula. Motivated by a problem in nutritional epidemiology, we use estimating equations to derive nonparametric estimators of a "parameter" depending on a predictor. The nonparametric component is estimated via local polynomials with loess or kernel weighting; asymptotic theory is derived for the latter. In keeping with the estimating equation paradigm, variances of the nonparametric function estimate are estimated using the sandwich method, in an automatic fashion, without the need (typical in the literature) to derive asymptotic formulas and plug-in an estimate of a density function. The same philosophy is used in estimating the bias of the nonparametric function; that is, an empirical method is used without deriving asymptotic theory on a case-by-case basis. The methods are applied to a series of examples. The application to nutrition is called "nonparametric calibration" after the term used for studies in that field. Other applications include local polynomial regression for generalized linear models, robust local regression, and local transformations in a latent variable model. Extensions to partially parametric models are discussed.

KEY WORDS: Asymptotic theory; Bandwidth selection; Local polynomial regression; Logistic regression; Measurement error; Missing data; Nonlinear regression; Partial linear models; Sandwich estimation.

1. INTRODUCTION

A general methodology that has found wide popularity recently, especially in biostatistics, is to estimate parameters via estimating equations. Maximum likelihood estimates, robust regression estimates (Huber 1981), variance function estimates (Carroll and Ruppert 1988), generalized estimating equation (GEE) estimates (Diggle, Liang, and Zeger 1994), marginal methods for nonlinear mixed-effects models (Breslow and Clayton 1993), and indeed most of the estimators used in non-Bayesian parametric statistics are all based on the same technology. If the data are independent observations $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$, with the $\mathbf{Y}$s possibly vector valued, then a parameter Θ is estimated by solving the estimating equation

$$0 = \sum_{i=1}^{n} \psi(\mathbf{Y}_i, \hat{\Theta}). \tag{1}$$

We allow Θ to be vector valued, and ψ must have the same dimension as Θ. For example, maximum likelihood estimates are versions of (1) when $\psi(\cdot)$ is the derivative of the log-likelihood function.

One of the reasons that estimating equation methodology has become so popular is that for most estimating equations, the covariance matrix of the parameter estimate can be consistently and nonparametrically estimated using the so-called "sandwich formula" (Huber 1967) described in detail in Section 3.2.

The combination of estimating equations and sandwich covariance matrix estimates thus form a powerful general methodology. In this article we pose the following simple question: How does one proceed if Θ depends in an unknown way on an observable variable Z, so that $\Theta = \Theta(Z)$? The question arises naturally in the context of calibration studies in nutritional epidemiology; Section 2 provides a detailed discussion.

Our aim is to provide methods with the same generality as parametric estimating equations and the sandwich method. Starting only from the parametric estimating equation (1), we propose to develop estimates of $\Theta(Z)$ and use the sandwich method to form consistent and nonparametric estimates of the covariance matrix.

The method that we proposed, called *local estimating equations,* essentially involves estimating $\Theta(Z)$ by local polynomials with local weighting of the estimating equation. The specific application in nutrition is called *nonparametric calibration* because of its roots in nutritional epidemiology calibration studies. This article is concerned primarily with the case where Z is scalar, although in Section 4.2 we describe extensions to the multivariate case and present a numerical example.

In practice, it is often the case that $\Theta(z)$ is a q-dimensional vector, whereas we are often interested in a scalar function of it, say $\alpha(z) = \mathcal{T}\{\Theta(z)\}$. For example, in the nutrition example motivating this research, $\Theta(z)$ is a $q = 6$ dimensional vector of conditional moments of $\mathbf{Y}$ given $Z = z$, and $\alpha(z)$ is the correlation between a component of $\mathbf{Y}$ and another, unobservable random variable.

Our basic method for estimating $\Theta(\cdot)$ involves local polynomials. With superscript (j) denoting a jth derivative with respect to z and with $\mathbf{b}_j = \Theta^{(j)}(z_0)/j!$, the

Raymond J. Carroll is Distinguished University Professor of Statistics and Professor of Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843. David Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. Alan H. Welsh is Reader, CMA, Australian National University, Canberra ACT 2601. Carroll's research was supported by grant CA-57030 from the National Cancer Institute. Carroll's research was partially completed while visiting the Institute for Statistics and Ökonometrie, Humboldt University zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award. Ruppert's research was supported by grants from the National Science Foundation and the National Security Agency. This research is an outgrowth of discussions with Laurence Freedman and Nancy Potischman of the National Cancer Institute concerning nutritional epidemiology; their suggestions are gratefully acknowledged. The authors also wish to acknowledge with thanks Donna Spiegelman and Walter Willett of the Harvard School of Public Health for making the Nurses' Health Trial calibration study data available, and two referees and an associate editor for making many useful suggestions.

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

local polynomial of order p in a neighborhood of z_0 is $\boldsymbol{\Theta}(z) \approx \sum_{j=0}^{p} \mathbf{b}_j(z - z_0)^j$. The local weight for a value of z near z_0 is denoted by $w(z, z_0)$. We then propose to solve in $(\mathbf{b}_0, \ldots, \mathbf{b}_p)$ the $q \times (p+1)$ equations

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_j(Z_i - z_0)^j\right\}\mathbf{G}_p^t(Z_i - z_0), \tag{2}$$

where $\mathbf{G}_p^t(v) = (1, v, v^2, \ldots, v^p)$. The final estimates are $\hat{\boldsymbol{\Theta}}(z_0) = \hat{\mathbf{b}}_0$ and $\hat{\alpha}(z_0) = \mathcal{T}\{\hat{\mathbf{b}}_0\}$.

Equations such as (2) are already in common use when $\boldsymbol{\Theta}(z)$ is scalar, although not at the level of generality given here (not being derived from estimating functions). Here are a few examples:

a. Ordinary multivariate-response Nadaraya–Watson kernel regression has $p = 0, \psi(\mathbf{Y}, \mathbf{v}) = \mathbf{Y} - \mathbf{v}$, and $w(z, z_0)$ chosen to be a kernel weight.

b. Local linear regression has $p = 1$ and $\psi(\mathbf{Y}, \mathbf{v}) = \mathbf{Y} - \mathbf{v}$, and if $w(z, z_0)$ is a nearest-neighbor weight, then the result is the loess procedure in S-PLUS (Chambers and Hastie 1992).

c. When the mean and variance of a univariate response Y are related through $E(Y|Z) = \mu\{\Theta(Z)\}$ and $\text{var}(Y|Z) = \sigma^2 V\{\Theta(Z)\}$ for known functions μ and V, local quasi-likelihood regression is based on

$$\psi(Y, x) = \{Y - \mu(x)\}\mu^{(1)}(x)/V(x). \tag{3}$$

With kernel weights, this is the method of Weisberg and Welsh (1994) when $p = 0$ and of Fan, Heckman, and Wand (1995) when $p \geq 1$.

This article is organized as follows. Section 2 describes in detail a problem from nutrition that motivated this work. This problem is easily analyzed in our general local estimating equation framework. Section 3 indicates that local polynomial methods usually have tuning constants that must be set or estimated. If they are to be estimated, then the typical approach is to minimize mean squared error (MSE) which in turn requires estimation of bias and variance functions. It is possible to derive asymptotic theoretical expressions for these functions (indeed, we do so for kernel regression in the Appendix) and then do a "plug-in" operation to obtain an estimate. But following this approach in practice requires density estimation, estimation of higher-order derivatives, and so on, and these complications would limit the range of applications. Instead, we estimate the bias and variance functions empirically, without explicit use of the asymptotic formulas. Bias estimation uses a modification of Ruppert's (1997) empirical bias method, whereas variance estimation can be done by adapting the sandwich formula of Huber (1967) to this context. That the sandwich formula provides consistent variance estimates in this context is not obvious, but in the Appendix we prove this to be the case.

Section 4 deals with a series of examples, including the analysis of nutrient intake data. Section 5 discusses modifications of the algorithm (2). Section 6 presents some concluding remarks. All theoretical details are collected in an Appendix.

Local estimation of parameters for likelihood problems has been previously considered in important work by such authors as Fan and Gijbels (1996), Fan et al. (1995), Hastie and Tibshirani (1990), Kauermann and Tutz (1997), Severini and Staniswallis (1994), Staniswallis (1989), and Tibshirani and Hastie (1987), and these techniques are implemented in S-PLUS for generalized linear models (GLMs). Our methods and this article differ from the local likelihood literature in several ways:

- We do not require a likelihood, but only an unbiased estimating function. Given the popularity of estimating functions in recent statistical work, such work would appear to be of some consequence. Estimating functions allow us to use such techniques as method of moments, robust mean and variance function estimation, Horvitz and Thompson (1952) adjustments for missing data, GEE-type mean and variance function modeling, and so on. A number of our examples, both numerical and theoretical, illustrate the use of nonlikelihood estimating functions.
- Our estimates of variance are straightforward, being nothing more than estimates based on the sandwich method from parametric problems. In particular, one need not compute asymptotic variances in each problem and then estimate the terms in the resulting (often complex) expressions. To the best of our knowledge the use of the parametric sandwich method in general nonparametric regression contexts has not been previously advocated, nor has it been shown theoretically to give consistent estimates of variances. We prove such consistency and derive expressions for bias and variance for kernel weighting. There have been earlier uses of the sandwich formula in special cases of nonparametric regression, however. For example, Ruppert and Wand (1994) gave a sandwich formula for the variance of local polynomial regression estimators, and Gozalo and Linton (1995) used the sandwich formula for an interesting approach to nonparametric regression—local nonlinear regression.
- Our methods allow for estimation of tuning constants such as the span in loess or local bandwidths in kernel weighting. The methods apply at least in principle to all local estimating function–based estimates and hence can be applied in new problems without the need to use asymptotic theory to derive a bias expression, to use additional nonparametric regressions to estimate this expression, or to develop case-by-case tricks to get started.

2. MOTIVATING EXAMPLE

In this section we demonstrate an important problem where $\boldsymbol{\Theta}(z)$ is a vector and $\psi(\cdot)$ arises from an estimating function framework. The assessment and quantification of an individual's usual diet is a difficult exercise but is fundamental to discovering relationships between diet and cancer and to monitoring dietary behavior among individuals and populations. Various dietary assessment instruments have been devised, of which three main types are most commonly used in contemporary nutritional research. The instrument of choice in large nutritional epidemiology studies is the food frequency questionnaire (FFQ). For proper interpretation of epidemiologic studies that use FFQs as the basic dietary instrument, one needs to know the relationship between reported intakes from the FFQ and true usual intake. Such a relationship is ascertained through a substudy, commonly called a calibration study.

The primary aim of a calibration study may vary from case to case. Here we focus on the estimation of the correlation between FFQ intake and usual intake. The variable we use is the % of calories from fat. This correlation can be of crucial interest if the FFQ has been modified extensively from previous versions or is to be used in a new population from which little previous data have been obtained. Very low correlations might persuade the investigators to postpone the main study, pending improvements in the design of the FFQ or in the way it is presented to study participants.

FFQs are thought to often involve a systematic bias (i.e., underreporting or overreporting at the level of the individual). The other two commonly used instruments are the 24-hour food recall and the multiple-day food record (FR). Each of these FRs is more work-intensive and more costly but is thought to involve considerably less bias than a FFQ. At the end of Section 4.1 we comment on this and other issues in nutrition data.

For the ith individual ($i = 1, \ldots, n$), let Q_i denote the intake of a nutrient reported on a FFQ. For the jth ($j = 1, \ldots, m$) replicate on the ith person, let F_{ij} denote the intake reported by a FR, and let T_i denote long-term usual intake for the ith person. A simple model (Freedman, Carroll, and Wax 1991) relating these three is a standard linear errors-in-variables model,

$$Q_i = \beta_0 + \beta_1 T_i + \varepsilon_i; \qquad (4)$$

$$F_{ij} = T_i + U_{ij}; \qquad j = 1, \ldots, m. \qquad (5)$$

In model (4) deviations from $\beta_0 = 0$ and $\beta_1 = 1$ represent the systematic bias of FFQs, and the U_{ij} are the within individual variation in FRs. All random errors (i.e., εs and Us) are uncorrelated for purposes of this article; see the end of Section 4.1 for more details and further comments.

In measurement error models one wishes to relate a response (in our case, Q) to a predictor (in our case, T). Because of measurement error and other sources of variability, one cannot observe T. Instead, one can observe only a variable (in our case, F) related to T. The measurement error model literature was recently surveyed by Carroll, Ruppert, and Stefanski (1995).

Two studies that we analyze herein fit exactly into this design. The Nurses' Health Study (Rosner, Willett, and Spiegelman 1989), hereafter denoted by NHS, is a calibration study of 168 women, all of whom completed a single FFQ and four multiple-day food diaries ($m = 4$ in our notation). The Women's Interview Survey of Health (WISH) is a calibration study with 271 participants who completed a FFQ and six 24-hour recalls on randomly selected days at least 2 weeks apart ($m = 6$ in our notation). Although different FFQs are used in the two studies, the major difference between them is that the diaries have considerably smaller within-person variability than the 24-hour recalls. For instance, using % calories from fat, a simple component-of-variance analysis suggests that the measurement error in the *mean* of the four diaries in the NHS has variance 3.43 and the variance of usual intake is $\sigma_t^2 = 14.7$; the numbers for the six 24-hour recalls in WISH are 12.9 and 10.8. Thus one can expect that the NHS data will provide considerably more power for estimating effects than the WISH data.

For an initial analysis, we computed ρ_{QT} for each subpopulation formed by the quintiles of age; Section 4.1 provides the computational details. The five estimated correlations were roughly .4, .6, .4, .5, and .8. The five estimated correlations are statistically significantly different ($p < .01$) using a weighted test for equality of means. Note that the highest quintile of age has the highest value of ρ_{QT}. The standard errors of the estimates are approximately .13, except for the highest quintile, for which it is approximately .07.

Such stratified analysis (i.e., defining the strata by age quintiles) can be considered from the viewpoint of nonparametric regression. In each stratum we are estimating a parameter Θ (often multidimensional) and through it a crucial parametric function such as ρ_{QT}. Because these both depend on the stratum, they are more properly labeled as $\Theta(Z_*)$ and $\rho_{QT}(Z_*)$, where Z_* is the stratum level for Z. *Looked at as a function of* Z, this method suggests that $\rho_{QT}(Z)$ is a *discontinuous* function of Z. To avoid the arbitrariness of the categorization, we propose to estimate $\rho_{QT}(Z)$ as a *smooth* function of Z. Our analysis suggests that at least for the NHS, the correlation between the FFQ and usual intake increases with age in a nonlinear fashion.

3. TUNING CONSTANTS

To implement (2), we need a choice of the weight function $w(z, z_0)$. Usually, this weight function will depend on a tuning constant h, and we will write it as $w(z, z_0, h)$. For example, in global bandwidth local regression, h is the bandwidth and $w(z, z_0, h) = h^{-1}K\{(z - z_0)/h\}$, where $K(\cdot)$ is the kernel (density) function. For nearest-neighbor local regression such as loess (Chambers and Hastie 1992, pp. 312–316), h is the span (the percentage of the data to be counted as neighbors of z_0), and $w(z, z_0, h) = K\{|z - z_0|/a(h)d(z_0)\}$, where $d(z_0)$ is the maximum distance from z_0 to the observations in the neighborhood of z_0 governed by the span and $a(h) = 1$ if $h < 1$ and $a(h) = h$ otherwise.

In practice one has two choices for the tuning constant: (a) fixed a priori or determined randomly as a function of the data, and (b) global (independent of z_0) or local. If the tuning constant is global, then one also has the choice of whether it is the bandwidth or the span; for local tuning constants, there is often no essential difference between using a bandwidth and a span. For example, in loess the span h is typically fixed and global; this makes sense, because the nearest-neighbor weighting of loess imposes locality indirectly. In kernel and local polynomial regression, there is a substantial literature for estimating a global bandwidth h, and some work on estimating local bandwidths.

For the purpose of specificity, here we consider local estimation of the tuning constant. If we could determine the bias and variance functions of $\hat{\alpha}(z_0)$, say $\text{bias}(z_0, h, \alpha)$ and $\text{var}(z_0, h, \alpha)$, then we might reasonably choose $h = h(z_0)$ to minimize the mean squared error (MSE) function $\text{MSE}(z_0, h, \alpha) = \text{var}(z_0, h, \alpha) + \text{bias}^2(z_0, h, \alpha)$. To implement this idea, one needs estimates of the bias and variance functions. An associate editor raised the question of whether one would have enough data to estimate a local bandwidth. The answer is often "strictly speaking, no," but there is a compromise between truly local bandwidths and a global bandwidth. Ruppert (1997) proposed smoothing of the MSE function before minimizing to obtain a local bandwidth and then smoothing the local bandwidth. This type of procedure was called a "partial local smoothing rule" by Hall, Marron, and Titterington (1995). Simulation studies by Ruppert (1997) for the smoothed empirical bias bandwidth selection local bandwidth and by Fan and Gijbels (1995, 1996) for another local bandwidth estimator show that local bandwidths can outperform global bandwidths even for moderately small datasets.

The kernel regression literature abounds with ways of estimating the bias and variance functions, usually based on asymptotic expansions. We digress here briefly to discuss this issue; the Appendix contains details of the algebraic arguments. In our general context, the bias and variance of $\hat{\Theta}(z)$ using kernel regression are qualitatively the same as for ordinary local polynomial regression. There are functions $\mathcal{G}_b\{z, K, \Theta(z), p\}$ and $\mathcal{G}_v\{z, K, \Theta(z), p\}$ with the property that in the interior of the support of Z,

$$\begin{aligned} \text{bias}\{\hat{\Theta}(z)\} &\sim h^{p+1}\mathcal{G}_b\{z, K, \Theta(z), p\} \quad \text{if } p \text{ is odd} \\ &\sim h^{p+2}\mathcal{G}_b\{z, K, \Theta(z), p\} \quad \text{if } p \text{ is even} \end{aligned}$$

and

$$\text{cov}\{\hat{\Theta}(z)\} \sim \{nhf_Z(z)\}^{-1}\mathcal{G}_v\{z, K, \Theta(z), p\}.$$

The function $\mathcal{G}_v$ does not depend on the design density. The same is true of $\mathcal{G}_b$ if p is odd, but not if p is even (see Ruppert and Wand 1994 for the case of local polynomial regression and also (A.4) in the Appendix). The actual formulas are given in the Appendix. Results similar to what is known to happen at the boundary in ordinary local polynomial regression can be derived in our context.

For example, if $p = 1$ and $\psi(\mathbf{y}, \mathbf{v}) = \mathbf{y} - \mathbf{v}$ (ordinary multivariate-response local linear regression), then

$$\mathcal{G}_b\{z, K, \Theta(z), 1\} = (1/2)\Theta^{(2)}(z) \int s^2 K(s)\, ds$$

and

$$\begin{aligned} &\mathcal{G}_v\{z, K, \Theta(z), 1\} \\ &\qquad = \left\{\int K^2(s)\, ds\right\} \{\mathbf{B}(z)\}^{-1}\mathbf{C}(z)\{\mathbf{B}^t(z)\}^{-1}, \end{aligned}$$

where

$$\mathbf{B}(z) = E\{(\partial/\partial\mathbf{v})\psi(\mathbf{Y}, \mathbf{v})|Z = z\}$$

and

$$\mathbf{C}(z) = E\{\psi(\mathbf{Y}, \mathbf{v})\psi^t(\mathbf{Y}, \mathbf{v})|Z = z\},$$

with both $\mathbf{B}(z)$ and $\mathbf{C}(z)$ evaluated at $\mathbf{v} = \Theta(z)$. In this specific example, if $\mathbf{I}$ is the identity matrix, then $\mathbf{B}(z) = -\mathbf{I}$ and $\mathbf{C}(z) = \text{cov}(\mathbf{Y}|Z = z)$.

We now return to tuning constant estimation. For local regression, one could in principle use the asymptotic expansions to derive bias and variance formulas for $\hat{\alpha}(z_0)$. This is complicated by the facts that (a) the bias depends on higher-order derivatives of $\Theta(z_0)$, (b) if p is even then the bias depends on the design density, and (c) the variance depends on the density of the Zs. Instead of carrying through this line of argument, we instead propose methods that avoid direct use of asymptotic formulas and that are applicable as well to methods other than local regression. Such a goal has already been achieved in the kernel literature for ordinary local polynomial estimation. [See Fan and Gijbels (1995) and Ruppert (1997), the latter of which we use in our more general context.]

3.1 Empirical Bias Estimation

Ruppert (1997) suggested a method of bias estimation that avoids direct estimation of higher-order derivatives arising in asymptotic bias formulas. He termed this the method empirical bias bandwidth selection (EBBS).

The basic idea is as follows. Fix h_0 and z_0, and use as a model for the bias a function $f(h, \gamma)$ known except for the parameters $\gamma = (\gamma_1, \ldots, \gamma_t)$; for example, $f(h, \gamma) = \gamma_1 h^{p+1} + \cdots + \gamma_t h^{p+t}$, where $t \geq 1$, for local pth degree polynomial kernel regression. The model $f(h, \gamma)$ comes from asymptotic theory, which shows that the asymptotic bias has an expansion in powers of h beginning with power $p+1$, assuming that Θ has at least $p+1$ continuous derivatives. For any h_0, form a neighborhood of tuning constants $\mathcal{H}_0$. On a suitable grid of tuning constants h in $\mathcal{H}_0$, say $\{h_1, \ldots, h_K\}$, where $K \geq t+1$, compute the local polynomial estimator $\hat{\alpha}(z_0, h)$, which should be well described as a function of h by $\hat{\alpha}(z_0, h) = \gamma_0 + f(h, \gamma) + o_P(h^{p+t})$, the value $\gamma_0 = \alpha(z_0)$ in the limit. Then let $(\hat{\gamma}_0, \hat{\gamma})$ minimize $\sum_{k=1}^{K}\{\hat{\alpha}(z_0, h_k) - (\hat{\gamma}_0 + f(h_k, \hat{\gamma}))\}^2$. Appealing to asymptotic theory, and if $\mathcal{H}_0$ is small enough, the bias should be well estimated at h_0 by $f(h_0, \hat{\gamma})$.

In practice, the algorithm is defined as follows. For any fixed z_0, set a range $[h_a, h_b]$ for possible local tuning constants. For example, h_a and h_b could be $d(z_0)$ corresponding to spans of .1 and 1.5. Our experience is that the optimal local bandwidth is generally in this range. Then form a geometrically spaced grid of M points,

$$\mathcal{H}_1 = \{h_j : j = 1, \ldots, M, h_1 = h_a, h_M = h_b\}.$$

We have not tried spacing other than geometric, because it seemed intuitive that smaller bandwidths should be more closely spaced.

Fix constants (J_1, J_2) such that $J_1 + J_2 \geq t$. For any $j = 1 + J_1, \ldots, M - J_2$, apply the procedure defined in the previous paragraph with $h_0 = h_j$ and $\mathcal{H}_0 = \{h_k, k =$

$j-J_1,\ldots,j+J_2\}$. This defines $\widehat{\text{bias}}\{\hat{\alpha}(z_0,h_j)\}$. For tuning constants not on the grid $\mathcal{H}_1$, interpolation via a cubic spline is used.

Note that we must set the limits of interesting tuning constants $[h_a,h_b]$ and the four tuning constants (t,M,J_1,J_2). Ruppert (1997) found that $J_1=1,(t,J_2)=(1,1)$ or $(2,2)$, and M between 12 and 20 give good numerical behavior in the examples that he studied using local polynomial kernel regression.

3.2 Empirical Variance Estimation: The Sandwich Method

It is useful to remember that q is the dimension of $\mathbf{\Theta}, p$ is the degree of the local polynomial, and $\mathbf{G}_p$ is defined just after (2). At this level of generality, the sandwich formula can be used to derive an estimate of the covariance matrix of $(\hat{\mathbf{b}}_0,\ldots,\hat{\mathbf{b}}_p)$. In parametric problems the solution $\hat{\mathbf{\Theta}}$ to (1) has sandwich (often called "robust") covariance matrix estimate $\mathbf{B}_n^{-1}\mathbf{C}_n(\mathbf{B}_n^t)^{-1}$, where

$$\mathbf{C}_n=\sum_{i=1}^{n}\psi(\mathbf{Y}_i,\hat{\mathbf{\Theta}})\psi^t(\mathbf{Y}_i,\hat{\mathbf{\Theta}})$$

and

$$\mathbf{B}_n=\sum_{i=1}^{n}(\partial/\partial\mathbf{\Theta}^t)\psi(\mathbf{Y}_i,\hat{\mathbf{\Theta}}).$$

The analogous formulas for the solution to (2) are defined as follows. In what follows, if $\mathbf{A}$ is $l\times q$ and $\mathbf{B}$ is $r\times s$, then $\mathbf{A}\otimes\mathbf{B}$ is the Kronecker product, defined as the $lr\times qs$ matrix formed by multiplying individual elements of $\mathbf{A}$ by $\mathbf{B}$; for example, if $\mathbf{A}$ is a 2×2 matrix, then

$$\mathbf{A}\otimes\mathbf{B}=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\otimes\mathbf{B}=\begin{bmatrix}a_{11}\mathbf{B} & a_{12}\mathbf{B}\\ a_{21}\mathbf{B} & a_{22}\mathbf{B}\end{bmatrix}.$$

Let $\chi(\mathbf{y},\mathbf{v})=(\partial/\partial\mathbf{v}^t)\psi(\mathbf{y},\mathbf{v})$. Then the asymptotic covariance matrix of $(\hat{\mathbf{b}}_0,\ldots\hat{\mathbf{b}}_p)$ is estimated by $\{\mathbf{B}_n(z_0)\}^{-1}$ $\mathbf{C}_n(z_0)\{\mathbf{B}_n^t(z_0)\}^{-1}$, where

$$\mathbf{C}_n(z_0)=\sum_{i=1}^{n}w^2(Z_i,z_0)\times[\{\mathbf{G}_p(Z_i-z_0)\mathbf{G}_p^t(Z_i-z_0)\}\otimes(\hat{\psi}_i\hat{\psi}_i^t)] \quad (6)$$

and

$$\mathbf{B}_n(z_0)=\sum_{i=1}^{n}w(Z_i,z_0)\times[\{\mathbf{G}_p(Z_i-z_0)\mathbf{G}_p^t(Z_i-z_0)\}\otimes\hat{\chi}_i], \quad (7)$$

where $\hat{\psi}_i=\psi\{\mathbf{Y}_i,\sum_{j=0}^{p}\hat{\mathbf{b}}_j(Z_i-z_0)^j\}$ and analogously for $\hat{\chi}_i$. In practice, we replace $\sum_{j=0}^{p}\hat{\mathbf{b}}_j(Z_i-z_0)^j$ by $\hat{\mathbf{\Theta}}(Z_i)$. An argument justifying these formulas is sketched in the Appendix. In practice, we multiply the sandwich covariance matrix estimate by $n/\{n-(p+1)q\}$, an empirical adjustment for loss of degrees of freedom. In a variety of problems that we have investigated (see, e.g., Simpson, Guth, Zhou, Carroll 1996), this empirical adjustment improves coverage probabilities of sandwich-based confidence intervals, when combined with t percentiles with $n-(p+1)q$ df. There is no theoretical justification for this adjustment, however. In specific problems, bias adjustments for the sandwich estimator may be more or less easy to construct. In the case for GLMs, covariance matrix estimators that automatically adjust for leverage and the like already exist (see Hastie and Tibshirani 1990, sec. 6.8.2, and Kauermann and Tutz 1997).

In some problems the sandwich term $\mathbf{C}_n(z_0)$ can be improved on because the covariance matrix of $\psi(\cdot)$ is known partially or fully. For example, if $\psi(\cdot)$ is given by (3), then $E(\psi\psi^t)=\sigma^2\{\mu^{(1)}\}^2/V$, and one would replace $(\hat{\psi}_i\hat{\psi}^t)$ in (6) by $\hat{\sigma}^2\{\hat{\mu}_i^{(1)}\}^2/\hat{V}_i$. In addition, using score-type arguments, one bases work on $\chi(\cdot)=-\{\mu^{(1)}(\cdot)\}^2/V$ and would replace $\hat{\chi}_i$ in (6) by $-\{\hat{\mu}_i^{(1)}\}^2/\hat{V}_i$. We suggest using such additional information when it is available, because the sandwich estimator can be considerably more variable than model-based alternatives. For example, in simple linear regression, sandwich-based estimates of precision are typically at least three times more variable than the usual precision estimates.

The sandwich method in parametric problems does not work in all circumstances, even asymptotically, the most notable exception being the estimate of the median. In this case, if Y is scalar, then $\psi(\mathbf{Y},x)=I(Y\le x)-1/2$, where I is the indicator function. This choice of $\psi(\cdot)$ has zero derivative, and thus (7) equals 0. Alternatives to the sandwich estimators do exist, however, although their implementation and indeed the theory itself needs further investigation. A sandwich-type method was described by Welsh, Carroll, and Ruppert (1994), who used a type of weighted differencing. Alternatively, one can use the so-called "m out of n" resampling method as defined by Politis and Romano (1994), although application of this latter technique requires that one know the rate of convergence of the nonparametric estimator, this being theoretically $(nh)^{1/2}$ for local linear regression. How to choose the level of subsampling m remains an open question.

4. EXAMPLES

In this section we present three example of local estimating equations. Other examples can be found in an early version of this article, available via anonymous ftp at stat.tamu.edu in the directory/pub/rjcarroll/nonparametric.calibration in the file npcal15.ps.

4.1 Nutrition Calibration: NHS and WISH

We used the NHS and WISH data described in Section 2 to understand whether the correlation between a FFQ and usual intake, ρ_{QT}, depends on age, based on the nutrient % calories from fat. Nutrition data with repeated measurements typically have the feature of time trends in total amounts and sometimes in percentages, so that, for example, one might expect reported caloric intake (energy) to decline over time. To take this into account, we ratio adjusted all measurements so that the mean of each FR equals the first. (For an example of ratio adjustment, see Nusser, Carriquiry, Dodd, and Fuller 1996.)

As described previously, i denotes the individual, Q_i and T_i are the nutrient intakes as reported on the FFQ and usual intake, and F_{ij} is the jth replicated FR for the ith individual. The mean of the replicated FRs is $\bar{F}_i$. The unknown parameters in the problem are conveniently characterized as $\Theta = (\theta_1, \ldots, \theta_6)$, where $\theta_1 = E(Q)$, $\theta_2 = E(F) = E(T)$, $\theta_3 = \mathrm{var}(Q)$, $\theta_4 = \mathrm{cov}(Q, F) = \mathrm{cov}(Q, T)$, $\theta_5 = \mathrm{var}(U)$, and $\theta_6 = \mathrm{var}(T)$. Note that for any two replicates F_{ij} and F_{ik} for $j \neq k$, $\theta_6 = \mathrm{cov}(F_{ij}, F_{ik})$. Letting $\mathbf{Y}_i = (Q_i, F_{i1}, \ldots, F_{im})$ be the observed data ($m = 6$ in WISH, $m = 4$ in NHS), the usual method-of-moments estimating function is

$$\psi(\mathbf{Y}_i, \Theta) = \begin{bmatrix} Q_i \\ \bar{F}_i \\ (Q_i - \theta_1)^2 \\ (Q_i - \theta_1)(\bar{F}_i - \theta_2) \\ (m-1)^{-1} \sum_{j=1}^{m} (F_{ij} - \bar{F}_i)^2 \\ \{m(m-1)\}^{-1} \sum_{j=1}^{m} \sum_{k \neq j}^{m} (F_{ij} - \theta_2)(F_{ik} - \theta_2) \end{bmatrix} - \Theta. \qquad (8)$$

Numerically, the solution to (2) is easily obtained. Local estimates of $\theta_1(z)$ and $\theta_2(z)$ use nothing more than direct local regression of Q_i and $\bar{F}_i$ on Z_i and once they are plugged into the third–sixth components of ψ, $\{\theta_3(z), \ldots, \theta_6(z)\}$ can also be computed by local least squares; for example, by regressing $(Q_i - \hat{\theta}_1)^2$ on Z_i to obtain $\hat{\theta}_3$. The main parameter of interest is the correlation between Q and T, $\rho_{QT}(z_0) = \theta_4(z_0)\{\theta_3(z_0)\theta_6(z_0)\}^{-1/2}$.

In this example we used nearest-neighbor weights based on the span, as described at the start of Section 3. As in the S-PLUS implementation of loess, we used the tricubed kernel function, which is proportional to $(1 - |v|^3)^3$ for $|v| \leq 1$ and equals 0 elsewhere. For a fixed value of the span, we assessed standard errors by two means. First, we obtained an estimated covariance matrix for $\hat{\Theta}(z_0)$ using the sandwich formula, and then used the delta method to obtain an estimated variance for $\rho_{QT}(z_0), \beta_1(z_0)$, etc. We based the second standard error estimates on the nonparametric bootstrap, with the pairs $(\mathbf{Y}, Z)$ resampled from the data with replacement; we used 500 bootstrap samples. For a range of spans and for a variety of datasets and nutrient variables, the sandwich delta and the bootstrap standard errors were very nearly the same. This is not unexpected, given that the spans used are fairly large. As a theoretical justification, note that if the span is bounded away from 0, then the estimator $\hat{\Theta}(z)$ converges at parametric rates (although to a biased estimate), and the bootstrap and sandwich covariance matrix estimates are asymptotically estimating the same quantity.

Figure 1 shows the value of ρ_{QT}(age) for the NHS % calories from fat for various spans in the range .6–.9 using local quadratic regression. To understand the age distribution in this study, we have also displayed the 10th, 25th, 50th, 75th, and 90th sample percentiles of age. Although there is some variation between the curves for the different values of the span, the essential feature is consistent—namely, that those under age 50 have significantly (in the practical sense) lower correlations than do those over age 50. The statistical significance of this finding can be assessed in various ways. The simplest is to split the data into two populations on the basis of age groups and simply compute $\hat{\rho}_{QT}$ for each population; the estimates are statistically significantly different at a significance level below .02.

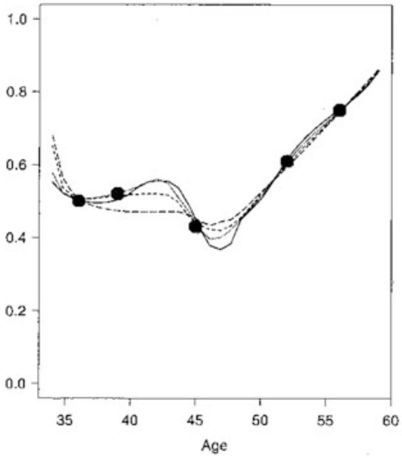


Figure 1. Nurses' Health Study Estimating ρ_{QT} According to Sensitivity to the Choice of Span. Percent calories from fat by using local quadratic regressions with 10th, 25th, 50th, 75th, and 90th percentiles of age. ——, span = .6; · · ·, span = .7; – – –, span = .8; — — —, span = .9.

A second test is slightly more involved. We computed the estimate of ρ_{QT}(age) for 16 equally spaced points on the range from 34–59, along with the bootstrap covariance matrix of these 16 estimates. We then tested whether the estimates were the same using Hotelling's T^2 test and tests for linear and quadratic trend using weighted least squares. As expected after inspection of Figure I, the linear and quadratic tests had significance levels below .05 for spans in [.7, .9].

We also estimated the span, in the following manner. For computational purposes, we used eight values of age, and using the methods of Section 3 we computed an estimate of the MSE using empirical bias estimation ($J_1 = J_2 = 5, M = 41, h_a = .6, h_b = 1.0$) and the sandwich method; we chose the estimated span to minimize the sum over the eight ages of the estimated MSE. The estimated span was .78 for local linear regression and .90 for local quadratic regression. We then bootstrapped this process, including the estimation of the span, and found that although the significance level was slightly greater than that for a fixed span, it was still below .05.

Because the empirical bias estimate has the tuning constants (M, J_1, J_2), there is still some art to estimating the span. We studied the sensitivity of the estimated span and the estimated average MSE to these tuning constants, and found that the results did not depend too heavily on them as long as J_1 and J_2 were increased with increasing values of M. For example, the estimated average MSEs for local linear regression in three cases—$(M, J_1, J_2) = (41, 5, 5)$, $(101, 13, 13)$, and $(201, 25, 25)$—were calculated, and there was little difference between the three MSEs. However, fixing J_1 and J_2 while increasing M resulted in quite variable bias estimates.

We repeated the estimation process for WISH. There is no evidence of an age effect on ρ_{QT} in WISH. This may be due to the different population or the different FFQ, but may just as well be due to the much larger measurement error in the FRs in WISH than in NHS.

Finally, we investigated local average, linear, quadratic, and cubic regression, with a span of .8; see Figure 2, where we also display the five estimates of ρ_{QT} based on the quintiles of the age distribution. Given the variability in the estimates, the main difference in the methods occurs for higher ages, where the local average regression is noticeably different from the others and from the quintile analysis. Our belief is that this difference arises from the well-known bias of local averages at endpoints.

We redid this analysis using kernel instead of loess weights with locally estimated bandwidths. The results of the two analyses were similar and are not displayed here.

Finally, we comment on issues specific to nutrition:

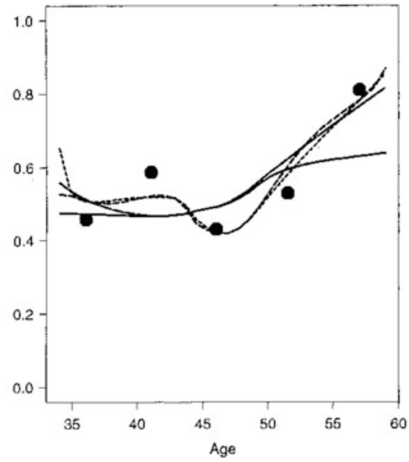


Figure 2. Nurses' Health Study (NHS) Estimating ρ_{QT} With Span .8 and for Local Average, Linear Quadratic, and Cubic Regression With Results From Quintile Analysis. ——, local average; — — —, local linear; - - -, local quadratic; – – –, local cubic.

- We have assumed that the errors ε_i are independent of U_{ij}. This appears to be roughly the case in these two datasets, although it is not true in other datasets that we have studied; for example, the Women's Health Trial data studied by Freedman et al. (1991). The model and the estimating equation are easily modified in general to account for such correlation when it occurs. Similarly, the model and the estimating equation can be modified to take into account a parametric model for correlation among the U_{ij}s; for example, an AR(1) model. Although such correlations exist in these datasets, they are relatively small and should not have a significant impact on the results.
- The method of moments (8) is convenient and easy to compute. In various asymptotic calculations and numerical examples, we have found that it is effectively equivalent to normal-theory maximum likelihood.
- There is emerging evidence from biomarker studies that food records such as those used in NHS are biased for total caloric intake, with those having high body mass index (BMI) underreporting total caloric intake by as much as 20% (see, e.g., Martin, Su, Jones, Lockwood, Tritchler and Boyd 1996). The bias is less crucial for log(total calories) and presumably even less so for the variable used in our analysis, % calories from fat, although no biomarker data exist to verify our conjecture. Despite our belief that this variable is not much subject to large biases explainable by BMI, we have performed various sensitivity analyses that allow for bias. For example, we changed the FFQ and food record data for those with $22 \leq$ BMI ≤ 28 by adding on average 4 to their % calories from fat (a 10% change), whereas for those with BMI > 28 we added on average 7 to their % calories from fat (a 20% increase). The adjustments were proportional to FFQs and food records, and the same adjustment was added to all food records of an individual. These adjustment in effect simulate adjustments to the data that would be made if a strong bias were found in % calories from fat for food records. The analysis of the modified data gave correlations very similar to those shown in our graphs; that is, the effect of bias on the correlation estimates was small.
- If one had replicated FFQs, then many modifications to the basic model could be made. One might conjecture an entirely different error structure; for example,

$$Q_{ij} = \beta_0 + \beta_1 T_i + r_i + \varepsilon_{ij}, \qquad F_{ij} = T_i + s_i + U_{ij},$$

$$\sigma_s^2 = \sigma_r^2.$$

This model is identifiable only if corr(r, s) is known. We have fit such models using local method of moments to a large $(n > 400)$ dataset with repeated FFQs and using 24-hour recalls for various choices of corr$(r, s) \leq .5$. The net effect was that such analyses are very different from those based on model (4)–(5); ρ_{QT} increased by a considerable amount, whereas the local estimates of var(T) as a function of age became much smaller. Of course, the point is that analyses of such complex models are relatively easy using our local estimating function approach.

4.2 Multivariate Z: Lung Cancer Mortality Rates

The methods of this article can be extended to the multivariate $\mathbf{Z}$ case. Suppose that $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{im})^t$, where the Z_{ij} are scalar. Then, following Ruppert and Wand (1994), local linear functions are $\boldsymbol{\Theta}(\mathbf{z}) = \mathbf{b}_0 + \mathbf{b}_1(\mathbf{z} - \mathbf{z}_0)$, where $\mathbf{b}_0$ is a $p \times 1$ vector and $\mathbf{b}_1$ is a $p \times m$ matrix. The generalization of (2) is to solve

$$0 = \sum_{i=1}^{n} w(\mathbf{Z}_i, \mathbf{z}_0)\psi\{\mathbf{Y}_i, \mathbf{b}_0 + \mathbf{b}_1(\mathbf{Z}_i - \mathbf{z}_0)\}\mathbf{G}_m(\mathbf{Z}_i - \mathbf{z}_0), \tag{9}$$

where $\mathbf{G}_m^t(\mathbf{v}) = (1, \mathbf{v}^t)$. When $\mathbf{Z}$ is multivariate and using kernel weights, the kernel K is multivariate and the bandwidth h is replaced by a positive-definite symmetric matrix $\mathbf{H}$. The simplest choice is to restrict $\mathbf{H}$ to equal $h\mathbf{I}$ for $h > 0$ and with $\mathbf{I}$ the identity matrix, and in this situation the methods we have discussed for empirical bias and variance estimation apply immediately to the estimates $\hat{\boldsymbol{\Theta}}(\mathbf{z}_0) = \hat{\mathbf{b}}_0$. The application of empirical bias modeling to more general bandwidth matrices is currently under investigation.

Extensions to higher-order local polynomials require more care. Completely nonparametric functional versions are easy in principle, but the notation is complex and practical implementation difficult (see Ruppert and Wand 1994, sec. 4). It is much easier to fit "local additive" models, so that if $\mathbf{z} = (z_1, \ldots, z_m)^t$ and $\mathbf{z}_0 = (z_{01}, \ldots, z_{0m})^t$, then $\boldsymbol{\Theta}(\mathbf{z}) = \mathbf{b}_0 + \sum_{k=1}^{m}\sum_{j=1}^{p} \mathbf{b}_{kj}(\mathbf{z}_k - \mathbf{z}_{0k})^j$; this is identical to (9) when $p = 1$, and the extension of (9) to $p > 1$ is immediate. We use the term "local additive" to warn the reader that we are not considering a globally additive model in the sense of Hastie and Tibshirani (1990), where $\boldsymbol{\Theta}(\mathbf{z}) = \boldsymbol{\Theta}_1(z_1) + \cdots + \boldsymbol{\Theta}_m(z_m)$ for all $\mathbf{z}$ and some functions $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_m$. Globally additive models are outside the scope of this article but would be quite useful when m is larger than 2 or 3 and the "curse of dimensionality" comes into play.

For an example of (9), we consider a problem in which $Y = 10 + \log[(R + .5)/(10^5 - R + .5)]$, where R is the mortality rate per 10^5 males for males dying of lung cancer, as a function of $\mathbf{Z} =$ (age class, year). We call Y the "adjusted" logit because of the .5 offset. The data come from the Australian Institute of Health and are publicly available. The age classes are represented by their midpoints, which are (2, 7, 12, 17, 22, 27, 32, 37, 42, 47, 52, 62, 67, 72, 77, 82, 87), and the years run from 1950–1992 inclusive. For each age class and year subpopulation, we can treat the number of deaths per 10^5 males as being $(d/N) \times 10^5$, where d, the total number of deaths in the subpopulation due to lung cancer, is binomial (N, π) with π the probability of death for an individual and N is the size of the relevant subpopulation. The values of the Ns are known and are used later. Because p is small, d is approximately $\text{Poisson}(Np)$ and $\text{var}(R) \approx (10^5/N)E(R)$. In this case, the logit and the log transformation are similar; we use the former to maintain comparability with other work currently being done on these data. We could model the variance of Y as a function of its mean and of N. Alternatively, we could model the variance of Y as a function of $\mathbf{Z}$. We start with the second possibility. If $\boldsymbol{\Theta} = (\theta_1, \theta_2)^t$, then the estimating function for mean and variance estimation is just $\psi(Y, \boldsymbol{\Theta}) = \{Y - \theta_1, (Y - \theta_1)^2 - \theta_2\}^t$. There are two good reasons for considering a robust analysis, however. First, there may be concern over the potential for outliers in the response; second, a robust analysis may be numerically more stable. We treat $\tau = \log(\theta_2)$ as the spread parameter (to ensure nonnegativity) and use the estimating equation

$$\psi(Y, \boldsymbol{\Theta}) = \begin{bmatrix} g\{(Y - \theta_1)/\exp(\tau)\} \\ g^2\{(Y - \theta_1)/\exp(\tau)\} - \int g^2(v)\phi(v)\,dv \end{bmatrix},$$

where $g(v) = g(-v) = v$ if $0 \le v \le c$ and $= c$ if $v > c$, $\phi(v)$ is the standard normal density function and c is a tuning constant controlling the amount of robustness desired; $c = 1.345$ is standard. In the robustness literature, the parameter estimator is known as "proposal 2" (Huber 1981). The spread estimating function can be rewritten as

$$g^2\{\exp(\log|Y - \theta_1| - \tau)\} - \int g^2(v)\phi(v)\,dv,$$

which expresses the spread equation in the form of a location equation. Consideration of the function $g^2\{\exp(x)\} - \int g^2(v)\phi(v)\,dv$ suggests that we simplify the procedure further by replacing it by the much simpler function g with $c = 2$ to increase the efficiency of spread estimation. This is in accordance with the procedure developed by Welsh (1996).

The response and spread surfaces, $\hat{\Theta}_1(z)$ and $\hat{\Theta}_2(z)$, for the lung cancer mortality data are shown in Figures 3a and 3b as surface plots and as contour plots. After some experimentation, the bandwidth matrix was restricted to be of the form h diag(2, 1), and then h was chosen empirically as in Section 3, with a backfitting modification to the basic algorithm (2) described in Section 5. But the results reported here are stable over a range of bandwidth matrices, the main effect of substantial increases in bandwidths being to reduce the ripple and peak in the response and spread surfaces at high ages and early years. Local linear fitting was used in Figures 3a and 3b; local quadratic estimates are similar but with somewhat higher peaks in the spread surface. It is clear that the logit of mortality increases nonlinearly with age class and that there is at best a very weak year effect that shows increased mortality in recent years in the highest age classes. The spread surface shows a ridge of high variability in age classes 20–40 with generally lower variability at both extremes. A delta-method analysis shows that this ridge is due to the logit transformation (with the .5 offset) and the near-Poisson variability of R; see the discussion in the final paragraph of this section. There is also high variability in the highest age classes for the earlier years. This is also the only evidence of a year effect on the variability. The roughness of the spread surface is due mostly to variation in the values of N.

We also modeled the variance of Y as a function of N and the mean of Y. Let N^* be the value of N for a given age class and year divided by the mean of all the N's. Let e^* be the "population size–adjusted residual," defined as the residual for that age class and year times $(N^*)^{1/2}$. Figure 3c plots the absolute values of the e^*s versus the fitted values. Figure 3d plots a local linear fit to the data in 3c. To estimate the spread for a given age class and year, one divides the fitted value from 3d by $(N^*)^{1/2}$. The peak in 3d corresponds to the ridge in 3a.

As mentioned earlier, if we assume that

$$\mathrm{var}(R) = (10^5/N)E(R), \tag{10}$$

then this ridge can be explained by a delta-method calculation showing that

$$\mathrm{SD}(Y) \approx \frac{10^5 + 1}{(E(R) + .5)(10^5 - E(R) + .5)} \{10^5 E(R)/N\}. \tag{11}$$

We checked (10) by dividing the residuals by the right side of (11), squaring, and then smoothing these squared "standardized residuals" against the fitted values and N. The resulting surface, not included here to save space, was nearly constantly equal to 1, supporting (10).

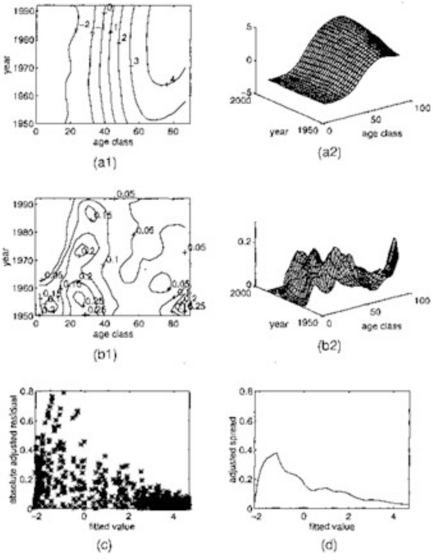


Figure 3. Lung Cancer Mortality Rates; Estimates of the Response Surface and Spread. (a1) and (a2) Adjusted logit of mortality; (b1) and (b2) spread; (c) absolute residuals multiplied by $(N^*)^{1/2}$; (d) local linear fit to the scatterplot in (c).

4.3 Variance Functions and Overdispersion

Problems involving count and assay data are often concerned with overdispersion. For example, if $\mathbf{Y} = (Y, X)$, then the mean of Y might be modeled as $\mu(\mathcal{B}, X)$ and its variance might have the form

$$\mathrm{var}(Y|X) = \exp[\theta_1 + \theta_2 \log\{\mu(\mathcal{B}, X)\}]. \tag{12}$$

Here we assume that the mean function is properly determined so that $\mathcal{B}$ is to be estimated parametrically. If $\theta_2 = 1$ and $\theta_1 > 1$, then we have overdispersion relative to the Poisson model, whereas $\theta_2 \neq 2$ means a departure from the gamma model. In general, we are asking how the variance function depends on the logarithm of the mean. For given θ_2, $\mathcal{B}$ is usually estimated by generalized least squares (quasi-likelihood). Consistent estimates of $\mathcal{B}$ can be obtained using quasi-likelihood assuming that θ_2 is a fixed value, even if it is not. This well-known fact is often referred to operationally by saying that (12) with fixed θ_2 is a "working" variance model (Diggle et al. 1994).

The problem then is one of variance function estimation, where if $\eta(\mathcal{B}, X) = \log\{\mu(\mathcal{B}, X)\}$, then we believe that the variances are of the form $\exp[\Theta\{\eta(\mathcal{B}, X)\}]$ for some function $\Theta(\cdot)$. Our objective now is to find a suitable estimator of $\Theta(\cdot)$. In a population the variance is $\exp(\Theta)$, which is estimated using the estimating function

$$\psi(\mathbf{Y}, \Theta, \hat{\mathcal{B}}) = \{Y - \mu(\hat{\mathcal{B}}, X)\}^2 \exp(-\Theta) - 1. \tag{13}$$

Estimating Θ as a function of $Z = \eta(\hat{\mathcal{B}}, X)$ is accomplished by using (2) in the obvious manner, namely

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi \times \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} b_j (Z_i - z_0)^j, \hat{\mathcal{B}} \right\} G_p^{\mathrm{t}}(Z_i - z_0, \hat{\mathcal{B}}), \tag{14}$$

with ψ given by (13). Because $\hat{\mathcal{B}}$ estimates $\mathcal{B}_0$ at parametric rates, asymptotically there is no effect due to estimating $\mathcal{B}_0$ on the estimate of $\Theta(z)$.

We applied this analysis to three datasets, the esterase assay and hormone assay datasets described by Carroll and Ruppert (1988, chap. 2) and a simulated dataset with $\Theta\{\eta(\mathcal{B}, X)\} = 1.6 + \sin\{\eta(\mathcal{B}, X)\}$, using the same Xs and estimates of $\mathcal{B}$ as in the esterase assay. The model for the mean in all three cases is linear. Previous analyses suggested that the esterase assay data were reasonably well described by a gamma model, with the hormone assay less well described as such because $\theta_2 \approx 1.6$. We used $\theta_2 = 2$ as our working variance model to obtain $\hat{\mathcal{B}}$ for these three datasets. We fit local linear models weighted using loess with the span allowed to take on values between .6 and 2.0 and estimated by the techniques of this article. Figure 4 compares the fitted variance functions divided by the gamma model variance function and rescaled on the horizontal axis to fit on the same plot. Through the range of the data, the deviation from the gamma function is only a factor of about 35% for the esterase assay, indicating a good fit for this model. The hormone assay deviates from the gamma model somewhat more, with variances ranging over factors of two. Both have estimated spans greater than 1.0, indicating that the linear model is a reasonable fit; the hormone

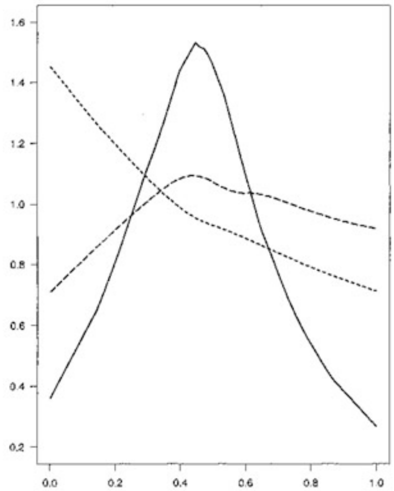


Figure 4. Esterase Assay (– – –), Hormone Assay (· · ·), and Simulated (——) Datasets. Estimated discrepancies from the gamma variance model, rescaled. On the vertical axis is the fitted variance function divided by the gamma model variance function. The horizontal axis is $Z = \eta(\mathcal{B}, X) = \log\{\mu(\mathcal{B}, X)\}$.

data simply have a value $\theta_2 < 2$. The simulated data show the sine-type behavior from which they were generated, and a much smaller estimated span (.7).

4.4 Partially Parametric Models

The overdispersion example in Sec. 4.3 contained a parametric part $\mathcal{B}_0$ and a nonparametric part $\Theta(\cdot)$. The "working" estimation method used for the parametric part was chosen so that $\hat{\mathcal{B}}$ was consistent and asymptotically normally distributed with variance of order n^{-1} *even if* $\Theta(\cdot)$ *was completely misspecified.* In other problems, an estimation method for $\mathcal{B}_0$ is chosen whose validity depends on correctly specifying or consistently estimating $\Theta(\cdot)$. An alternative estimator for $\mathcal{B}_0$ given a version $\hat{\Theta}(\cdot)$ is to solve in $\mathcal{B}$ the estimating equation $0 = \sum_{i=1}^{n} \Lambda\{\mathbf{Y}_i, \mathcal{B}, \hat{\Theta}(\cdot)\}$. The natural approach to use then is to solve the equations

$$0 = \sum_{i=1}^{n} \Lambda\{\mathbf{Y}_i, \mathcal{B}, \hat{\Theta}(\cdot)\} \qquad (15)$$

and

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi\{\mathbf{Y}_i, \mathcal{B}, \Theta(\cdot)\}\mathbf{G}_p^t(Z_i - z_0) \qquad (16)$$

for $z_0 = Z_1, \ldots, Z_n$, where $\Theta(z_0) = \sum_{j=0}^{p} \mathbf{b}_j(Z_i - z_0)^j$.

As we have described it, solving (15)–(16) simultaneously is a form of backfitting. One fixes the current estimate of $\mathcal{B}_0$ and obtains an updated estimate of $\Theta(\cdot)$, reverses the process, and then iterates. Asymptotically valid inferences for $\Theta(z)$ are obtained using only (16) and assuming that $\hat{\mathcal{B}}$ is fixed at its estimated value. Asymptotically valid estimates of the covariance matrix of $\hat{\mathcal{B}}$ remain an open problem, although in some cases they can be derived (see Carroll, Fan, Gijbels, and Wand 1997 for single-index models and Severini and Staniswallis 1994 for partial linear models).

The backfitting algorithm has a well-known feature. We confine our remarks to local regression, but these remarks hold for other types of fitting methods as well (Hastie and Tibshirani 1990, pp. 154–155). Specifically, in local linear regression, if the bandwidth is h, then $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B}_0)$ has variance of order 1 but has bias of the order $(nh^4)^{1/2}$, so that getting an asymptotic normal limit distribution with zero bias requires that $nh^4 \to 0$. Unfortunately, "optimal" kernel bandwidth selectors for given $\mathcal{B}$ are typically of the order $h \sim n^{1/5}$, in which case $nh^4 \to \infty$ and the bias in the asymptotic distribution of $\hat{\mathcal{B}}$ does not disappear. If one is even going to worry about this problem (we know of no commercial program that does, nor of any practical examples in which the bias problem is of real concern), then the usual solution is to undersmooth in some way.

Some problems allow for a somewhat more elegant solution to the bias problem, specifically when (15)–(16) are formed as the derivatives of a *single* optimization criterion. None of the estimators that we have described in this article has this form. Optimization of a single criterion basically means a likelihood specification. When this occurs, nonparametric likelihood as described by Severini and Wong (1992) can be applied to make the bias problem disappear, at least in principle, as follows. Let the data likelihood be $l\{\mathcal{B}, \Theta(\cdot)\}$. For fixed $\mathcal{B}$, let $\hat{\Theta}(\cdot, \mathcal{B})$ be the local estimator derived by maximizing the likelihood in Θ with $\mathcal{B}$ fixed. Nonparametric likelihood maximizes $l\{\mathcal{B}, \hat{\Theta}(\cdot, \mathcal{B})\}$ as a function of $\mathcal{B}$. In contrast, backfitting fixes the current $\hat{\Theta}(\cdot, \mathcal{B})$ and updates the estimate of $\mathcal{B}$ by maximizing $l\{\boldsymbol{\alpha}, \hat{\Theta}(\cdot, \mathcal{B})\}$ in $\boldsymbol{\alpha}$. Nonparametric likelihood can be more difficult to implement than backfitting, especially in our context when $\Theta(\cdot)$ is multivariate. But it is easy to implement if Θ is scalar, $\mathbf{Y} = (Y, X, Z)$, and Y follows a GLM with mean $f(\{\Theta(Z) + X^t\mathcal{B}\}$. (See Severini and Staniswallis 1994 for the ordinary kernel regression case.)

5. MODIFICATIONS OF THE ALGORITHM

The method suggested in (2) requires that all components of $\Theta(z_0)$ be estimated simultaneously. This may be undesirable in some contexts. For example, when estimating a variance function nonparametrically, one often would first estimate the mean function, say $\Theta_1(z)$, form squared residuals $\{\mathbf{Y} - \hat{\Theta}_1(Z_i)\}^2$, and then regress these squared residuals on Z nonparametrically to obtain $\hat{\Theta}_2(z_0)$, the variance estimate at a given z_0. In this context strict application of (2) is different, because it is based on squared pseudoresiduals $\{\mathbf{Y} - \sum_{j=0}^{p} \hat{\Theta}^{(j)}(z_0)(Z_i - z_0)^j/j!\}^2$. In addition, one would often use different tuning constants at each step, but (2) assumes use of the same tuning constant.

The aforementioned example, as well as the nonparametric calibration problem, are examples of a multistage process, where components of $\Theta(\cdot)$ are estimated first and then plugged into the estimating equation for further components. Such problems are easily handled by a slight modification of our approach.

We illustrate the idea in a two-stage context, so that $\mathbf{\Theta} = (\mathbf{\Theta}_1, \mathbf{\Theta}_2)$. By the two-stage process we mean that the first component can be estimated without reference to the second, with weight function w_1 and estimating function ψ_1, so that we solve

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0)\psi_1 \\ \times \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j \right\} \mathbf{G}_p^t(Z_i - z_0). \quad (17)$$

The estimate is $\hat{\mathbf{\Theta}}_1(z_0) = \hat{\mathbf{b}}_{0,1}(z_0)$.

At the second stage there is a second weight function w_2 and a second estimating function ψ_2, and we solve

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0)\psi_2 \\ \times \left\{ \mathbf{Y}_i, \hat{\mathbf{\Theta}}_1(Z_i), \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \mathbf{G}_p^t(Z_i - z_0). \quad (18)$$

The estimate is $\hat{\Theta}_2(z_0) = \hat{b}_{0,2}(z_0)$.

The asymptotic covariance matrix of $\{\hat{\Theta}_1(z_0), \hat{\Theta}_2(z_0)\}$ defined by (17)–(18) is estimated by applying the sandwich method to the estimating equation

$$0 = \sum_{i=1}^{n} \left[\begin{array}{l} w_1(Z_i, z_0)\psi_1 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j \right\} \\ w_2(Z_i, z_0)\psi_2 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \right. \\ \qquad\qquad \left. \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \end{array} \right] \\ \times \mathbf{G}_p^t(Z_i - z_0). \quad (19)$$

If $\mathbf{c}_{p \cdot i} = \mathbf{G}_p(Z_i - z_0)\mathbf{G}_p^t(Z_i - z_0)$, the sandwich formulas are

$$B_n(z_0) \\ = \sum_{i=1}^{n} \left\{ \begin{array}{ll} w_1(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\chi}}_{i11} & 0 \\ w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\chi}}_{i21} & w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\chi}}_{i21} \end{array} \right\}$$

and

$$C_n(z_0) = \sum_{i=1}^{n} \left\{ \begin{array}{l} w_1^2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\psi}_{i1}\hat{\psi}_{i1}^t \\ \quad w_1(Z_i, z_0)w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\psi}_{i1}\hat{\psi}_{i2}^t \\ w_1(Z_i, z_0)w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\psi}_{i2}\hat{\psi}_{i1}^t \\ \quad w_2^2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\psi}_{i2}\hat{\psi}_{i2} \end{array} \right\},$$

where $\boldsymbol{\chi}_i$ is made up of the elements $\boldsymbol{\chi}_{ijk}$ for $j, k = 1, 2$. In practice, one might replace $\sum_{j=0}^{p} \hat{\mathbf{b}}_{j,k}(Z_i - z_0)^j$ by $\hat{\mathbf{\Theta}}_k(Z_i)$.

Tuning constant estimation in multistage problems also may need adjustment. For example, using kernels with bandwidth h_k at stage k, for odd-powered polynomials the bias at stage 1 is of course of the order h_1^{p+1}, whereas at stage 2 it is $c_1(z_0)h_1^{p+1} + c_2(z_0)h_2^{p+1}$. Standard EBBS can be used to estimate h_1 at stage 1, whereas in general estimating h_2 requires a two-dimensional EBBS. But in both the variance function problem and nonparametric calibration, the effect on $\mathbf{\Theta}_2$ due to estimating $\mathbf{\Theta}_1$ is nil asymptotically, and standard EBBS can be used at each stage without modification.

In general problems, via backfitting one can use different weight functions and tuning constants to estimate each component of $\mathbf{\Theta}(z)$. For example, one might iterate between solving the two equations (with estimated tuning constants)

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0) \\ \times \psi_1 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \hat{\mathbf{\Theta}}_2(Z_i) \right\} \mathbf{G}_p^t(Z_i - z_0)$$

and

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0) \\ \times \psi_2 \left\{ \mathbf{Y}_i, \hat{\mathbf{\Theta}}_1(Z_i), \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \mathbf{G}_p^t(Z_i - z_0).$$

This is the procedure that we used in the lung cancer mortality example.

We conjecture that the asymptotic variance of these backfitted estimates can be estimated consistently by applying the sandwich formula to the equations

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0) \\ \times \psi_1 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \\ \times \mathbf{G}_p^t(Z_i - z_0)$$

and

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0) \\ \times \psi_2 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \\ \times \mathbf{G}_p^t(Z_i - z_0).$$

This idea can be shown to work in the case of robust estimation of a mean and variance function, as in the lung cancer mortality example.

6. DISCUSSION

We have extended estimating equation theory to cases where the parameter vector $\mathbf{\Theta}$ is not constant but rather depends on a covariate Z. The basic idea is to solve the estimating equation locally at each value of z using weights that for the ith case decrease with the distance between z and the observed Z_i. The weights depend on a tuning parameter; for example, a bandwidth h. A suitable value of h can be found by minimizing an estimate of the MSE. The latter if found by estimating variance using the "sandwich formula" (or more efficient modifications described earlier) and estimating bias empirically (as in Ruppert 1997).

We have applied this methodology to nonparametric calibration in nutritional studies, robust modeling of lung cancer mortality rates, and overdispersion. We have focused on local weighted polynomials. Regression splines could also be used in this context and appear to have considerable promise. Given a set of knots $(\xi_1, \dots, \xi_p)$, a regression cubic spline has the form

$$\begin{aligned}\Theta(z, \mathbf{b}_0, \dots, \mathbf{b}_{p+3}) \\ = \mathbf{b}_0 + \mathbf{b}_1 z + \mathbf{b}_2 z^2 + \mathbf{b}_3 z^3 + \sum_{j=1}^{p} \mathbf{b}_{j+3}(z - \xi_j)_+^3,\end{aligned}$$

where $v_+ = v$ if $v > 0$ and equals 0 otherwise. If regression splines are used, then (2) becomes

$$0 = \sum_{i=1}^{n} \psi\{\mathbf{Y}_i, \Theta(Z_i, \mathbf{b}_0, \dots, \mathbf{b}_{p+3})\} \mathbf{G}_{p,s}(Z_i),$$

where $\mathbf{G}_{p,s}^t(z) = (1, z, z^2, z^3, (z-\xi_1)_+^3, \dots, (z-\xi_p)_+^3)$. The interesting issue here is the selection of the knots, a problem of considerable interest in the broad context and one on which we are currently working for estimating functions. The regression splines outlined earlier may have an advantage, because the knots can be chosen on a componentwise basis. An alternative to knot selection would be to penalize the knot coefficients, as Eilers and Marx (1996) and Ruppert and Carroll (1997) have suggested for nonparametric regression.

The associate editor has noted that the estimating equation (2) is implicitly adapting to the component of $\Theta(z)$ that has the least amount of smoothness. In principle, one could allow different bandwidths for each component, or even different orders of the local polynomial, and the sandwich variance estimator would still apply. Also, in principle, the EBBS methodology can be used to estimate many different bandwidths. It is not at all clear to us, however, how to decide which components of $\Theta(z)$ are more or less smooth.

Finally, a referee has noted that local polynomial methods need not be range preserving. For example, consider the case where the Ys are all positive and thus the regression function of Y on Z is necessarily positive. Even in ordinary nonparametric local linear kernel estimation, the fitted regression function need not be positive, whereas for local averages the fitted function will be positive. In many cases, appropriate reformulation of the model will preserve ranges. For example, consider binary regression. If one runs an ordinary local linear regression of Y on Z ignoring the binary nature of Y, then it may happen that fitted probabilities do not fall in the unit interval. But if the binary regression is based on the likelihood score (3) where μ is the logistic function (i.e., local logistic regression as in Fan et al. 1995), then the fitted probabilities will necessarily fall in the unit interval. Similarly, for positive Ys, one could use the local model where the log of the mean function is a polynomial.

APPENDIX: ASYMPTOTICS

A.1 Bias and Variance for Local Polynomial Estimation

Here we give a brief derivation of bias and variance formulas for local polynomial estimation of order p in the interior of the support of Z. The methods use to derive the calculations roughly parallel those of Fan et al. (1995) and Ruppert and Wand (1994). The regularity conditions necessary include the smoothness conditions on $\Theta(z)$ and $f_Z(\cdot)$ of Fan et al. (1995), with the smoothness conditions on $\psi(\cdot)$ guaranteeing that it is at least twice continuously differentiable and the regularity conditions from estimating function theory assuring that a consistent sequence of solutions to (2) exists. A useful simplification is to let the unknown parameters be $\mathbf{a}_j = h^j \Theta^{(j)}(z_0)/j!$ (see the appendix of Fan et al. 1995).

For any $p \times q$ matrix $\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_\ell)^t$, where $\mathbf{c}_j$ is a $q \times 1$ vector, define $\operatorname{vec}(\mathbf{C}) = (\mathbf{c}_1^t, \dots, \mathbf{c}_\ell^t)^t$. Define $\mu_K(r) = \int z^r K(z)\,dz$ and $\gamma_K(r) = \int z^r K^2(z)\,dz$. Assume that K is symmetric about 0, so that $\mu_K(r) = \gamma_K(r) = 0$ if r is odd.

Let $\mathbf{C}(z_0) = E[\psi\{\mathbf{Y}, \Theta(z_0)\}\psi^t\{\mathbf{Y}, \Theta(z_0)\}|Z = z_0]$ and $\mathbf{B}(z_0) = E[\chi\{\mathbf{Y}, \Theta(z_0)\}|Z = z_0]$, where $\chi(\mathbf{Y}, \mathbf{v}) = (\partial/\partial\mathbf{v}^t)\,\psi(\mathbf{Y}, \mathbf{v})$. Define

$$\begin{aligned}\mathcal{L}_n(\mathbf{a}_0, \dots, \mathbf{a}_p) \\ = n^{-1} \sum_{i=1}^{n} K_h(Z_i - z_0) \\ \times \operatorname{vec}\left[\mathbf{G}_{p,h}(Z_i - z_0) \otimes \psi^t\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j / h^j\right\}\right],\end{aligned}$$

where $\mathbf{G}_{p,h}(v) = (1, v/h, v^2/h^2, \dots, v^p/h^p)^t$. We are solving $0 = \mathcal{L}_n(\hat{\mathbf{a}}_0, \dots, \hat{\mathbf{a}}_p)$, with $\hat{\mathbf{a}}_j = h^j \hat{\Theta}^{(j)}(z_0)/j!$ By a Taylor series expansion, we find that the estimates are asymptotically equivalent to

$$\begin{pmatrix} \hat{\mathbf{a}}_0 - \mathbf{a}_0 \\ \vdots \\ \hat{\mathbf{a}}_p - \mathbf{a}_p \end{pmatrix} \approx -\{\mathbf{B}_*(z_0)\}^{-1} \mathcal{L}_n(\mathbf{a}_0, \dots, \mathbf{a}_p), \qquad \text{(A.1)}$$

where

$$\mathbf{B}_*(z_0) = \frac{\partial}{\partial(\mathbf{a}_0^t, \dots, \mathbf{a}_p^t)} \mathcal{L}_n(\mathbf{a}_0, \dots, \mathbf{a}_p).$$

It is helpful to keep in mind the following aspect:

$$\begin{aligned}\mathcal{L}_n(\mathbf{a}_0, \dots, \mathbf{a}_p) \\ = n^{-1} \sum_{i=1}^{n} K_h(Z_i - z_0) \\ \times \begin{bmatrix} \psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j / h^j\right\} \\ \{(Z_i - z_0)/h\}\psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j / h^j\right\} \\ \vdots \\ \{(Z_i - z_0)/h\}^p\psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j / h^j\right\} \end{bmatrix}.\end{aligned} \qquad \text{(A.2)}$$

Also note that the as are vectors, of the same length as Θ and ψ. The calculations are easier to follow if this expanded form is used.

It is easily seen that

$$\mathbf{B}_*(z_0) \xrightarrow{P} f_Z(z_0)\{\mathbf{D}_p(\mu) \otimes \mathbf{B}(z_0)\}, \tag{A.3}$$

where $\mathbf{D}_p(\mu)$ is the $(p+1) \times (p+1)$ matrix with (j,k)th element $\mu_K(j+k-2)$. It is also easily shown that

$$\operatorname{cov}\{\mathcal{L}_n(\mathbf{a}_0, \ldots \mathbf{a}_p)\} \sim (nh)^{-1} f_Z(z_0)\{\mathbf{D}_p(\gamma) \otimes \mathbf{C}(z_0)\},$$

where $\mathbf{D}_p(\gamma)$ is the $(p+1) \times (p+1)$ matrix with (j,k)th element $\gamma_K(j+k-2)$.

Finally, note that because $E[\psi\{\mathbf{Y}, \boldsymbol{\Theta}(Z)\}|Z] = 0$,

$$\begin{aligned}
&E\mathcal{L}_n(\mathbf{a}_0, \ldots, \mathbf{a}_p) \\
&\quad = -\int K_h(z-z_0) f_{Y|Z}(\mathbf{y}|z) f_Z(z) \\
&\qquad \times \operatorname{vec}\Bigg(\mathbf{G}_{p,h}(z-z_0) \otimes \Bigg[\psi\{\mathbf{y}, \boldsymbol{\Theta}(z)\} \\
&\qquad\qquad - \psi\left\{\mathbf{y}, \sum_{j=0}^{p} \mathbf{a}_j (z-z_0)^j / h^j\right\}\Bigg]^t\Bigg)\, d\mathbf{y}\, dz \\
&\quad \approx -\int K_h(z-z_0) f_{Y|Z}(\mathbf{y}|z) f_Z(z) \\
&\qquad \times \operatorname{vec}\Bigg(\mathbf{G}_{p,h}(z-z_0) \otimes \Bigg[\chi\{\mathbf{y}, \boldsymbol{\Theta}(z)\} \\
&\qquad\qquad \times \left\{\boldsymbol{\Theta}(z) - \sum_{j=0}^{p} (z-z_0)^j \boldsymbol{\Theta}^{(j)}(z_0)/j!\right\}\Bigg]^t\Bigg)\, d\mathbf{y}\, dz.
\end{aligned}$$

But $\boldsymbol{\Theta}(z) - \sum_{j=0}^{p}(z-z_0)^j \boldsymbol{\Theta}^{(j)}(z_0)/j! = (z-z_0)^{p+1}\boldsymbol{\Theta}^{(p+1)}(z_0)/(p+1)! + (z-z_0)^{p+2}\boldsymbol{\Theta}^{(p+2)}(z_0)/(p+2)! + \mathcal{O}\{(z-z_0)^{p+3}\}$. Hence, to terms of order $\{1 + \mathcal{O}(h)\}$,

$$E\mathcal{L}_n(\mathbf{a}_0, \ldots, \mathbf{a}_p) \approx \mathbf{A}_{1h} + \mathbf{A}_{2h},$$

where

$$\begin{aligned}
&\mathbf{A}_{kh} \\
&\quad = -\frac{h^{p+k}}{(p+k)!} \int K(x) f_{Y|Z}(\mathbf{y}|z_0 + xh) f_Z(z_0 + xh) \\
&\qquad \times \operatorname{vec}(\mathbf{G}_{p,1}(x) \otimes [\chi\{\mathbf{y}, \boldsymbol{\Theta}(z_0 + xh)\}\boldsymbol{\Theta}^{(p+k)}(z_0) x^{p+k}]^t)\, d\mathbf{y}\, dx \\
&\quad = -\frac{h^{p+k}}{(p+k)!} \int K(x) f_Z(z_0 + xh) \\
&\qquad \times \operatorname{vec}[\mathbf{G}_{p,1}(x) \otimes \{B(z_0 + hx)\boldsymbol{\Theta}^{(p+k)}(z_0) x^{p+k}\}^t]\, dx.
\end{aligned}$$

Clearly,

$$\mathbf{A}_{2h} \approx \frac{-h^{p+2} f_Z(z_0)}{(p+2)!} \operatorname{vec}[\mathbf{D}_\mu(p+2) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+2)}(z_0)\}^t],$$

where $\mathbf{D}_\mu(L) = \{\mu_K(L), \mu_K(L+1), \ldots, \mu_K(L+p)\}^t$. If we define $\mathbf{Q}(z) = f_Z(z)\mathbf{B}(z)$ with first derivative $\mathbf{Q}^{(1)}(z)$, then it also follows that

$$\begin{aligned}
&\mathbf{A}_{1h} \\
&\quad \approx \frac{-h^{p+1}}{(p+1)!} \int K(x) \operatorname{vec}[x^{p+1}\mathbf{G}_{p,1}(x) \\
&\qquad\qquad \otimes \{\mathbf{Q}(z_0 + hx)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]\, dx \\
&\quad \approx -\frac{h^{p+1} f_Z(z_0)}{(p+1)!} \operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t] \\
&\qquad - \frac{h^{p+2}}{(p+1)!} \int K(x) \operatorname{vec}[x^{p+2}\mathbf{G}_{p,1}(x) \\
&\qquad\qquad \otimes \{\mathbf{Q}^{(1)}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]\, dx \\
&\quad \approx -\frac{h^{p+1}}{(p+1)!} \operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{f_Z(z_0)\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t] \\
&\qquad - \frac{h^{p+2}}{(p+1)!} \operatorname{vec}[\mathbf{D}_\mu(p+2) \otimes \{\mathbf{Q}^{(1)}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t].
\end{aligned}$$

Thus we have shown that asymptotically,

$$\begin{aligned}
&\operatorname{bias}(\hat{\mathbf{a}}_0^t, \hat{\mathbf{a}}_1^t, \ldots, \hat{\mathbf{a}}_0^t)^t \\
&\quad = h^{p+1}\{\mathbf{D}_p(\mu) \otimes \mathbf{B}(z_0)\}^{-1} \\
&\qquad \times \operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]/(p+1)! \\
&\qquad + h^{p+2}\{\mathbf{D}_p(\mu) \otimes \mathbf{B}(z_0)\}^{-1}\mathbf{s}(z_0) + \mathcal{O}(h^{p+3}),
\end{aligned} \tag{A.4}$$

where

$$\begin{aligned}
&\mathbf{s}(z_0) \\
&\quad = \operatorname{vec}\Bigg[\mathbf{D}_\mu(p+2) \\
&\qquad\qquad \otimes \left\{\frac{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+2)}(z_0)}{(p+2)!} + \frac{\mathbf{Q}^{(1)}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)}{f_Z(z_0)(p+1)!}\right\}^t\Bigg].
\end{aligned}$$

The variance is

$$\begin{aligned}
&\{nhf_Z(z_0)\}^{-1}\{\mathbf{D}_\mu(p) \otimes \mathbf{B}(z_0)\}^{-1} \\
&\qquad \times \{\mathbf{D}_\gamma(p) \otimes \mathbf{C}(z_0)\}\{\mathbf{D}_\mu(p) \otimes \mathbf{B}(z_0)\}^{-t}\{1 + o(1)\}.
\end{aligned}$$

The only thing left to show is that if p is even, then the bias is of order $\mathcal{O}(h^{p+2})$—that is, the first element in

$$\{\mathbf{D}_\mu(p) \otimes \mathbf{B}(z_0)\}^{-1}\operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]$$

equals 0, which is clearly the case because $\mu_K(r) = 0$ if r is odd. For z_0 on the boundary of the support of Z, the terms of order h^{p+1} dominate, and the bias is of that order.

It is useful for theoretical purposes to note that we have actually shown the following. Denote (A.2) by $\mathcal{F}(z_0)$, (A.3) by $\mathcal{T}(z_0)$ and (A.4) by $\mathcal{S}(z_0)$. Then we have shown that

$$\begin{pmatrix} \hat{\mathbf{a}}_0 - \mathbf{a}_0 \\ \vdots \\ \hat{\mathbf{a}}_p - \hat{\mathbf{a}}_p \end{pmatrix} \approx \mathcal{S}(z_0) - \{\mathcal{T}(z_0)\}^{-1}\mathcal{F}(z_0). \tag{A.5}$$

Remark. In the application of parametric estimating equations, unless the equations are linear in the parameter there is typically a bias of order n^{-1}, which, however, is negligible compared to the standard deviation. Similarly, there will be a bias of order $(nh)^{-1}$ here that stems from terms ignored in the linearizing approximation (A.1). Because h is chosen so that the *squared* bias from smoothing is of order $(nh)^{-1}$, bias terms of order $(nh)^{-1}$ are ignored here. (However, see Ruppert, Wand, Holst, and Hössjer 1995 for a method of correcting the order $(nh)^{-1}$ bias due to estimation of the mean when a variance function is estimated.)

A.2 The Sandwich Formula

Here we sketch a justification for the sandwich formula (6)–(7), using the notation established previously in this Appendix. We continue to work with the parameterization $(\mathrm{a}_0, \ldots, \mathrm{a}_p)$. Noting that $\mathbf{B}_*(z_0)$ in (A.1) equals $n^{-1}\mathrm{B}_n(z_0)$ in (7), it suffices to show that $n^{-1}\mathrm{C}_n(z_0)$ defined in (6) has limiting covariance matrix $(nh)^{-1} f_Z(z_0)\{\mathrm{D}_p(\gamma) \otimes \mathrm{C}(z_0)\}$, which is easily established. This completes the argument.

[Received July 1996. Revised August 1997.]

REFERENCES

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–36.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman and Hall.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth.

Diggle, P. J., Liang, K. Y., and Zeger, S. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties" (with discussion), *Statistical Science*, 11, 89–121.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

——— (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150.

Freedman, L. S., Carroll, R. J., and Wax, Y. (1991), "Estimating the Relationship Between Dietary Intake Obtained From a Food Frequency Questionnaire and True Average Intake," *American Journal of Epidemiology*, 134, 510–520.

Gozalo, P., and Linton, O. (1995), "Using Parametric Information in Nonparametric Regression," Working Paper 95-40, Brown University, Dept of Economics.

Hall, P., Marron, J. S., and Titterington, D. M. (1995), "On Partial Local Smoothing Rules for Curve Estimation," *Biometrika*, 82, 575–588.

Hastie, T. J., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the 5th Berkeley Symposium*, 1, 221–233.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Kauermann, G., and Tutz, G. (1997), "On Model Diagnostics and Bootstrapping in Varying Coefficient Models," unpublished manuscript.

Martin, L. J., Su, W., Jones, P. J., Lockwood, G. A., Tritchler, D. L., and Boyd, N. F. (1996), "Comparison of Energy Intakes Determined by Food Records and Doubly Labeled Water in Women Participating in a Dietary Intervention Trial," *American Journal of Clinical Nutrition*, 63, 483–490.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996), "A Semiparametric Transformation Approach to Estimating Usual Intake Distributions," *Journal of the American Statistical Association*, 91, 1440–1449.

Politis, D. N., and Romano, J. P. (1994), "Large Sample Confidence Regions Based on Subsamples Under Minimal Conditions," *The Annals of Statistics*, 22, 2031–2050.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error," *Statistics in Medicine*, 8, 1051–1070.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," submitted to *Journal of the American Statistical Association*, 92, 1049–1062.

Ruppert, D., and Carroll, R. J. (1997), "Penalized Regression Splines," manuscript. Available at hup://wwwone.corntll.edu/dav•dr/papers/index.html.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997), "Local Polynomial Variance Function Estimation," *Technometrics*, 39, 262–273.

Severini, T. A., and Staniswallis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Simpson, D. G., Guth, D., Zhou, H., and Carroll, R. J. (1996), "Interval Censoring and Marginal Analysis in Ordinal Regression," *Journal of Agricultural, Biological and Environmental Statistics*, 1, 354–376.

Staniswallis, J. G. (1989), "The Kernel Estimates of a Regression Function in Likelihood-Based Models," *Journal of the American Statistical Association*, 84, 276–283.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Weisberg, S., and Welsh, A. H. (1994), "Estimating the Missing Link Function," *The Annals of Statistics*, 22, 1674–1700.

Welsh, A. H. (1996), "Robust Estimation of Smooth Regression and Spread Functions and Their Derivatives," *Statistica Sinica*, 6, 347–366.

Welsh, A. H., Carroll, R. J., and Ruppert, D. (1994), "Fitting Heteroscedastic Regression Models," *Journal of the American Statistical Association*, 89, 100–116.

Aust. N. Z. J. Statist. **42**(2), 2000, 205–223

SPATIALLY-ADAPTIVE PENALTIES FOR SPLINE FITTING

David Ruppert[1]* and Raymond J. Carroll[2]

Cornell University and Texas A & M University

Summary

The paper studies spline fitting with a roughness penalty that adapts to spatial heterogeneity in the regression function. The estimates are pth degree piecewise polynomials with $p-1$ continuous derivatives. A large and fixed number of knots is used and smoothing is achieved by putting a quadratic penalty on the jumps of the pth derivative at the knots. To be spatially adaptive, the logarithm of the penalty is itself a linear spline but with relatively few knots and with values at the knots chosen to minimize the generalized cross validation (GCV) criterion. This locally-adaptive spline estimator is compared with other spline estimators in the literature such as cubic smoothing splines and knot-selection techniques for least squares regression. Our estimator can be interpreted as an empirical Bayes estimate for a prior allowing spatial heterogeneity. In cases of spatially heterogeneous regression functions, empirical Bayes confidence intervals using this prior achieve better pointwise coverage probabilities than confidence intervals based on a global-penalty parameter. The method is developed first for univariate models and then extended to additive models.

Key words: additive models; Bayesian inference; confidence intervals; hierarchical Bayesian model; regression splines.

1. Introduction

In this paper we study a variant of smoothing splines that we call penalized splines or, following Eilers & Marx (1996), p-splines. What is new is that we allow the penalty to vary spatially to adapt to possible spatial heterogeneity in the regression function. This spatial adaptivity can result in improved precision and also better confidence bounds on the regression function.

Suppose that we have data (x_i, y_i) where the x_i are univariate,

$$y_i = m(x_i) + \epsilon_i,$$

m is a smooth function equal to the conditional mean of y_i given x_i, and the ϵ_i are independent mean zero errors with a constant variance σ^2. The extension to additive models is

Received January 1999; revised July 1999; accepted September 1999.

* Author to whom correspondence should be addressed.

[1] School of Operations Research & Industrial Engineering, Cornell University, Ithaca, New York 14853-3801, USA. e-mail: davidr@orie.cornell.edu

[2] Statistics, Nutrition and Toxicology, Texas A & M University, College Station, TX 77843-3143, USA.

Acknowledgments. Ruppert's research was supported by an NSF grant. Carroll's research was supported by a grant from the National Cancer Institute and by the Texas A & M Center for Environmental and Rural Health through a grant from the National Institute of Environmental Health Sciences. Carroll's work was partially completed during visits to Sonderforschungsbereich 373 at the Humboldt Universität zu Berlin. Matt Wand kindly showed us his manuscript on a comparison of regression spline smoothing methods. We thank George Casella, Rob Kass, and Marty Wells for references to the empirical Bayes literature. This paper has benefited substantially from the constructive comments of two referees and an associate editor.

straightforward and is mentioned in Section 7. To estimate m we use a regression spline model

$$m(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - \kappa_k)_+^p , \tag{1}$$

where $p \geq 1$ is an integer, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_{p+K})^{\mathrm{T}}$ is a vector of regression coefficients, $(u)_+^p = u^p I(u \geq 0)$, and $\kappa_1 < \cdots < \kappa_K$ are fixed knots.

When fitting model (1) to noisy data, care is needed to prevent overfitting because that causes a rough fit tending interpolate the data. The traditional methods of obtaining a smooth spline estimate are knot selection, e.g. Friedman & Silverman (1989), Friedman (1991) and Stone *et al.* (1997), and smoothing splines (Eubank, 1988; Wahba, 1990). With the first set of methods, the knots are selected from a set of candidate knots by a technique similar to stepwise regression and then, given the selected knots, the coefficients are estimated by ordinary least squares. Smoothing splines have a knot at each unique value of x and control overfitting by using least squares estimation with a roughness penalty. The penalty is on the integral of the square of a specified derivative, usually the second. The penalized least squares estimator has the form of a ridge regression estimate. Luo & Wahba (1997) proposed a hybrid between knot selection and smoothing splines — they follow knot selection by penalized least squares estimation. Recently, there have appeared Bayesian methods that yield weighted averages of (essentially) least squares estimates. The averages are over the sets of possible knots, with a set's weight given by the posterior probability that the set is the 'true set' (Smith & Kohn, 1996; Denison, Mallick & Smith, 1998).

In this paper we use a penalty approach similar to smoothing splines but with fewer knots. We allow K in (1) to be large but typically far less than n. Unlike knot-selection techniques, we retain all candidate knots. As with smoothing splines, a roughness penalty is placed on $\{\beta_{p+k}\}_{k=1}^{K}$ which is the set of jumps in the pth derivative of $m(x;\boldsymbol{\beta})$. We could view this as a penalty on the $(p+1)$th derivative of $m(x;\boldsymbol{\beta})$ where that derivative is a generalized function. Eilers & Marx (1996) developed this method of p-splines, though they have traced the original idea to O'Sullivan (1986, 1988). Eilers and Marx use equally-spaced knots and they use the B-spline basis, whereas we use sample quantiles of x as knots and the truncated power function as basis. Also, they consider a somewhat more general class of penalties than we need here.

Because smoothness is controlled by a roughness penalty, once a certain minimum number of knots is reached, further increases in the number of knots cause little noticeable change in the fit given by a p-spline. In applications we have seen to actual data, use of between 5 and 40 knots works well. In some difficult problems used in simulation studies, more than 40 knots are needed; see Section 5. However, in the example of that section, use of more than 80 knots is not better than use of 80 knots, and 80 knots is still far less than the number of knots used by a smoothing spline which has the sample size of 400. We recommend letting κ_k be the $k/(K+1)$th sample quantile of the x_i, which we call 'equally-spaced sample quantiles'.

We treat the number of knots as a user-specified tuning parameter. Although the exact choice of the number of knots is often not crucial, for each dataset there exists a particular minimum number of knots needed to obtain a good fit. Therefore, some users may want a completely automatic algorithm, and we propose such a procedure in Section 3.

We define $\hat{\boldsymbol{\beta}}(\alpha)$ to be the minimizer of

$$\sum_{i=1}^{n}\big(y_i - m(x;\boldsymbol{\beta})\big)^2 + \sum_{k=1}^{K}\alpha(\kappa_k)\beta_{p+k}^2\,, \tag{2}$$

where $\alpha(\cdot)$ is a penalty function. Eilers & Marx (1996) use a constant α; that is, their α is the same for all knots, though its value depends on the data. A constant penalty weight is also used in the smoothing spline literature. We call a spline fit with a constant value of α a global-penalty spline. Local penalty splines are those with α varying across the knots.

A single penalty weight is not suitable for functions that rapidly oscillate in some regions and are rather smooth in other regions. In a simulation study, Wand (2000) shows that splines having a single smoothing parameter are inferior in terms of mean squared error (MSE). In that study, p-splines do not compete with knot-selection methods for regression spline fitting for regression functions with significant spatial inhomogeneity.

Another problem with having only a single smoothing parameter concerns inference. Smoothing splines and p-splines are both Bayes estimates for particular priors. A single smoothing parameter corresponds to a spatially homogeneous prior. For example, for a p-spline, the prior is that the $\{\beta_{p+k}\}_{k=1}^{K}$ are independent and identically distributed (iid) $\mathrm{N}(0,\tau^2)$ where τ^2 equals σ^2/α; see Section 4. The polynomial coefficients, $\beta_0,\dots,\beta_p$, are given an improper prior, uniform on $p+1$ dimensional space. Such priors on $\{\beta_{p+k}\}_{k=1}^{K}$ are not appropriate for spatially heterogeneous m. Consider confidence intervals based on the posterior variance of $m(\cdot)$ as in Wahba (1983) and Nychka (1988). As Nychka shows, the resulting confidence bands have good average (over x) coverage probabilities but do not have accurate pointwise coverage probabilities in areas of high oscillations of m or other ‘features’.

Section 2 describes our method. Section 3 presents a fully automatic estimator with all tuning parameters selected by the data. Section 4 discusses Bayesian inference and Section 5 gives Monte Carlo simulations. Section 6 contains an example using data from an experiment where atmospheric mercury is monitored by LIDAR. In Section 7 we extend our method to additive models, and we give final discussion and conclusions in Section 8.

2. A local-penalty method

Here is a simple approach to spatially varying α. Choose another set of the knots, $\{\kappa_k^*\}_{k=1}^{M}$, where M is smaller than K and such that $\{\kappa_1^* = \kappa_1 < \cdots < \kappa_M^* = \kappa_K\}$. The penalty at one of these ‘subknots’ (or ‘α-knots’), say κ_k^*, is controlled by a parameter α_k^*. The penalties at the original knots, $\{\kappa_k\}_{k=1}^{K}$, are determined by linear interpolation, say, of the penalties at the $\{\kappa_k^*\}_{k=1}^{M}$. The interpolation is on the log-penalty scale to ensure positivity of the penalties. Thus, we have a penalty $\alpha(\kappa_k)$ at each κ_k but these penalties depend only upon $\boldsymbol{\alpha}^* = (\alpha_1^*,\dots,\alpha_M^*)^{\mathrm{T}}$. Therefore, $(\alpha(\kappa_1),\dots,\alpha(\kappa_K))$ is a function of $\boldsymbol{\alpha}^*$. This function is not derived explicitly but rather is computed by using a linear interpolation algorithm; we used MATLAB’s built-in linear interpolator. One could, of course, use other interpolation methods, e.g. cubic interpolation. If linear interpolation is used, then $\log(\alpha(\cdot))$ is a linear spline with knots at $\{\kappa_k^*\}_{k=1}^{M}$.

Let $\boldsymbol{Y} = (y_1,\dots,y_n)^{\mathrm{T}}$ and $\boldsymbol{X}$ be the ‘design matrix’ for the regression spline so that the ith row of $\boldsymbol{X}$ is

$$\boldsymbol{X}_i = [\,1 \quad x_i \quad \cdots \quad x_i^p \qquad (x_i-\kappa_1)_+^p \quad \cdots \quad (x_i-\kappa_K)_+^p\,]\,.$$

Let $\boldsymbol{D}(\boldsymbol{\alpha}^*)$ be a diagonal matrix whose first $(1+p)$ diagonal elements are 0 and whose remaining diagonal elements are $\alpha(\kappa_1), \dots, \alpha(\kappa_K)$, which depend only on $\boldsymbol{\alpha}^*$. Then standard calculations show that $\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}^*)$ is given by

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}^*) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{D}(\boldsymbol{\alpha}^*)\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

This is a ridge regression estimator that shrinks the regression spline towards the least squares fit to a pth degree polynomial model (Hastie & Tibshirani, 1990 Section 9.3.6).

The smoothing parameter $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_M^*)$ can be determined by minimizing the generalized cross validation statistic

$$\mathrm{GCV}(\boldsymbol{\alpha}^*) = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}^*)\|^2}{(1 - \mathrm{df}(\boldsymbol{\alpha}^*)/n)^2}\,.$$

Here

$$\mathrm{df}(\boldsymbol{\alpha}^*) = \mathrm{tr}\left\{\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{D}(\boldsymbol{\alpha}^*)\right)^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\right\} \qquad (3)$$

is the degrees of freedom of the smoother which is defined to be the trace of the smoother matrix (Hastie & Tibshirani, 1990 Section 3.5). The right-hand side of (3) is suitable for computing because it is the trace of a matrix whose dimension is only $(1+p+K)^2$.

A search over an M-dimensional grid is not recommended because of computation cost. Rather, we recommend that one start with $\alpha_1^*, \dots, \alpha_M^*$ each equal to the best global value of α chosen by minimizing GCV. Then each α_k^* is varied, with the others fixed, over a one-dimensional grid centred at the current value of α_k^*. On each such step, α_k^* is replaced by the α-value minimizing GCV on this grid. This minimizing of GCV over each α_k^* is repeated a total of N_{iter} times. Although minimizing GCV over the α_k^*s one at a time in this manner does not guarantee finding the global minimum of GCV over $\alpha_1^*, \dots, \alpha_M^*$, our simulations show that this procedure is effective in selecting a satisfactory amount of local smoothing. The minimum GCV global α is a reasonably good starting value for the smoothing parameters and each step of our algorithm improves about this start in the sense of lowering GCV. Each α_k^* controls the penalty only over a small range of x, so the optimal value of one α_k^* should depend only slightly upon the other α_k^*s. We believe this is the reason that our one-at-a-time search strategy works effectively.

3. A completely automatic algorithm

The local-penalty method has three tuning parameters: the number of knots K, the number of subknots M, and the number of iterations N_{iter}. The exact values of the tuning parameters are not crucial provided they are within certain acceptable ranges. The crucial parameter is $\boldsymbol{\alpha}^*$ which is selected by GCV. However, users may want a completely automatic algorithm which requires no user-specified parameters, and which attempts to ensure that the tuning parameters are within acceptable ranges. It must choose tuning parameters that are large enough to obtain a good fit but not so large that the computation time is excessive. (Overfitting is not a concern because it is controlled by $\boldsymbol{\alpha}^*$.)

In this section, we propose such a procedure based on the following principle: as the complexity of m increases each of K, M, and N_{iter} should increase. The algorithm uses a sequence of values of (K, M, N_{iter}) where each parameter is non-decreasing in the sequence.

The algorithm stops when there is no appreciable decrease in GCV between two successive values of (K, M, N_{iter}). Monte Carlo experimentation discussed in Section 5.2 shows that the values of N_{iter} and M have relatively little effect on the fit, at least within the ranges studied. However, it seems reasonable to increase N_{iter} and M slightly with K. On the other hand, for a given K, computation time is roughly proportional to $M \times N_{iter}$, so we avoid $N_{iter} > 2$ and $M > 6$.

Specifically, we use this sequence of values of (K, M, N_{iter}): $(10,2,1)$, $(20,3,2)$, $(40,4,2)$, $(80,6,2)$, $(120,6,2)$. We compare GCV, minimized over $\boldsymbol{\alpha}^*$, using $(10,2,1)$ and $(20,3,2)$. If the value of GCV for $(20,3,2)$ is more than a constant C times the GCV value of $(10,2,1)$ we conclude that further increases in the tuning parameters will not appreciably decrease GCV. (In the simulations we used $C = 0.98$ and that choice worked well.) Therefore, we stop and use $(20,3,2)$ as the final value of the three tuning parameters. Otherwise, we fit using $(40,4,2)$ and compare its GCV value to that of $(20,3,2)$. If the value of GCV for $(40,4,2)$ is more than C times the GCV value of $(20,3,2)$, we stop and use $(40,4,2)$ as the final value of the three tuning parameters. Otherwise, we continue in this manner, comparing $(40,4,2)$ to $(80,6,2)$, etc. If very complex m were contemplated, then one could, of course, continue using increasingly larger values of the tuning parameters.

Note that the final tuning parameter vector is one of $(20,3,2)$, $(40,4,2)$, $(80,6,2)$ and $(120,6,2)$. The vector $(10,2,1)$ is used only to check if one can stop at $(20,3,2)$.

4. Bayesian inference

The p-spline method has an interpretation as a Bayesian estimator in a linear model. See Lindley & Smith (1972) and Box & Tiao (1973) for a discussion of Bayesian linear models. Suppose that $\epsilon_1, \dots, \epsilon_n$ are iid $\mathrm{N}(0, \sigma^2)$ and that the prior on $\boldsymbol{\beta}$ is $\mathbf{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\alpha}^*))$, where $\boldsymbol{\Sigma}(\boldsymbol{\alpha}^*)$ is a covariance matrix depending on $\boldsymbol{\alpha}^*$. Here $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean and covariance matrix $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For now, assume that σ^2 and $\boldsymbol{\alpha}^*$ are known. Then, up to an additive function of $\boldsymbol{Y}$ and $(\sigma^2, \boldsymbol{\alpha}^*)$, the posterior log density of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ is given by

$$-\tfrac{1}{2}\left\{\frac{1}{\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}(\boldsymbol{\alpha}^*)^{-1}\boldsymbol{\beta}\right\}. \tag{4}$$

The maximum *a posteriori* (MAP) estimator of $\boldsymbol{\beta}$, i.e. the mode of the posterior density, maximizes (4). Now let $\beta_0, \dots, \beta_p$ have improper uniform$(-\infty, \infty)$ priors and let $\{\beta_{p+k}\}_{k=1}^{K}$ be independent with β_{p+k} having an $\mathrm{N}(0, \sigma^2/\alpha_k)$ distribution. Then

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\alpha}^*) = \sigma^2 \operatorname{diag}(0, \dots, 0, \alpha_1, \dots, \alpha_K) \tag{5}$$

and the MAP estimator minimizes (2). (More precisely, we let $\beta_0, \dots, \beta_p$ have an $\mathrm{N}(0, \sigma_1^2)$ prior and then (5) holds in the limit as $\sigma_1 \to \infty$.)

The α_k are not known in practice. Empirical Bayes methods replace unknown 'hyperparameters' in a prior by estimates and then treat these hyperparameters as fixed. For example, if $\{\alpha_k^*\}_{k=1}^{M}$ are estimated by GCV and then considered fixed, one is using empirical Bayes inference. Standard calculations show that when $\boldsymbol{\alpha}^*$ and σ^2 are known, the posterior distribution of $\boldsymbol{\beta}$ is

$$\mathbf{N}\big(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}^*),\ \sigma^2\{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{\Sigma}(\boldsymbol{\alpha}^*)\}^{-1}\big). \tag{6}$$

Also, the posterior distribution of $\boldsymbol{m} = \{m(x_1), \dots, m(x_n)\}^{\mathsf{T}}$ is

$$\mathbf{N}\big(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}^*),\ \sigma^2\boldsymbol{X}\{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{\Sigma}(\boldsymbol{\alpha}^*)\}^{-1}\boldsymbol{X}^{\mathsf{T}}\big). \tag{7}$$

An approximate Bayes posterior replaces $\boldsymbol{\alpha}^*$ and σ^2 in (6) and (7) by estimates. If $\boldsymbol{\alpha}^*$ has been estimated by GCV, one need only estimate σ^2 by $\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}^*)\|^2/(n - \mathrm{df}(\hat{\boldsymbol{\alpha}}^*))$ where $\mathrm{df}(\boldsymbol{\alpha}^*)$ is defined by (6). Then the approximate posterior distribution for $\boldsymbol{m}$ is

$$\mathrm{N}\big(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}^*),\ \hat{\sigma}^2\boldsymbol{X}\{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}^*)\}^{-1}\boldsymbol{X}^{\mathsf{T}}\big). \tag{8}$$

The approximate Bayes $100(1-\gamma)\%$ confidence interval for $m(x_i)$ is

$$\hat{m}(x_i) \pm \Phi^{-1}(1 - \tfrac{1}{2}\gamma)\,\mathrm{se}(\hat{m}(x_i))$$

where $\hat{m}(x_i) = \boldsymbol{X}_i\hat{\beta}(\hat{\alpha})$ is the ith element of the posterior mean in (8), $\mathrm{se}(\hat{m}(x_i))$ is the square root of the ith diagonal of the posterior covariance matrix in (8), and Φ is the standard normal cumulative distribution function.

Hastie & Tibshirani (1990) make an interesting and cogent argument against using confidence bands about the regression line; instead they suggest plotting a sample of curves from the posterior distribution. By following their recommendation, one gets a much better sense of what the true regression curve might look like. Regardless of whether one samples from the posterior or looks at confidence intervals (or does both!), a posterior that reflects any spatial heterogeneity that may exist gives a more accurate picture of the true function.

These approximate Bayesian methods estimate hyperparameters but then pretend that the hyperparameters were known so they do not account for extra variability in the posterior distribution caused by the estimation of hyperparameters in the prior; for discussion see, e.g., Morris (1983), Laird & Louis (1987), Kass & Steffey (1989) or Carlin & Louis (1996). Everything else held constant, the underestimation of posterior variance should become worse as M increases, because each α_m^* is then determined by fewer data and is therefore more variable. As Nychka (1988) has shown empirically, this underestimation does not appear to be a problem for a global penalty which has only one hyperparameter. However, the local penalty has M hyperparameters. We have found for local-penalty splines that the pointwise approximate posterior variance of $\hat{m}$ is too small in the sense that it noticeably underestimates the frequentist's MSE.

A simple correction to this problem is to multiply the pointwise posterior variances of the local-penalty $\hat{m}$ from (8) by a constant so that the average pointwise posterior variance of $\hat{m}$ is the same for the global- and local-penalty estimators. The reasoning behind this correction is as follows. As stated above, the global penalty approximate posterior variance from (8) is nearly equal to the frequentist's MSE on average. The local-penalty estimate has an MSE that varies spatially but should be close, on average, to the MSE of the global-penalty estimate and therefore also close, on average, to the estimated posterior variance of the global-penalty estimator. We found that this adjustment is effective in guaranteeing coverage probabilities at least as large as nominal, though in extreme cases of spatial heterogeneity the adjustment can be conservative; see Section 5.4. The reason for the latter is that in cases of severe spatial heterogeneity the local penalty MSE is less, on average, than the MSE of the global-penalty estimate. Then, there is an over-correction and the local penalty MSE is overestimated by this adjusted posterior variance. The result is that confidence intervals constructed with this adjustment should be conservative. The empirical evidence in Section 5 supports this conjecture. In that section, we refer to these adjusted intervals as local-penalty, conservative.

Another correction would be to use a fully Bayesian hierarchical model, where the hyperparameters are given a prior. Deely & Lindley (1981) first considered such Bayesian empirical

Bayes methods. An exact Bayesian analysis for p-splines would seem to require Gibbs sampling or other computationally intensive techniques. Given the number of parameters involved and the model complexity, an accurate Monte Carlo Markov chain analysis (MCMC) could take days or weeks of computer time. In contrast, our algorithm with the adjustment above can be computed in a matter of seconds.

There are intermediate positions between the quick ad hoc conservative adjustment just proposed and an exact fully Bayesian analysis. One that we now describe is an approximate fully Bayesian method that uses a small bootstrap experiment and a delta-method correction adopted from Kass & Steffey's (1989) 'first order approximation'. (Kass and Steffey considered conditionally independent hierarchical models, which are also called empirical Bayes models, but their ideas apply directly to more general hierarchical Bayes models.)

The Kass and Steffey approximation is applied to p-splines as follows. Let $m_i = m(x_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. The posterior variance of m_i calculated from the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\alpha}^*)$ and by a standard identity is

$$\mathrm{var}(m_i) = \mathrm{E}\big(\mathrm{var}(m_i \mid \boldsymbol{\alpha}^*)\big) + \mathrm{var}\,\big(\mathrm{E}(m_i \mid \boldsymbol{\alpha}^*)\big).$$

$\mathrm{E}(\mathrm{var}(m_i \mid \boldsymbol{\alpha}^*))$ is well-approximated by the posterior variance of m_i when $\boldsymbol{\alpha}^*$ is treated as known and fixed at its posterior mode (Kass & Steffey, 1989). Thus, $\mathrm{var}(\mathrm{E}(m_i \mid \boldsymbol{\alpha}^*))$ is the extra variability in posterior distribution of m_i that the approximate posterior variance given by (8) does not account for. We estimate $\mathrm{var}(\mathrm{E}(m_i \mid \boldsymbol{\alpha}^*))$ by the following three steps and add this estimate to the posterior variance given by (8).

1. Use a parametric bootstrap to estimate $\mathrm{var}(\log(\hat{\boldsymbol{\alpha}}^*))$. (Here the log function is applied coordinate-wise to the vector $\boldsymbol{\alpha}^*$.)
2. Numerically differentiate $\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}^*)$ with respect to $\log(\boldsymbol{\alpha}^*)$ at $\boldsymbol{\alpha}^* = \hat{\boldsymbol{\alpha}}^*$. We use one-sided numerical derivatives with a step-length of 0.1.
3. Put the results from points 1 and 2 into the delta-method formula:

$$\mathrm{var}\,\big(\mathrm{E}(m_i \mid \boldsymbol{\alpha}^*)\big) \approx \Big(\frac{\partial \boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}^*)}{\partial \log(\boldsymbol{\alpha}^*)}\Big)^{\mathsf{T}} \mathrm{var}\,\big(\log(\hat{\boldsymbol{\alpha}}^*)\big)\Big(\frac{\partial \boldsymbol{X}\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}^*)}{\partial \log(\boldsymbol{\alpha}^*)}\Big). \qquad (9)$$

When (9) is added to the approximate posterior variance from (8), we call the corresponding confidence intervals 'local-penalty, corrected'. Since the correction, (9), is a relatively small portion of the corrected posterior variance, it need not be estimated by the bootstrap with as great a precision as when a variance is estimated entirely by a bootstrap. In our simulations, we used only 25 bootstrap samples in step 1.

In the simulations of the next section, the local-penalty, conservative intervals are close to the more computationally intensive local-penalty, corrected intervals. Since the latter have a theoretical justification, this closeness is some justification for the former.

5. Simulations

5.1. Mean squared error comparison

We performed a small Monte Carlo experiment using the 'spatial variability' scenario in Wand (2000) so that our results could be compared with his. The x s were equally spaced on

[0, 1], n was 400, and the ϵ_i were independent N$(0, 0.2^2)$. The regression function, whose spatial variability was controlled by a parameter j, was

$$m(x;\, j) = \sqrt{x(1-x)}\ \sin\left(\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right). \tag{10}$$

We used both $j = 3$ which gave low spatial variability and $j = 6$ which gave severe spatial variability; see panels (a) and (b) of Figure 1. We used both 40 and 80 knots. When we used 40 knots, $\{\kappa_{k(j)} : j = 1, 10, 20, 30, 40\}$ were the subknots used for the local penalty. For 80 knots, $\{\kappa_{k(j)} : j = 1, 20, 40, 60, 80\}$ were the subknots. In all cases, N_{iter} was 1 and quadratic splines were used. For each of the four combinations of j and K, we simulated 250 datasets and applied the global- and local-penalty function estimators to each. Figure 1 shows boxplots of

$$\log_{10}(\text{RMSE}) = \log_{10}\left(\frac{1}{n}\sum_{i=1}^{n}\bigl(\hat{m}(x_i) - m(x_i)\bigr)^2\right).$$

From the results in Figure 1 we can draw the following conclusions.

- Locally varying penalties are as effective as a global penalty when there is little spatial variability. There appears to be little or no cost in statistical efficiency when a local penalty is used but not needed.
- For severe spatial variability, the local-penalty approach is far superior to a global penalty.
- There is little difference between using 40 and 80 knots except in one important situation. If one uses a local penalty and $j = 6$, 80 knots is significantly better than 40 because 80 knots allows the spline to track the rapid oscillations on the left, but only if a local penalty is used.

Also, comparing the results in Figure 1 to the results in Wand (2000) for $j = 6$, the local-penalty approach is somewhat better than the Bayesian method of Smith & Kohn (1996) and the stepwise selection method of Stone *et al.* (1997). However, Wand's simulations used code provided by Smith that had 35 knots 'hard-wired' into it (Wand, personal communication). With more knots, the Smith and Kohn method could very well be competitive with the local-penalty method.

We have also looked at moderate spatial variability ($j = 4$ or 5). There the local-penalty estimator is better than the global-penalty estimator, and again the local-penalty estimator is as good as the Bayesian and stepwise methods studied by Wand.

To compare the local- and global-penalty splines with other smoothers besides those in Wand's (2000) study, we used one of the sampling situations in Luo & Wahba (1997 Case 6). The regression function there is

$$m(x) = \sin\bigl(2(4x-2)\bigr) + 2\exp\bigl(-16^2(x-0.5)^2\bigr).$$

We used the same values of σ^2 and n as Luo and Wahba ($n = 256$ and $\sigma = 0.3$) and used equally spaced xs on $[0, 1]$ as they did. Table 1 gives results for local- and global-penalty splines, and results of Luo and Wahba for their hybrid adaptive spline (HAS), smoothing splines (SS), the SureShrink of Donoho & Johnstone (1995), and MARS of Friedman (1991).

Denison *et al.* (1998) tested their Bayesian splines on the same example as Luo & Wahba (1997), but they used $n = 200$ instead of 256 and reported MSE values instead of medians of squared errors. They found MSE values of 0.0096 and 0.0087 for linear and quadratic splines,

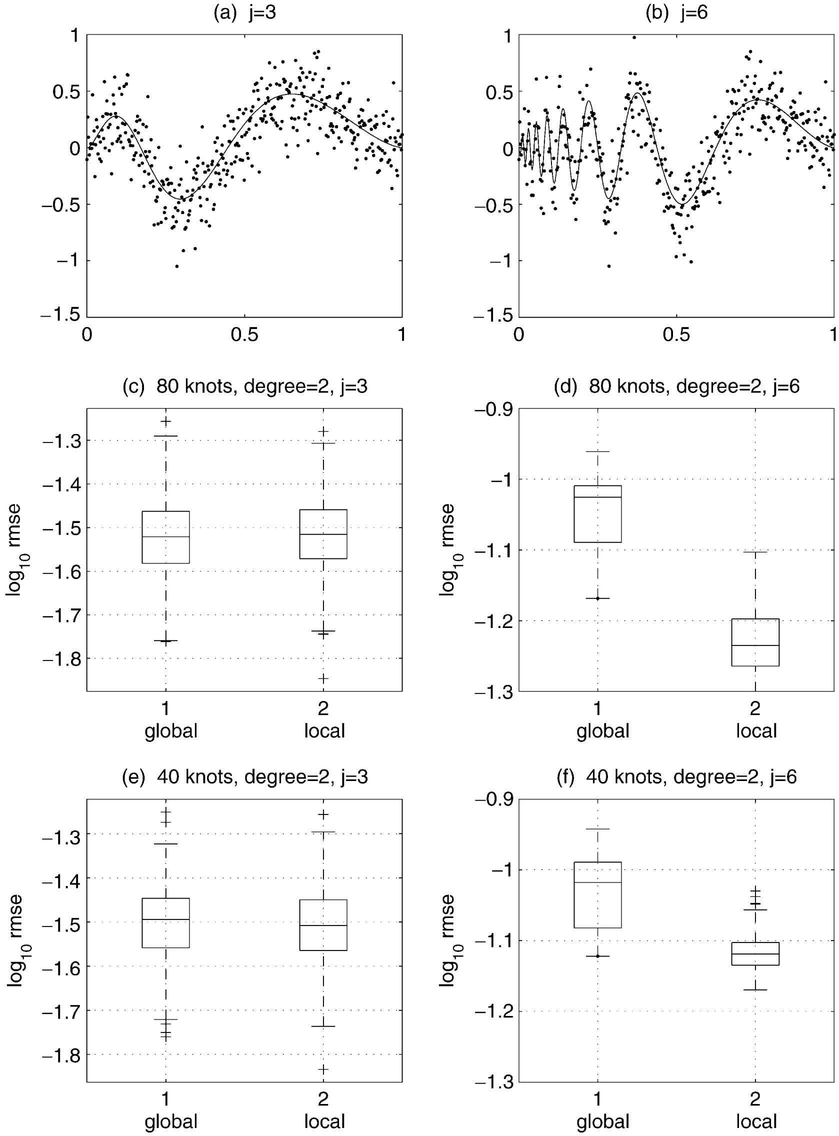


Figure 1. Comparison of global- and local-penalty parameters under low ($j = 3$) and severe ($j = 6$) spatial variability in the oscillations of the regression function: (a) the regression function (solid) and one sample (dots) when $j = 3$; (b) same as (a) but $j = 6$; (c) boxplots of $\log_{10}$(RMSE) for 250 simulated samples using global- and local-penalty parameters, 80 knots, quadratic splines, and $j = 3$; (d) same as (c) but $j = 6$; (e) same as (c) but 40 knots. (f) Same as (e) but $j = 6$.

TABLE 1

Median of squared errors (interquartile range of squared errors) for six smoothers. The results for HAS, SS, *SureShrink, and* MARS *are from Luo & Wahba (1997).*

HAS	SS	SureShrink	MARS	Local PS	Global PS
0.007 (0.006)	0.006 (0.003)	0.018 (0.004)	0.007 (0.004)	0.0053 (0.0035)	0.0061 (0.0029)

while for the same sampling situation we found MSE values of 0.0075 and 0.0083 for the local- and global-penalty quadratic splines.

We also tried cubic p-splines with both global and local penalties, but we found cubic p-splines somewhat inferior to quadratic p-splines.

5.2. Effects of the tuning parameters

We conducted a Monte Carlo experiment to learn further how the tuning parameters affect the accuracy of the local p-spline. The regression function (10) was used with j varying as a factor with levels 3, 4, 5 and 6. The sample size and values of x and σ were the same as in Section 5.1. There were three other factors: K with levels 20, 40, 80 and 120; M with levels 3, 4, 6 and 8; and N_{iter} with levels 1, 2 and 3. A full factorial design was used with two replications for a total of 384 runs.

The response was log(MSE). First, a quadratic response surface with two-way interactions was fitted to all four factors. Then, to look at the data from a slightly different perspective, we fitted quadratic response surfaces in the three tuning parameters with j fixed at each of its four levels. This second perspective was more illuminating. We found that for $j = 3$, 4, or 5, the tuning parameters had no appreciable effects on log(MSE). For $j = 6$, only the number of knots, K, had an effect on log(MSE). That effect was nonlinear — log(MSE) decreased rapidly as K increased up to about 80 but then log(MSE) levelled off.

In summary, for the scenario we simulated, of three tuning parameters only K has a detectable effect on log(MSE). It is important that K be at least a certain minimum value depending on the regression function, but after K is sufficiently large further increases in K do not affect accuracy.

5.3. The automatic algorithm

We tested the algorithm in Section 3 that chooses all tuning parameters automatically. As just mentioned, it is important that the number of knots, K, is large enough that all significant features of the regression function can be modelled. Thus, the main function of the automatic algorithm is to ensure that K is sufficiently large. As reported in Section 5.2 the number of subknots and the number of iterations did not appear to affect accuracy, but in our proposed algorithm we allowed them to increase slightly with K.

For each of $j = 3$ and 6 we used the algorithm on 250 datasets, with $n = 400$ and the standard deviation of the ϵ equal to 0.2 as before. Recall that the algorithm can choose as the final value of (K, M, N_{iter}) one of the vectors $(20, 3, 2)$, $(40, 4, 2)$, $(80, 6, 2)$ and $(120, 6, 2)$. With $j = 3$, the first vector was chosen 249 times and the second vector once. The tuning parameter vector $(20, 3, 2)$ gave MSE values quite similar to larger tuning parameter values, so stopping at $(20, 3, 2)$ was clearly appropriate. With $j = 6$ the fourth tuning parameter vector was chosen 247 times, while the third vector was chosen the remaining three times. As we saw in the last section, 80 knots was preferable here to a lesser number of

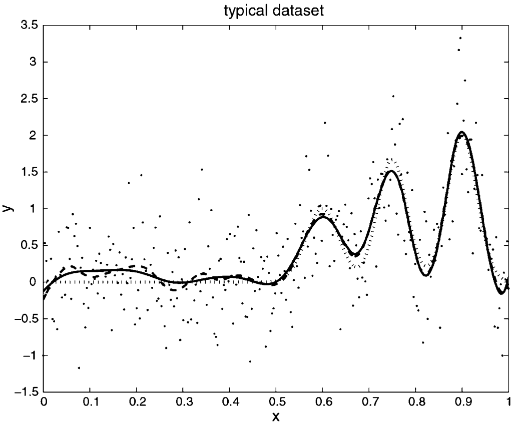


Figure 2. Typical data in the Bayesian inference study; local- and global-penalty splines with 95% confidence intervals based on the local spline (dots = data; dashed curve = global-penalty estimator; solid = local-penalty estimator; dotted = true function)

knots. However, use of 120 knots offerred no improvement over 80 (but was no worse either). Therefore, selection of either of the two largest possible values of the tuning parameters, as happened in all 250 trials, was the appropriate choice in this situation.

We conclude that the automatic algorithm can supply reasonable values of the tuning parameters when the user has little idea of how to choose them. The automatic algorithm is, of course, slower than using a fixed, user-specified tuning parameter vector because the automatic algorithm can require up to five fits. This slowness is not a serious problem when fitting a few datasets, but it slows down Monte Carlo simulations. Therefore, for the remainder of the study we use fixed values of (K, M, N_{iter}).

5.4. Bayesian inference

To compare posterior distributions with and without a local penalty, we used a spatially heterogeneous regression function

$$m(x) = \exp\left(-400(x-0.6)^2\right) + \tfrac{5}{3}\exp\left(-500(x-0.75)^2\right) + 2\exp\left(-500(x-0.9)^2\right). \quad (11)$$

The x_i were equally spaced on $[0, 1]$, the sample size was $n = 300$, and the ϵ_i were normally distributed with $\sigma = 0.5$. We used quadratic splines with $K = 40$ knots and we had $M = 4$ subknots, and the number of iterations to minimize GCV using the local penalty was $N_{iter} = 1$.

Figure 2 shows a typical dataset and the global- and local-penalty estimates. The global-penalty estimate has a small penalty chosen by GCV to accommodate the oscillations on the right, but the unfortunate side effect is the undersmoothing on the left. The local penalty removes this problem. Figure 3 shows the pointwise MSE and squared bias of the global-penalty estimator calculated from 500 Monte Carlo samples. Also shown is the pointwise posterior variance given by (8) averaged over the 500 repetitions. The posterior variance should be estimating the MSE. We see that the posterior variance is constant, except for boundary effects, and cannot detect the spatial heterogeneity in the MSE. Figure 4 is a similar figure for the local-penalty estimator. Two posterior variances are shown, the conservative adjustment and

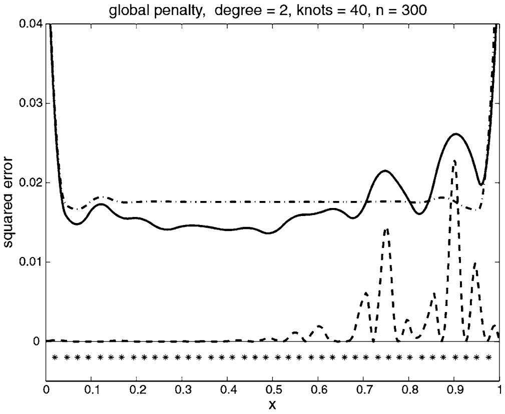


Figure 3. Bayesian inference study: behaviour of the global-penalty estimator — graphs of pointwise MSE, squared bias, and average (over Monte Carlo trials) posterior variance; the MSE and the posterior variance have been smoothed to reduce Monte Carlo variance; the posterior variance assumes that α^* is known, so the variability in $\hat{\alpha}^*$ is not taken into account (solid line = MSE; dashed = squared bias; dashed-and-dotted = posterior variance; asterisks = knot locations)

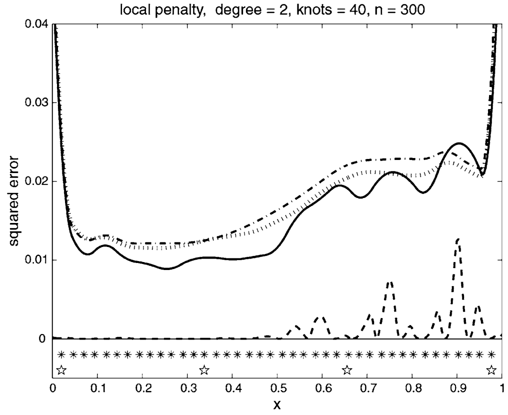


Figure 4. Bayesian inference study: behaviour of the local-penalty estimator — graphs of pointwise MSE, squared bias, and average (over Monte Carlo trials) posterior variance; the MSE and the posterior variance have been smoothed to reduce Monte Carlo variance (solid line = MSE; dashed = squared bias; dashed-and-dotted = posterior variance with conservative adjustment; dotted = posterior variance with Kass–Steffey correction; asterisks = knot locations)

the Kass–Steffey type correction. The MSE looks somewhat different from Figure 3 because the estimator adapts to spatial heterogeneity. Also, the posterior variance tracks the MSE better than for the global-penalty estimator and the corrected version of the posterior variance tends to be a little closer to the MSE than the adjusted version.

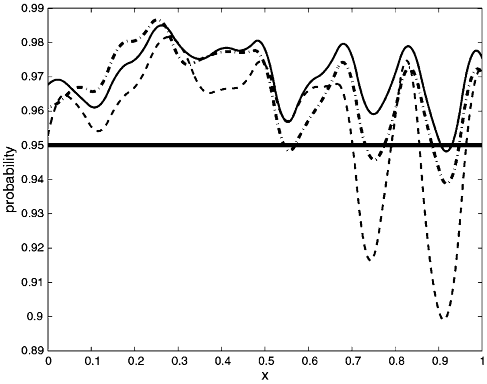


Figure 5. Bayesian inference study using function (11): pointwise coverage probabilities of 95% Bayesian confidence intervals for $m(x_i)$ using global and local penalties — the probabilities have been smoothed to remove Monte Carlo variability; the local-penalty intervals use the conservative adjustment to the posterior variance (solid) and the Kass–Steffey correction (dashed-and-dotted); the dashed curve is the global-penalty estimate

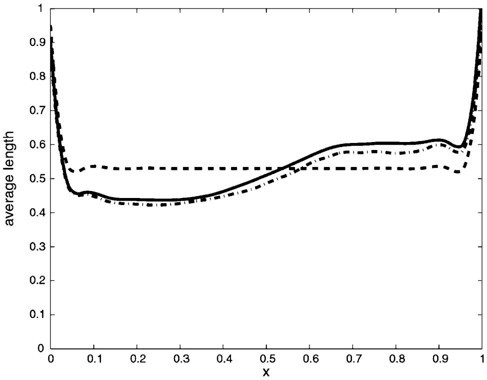


Figure 6. Bayesian inference study using function (11); expected lengths of 95% Bayesian confidence intervals for $m(x_i)$ using global and local penalties — the average (over Monte Carlo trials) lengths have been smoothed to remove Monte Carlo variability; the local-penalty intervals use the conservative adjustment to the posterior variance (solid) and the Kass–Steffey correction (dashed-and-dotted); the dashed curve is the global-penalty estimate

In Figures 5 and 6 we present the Monte Carlo estimates of the pointwise coverage probabilities and average lengths of nominal 95% Bayesian confidence intervals based on the global- and local-penalty estimators. These coverage probabilities have been smoothed by p-splines to remove some of the Monte Carlo variability. All three confidence interval procedures achieve pointwise coverage probabilities close to 95%. Because the local-penalty methods are somewhat conservative, the global-penalty method is, on average, the closest to 95%, but the local-penalty methods avoid low coverage probabilities around features in m.

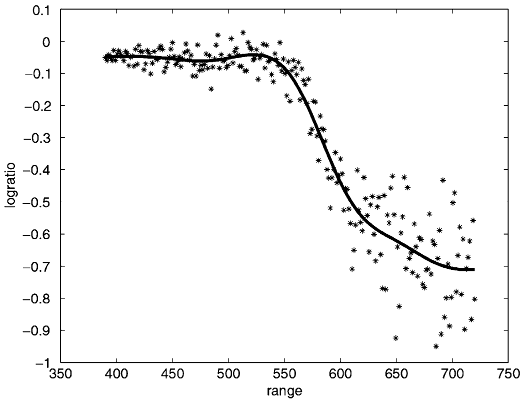


Figure 7. LIDAR data: a global-penalty quadratic spline fit has been added

6. An example

The LIDAR (LIght Detection And Ranging) uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere; see Sigrist (1994).

A typical LIDAR dataset, shown in Figure 7, was taken from Holst *et al.* (1996). The horizontal variable, range, is the distance travelled before the light is reflect back to its source. The vertical variable, log-ratio, is the logarithm of the ratio of received signals at frequencies on and off the resonance frequency of the chemical species of interest, which is mercury in this example.

For this example there is scientific interest in the first derivative, m', as well as our interest in m itself, because $-m'(x)$ is proportional to concentration at range x; see Ruppert *et al.* (1997) for further discussion. For the estimation of m a global penalty works satisfactorily. Figure 7 shows the global-penalty fit. The local-penalty fit was not included in that figure, because it would be difficult to distinguish from the global-penalty fit.

However, for the estimation of m', a local penalty appears to improve upon a global penalty. Figure 8 shows the derivatives (times -1) of fitted splines and their confidence intervals using global and local (spatially-adaptive) penalties. Notice that the confidence intervals using the local penalty are generally narrower than for the global penalty, except at the peak where the extra width should be reflecting real uncertainty. The local-penalty estimate has a sharper peak and less noise in the flat areas.

A referee has made the valid point that it is a common, but questionable practice to choose α^* to estimate m and then to use this α^* to estimate m'. We intend to investigate methods that choose α^* to estimate m', but not in this paper. The GCV, though it targets m, seems to be effective in choosing the right amount of smoothing for estimating m' in this example.

7. Additive models

7.1. An algorithm for additive models

Until now, we have confined our attention to univariate splines, but our ideas can be easily extended to additive models. Suppose we have L predictor variables and that $\boldsymbol{x}_i =$

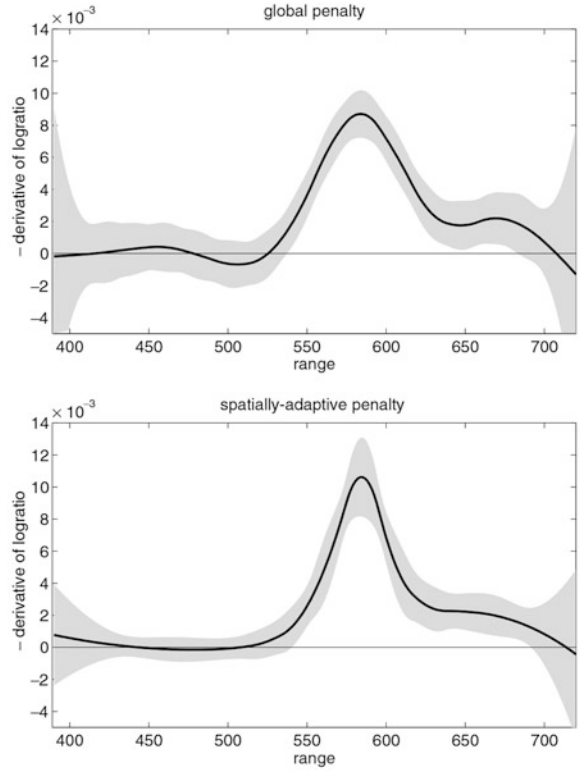


Figure 8. Estimates of $-m'$(range) with global and spatially-adaptive penalties the shaded regions show pointwise 95% confidence intervals

$(x_{i1}, \ldots, x_{iL})^{\mathsf{T}}$ is the vector of predictor variables for the ith case. The additive model is

$$y_i = \beta_0 + \sum_{\ell=1}^{L} m_\ell(x_{i\ell}) + \epsilon_i .$$

If the ℓth predictor variable has K_ℓ knots, $\kappa_{1\ell} \ldots, \kappa_{K_\ell \ell}$, the additive spline model is

$$m(\boldsymbol{x}, \boldsymbol{\beta}) = \beta_0 + \sum_{\ell=1}^{L}\biggl(\beta_{1\ell}x_\ell + \cdots + \beta_{p\ell}x_\ell^p + \sum_{k=1}^{K_\ell}\beta_{p+k,\ell}(x_\ell - \kappa_{k\ell})_+^p\biggr).$$

The parameter vector is $\boldsymbol{\beta} = (\beta_0, \beta_{11}, \ldots, \beta_{p+K_\ell,1}, \ldots, \beta_{p+K_L,L})^{\mathsf{T}}$. Let $\alpha_\ell(\cdot)$ be the penalty function for the ℓth predictor. Then the penalized criterion to minimize is

$$\sum_{i=1}^{n}\bigl(y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})\bigr)^2 + \sum_{\ell=1}^{L}\alpha_\ell(\kappa_{k\ell})\beta_{p+k,\ell}^2 .$$

As discussed in Marx & Eilers (1998), one need not use backfitting to fit an additive spline model. Rather, all L components can be estimated simultaneously.

Consider three levels of complexity of the penalty:

1. $\alpha_\ell(\cdot) \equiv \alpha$ (a common global penalty),
2. $\alpha_\ell(\cdot) \equiv \alpha_\ell$ (separate global penalties),
3. $\alpha_\ell(\cdot)$ is a linear spline (separate local penalties).

The following algorithm allows one to fit separate local penalties using only one-dimensional grid searches for minimizing GCV. First one minimizes GCV using a common global penalty. For this penalty to be reasonable, one must standardize the predictors so that they have common standard deviations, say. Then using the common global penalty as a starting value, one minimizes GCV over separate global penalties. During minimization, the L penalty parameters are varied one at a time, with the rationale that the optimal value of α_ℓ depends only slightly on the $\alpha_{\ell'}, \ell' \neq \ell$. Finally, using separate global penalties as starting values, one minimizes GCV over separate local penalties. The ℓth local penalty has M_ℓ parameters so there are a total of $M_1 + \cdots + M_L$ penalty parameters. These are varied in succession to minimize GCV.

7.2. Simulations of an additive model

To evaluate the practicality of this algorithm we used a variation of the simulation example in Section 5.4 where we added two spatially homogeneous component functions to the spatially heterogeneous function (11). Thus, there were three predictor variables which for each case were independently distributed as uniform(0,1) random variables. The components of m were $m_1(x_1) = \sin(4\pi x_1)$ and $m_2(x_2) = x_2^3$, and $m_3(x_3)$ was the same as $m(x)$ in (11). As in Section 5.4, $n = 300$ and the ϵs were iid $\mathrm{N}(0, 0.25)$. We used quadratic splines and 10, 10 and 40 knots for m_1, m_2 and m_3, respectively. The local-penalty estimate had four subknots for all four functions.

First consider computation time. For a single dataset and using our MATLAB program on a SUN Ultra 1, the common global-penalty estimate took 2.1 seconds to compute, the separate global-penalty estimate took an additional 1.5 seconds to compute, and then separate local-penalty estimates took an additional 10.4 seconds to compute. Thus, local penalties are more computationally intensive than global penalties, but still feasible for small L. Now consider larger values of L. Everything else held constant, the number of parameters of an additive model grows linearly in L and, since matrix inversion time is cubic in dimension, the time for a single fit should grow cubically in L. The number of fits needed for the sequential grid searching described above grows linearly in L, so the total computation time for local penalties should be roughly proportional to L^4. To test this rough calculation empirically, we found the computation time for fitting additive models with 300 data points, 10 knots per variable, and four subknots per variable. L took five values from 1 to 15. Figure 9 is a log–log plot of computation time versus L. A linear fit on the log-scale is also shown; its slope is 2.45, not 4 as the quartic model predicts. The actual data show log-times that are nonlinear in $\log(L)$ with an increasing slope. Thus, a quartic model of time as a function of L may work for large values of L, but a quadratic or cubic model would be better for L in the 'usual' range of 1 to 15. It is possible that the quartic model's poor fit for smaller L is because the quartic model ignores parts of the computation that are linear, quadratic and cubic in L. The computation time for eight variables is about 1.5 minutes, but for 15 variables it is about 10.5

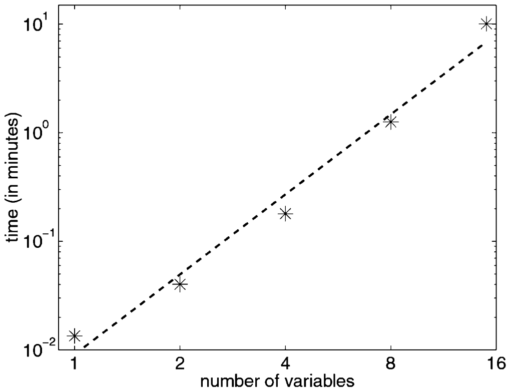


Figure 9. Log–log plot of the computation time for fitting an additive model with local penalties as a function of the number of variables L; a linear fit is also shown; the slope of the linear fit is 2.45

minutes. It seems clear that local additive fitting is feasible up to at least 8–10 variables and maybe 15 variables, but is only 'interactive' up to three variables.

An important point to keep in mind is that computation times are largely independent of the sample size n. The reason for this is that once $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}Y$ have been computed, all computation times are independent of n and the computation of $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}Y$ is quite fast unless n is enormous.

Now consider statistical efficiency. The MSEs computed over 500 Monte Carlo samples for the separate local penalties estimator were 0.010, 0.0046, and 0.0165 for m_1, m_2 and m_3, respectively. Thus, m_2 is relatively easy to estimate and m_3 is slightly more difficult to estimate than m_1. The ratios of the MSE for common global penalties to separate local penalties were 1.26, 2.36, and 1.23 for m_1, m_2 and m_3, respectively. The ratios of the MSE for separate global penalties to separate local penalties were 0.85, 0.88 and 1.20 for m_1, m_2 and m_3, respectively. Thus, for all three component functions, the common global-penalty estimator with a single smoothing parameter is less efficient than the fully-adaptive estimator with separate local penalties. For the spatially homogeneous functions m_1 and m_2, there is some loss of efficiency when using local penalties rather than separate global penalties, but the spatially heterogeneous m_3 is best estimated by a local penalty. These findings are somewhat different from those we found for univariate regression, where no efficiency loss was noticed when a local penalty was used where a global penalty would have been adequate. There may be practical situations where one knows that a certain component function is spatially heterogeneous but the other component functions are not. Then greater efficiency should be achievable by using global penalties for the spatially homogeneous component functions and local penalties for the spatially heterogeneous ones.

The results in this section provide evidence that sequential one-dimensional grid searches to find the smoothing parameter vector are effective. This is because the optimal value of one tuning parameter depends only weakly upon the other tuning parameters. The result is that searches over a rather large number of tuning parameters (up to 60 when L is 15 and there are four subknots per variable) appear feasible.

A study of Bayesian inference for additive models is beyond the scope of the present paper.

8. Summary and conclusions

Spatial adaptivity is important for improved precision of point estimators and improved accuracy of confidence intervals.

The local-penalty spline is effective in increasing efficiency as measured by MSE when the regression function is spatially heterogeneous in complexity. The local-penalty method of Bayesian inference has good coverage probability throughout the range of the predictor variable, though it is somewhat conservative with coverage probabilities typically a bit higher than nominal. This conservativeness may be due to the ad hoc 'adjustment' we make for estimation of multiple smoothing parameters. The adjustment is to multiply the pointwise posterior variance of the local-penalty $\hat{m}$ by a constant so that its average posterior variance is the same as the global-penalty spline. We also considered a more theoretically justified 'correction' based on the work of Kass & Steffey (1989). This correction is slightly less conservative than the ad hoc adjustment and would be recommended in preference to the adjustment except that the correction increases computation cost considerably because one step involves a small bootstrap.

For the test cases we have studied that have a moderate spatial heterogeneity, local-penalty splines with knots at equally-spaced quantiles of x perform as well as, and perhaps a bit better than, estimators using sequential knot selection.

In practice, reasonable values of the tuning parameters (K, M, N_{iter}) can often be specified by the user. However, an automatic algorithm that selects these tuning parameters by GCV has proved effective.

When a global penalty is appropriate, there seems to be little or no loss of efficiency in using local penalties, at least in the univariate case. For additive models, there can be some loss of efficiency when using a local penalty where a global penalty is appropriate.

References

BOX, G.E.P. & TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

CARLIN, B.P. & LOUIS, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall.

DEELY, J.J. & LINDLEY, D.V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76**, 833–841.

DENISON, D.G.T., MALLICK, B.K. & SMITH, A.F.M. (1998). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. Ser. B* **60**, 333–350.

DONOHO, D.L. & JOHNSTONE, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200–1224.

EILERS, P.H.C. & MARX, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11**, 89–121.

EUBANK, R.L. (1988). *Spline Smoothing and Nonparametric Regression.* New York and Basel: Marcel Dekker.

FRIEDMAN, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1–141.

FRIEDMAN, J.H. & SILVERMAN, B.W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3–39.

HASTIE, T.J. & TIBSHIRANI, R. (1990). *Generalized Additive Models.* London: Chapman & Hall.

HOLST, U., HÖSSJER, O., BJÖRKLUND, C., RAGNARSON, P. & EDNER, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Environmetrics* **7**, 401–416.

KASS, R.E. & STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **84**, 717–726.

LAIRD, N.M. & LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *J. Amer. Statist. Assoc.* **82**, 739–757.

LINDLEY, D.V. & SMITH, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 1–41.

LUO, Z. & WAHBA, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* **92**, 107–116.

MARX, B.D. & EILERS, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. *Comput. Statist. Data Anal.* **28**, 193–209.

MORRIS, C.N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78**, 47–65.

NYCHKA, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83**, 1134–1143.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statist. Sci.* **1**, 505–527.

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9**, 363–379.

RUPPERT, D., WAND, M.P., HOLST, U. & HÖSSJER, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262–273.

SIGRIST, M. (1994). *Air Monitoring by Spectroscopic Techniques.* (Chemical Analysis Series, Vol. 127) Wiley.

SMITH, M. & KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–344.

STONE, C.J., HANSEN, M., KOOPERBERG, C. & TRUONG, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371–1470.

WAHBA, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45**, 133–150.

WAHBA, G. (1990). *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.

WAND, M.P. (2000). A comparison of regression splines smoothing procedures. *Comput. Statist.* (to appear).

Chapter 5
Nonparametric and Semiparametric Regression for Dependent Data

By Yehua Li and Naisyin Wang

About the Authors. Yehua Li is an Associate professor of statistics at the Iowa State University. He received his PhD in statistics from Texas A&M University in 2006 under the direction of Raymond Carroll and Tailen Hsing. Thereafter, Yehua was appointed as an assistant professor of statistics at the University of Georgia, where he served for six years. He received a CAREER Award from the National Science Foundation in 2012. He has coauthored four articles with Ray and Naisyin Wang on functional data analysis and measurement error problems.

Naisyin Wang is a professor of statistics and biostatistics at the University of Michigan. She obtained a PhD in Statistics from Cornell University in 1992 under the direction of David Ruppert. She met Ray as a graduate student while both David Ruppert and Ray were visiting Peter Hall in Canberra, Australia. Naisyin joined the faculty at Texas A&M University in 1992 and had an office right next to Ray's from 1992 to 2004. She collaborated with Ray on measurement error problems in mixed effects models and on various topics in nonparametric and semiparametric estimation. Working with Ray, she co-advised graduate students Jeffrey S. Morris and Yehua Li (working with Yehua after his co-advisor, Tailen Hsing, moved to The Ohio State University). While at Texas A&M University, Naisyin and Ray enjoyed long-term collaborative relationships with faculty members in the areas of toxicology and nutrition.

Selected Papers on Nonparametric and Semiparametric Regression for Dependent Data

[NSRD-1]-[140] Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured with/without error. *Journal of the American Statistical Association*, 95, 520–534.

[NSRD-2]-[113] Lin, X. and Carroll, R. J. (2001a). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96, 1046–1056.

[NSRD-3]-[38] Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, 68, 68–88.

[NSRD-4]-[64] Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68, 179–199.

We are honored to include this article to celebrate the publication of Ray's collected works by Springer. These collected works mark another milestone in Ray's distinguished career, as evident by the wide range of statistics publications he has contributed to the literature over the years. The enclosed papers have had far-reaching impacts on the field of statistics, in addition to other disciplines, including AIDS research, cancer studies, finance, image analysis, proteomics, and genomics. We anticipate future new applications of the methods and theory Ray developed

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_5,

that are represented in these papers. We appreciate this opportunity to discuss Ray's contributions to non- and semiparametric longitudinal and functional data analysis.

Nonparametric and Semiparametric Regression for Longitudinal/Clustered Data

Ray's work on nonparametric regression problems for clustered data started with the paper he developed with Xihong Lin, which was published in the *Journal of the American Statistical Association* (*JASA*) in 2000. As Ray later shared with us, this work was motivated by a measurement error problem in nonparametric regression. Their original plan was to apply Len Stefanski's SIMEX method to some existing local polynomial estimators to reduce the bias caused by the measurement errors. The data they considered happened to be repeatedly measured. As they worked out the technical proofs, they found an astonishing result, which showed that the classical kernel method performed worse than the method that totally ignored the within-cluster correlation. This finding opened the door to a series of research papers on how to efficiently perform nonparametric regression when correlation among the data exists. In the 10 years since their publication, these two papers (2000 [NSRD-1], 2001a [NSRD-2]) have been cited 155 and 130 times, respectively, which are excellent citation rates in statistics.

The data considered in Lin and Carroll (2000 [NSRD-1]) are from n independent clusters with m_i observations in the ith cluster. The jth observation within the ith cluster is a pair (Y_{ij}, Z_{ij}), where Y_{ij} is the response variable and Z_{ij} is a covariate. The conditional mean of Y_{ij} given Z_{ij} is given by $\mu_{ij} = E(Y_{ij}|Z_{ij})$ with

$$\mu_{ij} = \mu\{\theta(Z_{ij})\}, \tag{5.1}$$

where $\mu(\cdot)$ is a known link function and $\theta(\cdot)$ is an unknown nonparametric function. The response variables within the ith cluster are correlated, with the conditional covariance matrix defined as $\Sigma_i = \mathrm{cov}(\tilde{Y}_i \mid \tilde{Z}_i)$, where $\tilde{Y}_i$ and $\tilde{Z}_i$ are vectors collecting Y and Z's within the ith cluster. The covariate Z_{ij} can be observed either with or without measurement error. When Z is measured with error, Lin and Carroll show that the biases caused by measurement errors can be overcome by SIMEX methods.

In the case where Z_{ij} is measured without error, Lin and Carroll (2000 [NSRD-1]) consider two sets of kernel estimating equations to estimate $\theta(x)$,

$$\sum_{i=1}^{n} G_{ip}(x)^T \Delta_i(x) V_{2i}(x)^{-1} K_{ih}(x)\{\tilde{Y}_i - \mu_i(x)\} = 0, \tag{5.2}$$

$$\text{or} \quad \sum_{i=1}^{n} G_{ip}(x)^T \Delta_i(x) K_{ih}^{1/2}(x) V_{2i}(x)^{-1} K_{ih}^{1/2}(x)\{\tilde{Y}_i - \mu_i(x)\} = 0, \tag{5.3}$$

where $V_{2i} = S_i^{1/2} R_{2i} S_i^{1/2}$, R_{2i} is the working correlation matrix, S_i is a diagonal matrix of the conditional variances, K_{ih} is a diagonal matrix containing the kernel weights, $G_{ip}(x)$ is the design matrix for the local polynomial, and Δ_i contains the derivatives of the link function. The most striking result derived by Lin and Carroll

(2000 [NSRD-1]) is that, for both estimators defined in the Eqs. (5.2) and (5.3), the asymptotic variance of $\hat{\theta}(x)$ is minimized when correlation is ignored, i.e., $R_{2i} = I$.

Lin and Carroll (2001a [NSRD-2]) further extend this investigation to the semiparametric setting, where they considered a generalized partially linear model

$$g(\mu_{ij}) = X_{ij}^T \beta + \theta(Z_{ij}), \tag{5.4}$$

with $g(\cdot) = \mu^{-1}(\cdot)$ and X_{ij} being a covariate vector. They proposed a profile estimation scheme. For a fixed value of β, an estimator for the nonparametric component $\hat{\theta}(z;\beta)$ is obtained by solving either Eq. (5.2) or (5.3). An estimator of the parametric component $\hat{\beta}$ is obtained by solving

$$\sum_{i=1}^{n} \frac{\partial \mu\{\tilde{X}_i \beta + \hat{\theta}(\tilde{Z}_i;\beta)\}^T}{\partial \beta} V_{1i}^{-1}(\tilde{X}_i, \tilde{Z}_i)[\tilde{Y}_i - \mu\{\tilde{X}_i \beta + \hat{\theta}(\tilde{Z}_i;\beta)\}] = 0, \tag{5.5}$$

where $\tilde{X}_i = (X_{i1}^T, \ldots, X_{i,m_i}^T)^T$. The working covariance matrix V_{1i} could be different from V_{2i} which is used in nonparametric estimation. Under this estimation scheme, Lin and Carroll (2001a [NSRD-2]) show that $\hat{\beta}$ is in general root-n inconsistent, unless working independence is assumed or $\hat{\theta}$ is undersmoothed. They also investigated the semiparametric efficient score for this problem, and showed that, even when undersmoothing is applied, $\hat{\beta}$ is semiparametrically inefficient. However, in the special case when X and Z are independent, R_{1i} is equal to the true correlation and $R_{2i} = I$, $\hat{\beta}$ is root-n consistent and semiparametrically efficient. Lin and Carroll (2001b) study the same framework, but concentrate on the special case in which Z is a cluster-level variable, i.e., $Z_{ij} = Z_i$. Under this special case, they show that the estimators above are semiparametrically efficient.

The work of Lin and Carroll became a heated topic of discussion after it was presented at the second Seattle Symposium of Biostatistics. It inspired a lot of research on the use of within-cluster correlation to improve the efficiency of a nonparametric kernel estimator. Among others, Wang (2003) proposed a kernel regression estimator that leads to a fully efficient estimator in the setup of Lin and Carroll (2000) when the true correlation structure is used. Lin et al. (2004) study this estimator further. Ray named this estimator as the seemingly unrelated (SU) kernel estimator to indicate the underlying principle of how correlated errors are utilized. They show that this estimator is asymptotically equivalent to a smoothing spline estimator, and the equivalent kernels of both estimators are nonlocal. The original proposal of this estimator was an iterative procedure that is applicable to generalized linear models, and Lin et al. (2004) show that a non-iterative closed-form solution to this estimator exists when the link function is an identity link function. Wang et al. (2005) revisit the semiparametric regression problem in Lin and Carroll (2001a [NSRD-2]) and propose an estimation procedure that combines the SU kernel estimator with the profiling technique. They show that this profile procedure is semiparametrically efficient.

Lin and Carroll (2006 [NSRD-3]) extend the techniques invented in this series of works to a general framework, which covers an impressively wide range of semiparametric regression problems for data with repeated measures. In summary, the examples under this new framework covered problems from matched case–control studies, finance, and many widely used statistical models in biological and medical research, such as generalized linear mixed models and generalized partially linear models. The data are described in terms of a *criterion function* $\mathscr{L}(\tilde{Y},\tilde{X},\tilde{\eta},\mathscr{B})$, where $\tilde{Y}$ and $\tilde{X}$ are the vector and matrix collecting the m repeatedly measured responses and covariates within the same cluster, and $\tilde{\eta}=\{\theta(Z_1),\ldots,\theta(Z_m)\}^T$ and $\mathscr{B}$ are the nonparametric and parametric components in the model, respectively. The criterion function needs to satisfy

$$E[\{\partial\mathscr{L}(\tilde{Y},\tilde{X},\tilde{\eta}_0,\mathscr{B}_0)/\partial(\tilde{\eta},\mathscr{B})\}\mid\tilde{X},\tilde{Z}]=0.$$

Examples of such criterion functions include the quasi-likelihood and the likelihood functions as special cases.

For a given $\mathscr{B}$, Lin and Carroll (2006 [NSRD-3]) propose to estimate $\theta(z)$ by $\hat{\theta}(z,\mathscr{B})$ which is the solution of

$$\begin{aligned}0=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(Z_{ij}-z)G_{ij}(z,h)\mathscr{L}_{j\theta}\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}(Z_{i1},\mathscr{B}),\ldots,\\ \hat{\theta}(z,\mathscr{B})+\hat{\theta}^{(1)}(z,\mathscr{B})(Z_{ij}-z),\ldots,\hat{\theta}(Z_{im},\mathscr{B}),\mathscr{B}\},\quad(5.6)\end{aligned}$$

$K(\cdot)$ is a kernel function, $G_{ij}(z,h)$ is the design matrix of the local polynomial and $\mathscr{L}_{j\theta}=\partial\mathscr{L}/\partial\eta_j$. The estimating equation is solved in an iterative fashion.

To estimate the parametric component, they proposed two schemes—profiling and backfitting. The profile estimator $\hat{\mathscr{B}}_p$ maximizes

$$\sum_{i=1}^{n}\mathscr{L}_i\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}(Z_{i1},\mathscr{B}),\ldots,\hat{\theta}(Z_{im},\mathscr{B}),\mathscr{B}\}.$$

The backfitting algorithm requires updating $\mathscr{B}$ iteratively by maximizing

$$\sum_{i=1}^{n}\mathscr{L}_i\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}(Z_{i1},\mathscr{B}_*),\ldots,\hat{\theta}(Z_{im},\mathscr{B}_*),\mathscr{B}\},$$

where $\mathscr{B}_*$ is the value of $\mathscr{B}$ from the previous iteration. At convergence, the final estimator is denoted as $\hat{\mathscr{B}}_b$.

Lin and Carroll (2006 [NSRD-3]) study the asymptotic properties for $\hat{\theta}(z)$, $\hat{\mathscr{B}}_p$ and $\hat{\mathscr{B}}_b$. They investigate the semiparametric efficiency bound for these general semiparametric regression problems and show that the asymptotic variance of $\hat{\mathscr{B}}_p$ can achieve this information bound if $\mathscr{L}$ is the likelihood function. They also show that, under an undersmoothing scheme, the backfitting estimator $\hat{\mathscr{B}}_b$ is asymptotically equivalent to the profile estimator $\hat{\mathscr{B}}_p$. In addition, they discuss a variety of statistical issues that may occur when working with real data; for example, the existence of nuisance parameters that need to be estimated, and extension to measurement error problems.

Ray has had some important work in this area published more recently. In Carroll et al. (2009), Ray and his coauthors investigate additive models for repeatedly measured data. They propose a smooth backfitting algorithm, which allows for a working covariance. They show the method is most efficient when the true covariance matrix is used, and in that case each component function in the additive model achieves the known asymptotic variance lower bound for one-dimensional smoothing. In Maity et al. (2009), Ray and his coauthors propose a new class of semiparametric models that can be used to model gene–environment interactions. They investigated the cases with or without repeated measures and show that some of the general methods and results in Lin and Carroll (2006 [NSRD-3]) are applicable. The focus of this paper is on developing a score test to test the interaction between gene expression levels and environmental variables.

Functional Data Analysis

Functional data analysis is one of the fastest-growing fields in statistics at this time. Technological advances in scientific and medical research have resulted in large data sets, in which each datum is a function (e.g., a curve or an image) and the whole data set consists of collections of such functions sampled on a fine grid. Ray and his collaborators developed theory and methods for such data that have made high-impact contributions to functional data analysis.

Ray's collaboration with Jeffrey Morris on hierarchical functional data started in Morris et al. (2003) as a discussion paper in *JASA*. The proposed functional mixed model framework was later perfected in Morris and Carroll (2006 [NSRD-4]). The model they consider is given by

$$Y(t) = XB(t) + ZU(t) + E(t), \tag{5.7}$$

where $Y(t) = \{Y_1(t), \dots, Y_N(t)\}^T$ is a vector of observed functions, $B(t) = \{B_1(t), \dots, B_p(t)\}^T$ is a vector of fixed effect functions with design matrix X, $U(t) = \{U_1, \dots, U_m(t)\}^T$ is a vector of random effect functions with design matrix Z, and $E(t) = \{E_1(t), \dots, E_N(t)\}^T$ is a vector of error processes. Both $U(t)$ and $E(t)$ are assumed to be independent multivariate Gaussian processes with distribution $MGP(P, Q)$ and $MGP(R, S)$, where P and R are covariance matrices and $Q(t_1, t_2)$ and $S(t_1, t_2)$ are covariance surfaces.

The proposed modeling framework is very flexible and can be applied to a variety of problems. The fixed effects may include the mean function, functional main effects, functional interactions, functional liner coefficients for continuous covariates, and interactions of functional coefficient with other effects. The functional random effects induce different covariance structure among the data and can be used to accommodate functional data from nested designs, split-plot designs, subsampling designs, and other designs involving repeated measures over time or space.

Morris and Carroll (2006 [NSRD-4]) propose modeling all functions in Eq. (5.7) by wavelets, which are popular basis functions for nonparametric regression that are especially suitable for spatially adaptive smoothing. Wavelets are a set of orthonormal basis function indexed by location and scale. By taking wavelet transformation on both sides of Eq. (5.7), the model is translated into a mixed model in wavelet coefficients

$$D = XB^* + ZU^* + E^*, \tag{5.8}$$

where $D = YW^T$, $U^* = UW^T$, and $E^* = EW^T$, with W being the projection matrix corresponding to the discrete wavelet transformation. The random terms in the wavelet domain model (5.8) are the zero-mean Gaussian matrices, $D \sim MN(P, Q^*)$ and $E^* \sim MN(R, S^*)$, where $Q^* = WQW^T$ and $S^* = WSW^T$. One great benefit of using wavelets is that one can take advantage of whitening property of the wavelet transformation Johnstone and Silverman (1997). Even though the processes $U(t)$ and $E(t)$ may have serial correlation in t, the corresponding wavelet coefficients tend to be decorrelated. Thus, one can model Q^* and S^* as diagonal matrices and achieve parsimony in the model while still accommodating a flexible class of nonstationary covariance matrices Q and S. This results in huge gains in computational efficiency, especially when the data set is gigantic.

Morris and Carroll (2006 [NSRD-4]) adapt this problem into a Bayesian framework, where the fixed effects B^* are assigned mixture priors with shrinkage effects, and variance components in P, R, Q^*, and S^* are assigned with vague proper priors. To achieve adaptive smoothing of the functions, Morris and Carroll propose estimating the hyperparameters by using an empirical Bayes method, where the amount of smoothing is estimated directly from the data. To estimate the parameters in the model, they proposed an efficient MCMC algorithm. The Bayesian framework also provides automatic devices for inference and prediction in the model.

The work of Morris and Carroll (2006 [NSRD-4]) introduces a very successful framework, which has had a huge impact through its eventual application by Jeffrey Morris and collaborators to the realms of proteomics, genomics, forestry, neuroimaging, and ophthalmology. Ray and his coauthors have continued to work on extensions and generalization of hierarchical functional modeling. Baladandayuthapani et al. (2008) study spatially correlated hierarchical functional data using penalized splines, where a more flexible Matérn spatial correlation is considered. The assumption of Q^* being a diagonal matrix from Morris and Carroll (2006 [NSRD-4]) has provided a large computational advantage, but may not be suitable for all functional data. The latest advance in related methodology is to model the within-function covariance structure by functional principal component analysis (FPCA). Unlike Morris and Carroll (2006 [NSRD-4]) and Baladandayuthapani et al. (2008), who use pre-determined basis functions (wavelets or splines), FPCA provides a set of data-driven orthonormal basis functions. Zhou et al. (2008) study joint modeling of paired functional data using spline-based FPCA. Staicu et al. (2010) propose a multi-level FPCA approach for spatially correlated hierarchical functional data. However, like Morris and Carroll (2006 [NSRD-4]) and Baladandayuthapani et al.

(2008), Staicu et al. (2010) assume a separable structure for the within-curve covariance and the between-curve spatial correlation. The latest work from Ray's research group is Zhou et al. (2010), which so far provides the most general framework for hierarchical functional modeling. In this work, the data are modeled by multi-level FPCA, and FPCA scores are spatially correlated. The within-curve covariance and the between-curve correlation are not necessarily separable and can thus accommodate more general data.

In addition to hierarchical functional modeling, Ray made other important contributions to the literature on functional data analysis. Li et al. (2010) propose a new class of semiparametric regression models for a scalar response, with multivariate and functional predictors. The model can accommodate interactions between multivariate and functional predictors. To avoid the curse of dimensionality, the authors proposed an innovative single-index structure for the interactions. This approach strikes a nice compromise between model flexibility and practical feasibility in computation and stability.

This article is limited to a brief summary of only part of Ray's ample contributions to this important research area. We are confident that Ray's high energy and devotion will propel him to generate new directions of research focus and will intrigue more researchers of all levels to be involved in this area of work. We look forward to seeing continuous growth of this research area that Ray has inspired.

References

Other publications by Ray Carroll cited in this chapter.

Baladandayuthapani, V., Hong, M. Y., Mallick, B. K., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64, 64–73.

Carroll, R. J., Maity, A., Mammen, E., and Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika*, 96, 383–398.

Johnstone, I. M., and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, 59, 319–351.

Li, Y., Wang, N., and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions, *Journal of the American Statistical Association*, 105, 621–633.

Lin, X. and Carroll, R. J. (2001b). Semiparametric regression for clustered data *Biometrika*, 88, 1179–1185.

Lin, X., Wang, N., Welsh, A. H., and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for longitudinal/clustered data. *Biometrika*, 91, 177–194.

Maity, A., Carroll, R. J., Mammen, E., and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene-environment interactions. *Journal of the Royal Statistical Society, Series B*, 71, 75–96.

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98, 573–597 (Editor's Invited Paper for 2003).

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11, 177–194.

Wang, N., Carroll, R. J., and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100, 147–157.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modeling of paired sparse functional data using principal components. *Biometrika*, 95, 601–619.

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J., (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, 105, 390–400.

Publications by other authors cited in this chapter.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, 90, 43–52.

Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error

Xihong LIN and Raymond J. CARROLL

We consider local polynomial kernel regression with a single covariate for clustered data using estimating equations. We assume that at most $m < \infty$ observations are available on each cluster. In the case of random regressors, with no measurement error in the predictor, we show that it is generally the best strategy to ignore entirely the correlation structure within each cluster and instead pretend that all observations are independent. In the further special case of longitudinal data on individuals with fixed common observation times, we show that equivalent to the pooled data approach is the strategy of fitting separate nonparametric regressions at each observation time and constructing an optimal weighted average. We also consider what happens when the predictor is measured with error. Using the SIMEX approach to correct for measurement error, we construct an asymptotic theory for both the pooled and the weighted average estimators. Surprisingly, for the same amount of smoothing, the weighted average estimators typically have smaller variances than the pooling strategy. We apply the proposed methods to analysis of the AIDS Costs and Services Utilization Survey.

KEY WORDS: AIDS; Asymptotic bias and variance; Clustered data; Efficiency; Errors in variables; Estimating equations; Generalized linear models; Kernel regression; Longitudinal data; Measurement error; Nonparametric regression; Panel data; SIMEX.

1. INTRODUCTION

A vast literature has developed in the past decade on parametric regression for clustered data using estimating equations (Liang and Zeger 1986), where generalized linear models are a special case. Such parametric assumptions may not always be desirable, as appropriate functional forms of the covariates may not be known in advance, and the outcome may depend on the covariates in a complicated manner. There has been substantial recent interest in extending the existing parametric models to allow for nonparametric covariate effects (Severini and Staniswalis 1994; Wild and Yee 1996; Zeger and Diggle 1994). Such nonparametric regression allows for more flexible functional dependence of the outcome variable on the covariates and also can be used to investigate whether an appropriate parametric function can be developed to describe the data well.

Another complication in the analysis of clustered data is the presence of covariate measurement error. For example, it has been well documented in the literature that covariates such as blood pressure (Carroll, Ruppert, and Stefanski 1995) and CD4 count (Tsiatis, Degruttola, and Wulfsohn 1995) are often subject to measurement error. We consider here data from the AIDS Costs and Services Utilization Survey (ACSUS) (Berk, Maffeo, and Schur 1993). The ACSUS sampled 2,487 subjects in 10 randomly selected U.S. cities with the highest AIDS rates. A series of six interviews were conducted for each respondent every 3 months from 1991 to 1992. A main outcome of interest was whether an interviewee had had hospital admissions (yes/no) during the past 3 months. The collected covariates included demographic variables, HIV status, CD4 count, and treatments.

A question of interest in this study is how CD4 count affects the risk of hospitalization. Analysis of this dataset entails two major complications. The first complication is that even though it is believed that a lower CD4 count is associated with a greater risk of hospitalization, the functional form of this relationship is not known. We are interested in whether the relationship is simply linear, or whether there is a changepoint, or whether the relationship has a complex form. The second complication is that CD4 count was measured with error. One source of error came from its substantial variability; for example, the coefficient of variation could be as large as 50% (Tsiatis et al. 1995). The other source of error came from the fact that CD4 count was not measured at the time of each interview, but rather the most recent CD4 count was abstracted from each respondent's medical record using his or her usual source of care. In view of these complications, we are interested in modeling the effect of CD4 count nonparametrically and accounting for its measurement error. Our nonparametric approach allows us to model the relationship between hospitalization and CD4 count using a flexible function without restricting any particular functional form and to investigate whether we can identify a simple parametric function to capture this relationship. Another advantage is that nonparametric regression can often help recover unexpected patterns of the relationship.

We consider nonparametric regression estimation for clustered data with a single covariate using estimating equations when the covariate is measured accurately or with error. We estimate the nonparametric function using the local

Xihong Lin is Associate Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *xlin@sph.umich.edu*). Raymond J. Carroll is Distinguished Professor, Departments of Statistics and Biostatistics & Epidemiology, Texas A&M University, College Station, TX 77843 (E-mail: *carroll@stat.tamu.edu*).

Her research was supported by National Cancer Institute grant CA-76404. His research was supported by National Cancer Institute grant CA-57030 and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30-ES09106.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

polynomial kernel methods and extend these methods to the measurement error case using the simulation–extrapolation (SIMEX) method (Cook and Stefanski 1994). We study the asymptotic biases and variances of the proposed estimators.

We develop two main striking results:

When the Covariate is Measured Accurately. Several authors have tried to account for within correlation when constructing an estimator for the nonparametric function (Severini and Staniswalis 1994; Wild and Yee 1996; Verbyla, Cullis, Kenward, and Welham 1999). We show that generally the best strategy is to ignore the correlation structure entirely, and pretend as if the data within a cluster were independent (i.e., the working independence model in generalized estimating equation terminology). Furthermore, correctly specifying the correlation structure in estimating the nonparametric function in fact has adverse effects; that is, it results in an asymptotically less efficient estimator. This result is dramatically different from the parametric regression situation for clustered data, where correctly specifying the correlation structure gives the most efficient estimators of regression coefficients (Liang and Zeger 1986). Although the result was a surprise to us, it may result from the local property of local polynomial estimation. As the bandwidth becomes smaller, the chance that correlated observations from the same cluster fall in the same bandwidth vanishes and the observations essentially behave independently.

"Panel Data" With Measurement Error. In "panel data," observations for different subjects are obtained at a series of common time points during a longitudinal follow-up. We show that it is preferable to fit separate functions to each time period and then combine the methods via weighted averaging, rather than try to perform a single measurement error analysis by pooling all of the data from different panels. This result is also dramatically different from parametric measurement error regression, where pooled analysis gives an asymptotically efficient estimator.

The article is organized as follows. In Section 2 we introduce the model. In Section 3 we consider local polynomial methods for nonparametric regression in clustered data when the predictor is observed exactly. We study the asymptotic biases and variances of the local polynomial kernel estimators. Ruckstuhl, Welsh, and Carroll (1999) have investigated this issue in the Gaussian model when the covariance structure of observations within a cluster is that of the usual one-way random-effects analysis of variance model. One part of this article consists of extending their work to generalized linear models, allowing for an arbitrary correlation structure and working correlation models. The results of the generalization are surprising to us and much in line with those of Ruckstuhl et al. Specifically, we show that the asymptotically most efficient estimator of the nonparametric function is obtained by entirely ignoring the correlation within each cluster. This result has by the way been conjectured in the Gaussian case by Hoover, Rice, Wu, and Yang (1998) and Wu, Chiang, and Hoover (1998) and used as the basis for their methods.

Two methods emerge from our analysis. The first simply pools the data and runs a standard nonparametric regression analysis, possibly with weighting for variability. The second method applies to the "panel data" problem, in which case it makes sense to compute regression estimates separately for each time point and form a weighted average of the resulting estimates. We show in Section 3 that the methods of pooling and weighted averaging yield asymptotically equivalent estimates.

In Section 4 we take up the issue of measurement error. We consider the behavior of the SIMEX methodology (Cook and Stefanski 1994) for correcting measurement error, obtaining asymptotic theory for the pooling method and for the weighted average method. Surprisingly, the two methods are no longer asymptotically equivalent in the "panel data" context, where the weighted average method can have a smaller variance. We apply the proposed methods to the analysis of the ACSUS data in Section 5, followed by discussion in Section 6.

2. THE MODEL

Suppose that the data consist of n clusters with the ith $(i = 1, \dots, n)$ cluster having m_i observations. Let Y_{ij} and (X_{ij}, W_{ij}) be the response variable, the true unobserved covariate, and the observed X-related error-prone covariate of the jth $(j = 1, \dots, m_i)$ observation in the ith cluster. The observations within the same cluster might be correlated. Given the true covariate X_{ij}, the mean and variance of Y_{ij} are $E(Y_{ij}|X_{ij}) = \mu_{ij}$ and $\text{var}(Y_{ij}|X_{ij}) = \phi_j w_{ij}^{-1} V(\mu_{ij})$, where ϕ_j is a scale parameter, w_{ij} is a weight, and $V(\cdot)$ is a variance function. The marginal mean μ_{ij} depends on X_{ij} through a known monotonic link function $\mu(\cdot)$,

$$\mu_{ij} = \mu\{\theta(X_{ij})\}, \tag{1}$$

where $\theta(\cdot)$ is an unknown smooth function and the link function $\mu(\cdot)$ is differentiable. Note that so far we have not specified a within-cluster correlation structure for the observations Y_{ij}.

The model is completed by assuming that the unobserved covariate X_{ij} is related to the observed covariate W_{ij} by an additive measurement error model

$$W_{ij} = X_{ij} + U_{ij}, \tag{2}$$

where U_{ij} is a measurement error and $\mathbf{U}_i = (U_{i1}, \dots, U_{im_i})^T$ follows normal$(0, \mathbf{\Sigma}_{i,uu})$. Note that we have not assumed a distribution for the X_{ij}, and they may be correlated within the same cluster.

In some examples, the index j takes on a special meaning. For example, there could be $j = 1, \dots, m$ sampling times at which an individual is measured (e.g., in a panel study), or j could refer to a family member (e.g., mother, daughter). With some abuse of terminology, we call such situations "panel data" problems. In this special case it makes sense to distinguish among the values of j; for example, allowing different scale parameters, different density functions for the X's, or even different measurement error variances. Outside of this special case, with no meaning attached to j, it makes more sense to let the scale parameters, densities, and so on be independent of j. In what follows we do our calculations as if special meaning was attached to j, but all calculations cover the general case.

3. ESTIMATION WHEN THERE IS NO MEASUREMENT ERROR

3.1 Local Polynomial Kernel Estimators

For independent data, local polynomial kernel smoothing has been widely used in nonparametric regression. We now extend local polynomial kernel smoothing to model (1) for clustered data. To motivate the estimating equations for the kernel estimators of the nonparametric function $\theta(\cdot)$, we first consider estimating equations when $\theta(\cdot)$ is a parametric pth polynomial function $\theta(\cdot) = \mathbf{G}_p(\cdot)^T\boldsymbol{\beta}$, where $\mathbf{G}_p(z) = (1, z, \ldots, z^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ and $G_{ip} = \{\mathbf{G}_p(X_{i1}), \ldots, \mathbf{G}_p(X_{im_i})\}^T$. The regression coefficients $\boldsymbol{\beta}$ can be estimated using the conventional generalized estimating equations (GEEs) (Liang and Zeger 1986),

$$\sum_{i=1}^{n} \mathbf{G}_{ip}^T \boldsymbol{\Delta}_i \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \qquad (3)$$

where $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$ with its jth component $\mu_{ij} = \mu\{\mathbf{G}_p^T(X_{ij})\boldsymbol{\beta}\}, \boldsymbol{\Delta}_i = \text{diag}[\mu^{(1)}\{\mathbf{G}_p^T(X_{ij})\boldsymbol{\beta}\}], \mu^{(1)}(\cdot)$ is the first derivative of $\mu(\cdot), \mathbf{V}_i = \mathbf{S}_i^{1/2}\mathbf{R}_i(\boldsymbol{\delta})\mathbf{S}_i^{1/2}, \mathbf{S}_i = \text{diag}[\phi_j w_{ij}^{-1} V\{\mu_{ij}\}]$, and $\mathbf{R}_i$ is an invertible working correlation matrix, possibly depending a parameter vector $\boldsymbol{\delta}$, which can be estimated using the method of moments. Liang and Zeger (1986) showed that the GEE estimator $\hat{\boldsymbol{\beta}}$ is asymptotically consistent if the mean function μ_{ij} is correctly specified even when the working correlation matrix $\mathbf{R}_i$ is misspecified. The most efficient estimator of $\boldsymbol{\beta}$ is obtained by correctly specifying $\mathbf{R}_i$.

We now consider how to extend the parametric GEE (3) to (1) when $\theta(\cdot)$ is a nonparametric function using the kernel method. In what follows, the order of the local polynomial is p, the bandwidth is h, and the symmetric kernel density function is $K(\cdot)$, normalized without loss of generality to have unit variance. Let $K_h(v) = h^{-1}K(v/h)$. The idea is to approximate $\theta(\cdot)$ at any given x using a local polynomial satisfying $\theta(X) = \{\mathbf{G}_p(X - x)\}^T\boldsymbol{\beta}$, where $\mathbf{G}_p(\cdot)$ and $\boldsymbol{\beta}$ are as defined earlier. Having estimated $\boldsymbol{\beta}$ at x, the estimated $\theta(x)$ satisfies $\hat{\theta}(x) = \hat{\beta}_0$.

Let $\mathbf{G}_{ip}(x) = \{\mathbf{G}_p(X_{i1} - x), \ldots, \mathbf{G}_p(X_{im_i} - x)\}^T$. Kernel estimation of the nonparametric function $\theta(\cdot)$ at any given x requires incorporating the kernel weight function $K_h(\cdot)$ in GEE (3). Two ways are possible, and they give two sets of kernel estimating equations for $\theta(x)$,

$$\sum_{i=1}^{n} \mathbf{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x) \mathbf{V}_i(x)^{-1} \mathbf{K}_{ih}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\} = 0 \qquad (4)$$

or

$$\sum_{i=1}^{n} \mathbf{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x) \mathbf{K}_{ih}^{1/2}(x) \mathbf{V}_i^{-1}(x) \times \mathbf{K}_{ih}^{1/2}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\} = 0, \qquad (5)$$

where $\mathbf{K}_{ih}(x) = \text{diag}\{K_h(X_{ij} - x)\}$, and $\{\boldsymbol{\mu}_i(x), \boldsymbol{\Delta}_i(x), \mathbf{V}_i(x), \mathbf{S}_i(x)\}$ are the same as those in (3) except that they are evaluated at $\mu_{ij}(x) = \mu\{\mathbf{G}_p^T(X_{ij} - x)\boldsymbol{\beta}\}$. The working correlation matrix $\mathbf{R}_i$ in $\mathbf{V}_i(x)$ may depend on a parameter vector $\boldsymbol{\delta}$, which again can be estimated using the method of moments.

One can easily see that the two estimating equations (4) and (5) are often different except when $\mathbf{V}_i(x)$ is a diagonal matrix (assuming independence). Equation (5) weights the residuals $\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}$ symmetrically, whereas (4) does not. They hence often give different estimators of $\theta(x)$. We let $\hat{\theta}_p(x; h)$ denote the local pth-order kernel estimator using (4) and let $\hat{\theta}_p^*(x; h)$ denote the local pth-order kernel estimator using (5). These two estimators are identical when $\mathbf{V}_i(x)$ is a diagonal matrix. We show in Sections 3.2 and 3.3 that the symmetric property of (4) and (5) results in different asymptotic properties of $\hat{\theta}_p(x; h)$ and $\hat{\theta}_p^*(x; h)$.

We have allowed the scale parameters ϕ_j to depend on j. In many problems it is reasonable to suppose that they do not depend on j; then we can set $\mathbf{S}_i(x) = \text{diag}[w_{ij}^{-1} V\{\mu_{ij}(x)\}]$. If the ϕ_j do depend on j, then they will have to be estimated, again by the method of moments.

Application of the Fisher scoring algorithm to (4) shows that the estimator $\hat{\boldsymbol{\beta}}$ can be updated using iteratively reweighted least squares,

$$\left[\sum_{i=1}^{n} \mathbf{G}_{ip}(x)^T \mathbf{C}_i(x) \mathbf{G}_{ip}(x)\right] \hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{G}_{ip}(x)^T \mathbf{C}_i(x) \mathbf{y}_i, \qquad (6)$$

where $\mathbf{C}_i(x) = \boldsymbol{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\boldsymbol{\Delta}_i(x)$ is a working weight matrix and $\mathbf{y}_i = \mathbf{G}_{ip}(x)^T\boldsymbol{\beta} + \boldsymbol{\Delta}_i^{-1}(x)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}$ is a working vector. The variance of $\hat{\theta}_p(x; h)$ is equal to $\text{var}\{\hat{\beta}_0(x)\}$ and can be estimated using a sandwich estimator, which takes the form $\text{cov}\{\hat{\theta}_p(x; h)\} = \mathbf{e}^T\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{-1}\mathbf{e}$, where $\mathbf{e} = (1, 0, \ldots, 0)^T$ and

$$\boldsymbol{\Omega}_1 = \sum_{i=1}^{n} \mathbf{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x) \mathbf{V}_i^{-1}(x) \mathbf{K}_{ih}(x) \boldsymbol{\Delta}_i(x) \mathbf{G}_{ip}(x)$$

and

$$\boldsymbol{\Omega}_2 = \sum_{i=1}^{n} \mathbf{G}_{ip}(x)^T \boldsymbol{\Delta}_i(x) \mathbf{V}_i^{-1}(x) \mathbf{K}_{ih}(x) \{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\} \times \{\mathbf{Y}_i - \boldsymbol{\mu}_i(x)\}^T \mathbf{K}_{ih}(x) \mathbf{V}_i^{-1}(x) \boldsymbol{\Delta}_i(x) \mathbf{G}_{ip}(x).$$

A similar Fisher scoring algorithm can be constructed to solve (5) for $\hat{\theta}_p^*(x; h)$ and to calculate its variance. Specifically, one simply replaces $\mathbf{V}_i(x)\mathbf{K}_{ih}(x)$ by $\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i(x)\mathbf{K}_{ih}^{1/2}(x)$ in $(\mathbf{C}_i, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$.

Some versions of (4) have been proposed earlier. There are three obvious choices: (I) let $\mathbf{R}_i$ be an estimator of the actual correlation matrix; (II) let $\mathbf{V}_i^{-1}$ be the diagonal values of the inverse of the covariance matrix of $\mathbf{Y}_i$; and (III) let $\mathbf{R}_i$ be the identity matrix, thus effectively ignoring the correlation structure within clusters. We call method (III) the *weighted pooled estimator.* Method (I) was proposed by Severini and Staniswalis (1994) in their equation (18) for average kernel $p = 0$. Method (II) is a generalization, from the Gaussian case to generalized linear models, of the modified quasi-likelihood proposal of Ruckstuhl et al. (1999). Method (III) is a generalization and modification, from the Gaussian case to generalized linear models, of the "pooled" method of Ruckstuhl et al. (1999), allowing for different values of ϕ depending on the value of j. In Section 3.5 we consider another estimator called the "weighted average estimator."

Ruckstuhl et al. (1999) considered (4) for Gaussian data; that is, $w_{ij} \equiv 1$ and $V(\mu_{ij}) \equiv 1$. They showed that under the simple variance component model $\mathbf{V}_i = \phi\mathbf{I} + \delta\mathbf{J}$, where $\mathbf{I}$ is an identity matrix and $\mathbf{J}$ is a matrix of 1's, when $m_i \equiv m, p = 1$ so that local linear regression is used and the X_{ij}'s are iid, methods (II) and (III) are asymptotically equivalent and have uniformly smaller asymptotic mean squared errors than method (I). Methods (II) and (III) can also be shown to have faster rates of convergence for local quadratics, $p = 2$.

We study in the next two sections the asymptotic biases and variances of the general kernel estimators $\hat{\theta}_p(x;h)$ and $\hat{\theta}_p^*(x;h)$ under the kernel GEEs (4) and (5). This investigation will allow us to compare the asymptotic performance of methods (I)–(III) and to identify an *optimal* working correlation matrix $\mathbf{R}_i$. Our main conclusions from the asymptotic analyses are as follows:

1. The two kernel estimators $\hat{\theta}_p(x;h)$ and $\hat{\theta}_p^*(x;h)$ often have different asymptotic properties, and the asymptotic properties of $\hat{\theta}_p(x;h)$ is much harder to study.
2. Unlike the parametric GEE estimator in (3), if $\theta(x)$ is a nonparametric function, the asymptotically most efficient estimators of both $\hat{\theta}_p(x;h)$ and $\hat{\theta}_p^*(x;h)$ are obtained when ignoring the within-cluster correlation entirely; that is, assuming working independence $\mathbf{R}_i = \mathbf{I}_i$. Correctly specifying the correlation matrix in fact results in an asymptotically *less* efficient estimator of $\theta(x)$.

3.2 Asymptotic Theory for the Kernel Estimator $\theta_p(x;\ h)$ From (4)

Asymptotic bias and variance analysis of $\hat{\theta}_p(x;h)$ under (4) is often difficult for general local pth polynomial estimation, a general working correlation matrix $\mathbf{R}_i$ and non-Gaussian data. Hence for general working correlation matrix $\mathbf{R}_i$, we first focus on average kernel estimation ($p = 0$) for both Gaussian and non-Gaussian data (Theorem 1), and then study local linear kernel estimation ($p = 1$) for Gaussian data (Theorem 2). If working independence $\mathbf{R}_i = \mathbf{I}$ is assumed, then asymptotic bias and variance analysis of general local pth polynomial estimation for both Gaussian and non-Gaussian data is simple and is presented in Theorem 3.

In what follows, let $m_i = m < \infty$. We allow $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})^T$ to be correlated unless stated otherwise, and let $f_j(\cdot)$ denote the marginal density of X_{ij}. We further assume that the $(\mathbf{Y}_i, \mathbf{X}_i)$ $(i = 1, \ldots, n)$ are iid pairs with a continuous density function, and $\mathbf{V}_i(\boldsymbol{\mu}_i, \boldsymbol{\delta}) = \mathbf{V}(\boldsymbol{\mu}_i, \boldsymbol{\delta})$. Let $g^{(r)}(\cdot)$ denote the rth derivative of $g(\cdot)$, and let v^{jk} denote the (j,k)th element of $\mathbf{V}^{-1}(\cdot)$. Let $c_K(r) = \int z^r K(z)\, dz$, with $c_K(0) = c_K(2) = 1, \gamma_K(r) = \int z^r K^2(z)\, dz, \mathbf{E}_c(L) = \{c_K(L), c_K(L+1), \ldots, c_K(L+p)\}^T$, and $\mathbf{E}_p(c)$ and $\mathbf{E}_p(\gamma)$ the $(p+1)\times(p+1)$ matrices with (j,k) element $c_K(j+k-2)$ and $\gamma_K(j+k-2)$. We further assume that $nh \to \infty$ as $n \to \infty$ and $h \to 0$.

Theorem 1. Let $\hat{\theta}_0(x;h)$ be the solution of (4) for $p = 0$ and for any given weight matrix $\mathbf{V}_i$.

a. The asymptotic bias and variance of $\hat{\theta}_0(x;h)$ are given by

$$\text{bias}\{\hat{\theta}_0(x;h)\} \approx h^2 \left\{ \theta^{(1)}(x) \frac{\sum_{j=1}^m v^{j\cdot}(x) f_j^{(1)}(x)}{\sum_{j=1}^m v^{j\cdot}(x) f_j(x)} + \frac{\theta^{(2)}(x)}{2} \right\}$$

and

$$\text{var}\{\hat{\theta}_0(x;h)\} \approx \frac{\gamma_K(0)}{nh} \frac{\sum_{j=1}^m \{v^{j\cdot}(x)\}^2 \sigma_{jj}(x) f_j(x)}{[\mu^{(1)}\{\theta(x)\}]^2 \left\{\sum_{j=1}^m v^{j\cdot}(x) f_j(x)\right\}^2},$$

where $\sigma_{jj}(x) = \text{var}(Y_{ij}|X_{ij} = x) = w_j^{-1}\phi_j V[\mu\{\theta(x)\}]$, and $v^{j\cdot}(x) = \sum_{l=1}^m v^{jl}(x)$. If $f_j(\cdot) = f(\cdot)$, the bias of $\hat{\theta}_0(x;h)$ is free of $\mathbf{V}$.

b. The asymptotic variance of $\hat{\theta}_0(x;h)$ is minimized when one assumes the working correlation matrix $\mathbf{R} = \mathbf{I}$ (independence), and is equal to

$$\min_{\mathbf{V}}[\{\text{var}\{\hat{\theta}_0(x;h)\}] \approx \{\gamma_K(0)/nh\} \times \left([\mu^{(1)}\{\theta(x)\}]^2 \sum_{j=1}^m \{f_j(x)/\sigma_{jj}(x)\} \right)^{-1}.$$

The proof of Theorem 1 is given in Appendix A.1. We discuss the implication of Theorem 1 after presenting Theorem 2. For linear kernel estimation ($p = 1$), it is difficult to study asymptotic properties for general $\mathbf{V}$ and non-Gaussian data. This is because for any given weight matrix $\mathbf{V}$, asymptotic bias and variance analysis depends on the forms of $\mu(\cdot)$ and $V(\cdot)$. We hence concentrate on the Gaussian case and study in Theorem 2 its asymptotic bias and variance. The proof of Theorem 2 is given in Appendix A.2.

Theorem 2. Let $\hat{\theta}_{1,G}(x;h)$ be the solution of (4) for Gaussian data with $V(\cdot) = 1, w_{ij} = 1$, and $p = 1$ and any given weight matrix $\mathbf{V}$.

a. The asymptotic bias and variance of $\hat{\theta}_{1,G}(x;h)$ are $h^2\theta^{(2)}(x)/2$ and $c/(nh)$, where the expression of c is complicated and is given in Appendix A.2 and Ruckstuhl et al. (1999). Note that the asymptotic bias of $\hat{\theta}_{1,G}(x;h)$ is free of the distribution of the X_{ij} and $\mathbf{V}$.
b. If the $\mathbf{X}_{ij}$ are iid with common density $f(\cdot)$, the asymptotic variance of $\hat{\theta}_{1,G}(x;h)$ is minimized when one assumes the working correlation matrix $\mathbf{R} = \mathbf{I}$

(independence) and is equal to

$$\min_{\mathbf{V}}[\text{var}\{\hat{\theta}_{1,G}(x;h)\}]$$

$$\approx \{\gamma_K(0)/nh\}\left[f(x)\sum_{j=1}^{m}\{1/\sigma_{jj}\}\right]^{-1}.$$

where $\sigma_{jj} = \text{var}(Y_{ij}|X_{ij} = x) = \phi_j$.

Part (b) of Theorems 1 and 2 are the most important results. They suggest that at least for average kernel estimation $p = 0$ (Gaussian and non-Gaussian data) and for local linear kernel estimation $p = 1$ (Gaussian), it is optimal to simply assume independence for kernel regression using (4) for clustered data, and method (III) dominates methods (I) and (II). In other words, the asymptotically most efficient estimators $\hat{\theta}_0(x;h)$ and $\hat{\theta}_{1,G}(x;h)$ are obtained by completely ignoring the within-cluster correlation and correctly specifying the correlation results in less efficient estimators.

Study of the asymptotic properties of general local pth polynomial estimation under a general working correlation matrix $\mathbf{R}$ in (4) is difficult, even for Gaussian data. However, such calculations are possible when assuming independence $\mathbf{R} = \mathbf{I}$—that is, for the weighted pooled estimator [method (III)]. These results are stated in Theorem 3, whose proof is given in Appendix A.3.

Theorem 3. Let $\hat{\theta}_{p,\text{wpe}}(x;h)$ be the weighted pooled estimator; that is, the solution of (4) for any given p and $\mathbf{R} = \mathbf{I}$ (working independence). Then

a. The asymptotic bias of $\hat{\theta}_{p,\text{wpe}}(x;h)$ is

if $p = 0$,

$$\text{bias}\{\hat{\theta}_{0,\text{wpe}}(x;h)\}$$

$$\approx h^2\left\{\theta^{(1)}(x)\,\frac{\sum_{j=1}^{m} f_j^{(1)}(x)/\sigma_{jj}(x)}{\sum_{j=1}^{m} f_j(x)/\sigma_{jj}(x)} + \frac{\theta^{(2)}(x)}{2}\right\};$$

if $p =$ odd,

$$\text{bias}\{\hat{\theta}_{p,\text{wpe}}(x;h)\}$$

$$= h^{p+1}\,\frac{\theta^{(p+1)}(x)}{(p+1)!}\,\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1);$$

if $p =$ even and $p > 0$,

$$\text{bias}\{\hat{\theta}_{p,\text{wpe}}(x;h)\}$$

$$\approx h^{p+2}\left\{\frac{\theta^{(p+1)}(x)}{(p+1)!}\,\frac{\sum_{j=1}^{m}\partial\{L_j(x)f_j(x)\}/\partial x}{\sum_{j=1}^{m} L_j(x)f_j(x)} + \frac{\theta^{(p+2)}(x)}{(p+2)!}\right\}\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+2),$$

where $L_j(x) = [\mu^{(1)}\{\theta(x)\}]^2/\sigma_{jj}(x)$ and $\sigma_{jj}(x) = \text{var}(Y_{ij}|X_{ij} = x) = w_j^{-1}\phi_j V[\mu\{\theta(x)\}]$.

b. The asymptotic variance of $\hat{\theta}_{p,\text{wpe}}(x;h)$ is

$$\text{var}\{\hat{\theta}_{p,\text{wpe}}(x;h)\}$$

$$\approx \frac{\gamma_K(0)}{nh}\left([\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m} f_j(x)/\sigma_{jj}(x)\right)^{-1}$$

$$\times\, \mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\mathbf{e}. \tag{7}$$

Using the results in Theorem 3, one can easily show that, for example, the asymptotic bias and variance of the weighted pooled local linear kernel estimator $\hat{\theta}_{1,\text{wpe}}(x;h)$ are

$$\text{bias}\{\hat{\theta}_{1,\text{wpe}}(x;h)\} \approx h^2\theta^{(2)}(x)/2$$

and

$$\text{var}\{\hat{\theta}_{1,\text{wpe}}(x;h)\} \approx (\gamma_K(0)/nh)$$

$$\times\left([\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m}\{f_j(x)/\sigma_{jj}(x)\}\right)^{-1}.$$

3.3 Asymptotic Theory of the Kernel Estimator $\theta_p^*(x;\, h)$ Using (5)

We study in Theorem 4 the asymptotic bias and variance of $\hat{\theta}_p^*(x;h)$, the solution of the estimating equation (5), for a general local pth-order polynomial and a general weight matrix $\mathbf{V}_i$ for both Gaussian and non-Gaussian cases. Unlike $\hat{\theta}_p(x;h)$, whose bias and variance analysis under this general condition is difficult, such a general analysis is feasible for $\hat{\theta}_p^*(x;h)$, and the results are much simpler and are different from those of $\hat{\theta}_h(x;h)$. This is due to the symmetric nature of the estimating equation (5). These results allow us to easily study the *optimal* choice of the working correlation matrix $\mathbf{R}_i$.

The key result in Theorem 4 is given in part c; that is, the asymptotically most efficient estimator $\hat{\theta}_p^*(x;h)$ is obtained by entirely ignoring the within-cluster correlation and assuming that the data were independent. Note that under working independence, the two kernel estimators $\hat{\theta}_p(x;h)$ and $\hat{\theta}_p^*(x;h)$ are identical and have the same asymptotic properties. The proof of Theorem 4 is given in Appendix A.4.

Theorem 4. Suppose that $\int s^r K^{1/2}(s)\,ds < \infty$ for integers $r \le p$. Let $\hat{\theta}_p^*(x;h)$ be the solution of (5) for any given p and any given weight matrix $\mathbf{V}$.

a. The asymptotic bias of $\hat{\theta}_p^*(x;h)$ is

if $p = 0$,

$$\text{bias}\{\hat{\theta}_0(x;h)\}$$

$$\approx h^2\left\{\theta^{(1)}(x)\,\frac{\sum_{j=1}^{m} v^{jj}(x)f_j^{(1)}(x)}{\sum_{j=1}^{m} v^{jj}(x)f_j(x)} + \frac{\theta^{(2)}(x)}{2}\right\};$$

if $p =$ odd,

$$\text{bias}\{\hat{\theta}_p^*(x;h)\} \approx h^{p+1}\,\frac{\theta^{(p+1)}(x)}{(p+1)!}\,\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1);$$

(note that bias $\{\hat{\theta}_p^*(x;h)\}$ for odd p is free of $\mathbf{V}_i$ and $f_j(x)$); and

if $p =$ even and $p > 0$,

$$\begin{aligned}\text{bias}\{\hat{\theta}_p^*(x;h)\} \\ \approx h^{p+2}\left\{\frac{\theta^{(p+1)}(x)}{(p+1)!}\,\frac{\sum_{j=1}^m \partial\{T_j(x)f_j(x)\}/\partial x}{\sum_{j=1}^m T_j(x)f_j(x)}\right. \\ \left. + \frac{\theta^{(p+2)}(x)}{(p+2)!}\right\}\ \mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+2);\end{aligned}$$

where $T_j(x) = [\mu^{(1)}\{\theta(x)\}]^2 v^{jj}(x)$.

b. The asymptotic variance of $\hat{\theta}_p^*(x;h)$ is

$$\begin{aligned}\text{var}\{\hat{\theta}_p^*(x;h)\} \\ \approx \frac{\gamma_K(0)}{nh}\ \frac{\sum_{j=1}^m \{v^{jj}(x)\}^2\sigma_{jj}(x)f_j(x)}{[\mu^{(1)}\{\theta(x)\}]^2\left[\sum_{j=1}^m v^{jj}(x)f_j(x)\right]^2} \\ \times\ \mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\mathbf{e}.\end{aligned}$$

c. The asymptotic variance of $\hat{\theta}_p^*(x;h)$ is minimized when one assumes the working correlation matrix $\mathbf{R} = \mathbf{I}$ (independence), and is given in (7).

Part (c) of Theorem 4 gives the most important results. It suggests that under estimating equation (5), for any given local pth-order polynomial, the most efficient kernel estimator $\hat{\theta}_p^*(x;h)$ is obtained by simply assuming independence for kernel regression, and method (III) dominates methods (I) and (II). It is also interesting to notice that unlike estimating equation (4), methods (I) and (II) behave the same asymptotically under estimating equation (5). If $\mathbf{R} = \mathbf{I}$, then the results in Theorem 4 reduce to those in Theorem 3.

It is of interest to compare the asymptotic performance of $\hat{\theta}_p(x;h)$ and $\hat{\theta}_p^*(x;h)$ when a general weight matrix $\mathbf{V}$ is specified. Such a comparison is difficult for any given local pth-order polynomial. We hence restrict our attention to average kernel estimation ($p = 0$). The results in Theorems 1 and 4 suggest that $\hat{\theta}_0^*(x;h)$ replaces $v^{j\cdot}(x)$ in the bias and variance expressions of $\hat{\theta}_0(x;h)$ by $v^{jj}(x)$. Consider the case when $f_j(\cdot) = f(\cdot)$ and $\sigma_{jj}(x) = \sigma(x)$. If $\sum_{j=1}^m\{v^{jj}(x)\}^2/\{\sum_{j=1}^m v^{jj}(x)\}^2 < \sum_{j=1}^m\{v^{j\cdot}(x)\}^2/\{\sum_{j=1}^m v^{j\cdot}(x)\}^2$, then it is better to use $\hat{\theta}_0^*(x;h)$, which has a smaller variance. If $v^{jj}(x)$ and $v^{j\cdot}(x)$ do not depend on j (e.g., under exchangeable working correlation or AR(q) working correlation assumption), then $\theta_0(x;h)$ and $\hat{\theta}_0^*(x;h)$ have the same asymptotic variance. Furthermore, when $\mathbf{V}$ is a diagonal matrix [e.g., under methods (II) and (III)], $\hat{\theta}_p(x;h) = \hat{\theta}_p^*(x;h)$, and they have the same asymptotic properties.

3.4 Selection of the Bandwidth Parameter

An important step in kernel smoothing is to choose the bandwidth parameter h. One approach is to use cross-validation by deleting one cluster datum at a time, and then choose h to minimize

$$\text{CV}(h) = \sum_{i=1}^n\sum_{j=1}^{m_i}\frac{\{Y_{ij} - \hat{\mu}_{ij}^{(-i)}(X_{ij})\}^2}{\hat{\phi}_j w_{ij}^{-1}V\{\hat{\mu}_{ij}^{(-i)}(X_{ij})\}},$$

where $\hat{\mu}_{ij}^{(-i)}(\cdot)$ is the estimate of $\mu_{ij}(\cdot)$ calculated from the data leaving out the ith cluster. A difficulty in using cross-validation is that it is computationally intensive.

An alternative approach is to extend the Ruppert (1997) empirical bias bandwidth selection (EBBS) method to clustered data. Specifically, one calculates the empirical mean squared errors EMSE(x,h) of $\hat{\theta}(x;h)$ (either $\hat{\theta}_p(x;h)$ or $\hat{\theta}_p^*(x;h)$) at a series of values of x and h and chooses h to minimize EMSE(x,h) for each x. Calculations of EMSE(x_0,h_0) at any given value of x_0 and h_0 proceed by EMSE$(x_0,h_0) = \text{bias}^2\{\hat{\theta}(x_0;h_0)\} + \widehat{\text{var}}\{\hat{\theta}(x_0;h_0)\}$. Here $\widehat{\text{bias}}\{\hat{\theta}(x_0;h_0)\}$ denotes the empirical bias of $\hat{\theta}(x_0;h_0)$ at x_0 and h_0 and is estimated by fitting a polynomial regression,

$$E\{\hat{\theta}(x_0;h)\} = \nu_0 + \nu_1 h^{p+1} + \cdots + \nu_t h^{p+t}, \qquad (8)$$

using the "data" $\{h, \hat{\theta}(x_0;h)\}$ in a neighborhood of h_0 for a given integer t (e.g., $t = 1$ or 2). The empirical bias, $\widehat{\text{bias}}\{\hat{\theta}(x_0;h_0)\}$, is calculated as the estimated value of $\nu_1 h_0^{p+1} + \cdots + \nu_t h_0^{p+t}$. The variance, $\widehat{\text{var}}\{\hat{\theta}(x_0;h_0)\}$, can be easily calculated using the sandwich estimator in Section 3.1. We use this method in Section 5 to choose h when analyzing the ACSUS data.

3.5 Summary of Nonparametric Regression for Clustered Data

Our results in Sections 3.2 and 3.3 suggest that it is the best strategy to use (4) or (5) with $\mathbf{R} = \mathbf{I}$; that is, entirely ignoring the within-cluster correlation. The proposal is extremely easy to compute: simply pool the data and compute a standard local polynomial kernel estimator in generalized linear models (GLIMs), with weights depending on the cluster if the scale parameters ϕ_j are not constant.

In the "panel data" problem with $m_i \equiv m$, another estimator can be considered—to compute $\hat{\theta}_j(x;h)$ based only on the (Y_{ij}, X_{ij}) for fixed j, and then construct an optimal weighted average of the resulting estimators, where the optimal weights are the reciprocal of the var$\{\hat{\theta}_j(x;h)\}$. We call such an estimator the *weighted average estimator.* A simple generalization of the results of Ruckstuhl et al. (1999) shows that this estimator is asymptotically equivalent to method (III), the weighted pooled estimator. The key step in proving this result is to show that cov$\{\hat{\theta}_j(x;h), \hat{\theta}_{j'}(x;h)\} = O(n^{-1})$ $(j \neq j')$ is of smaller order compared to var$\{\hat{\theta}_j(x;h)\} = O\{(nh)^{-1}\}$. In other words, for asymptotic arguments, the individual estimators $\hat{\theta}_j(x;h)$ are independent.

It seems that the technique of constructing separate estimators and then pooling them could be complex, because asymptotically the optimal weights depend on the density functions of X_{ij} for $j = 1,\ldots,m$, which must then be estimated separately. In practice, this is not really that important an issue, because standard kernel methods allow estimation of variances (and hence weights) via such techniques as the sandwich method. As we show later, the extra complication in the no measurement error case of having to estimate weights can be worthwhile when there is measurement error, as the weighted average estimator is asymptotically more efficient than the weighted pooled estimator.

4. SIMEX LOCAL POLYNOMIAL ESTIMATION WHEN THERE IS MEASUREMENT ERROR

In this section we discuss extending the kernel methods in Section 3 to the case when the covariate X is measured with error under the additive measurement error model (2). We use the SIMEX method (Cook and Stefanski 1994) to correct measurement error. The results in Section 3 show that when X is accurately measured, it is the best strategy to entirely ignore the correlation and assume independence when calculating the kernel estimator of $\theta(x)$. In view of this result, we propose calculating the naive kernel estimator by assuming independence in the simulation step of the SIMEX method.

This approach leads to two SIMEX estimators of $\theta(x)$: the SIMEX weighted pooled estimator and the SIMEX weighted average estimator. The former calculates the naive weighted pooled estimators in the simulation step, whereas the latter calculates the naive weighted average estimators in the simulation step and can be applied only to the "panel data" case. The most interesting result we have found is that unlike in the no measurement error case, where the two estimators have the same asymptotic properties, the SIMEX weighted average estimator has a smaller asymptotic variance than the SIMEX weighted pooled estimator in the presence of measurement error. We describe local polynomial kernel estimation using SIMEX and propose the SIMEX weighted pooled estimator in Section 4.1, and study the asymptotic properties of this estimator in Section 4.2. We discuss the SIMEX weighted average estimator in Section 4.3.

4.1 The SIMEX Kernel Estimator

The SIMEX estimator was developed by Cook and Stefanski (1994). The idea behind the SIMEX method is seen most clearly in simple linear regression when the independent variable is subject to measurement error. Suppose that the regression model is $E(Y|X) = \alpha + \beta X$ and that $W = X + U$, rather than X, is observed where U has mean 0 and variance σ_u^2 and the measurement error variance σ_u^2 is known. It is well known that the ordinary least squares estimate of the slope from regressing Y on W converges to $\beta\sigma_x^2(\sigma_x^2 + \sigma_u^2)^{-1}$, where $\sigma_x^2 = \mathrm{var}(X)$.

For any fixed $\lambda > 0$, suppose that one repeatedly "adds on," via simulation, additional error with mean 0 and variance $\sigma_u^2\lambda$ to W, computes the ordinary least squares slope each time, and then takes the average. This simulation estimator consistently estimates $g(\lambda) = \beta\sigma_x^2/\{\sigma_x^2 + \sigma_u^2(1+\lambda)\}$. Because, formally at least, $g(-1) = \beta$, the idea is to plot $g(\lambda)$ against $\lambda \geq 0$, fit a model to this plot, and then extrapolate back to $\lambda = -1$. Cook and Stefanski (1994) showed that this procedure will yield a consistent estimate of β if one fits the model $g(\lambda) = \gamma_0 + \gamma_1(\gamma_2 + \lambda)^{-1}$.

The SIMEX estimator for nonparametric regression is constructed as follows. We discuss only the case where measurement error covariance matrices $\boldsymbol{\Sigma}_{i,uu}$ are known, and we keep track of these variances by means of the shorthand "$\boldsymbol{\Sigma}_{uu}$." In practice, the $\boldsymbol{\Sigma}_{i,uu}$ will have to be estimated, but estimating such parameters occurs at a parametric rate faster than the rate of convergence of any nonparametric estimator. Thus the theory is unchanged by estimating $\boldsymbol{\Sigma}_{uu}$.

Fix $D > 0$ to be a large but finite integer (50–200 in practice), and consider estimation of $\theta(x)$ in (1). For $d = 1, \ldots, D$ and any $\lambda > 0$, let $(\varepsilon_{ijd})_1^n$ be a set of independent standard normal random variables that are then transformed to have sample mean 0 and variance 1 and to be uncorrelated with the Y's and the W's. Let $\boldsymbol{\Sigma}_{i,uu}^{1/2}$ be the matrix square root of $\boldsymbol{\Sigma}_{i,uu}$. Define $\{W_{i1d}(\lambda), \ldots, W_{im_id}(\lambda)\}^T = \{W_{i1}, \ldots, W_{i,m_i}\}^T + \lambda^{1/2}\boldsymbol{\Sigma}_{i,uu}^{1/2}\{\varepsilon_{i1d}, \ldots, \varepsilon_{im_id}\}^T$. We calculate the GEE kernel estimator, which solves either (4) or (5), from these simulated data and denote it by $\hat{\theta}_d\{x, (1+\lambda)\boldsymbol{\Sigma}_{uu}\}$. The average of these estimates over $d = 1, \ldots, D$ is denoted by $\hat{\theta}\{x, (1+\lambda)\boldsymbol{\Sigma}_{uu}\}$. We run the SIMEX algorithm with D simulation replications at each value λ in a finite set $\boldsymbol{\Lambda}$. We extrapolate $\hat{\theta}\{x, (1+\lambda)\boldsymbol{\Sigma}_{uu}\}$ using a polynomial of order q_s back to $\lambda = -1$. This gives the SIMEX local polynomial estimator $\hat{\theta}_{sx}(x)$.

In view of the results in Section 3, we propose calculating the naive estimators $\hat{\theta}_d\{x, (1+\lambda)\boldsymbol{\Sigma}_{uu}\}$ using the weighted pooled estimator by assuming independence of observations within a cluster. The resulting estimator, called the SIMEX weighted pooled estimator, is denoted by $\hat{\theta}_{sx,\mathrm{wpe}}(x)$.

4.2 Asymptotic Theory for the SIMEX Weighted Pooled Estimator

The SIMEX estimator has a more complex theory for the weighted pooled estimator than in the independent data case where $m_i \equiv 1$, because the marginal distributions of W_{ij} and the conditional distributions of X_{ij} given W_{ij} may depend on j, for example, because the distributions of X_{ij} or the measurement error may depend on j. This means that the "naive" regression for Y_{ij} on W_{ij} ignoring measurement error may have a mean $\zeta_j(w, \boldsymbol{\Sigma}_{uu}) = E(Y_{ij}|W_{ij} = w)$ depending on j.

In the case where $m_i \equiv m$, the following is easily shown. Let $f_{jW}(\cdot, \boldsymbol{\Sigma}_{uu})$ be the marginal density of W_{ij}. Let $\phi_j(\boldsymbol{\Sigma}_{uu})$ be the limiting value of estimates of ϕ_j ignoring measurement error. Then the naive estimate of $\theta(w)$ converges to $\theta_N(w, \boldsymbol{\Sigma}_{uu})$ given by

$$\mu\{\theta_N(w, \boldsymbol{\Sigma}_{uu})\} = \left\{\sum_{j=1}^{m} \zeta_j(w, \boldsymbol{\Sigma}_{uu}) f_{jW}(w, \boldsymbol{\Sigma}_{uu})/\phi_j(\boldsymbol{\Sigma}_{uu})\right\} \times \left\{\sum_{j=1}^{m} f_{jW}(w, \boldsymbol{\Sigma}_{uu})/\phi_j(\boldsymbol{\Sigma}_{uu})\right\}^{-1}. \quad (9)$$

Let $\mathrm{s}(\lambda)$ be the (q_s+1)-vector with jth element λ^{j-1}, let $\mathbf{E}_s$ be the $(q_s+1) \times (q_s+1)$ matrix whose elements are 0 except that the first element equals 1, and let $\mathbf{q}^T(\mathbf{\Lambda}) = \mathbf{s}(-1)^T\{\sum_{\lambda\in\Lambda}\mathbf{s}(\lambda)\mathbf{s}^T(\lambda)\}^{-1}$. The results are unchanged, and the theory simplifies tremendously, if we assume that for each λ, the same bandwidth h_λ for all SIMEX replicates. In our theory we also require that the polynomial extrapolation is exact; that is, $\mathbf{q}^T(\mathbf{\Lambda})\sum_{\lambda\in\Lambda}\theta_N\{x;(1+\lambda)\mathbf{\Sigma}_{uu}\}\mathbf{s}(\lambda) = \theta(x)$. Hence the extrapolation results in a consistent estimate of $\theta(x)$. This is exactly true, of course, only in special cases. The bias that results from the extrapolation changes only the bias expression in the results given later, but not the variance expression.

Let the SIMEX weighted pooled estimator at x be denoted by $\hat\theta_{sx,\mathrm{wpe}}(x)$. The naive weighted pooled estimator that ignores measurement error is given by $\hat\theta_{N,\mathrm{wpe}}(x)$. Finally, define

$$Q(w,\mathbf{\Sigma}_{uu}) = \frac{\sum_{j=1}^m\{\mathcal{U}_j(w,\mathbf{\Sigma}_{uu})+\Gamma_j(w,\mathbf{\Sigma}_{uu})\} \times f_{jW}(w,\mathbf{\Sigma}_{uu})/\phi_j^2(\mathbf{\Sigma}_{uu})}{[\mu^{(1)}\{\theta_N(w,\mathbf{\Sigma}_{uu})\}\sum_{j=1}^m f_{jW}(w,\mathbf{\Sigma}_{uu})/\phi_j(\mathbf{\Sigma}_{uu})]^2},$$

where $\mathcal{U}_j(w,\mathbf{\Sigma}_{uu}) = [\zeta_j(w,\mathbf{\Sigma}_{uu}) - \mu\{\theta_N(w,\mathbf{\Sigma}_{uu})\}]^2$ and $\Gamma_j(w,\mathbf{\Sigma}_{uu}) = \mathrm{var}(Y_{ij}|W_{ij} = w)$. In Appendix A.5 we sketch an argument that gives the following approximate bias and variance expansions, assuming that the number of SIMEX replicates D is large. For simplicity, the bias expressions given here assume that p is odd:

$$\mathrm{bias}\{\hat\theta_{N,\mathrm{wpe}}(x)\} \approx \theta_N(x,\mathbf{\Sigma}_{uu}) - \theta(x) + h_0^{p+1}\theta_N^{(p+1)}\{x,\mathbf{\Sigma}_{uu}\} \times \{\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\},$$

$$\mathrm{bias}\{\hat\theta_{sx,\mathrm{wpe}}(x)\} \approx \frac{\mathbf{q}^T(\mathbf{\Lambda})}{(p+1)!}\sum_{\lambda\in\mathbf{\Lambda}} h_\lambda^{p+1}\theta_N^{(p+1)}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}\mathbf{s}(\lambda) \times \{\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}, \tag{10}$$

$$\mathrm{var}\{\hat\theta_{N,\mathrm{wpe}}(x)\} \approx (nh_0)^{-1}Q(x,\mathbf{\Sigma}_{uu})\{\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\mathbf{e}\}, \tag{11}$$

and

$$\mathrm{var}\{\hat\theta_{sx,\mathrm{wpe}}(x)\} \approx (nh_0)^{-1}Q(x,\mathbf{\Sigma}_{uu}) \times \{\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\mathbf{e}\mathbf{q}^T(\mathbf{\Lambda})\mathbf{E}_s\mathbf{q}(\mathbf{\Lambda})\}. \tag{12}$$

Equations (11) and (12) are the most surprising, because they say that the variance of the SIMEX estimate is asymptotically the same as if measurement error were ignored, but multiplied by the factor $\mathbf{q}^T(\mathbf{\Lambda})\mathbf{E}_s\mathbf{q}(\mathbf{\Lambda})$, a factor that is independent of the problem. Thus we can easily compare the various extrapolants on the basis of variance. For instance, suppose that the set of possible values of $\mathbf{\Lambda} = \{0, .5, 1.0, 1.5, 2.0\}$. Then direct calculation shows that use of the quadratic extrapolant leads to an estimator that is 9 times more variable than that based on the linear extrapolant, whereas the cubic extrapolant is 52 times more variable than the linear extrapolant. Of course, such increases in variance have to be balanced by decreases in bias, and it is our experience in other problems (Carroll, Maca, and Ruppert 1999) that the excess bias of the linear extrapolant is sufficiently large so that many times the quadratic extrapolant is preferred in terms of mean squared error.

Variance estimation of the SIMEX regression function can be performed in two ways. First, one can use the sandwich formula described previously to estimate the variance for the naive estimator which ignores measurement error, and then multiply it by the factor $\mathbf{q}^T(\mathbf{\Lambda})\mathbf{E}_s\mathbf{q}(\mathbf{\Lambda})$ in (12) to account for the extrapolation. An alternative method uses the sandwich formula and the SIMEX replicates (see Stefanski and Cook 1995, sec. 5.4).

4.3 The SIMEX Weighted Average Estimator

The weighted pooled estimator in Section 4.2 is applicable in great generality. In particular, different cluster sizes are easily accommodated, and a natural ordering is not required, so that the jth observation in one cluster is somehow linked with the j observation in any other cluster. However, when such a natural ordering exists, the fact is that the variance of the SIMEX weighted pooled estimator is inflated by the terms $\mathcal{U}_j(\cdot)$. These terms are an artifact, arising only because that although the regression of Y_{ij} on X_{ij} does not depend on j in the presence of measurement error, the regression of Y_{ij} on W_{ij} may exhibit such a dependence. It seems sensible, therefore, to explore circumstances under which less variable methods can be constructed.

One such circumstance occurs in the "panel data" problem with $m_i \equiv m$; for example, in a panel study where subjects are observed at the same time points. In such a situation, one could instead estimate the regression function $\theta(x)$ separately using SIMEX for each of $j = 1,\ldots,m$, and then average the estimates using some weights. Because each SIMEX estimate is an approximately consistent estimate, this device should in principle help us avoid an artificial variance inflation. We term the resulting estimator the SIMEX weighted average estimator and denote it by $\hat\theta_{sx,\mathrm{wae}}(x)$.

To see how this might work, suppose that the bandwidths in the jth observation are h_λ, the same as for the weighted pooled estimator. Then applying (12) but for a single observation, the asymptotic variance in the jth observation of the SIMEX estimate $\hat\theta_{sx,j}(x)$, is proportional to $(nh_0)^{-1}\Gamma_j(x,\mathbf{\Sigma}_{uu})\{[\mu^{(1)}\{\theta_j(x,\mathbf{\Sigma}_{uu})\}]^2 f_{jW}(x,\mathbf{\Sigma}_{uu})\}^{-1}$, where $\theta_j(x,\mathbf{\Sigma}_{uu}) = \mu^{-1}\{\xi_j(x,\mathbf{\Sigma}_{uu})\}$. The constant of proportionality is enclosed in brackets in (12). We construct the SIMEX weighted average estimator $\hat\theta_{sx,\mathrm{wae}}(x)$ as the optimal linear combination of the individual estimators as

$$\hat\theta_{sx,\mathrm{wae}}(x) = \sum_{j=1}^m \alpha_j\hat\theta_{sx,j}(x), \tag{13}$$

where $\alpha_j \propto \{[\mu^{(1)}\{\theta_j(x,\ \mathbf{\Sigma}_{uu})\}]^2 f_{jW}(x,\ \mathbf{\Sigma}_{uu})\}\{\Gamma_j(x,\mathbf{\Sigma}_{uu})\}^{-1}$ and $\sum_{j=1}^m\alpha_j = 1$. Assuming that the polynomial extrapolation is exact for each j—that is, $\mathbf{q}^T(\mathbf{\Lambda})\sum_{\lambda\in\mathbf{\Lambda}}\theta_j\{x;(1+\lambda)\mathbf{\Sigma}_{uu}\}s(\lambda) = \theta(x)$—the asymptotic bias of $\hat{\theta}_{sx,\mathrm{wae}}(x)$ is

$$\begin{aligned}\text{bias}\{\hat{\theta}_{sx,\mathrm{wae}}(x)\} \\ \approx \frac{\mathbf{q}^T(\mathbf{\Lambda})}{(p+1)!}\sum_{j=1}^{m}\sum_{\lambda\in\mathbf{\Lambda}}\alpha_j h_\lambda^{p+1}\theta_j^{(p+1)} \\ \times \{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}\mathbf{s}(\lambda)\{\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}. \quad (14)\end{aligned}$$

It is difficult to compare its bias with the bias of the SIMEX weighted pooled estimator $\hat{\theta}_{sx,\mathrm{wpe}}(x)$. However, if $h_\lambda = h$, assuming that the q_sth order polynomial extrapolation is exact for both $\hat{\theta}_{sx,\mathrm{wpe}}(x)$ and $\hat{\theta}_{sx,\mathrm{wae}}(x)$, then (10) and (14) are identical and are simplified as

$$\begin{aligned}\text{bias}\{\hat{\theta}_{sx,\mathrm{wpe}}(x)\} = \text{bias}\{\hat{\theta}_{sx,\mathrm{wae}}(x)\} \\ \approx \frac{h^{p+1}}{(p+1)!}\theta^{(p+1)}(x)\{\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)\}.\end{aligned}$$

This means that the asymptotic bias of the SIMEX estimators $\hat{\theta}_{sx,\mathrm{wpe}}(x)$ and $\hat{\theta}_{sx,\mathrm{wae}}(x)$ is the same as that when X is observed.

The variance of the weighted average estimator $\hat{\theta}_{sx,\mathrm{wae}}(x)$ is proportional to

$$\begin{aligned}\text{var}\{\hat{\theta}_{sx,\mathrm{wae}}(x)\} \\ \propto (nh_0)^{-1}\left\{\sum_{j=1}^{m}[\mu^{(1)}\{\theta_j(x,\mathbf{\Sigma}_{uu})\}]^2 \right. \\ \left. \times f_{jW}(x,\mathbf{\Sigma}_{uu})[\Gamma_j(x,\mathbf{\Sigma}_{uu})]^{-1}\right\}^{-1}, \quad (15)\end{aligned}$$

where again the constant of proportionality is enclosed in brackets in (12). The proof of (15) again has used the fact that the covariance $\text{cov}\{\hat{\theta}_{sx,j}(x),\hat{\theta}_{sx,j'}(x)\} = O(n^{-1})$ for $(j \neq j')$, which is of smaller order than $\text{var}\{\hat{\theta}_{sx,j}(x)\} = O\{(nh_0)^{-1}\}$. In other words, the individual SIMEX estimates $\hat{\theta}_{sx,j}(x)$ are independent asymptotically. In Appendix A.6 we show that the variance of the SIMEX weighted pooled estimator $\hat{\theta}_{sx,\mathrm{wpe}}(x)$ is greater than or equal to the variance of the SIMEX weighted average estimator $\hat{\theta}_{sx,\mathrm{wae}}(x)$. Of course, in the case that the distribution of (Y, W, X) is independent of j, the two expressions are equal.

Because of the complex nature of the bias expressions for SIMEX estimators, it is generally not possible to compare the SIMEX weighted pooled estimator and the SIMEX weighted average estimator in terms of mean squared error. However, when $h_\lambda = h$, such a comparison is possible, and our calculations suggest that the latter should be used if there are major observed differences as a function of j in the regression functions.

Because the weights used to calculate the SIMEX weighted average estimate $\hat{\theta}_{sx,\mathrm{wae}}(x)$ depend on the unknown density functions $f_{iW}(x,\mathbf{\Sigma})$ and the unknown conditional variances $\Gamma_j(x,\mathbf{\Sigma}_{uu})$, it is difficult to calculate $\hat{\theta}_{sx,\mathrm{wae}}(x)$ using (13) in practice. We hence propose the following procedure, which yields an asymptotically equivalent estimate. For the dth simulated SIMEX dataset, we first calculate the naive weighted average estimate $\hat{\theta}_{N,\mathrm{wae},d}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\} = \sum_{j=1}^{m}\alpha_{jd}\hat{\theta}_{N,jd}(x)\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$, where $\hat{\theta}_{N,jd}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$ is the naive kernel estimate using the simulated jth observation data $W_{ijd}(\lambda)$, and α_{jd} is the reciprocal of the variance of $\hat{\theta}_{N,jd}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$ obtained from standard kernel regression (e.g., the sandwich estimate). We then calculate the average of these estimates over $d = 1,\ldots,D$ and extrapolate it back to $\lambda = -1$. To compute the variance of the resulting estimate, we only need to calculate the variance of the weighted average estimate $\hat{\theta}_{N,\mathrm{wae},d}\{x,(1+\lambda)\mathbf{\Sigma}_{uu}\}$ using the sandwich method (see Sec. 3.1) and then apply the SIMEX standard error method of Stefanski and Cook (1995).

5. APPLICATION TO THE ACSUS DATA

We applied the proposed SIMEX local polynomial kernel method to analyzing the ACSUS data described in Section 1. Because the risk of hospitalization depends on various covariates, such as HIV status, treatments, race, and gender, and we allow only a single covariate in model (1), we limited our analysis to a subset of homogeneous subjects. Specifically, we restricted our attention to 273 white male patients who were HIV positive at entry into the study and were treated with antiretroviral drugs. The study participants were interviewed about every 3 months for about 18 months and were asked whether they had had hospital admissions (yes/no) during the interviews. The question of main interest was how the CD4 counts affected the risk of hospitalization. The total number of observations was 1,059, with each patient contributing from 1 to 6 observations over time. The major covariate of interest, CD4 count, ranged from 1 to 2,131, and 90% of these patients had CD4 count below 500. As discussed in Section 1, the CD4 counts were measured with error, because the most recent CD4 counts prior to each interview were subject to substantial lab errors. Because the investigator does not know in what fashion the risk of hospitalization decreases with the CD4 counts and is interested in identifying the form of the functional dependence (see Sec. 1 for discussion), we would like to make such dependence as flexible as possible by assuming a nonparametric function to properly identify the functional form. Note that the other covariates included interview time and age. Examination of the data suggests only slight dependence of the risk of hospitalization on time and age, and we did not include them in the model.

We fit model (1) using the logit link with a single covariate W defined as $W = \log$ (CD4/100), a transformation that reduces the marked skewness of CD4 counts. We assumed the measurement errors U_{ij} were independent and normally distributed with mean 0 and variance σ_u^2. However, W itself is left skewed, and so an assumption that X is normally distributed would be inappropriate. The power of the SIMEX idea is that no assumptions need be made about the distribution of X. To estimate the measurement error variance σ_u^2, one needs either a validation study or replicates of CD4 count measures. But these were not available in the ACSUS, and hence we were not able to estimate σ_u^2 using the ACSUS. We thus conducted a sensitivity analysis by assuming σ_u^2 equal to 1/4 and 1/2 of the variance of W; that is, assuming $\sigma_u^2 = .34$ and $\sigma_u^2 = .68$. Wulfsohn and Tsiatis (1995) estimated the measurement error variance of log(CD4) as $\sigma_u^2 = .39$ using data from a clinical trial conducted by Burroughs-Wellcome. Our assumption of $\sigma_u^2 = .34$ is similar to their finding. Following Wulfsohn and Tsiatis (1995), we assumed that the measurement errors were independent and the measurement error variance σ_u^2 was a constant. If we had validation or replication data, then we could of course assess the possibility of correlated measurement errors, additivity, constant measurement variance, and whether a different transformation of CD4 counts is required by using the techniques of Nusser, Carriquiry, Dodd, and Fuller (1996) and Eckert, Carroll, and Wang (1997).

Because different subjects had different numbers of observations, calculation of the weighted average estimate of $\theta(x)$ was difficult. We calculated the SIMEX weighted pooled estimate of $\theta(x)$, letting $\lambda = (0, 1.0, 1.5, 2.0)$. We used the EBBS method discussed in Section 2.4 to select the bandwidth parameter h for each simulated dataset and assumed $t = 2$ in (8). We further treated σ_u^2 as fixed and known. We used a quadratic extrapolation function in the SIMEX procedure and calculated the standard errors of the SIMEX estimates $\hat{\theta}_{sx,\mathrm{wpe}}(x)$ using the standard error estimation method of Stefanski and Cook (1995). The SIMEX method was applied with $D = 100$. Analysis of each simulated dataset including estimating the bandwidth parameter h took only 16 seconds on a SPARC Ultra.

Figures 1–3 plot the estimated $\theta(x)$ against x = log(CD4/100) with it 95% confidence intervals assuming $\sigma_u^2 = 0$ (naive estimate ignoring measurement error), $\sigma_u^2 =$ .34 and .68. The results suggest that the risk of hospitalization decreases as the CD4 count increases, but not in a linear fashion. It decreases more quickly when CD4 count is relatively low (CD4 < 14, log(CD4/100) < −2) or high (CD4 > 100, log(CD4/100) > 0) and is fairly stable when CD4 count takes middle values, for example, between 14 and 100. Ignoring measurement error clearly affects the estimated risk of hospitalization. The naive curve is attenuated toward 0 compared to the SIMEX curves, especially for small and large values of CD4 counts. As expected, an increase in the measurement error variance leads to more change in the SIMEX estimate.

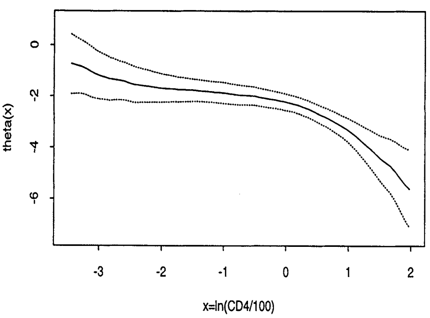


Figure 2. Estimated SIMEX/Kernel Estimate $\hat{\theta}(x)$ Assuming σ_u^2 = .34, Where x = ln(CD4/100) for the ACSUS Data and Its 95% Pointwise Confidence Intervals. ——— $\hat{\theta}(x)$; - - - CI.

As a further check on the results, instead of kernel regression, we fit the model by smoothing splines with the GAM procedure in S-PLUS by assuming independence for each simulated SIMEX data and calculated the SIMEX estimate of $\theta(x)$. The fitted model ignoring measurement error, as well as the two SIMEX fits, were well within accord with Figures 1–3.

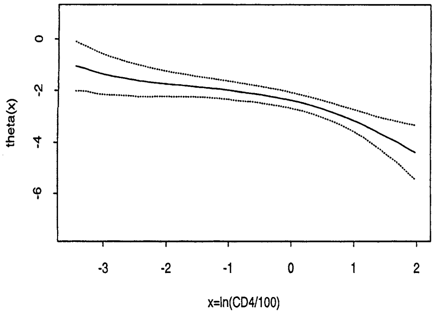


Figure 1. Estimated Naive Kernel Estimate $\hat{\theta}(x)$ That Ignores Measurement Error, Where x = ln(CD4/100) for the ACSUS Data and Its 95% Pointwise Confidence Intervals. ——— $\hat{\theta}(x)$; - - - CI.

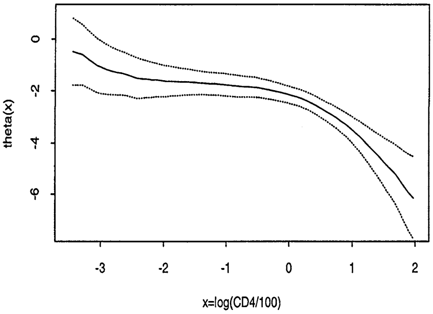


Figure 3. Estimated SIMEX/Kernel Estimate $\hat{\theta}(x)$ Assuming σ_u^2 = .68, Where x = ln(CD4/100) for the ACSUS Data and Its 95% Pointwise Confidence Intervals. ——— $\hat{\theta}(x)$; - - - CI.

To examine whether a simple parametric model can fit the data as well as the nonparametric model, we fit a simple linear model and a quadratic model using the GEE method assuming working independence (Liang and Zeger 1986) and calculated the SIMEX estimates to account for measurement error. For illustration, Figure 4 compares the SIMEX kernel estimate with the SIMEX linear and quadratic estimates when $\sigma_u^2 = .34$. Figure 4 shows that the SIMEX local polynomial kernel estimator seems to have nonlinearity detected neither by the linear model nor by the quadratic model. To test whether this extra nonlinearity is simply a figment of noise, we fit a cubic model to the data. Table I shows the naive and the SIMEX regression coefficient estimates of the cubic model assuming $\sigma_u^2 = (0, .34, .68)$, along with 95% bootstrap confidence intervals based on 2,000 bootstrap samples. The coefficient of the cubic term is marginally statistically significant in naive regression when measurement error is ignored, and it is statistically significant for both SIMEX analyses after accounting for measurement error.

6. DISCUSSION

We have discussed local polynomial kernel regression methods for clustered data in the absence/presence of measurement error. We have emphasized that our work is specific to the case of random regressors with a bounded number of observations per cluster, while the number of clusters becomes large. We developed two main results. First, in the absence of measurement error, methods based on ignoring within-cluster correlations generally improve on methods that attempt to use these correlations. Furthermore, correctly specifying correlation in estimation results in an asymptotically less efficient estimator. This is due mainly to the fact that kernel methods, being local, then essentially act as if the data were independent. A referee suggested that one might gain additional insight into the explanation of this result by considering a sequence of models wherein the within-cluster correlation approaches 1 as $n \to \infty$. It should be noted that our results in this article assume that the working covariance matrix V is invertible and distributions of $\mathbf{Y}_i$ and $\mathbf{X}_i$ are continuous, and they might not be applied directly to this situation. Our second main result is in the "panel data" context with measurement error, where it can be preferable to fit separate functions to each time period and then combine the methods via weighted averaging, rather than try to perform a single pooled measurement error analysis. For simplicity, we assume a single nonparametric function. We conjecture that our results are applicable to models involving several continuous nonparametric functions; for example, in the generalized additive model context.

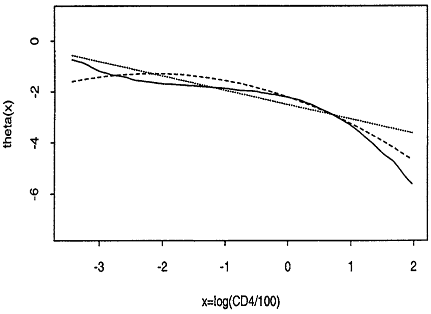


Figure 4. Comparison of the SIMEX Kernel Curve Estimate and the Linear and Quadratic Curve Estimates When σ_u^2 = .34. ——— SIMEX kernel estimate; --- SIMEX linear curve estimate; — — — SIMEX quadratic curve estimate.

Table 1. Naive and SIMEX Estimates of the Regression Coefficients of the Cubic Models for the ACSUS Data

	Naive	*SIMEX(σ_u^2 = .34)*	*SIMEX(σ_u^2 = .68)*
Intercept	−2.19	−1.84	−1.66
Linear	−.54	−.65	−.75
Quadratic	−.29	−.65	−.78
Cubic	−.06	−.14	−.17
Cubic, 2.5% bootstrap quantile	−.17	−.38	−.45
Cubic, 97.5% bootstrap quantile	.007	−.02	−.02

NOTE: The 4% and 96% bootstrap quantiles of the naive cubic term are −.153 and −.001.

Our results may have implications outside the realm of kernel smoothing—for example, to spline smoothing—because of the well-known "equivalent kernel" results of Silverman (1984). These results say that linear and cubic smoothing splines behave away from the boundary like a Nadaraya–Watson kernel regression estimator with a locally chosen bandwidth and a higher-order kernel. Using this equivalent kernel, our results on kernel smoothing suggest that even for splines, it may be more efficient statistically, and is certainly easier computationally, to ignore the correlation structure within clusters and simply compute a weighted smoothing spline for GLIMs with weights inversely proportional to the ϕ_j.

Our results thus may have a direct impact on recent very active developments in modeling longitudinal curve data using smoothing splines via a linear mixed-effects model formulation (Brumback and Rice 1998; Verbyla et al. 1999; Wang 1998). These authors account for the within-cluster correlation using random effects while estimating the nonparametric function using a smoothing spline. An advantage of this approach is that the smoothing spline estimators can be written as a linear combination of fixed effects and random effects, and hence an enlarged linear mixed model can be used to fit a linear random-effects smoothing spline model. But our results show that the smoothing spline estimator obtained in this way possibly could be asymptotically less efficient than that obtained by ignoring correlation. These suggestions are, of course, all conjectures, based on an equivalence in the nonclustered framework between local polynomial estimation and smoothing spline estimation. But it would appear important for smoothing spline methodologists to show explicitly that accounting for correlation within clusters is a worthwhile endeavor. We would not expect our results to apply to nonlinear random-effects smoothing spline models, such as generalized additive mixed models (Lin and Zhang 1999).

Our results do not, of course, apply to the time series context, where the predictors are the fixed observation times, with the number of such times converging to infinity. It is well known that one can construct estimators that take advantage of the autocorrelation structure in this case (Hart 1991), and the asymptotic variance of the estimator of the nonparametric function depends on the correlation function.

In view of the no measurement error results, we have considered in the measurement error case estimation of the nonparametric function using the SIMEX approach by ignoring the within-cluster correlation in calculating the naive kernel estimators in the simulation step. It is unclear whether this strategy is the best strategy; that is, whether ignoring correlation yields the most efficient SIMEX estimator. More research is needed, although we expect the theory to be extremely difficult.

An advantage of the SIMEX method is that it makes no distributional assumption on the unobserved covariate X. It is clearly of substantial interest for future work to develop methods that allow for an assumed parametric distribution for X. It is known (in models without correlated responses) that correct specification of a distribution for X can allow substantial gains in efficiency (Carroll et al. 1999), albeit at the price of a loss of robustness to misspecification of the distributions of X.

APPENDIX: THEORY FOR KERNEL METHODS

A.1 Proof of Theorem 1

For $p = 0$, a simple Taylor expansion of (4) shows that its solution $\hat{\beta}_0 = \hat{\theta}_0(x;h)$ satisfies $\hat{\theta}_0(x;h) - \theta(x) \approx B_n^{-1}A_n$, where

$$B_n = n^{-1}\sum_{i=1}^{n} \mathbf{1}^T \boldsymbol{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)\boldsymbol{\Delta}_i(x)\mathbf{1}$$

and

$$A_n = n^{-1}\sum_{i=1}^{n} \mathbf{1}^T \boldsymbol{\Delta}_i(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}(x)[\mathbf{Y}_i - \mu\{\theta(x)\}\mathbf{1}],$$

and $\mathbf{1}$ is an $m \times 1$ vector of 1's. Let $B = \lim_{n\to\infty} B_n$. The asymptotic bias of $\hat{\theta}_0(x;h)$ is $B^{-1}E(A_n)$ and the asymptotic variance of $\hat{\theta}_0(x;h)$ is $\mathrm{var}(A_n)/B^2$.

Specifically, some calculations give

$$\begin{aligned} B &= E\left\{\sum_{j=1}^{m} [\mu^{(1)}\{\theta(x)\}]^2 v^{j\cdot} K_h(X_j - x)\right\} \\ &= [\mu^{(1)}\{\theta(x)\}]^2 \sum_{j=1}^{m} v^{j\cdot} f_j(x) + O(h), \\ E(A_n) &= E\left\{\sum_{j=1}^{m} \mu^{(1)}\{\theta(x)\} v^{j\cdot} K_h(X_j - x)[\mu\{\theta(X_j)\} \right. \\ &\qquad \left. - \mu\{\theta(x)\}]\right\} \\ &= h^2[\mu^{(1)}\{\theta(x)\}]^2 \sum_{j=1}^{m} v^{j\cdot}\{f_j^{(1)}(x)\theta^{(1)}(x) \\ &\qquad + f_j(x)\theta^{(2)}(x)/2\} + o(h^2), \end{aligned}$$

and

$$\begin{aligned} \mathrm{var}(A_n) &\approx n^{-1}E\left\{\sum_{j=1}^{m}\sum_{l=1}^{m} [\mu^{(1)}\{\theta(x)\}]^2 v^{j\cdot} v^{l\cdot} \sigma_{jl} \right. \\ &\qquad \left. \times K_h(X_j - x)K_h(X_l - x)\right\} \\ &= \frac{\gamma_K(0)[\mu^{(1)}\{\theta(x)\}]^2}{nh} \sum_{j=1}^{m} \{v^{j\cdot}\}^2 \sigma_{jj} f_j(x) \\ &\qquad + o\{(nh)^{-1}\}, \end{aligned}$$

where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{Y}_i)$ and σ_{jl} is the (j,l)th element of $\boldsymbol{\Sigma}$. Part (a) follows immediately. A direct application of the Cauchy–Schwartz inequality leads to part (b).

A.2 Proof of Theorem 2

For part (a), see theorem 2 of Ruckstuhl et al. (1999). Here we prove part (b). The results of appendix A.3 of Ruckstuhl et al. (1999) show that when the X_j are independent with density $f_j(\cdot)$, the asymptotic variance of $\hat{\theta}_{1,G}(x)$ is the first diagonal element of $\mathbf{B}^{-1}\mathrm{cov}(\mathbf{A}_n)(\mathbf{B}^{-1})^T$, where

$$\mathbf{B} = \begin{bmatrix} B_{00} & hB_{01} \\ h^{-1}B_{01} & B_{11} \end{bmatrix}$$

and

$$\mathrm{cov}(\mathbf{A}_n) \approx \frac{\gamma_0}{nh}\begin{bmatrix} A_{00} & h^{-1}A_{01} \\ h^{-1}A_{01} & h^{-2}A_{11} \end{bmatrix}.$$

Here

$$\begin{aligned} B_{00} &= \sum_{j=1}^{m} v^{j\cdot} f_j(x) \qquad B_{01} = \sum_{j=1}^{m} v^{j\cdot} f_j^{(1)}(x) \\ B_{10} &= \sum_{j=1}^{m}\sum_{l\neq j} v^{jl}E(X_l - x)f_j(x) \\ B_{11} &= \sum_{j=1}^{m}\sum_{l\neq j} v^{jl}E(X_l - x)f_j^{(1)}(x) + \sum_{j=1}^{m} v^{jj} f_j(x) \end{aligned}$$

and

$$\begin{aligned} A_{00} &= \sum_{j=1}^{m} \{v^{j\cdot}\}^2 \sigma_{jj} f_j(x) \\ A_{01} &= \sum_{j=1}^{m} v^{j\cdot} \sigma_{jj} f_j(x) \sum_{l\neq j} v^{jl}E(X_l - x) \\ A_{11} &= \sum_{j=1}^{m} \sigma_{jj} f_j(x) \left[\sum_{l\neq j} \{v^{jl}\}^2 E(X_l - x)^2 \right. \\ &\qquad \left. + \left\{\sum_{l\neq j} v^{jl}E(X_l - x)\right\}^2\right]. \end{aligned}$$

Some calculations show that $\mathrm{var}\{\hat{\theta}_{1,G}(x;h)\} \approx M\gamma_K(0)/nhH^2$, where $M = A_{00}B_{11}^2 - 2A_{01}B_{11}B_{01} + A_{11}B_{01}^2$ and $H = B_{00}B_{11} -$

$B_{10}B_{01}$, which is

$$H = \left\{\sum_{j=1}^{m} v^{j\cdot} f_j\right\}\left\{\sum_{j=1}^{m} v^{jj} f_j\right\} + \left\{\sum_{j=1}^{m} v^{j\cdot} f_j^{(1)}\right\}$$
$$\times \left\{\sum_{j=1}^{m} v^{jj} f_j\right\} - \left\{\sum_{j=1}^{m} v^{j\cdot} f_j\right\}\left\{\sum_{j=1}^{m} v^{jj} f_j^{(1)}\right\}.$$

If the X_j are iid with common density $f(\cdot)$, some calculations show that H and M can be simplified as

$$H = \left(\sum_{j=1}^{m} v^{j\cdot}\right)\left(\sum_{j=1}^{m} v^{jj}\right) f^2(x),$$

$$M = f(x)\left[\left(\sum_{j=1}^{m} \{v^{j\cdot}\}^2\sigma_{jj}\right)\left(\sum_{j=1}^{m} v^{jj}\right)^2 \{f(x) - af^{(1)}(x)\}^2\right.$$
$$+ \left(\sum_{j=1}^{m} \{v^{jj}\}^2\sigma_{jj}\right)\left(\sum_{j=1}^{m} v^{j\cdot}\right)^2 \{af^{(1)}(x)\}^2$$
$$- 2\left(\sum_{j=1}^{m} v^{j\cdot}\right)\left(\sum_{j=1}^{m} v^{jj}\right)\left(\sum_{j=1}^{m} v^{j\cdot}v^{jj}\sigma_{jj}\right)$$
$$\times \{f(x) - af^{(1)}(x)\}af^{(1)}(x),$$
$$+ \left(\sum_{j=1}^{m} \sigma_{jj}\sum_{l\neq j} \{v^{jl}\}^2\right)\left(\sum_{j=1}^{m} v^{j\cdot}\right)^2$$
$$\left.\times \{f^{(1)}(x)\}^2 E(X-x)^2\right],$$

where $a = E(X - x)$. Noting that the last term of M is nonnegative, and using the Cauchy–Schwartz inequality for the first three terms, some calculations show that $M \geq f^3(x)|\sum_{j=1}^{m} v^{jj}||\sum_{j=1}^{m} v^{j\cdot}|\{\sum_{j=1}^{m} 1/\sigma_{jj}\}^{-1}$. It follows that $\operatorname{var}(\hat{\theta}_{1,G}(x;h)) \geq \gamma_K(0)\{nhf(x)\sum_{j=1}^{m} 1/\sigma_{jj}\}^{-1}$. This completes the proof of part (b).

A.3 Proof of Theorem 3

For simplicity, we provide the proof by assuming that p is odd. When p is even, bias calculations are similar but more complex (see App. A.4 and Carroll, Ruppert, and Welsh 1998). Let $\Psi(Y,s) = \{Y - \mu(s)\}\mu^{(1)}(s)/V(s)$. Then, using the techniques of Carroll et al. (1998), it can be shown that $\hat{\theta}_{p,\mathrm{wpe}}(x;h)$ has the expansion

$$\hat{\theta}_{p,\mathrm{wpe}}(x,h) - \theta(x)$$
$$\approx h^{p+1}\frac{\theta^{(p+1)}(x)}{(p+1)!}\,\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)$$
$$+ \left\{[\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m} f_j(x)/\sigma_{jj}\right\}^{-1}$$
$$\times n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{e}^T\mathbf{E}_p^{-1}(c)\phi_j^{-1}K_h(X_{ij}-x)$$
$$\times \mathbf{G}_p\{(X_{ij}-x)/h\}\Psi\{Y_{ij},\theta(X_{ij})\}, \qquad \text{(A.1)}$$

where $\sigma_{jj} = \phi_j w_j^{-1}V[\mu\{\theta(x)\}]$. The bias of $\hat{\theta}_{p,\mathrm{wpe}}(x,h)$ is the first term in (A.1), and the variance is

$$\operatorname{var}\{\hat{\theta}_{p,\mathrm{wpe}}(x,h)\}$$
$$\approx \gamma_K(0)(nh)^{-1}\left\{[\mu\{\theta(x)\}]^2\sum_{j=1}^{m} f_j(x)/\sigma_{jj}\right\}^{-1}$$
$$\times \mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\mathbf{e}.$$

A.4 Proof of Theorem 4

Reparameterize $\mathbf{G}_p(X_{ij}-x)$ in (5) as $\mathbf{G}_p\{(X_{ij}-x)/h\}$ and $\boldsymbol{\beta}$ as $\boldsymbol{\alpha}$ whose jth component $\alpha_j = h^j\theta^{(j)}(x)/j!$. Then $\hat{\theta}_p^*(x;h) = \hat{\alpha}_0$. A Taylor expansion of (5) gives $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \mathbf{B}_n^{-1}\mathbf{A}_n$, where

$$\mathbf{B}_n = n^{-1}\sum_{i=1}^{n}\mathbf{G}_{ip}^T(x)\boldsymbol{\Delta}_i(x)\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i^{-1}(x)$$
$$\times \mathbf{K}_{ih}^{1/2}(x)\boldsymbol{\Delta}_i(x)\mathbf{G}_{ip}(x)$$

and

$$\mathbf{A}_n = \sum_{i=1}^{n}\mathbf{G}_{ip}^T(x)\boldsymbol{\Delta}_i(x)\mathbf{K}_{ih}^{1/2}(x)\mathbf{V}_i^{-1}(x)\mathbf{K}_{ih}^{1/2}(x)(\mathbf{Y}_i - \boldsymbol{\mu}_i).$$

Because $(\mathbf{Y}_i, \mathbf{X}_i)$ are iid, we suppress the subscript i. Let $\mathbf{B} = \lim_{n\to\infty}\mathbf{B}_n = E\{\mathbf{G}_p^T(x)\boldsymbol{\Delta}(x)\mathbf{K}_h^{1/2}(x)\mathbf{V}^{-1}(x)\mathbf{K}_h^{1/2}(x)\boldsymbol{\Delta}(x)$ $\mathbf{G}_p(x)\}$. The (r_1, r_2)th component of $\mathbf{B}$ is

$$B_{r_1,r_2} = E\left\{\sum_{j=1}^{m}\sum_{l=1}^{m}\mu_j^{(1)}\mu_l^{(1)}v^{jl}K_h^{1/2}(X_j-x)K_h^{1/2}(X_l-x)\right.$$
$$\left.\times \left(\frac{X_j-x}{h}\right)^{r_1-1}\left(\frac{X_l-x}{h}\right)^{r_2-1}\right\},$$

where $\mu_j^{(1)} = \mu_j^{(1)}[\mathbf{G}_p^T\{(X_j-x)/h\}\boldsymbol{\alpha}]$. Some calculations give

$$B_{r_1,r_2} = \sum_{j=1}^{m}\int\{\mu_j^{(1)}\}^2 v^{jj}K(s_j)f_j(x+s_jh)s_j^{r_1+r_2-2}\,ds_j + o(h)$$
$$= [\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m} v^{jj}f_j(x)c_K(r_1+r_2-2) + o(h).$$

It follows that $\mathbf{B} = [\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m} v^{jj}f_j(x)\mathbf{E}_p(c) + o(h)$.

The rth component of $E(A_n)$ is

$$E\left[\sum_{j=1}^{m}\sum_{l=1}^{m}\{\mu_j^{(1)}\}^2 v^{jl}K_h^{1/2}(X_j-x)K_h^{1/2}(X_l-x)\left(\frac{X_j-x}{h}\right)^{r-1}\right.$$
$$\times \left\{\frac{h^{p+1}\theta^{(p+1)}(x)}{(p+1)!}\left(\frac{X_l-x}{h}\right)^{p+1}\right.$$
$$\left.\left.+ \frac{h^{p+2}\theta^{(p+2)}(x)}{(p+2)!}\left(\frac{X_l-x}{h}\right)^{p+2}\right\}\right] + o(h^{p+2})$$
$$= \sum_{j=1}^{m}\int\{\mu_j^{(1)}\}^2 v^{jj}K(s_j)f_j(x+s_jh)$$
$$\times \left\{\frac{h^{p+1}\theta^{(p+1)}(x)}{(p+1)!}s_j^{r+p} + \frac{h^{p+2}\theta^{(p+2)}(x)}{(p+2)!}s_j^{r+p+1}\right\}$$
$$+ o(h^{p+2}). \qquad \text{(A.2)}$$

If $p = 0$, then, noting that $c_K(2) = 1$, some calculations show that (A.2) becomes

$$h^2[\mu^{(1)}\{\theta(x)\}]^2 \times \left\{\theta^{(1)}(x)\sum_{j=1}^{m} v^{jj} f_j^{(1)}(x) + \frac{\theta^{(2)}(x)}{2}\sum_{j=1}^{m} v^{jj} f_j(x)\right\} + o(h^2).$$

If $p > 0$, then (A.2) becomes

$$h^{p+1}\,\frac{\theta^{(p+1)}(x)}{(p+1)!}\,[\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m} v^{jj} f_j(x) c_K(r+p) + h^{p+2}\left\{\frac{\theta^{(p+1)}(x)}{(p+1)!}\sum_{j=1}^{m}\frac{\partial[T_j(x)f_j(x)]}{\partial x} + \frac{\theta^{(p+2)}(x)}{(p+2)!}\sum_{j=1}^{m} T_j(x) f_j(x)\right\} c_K(r+p+1),$$

where $T_j(x) = [\mu^{(1)}\{\theta(x)\}]^2 v^{jj}(x)$. Noting that $c_K(s) = 0$ and that the $(1, s+1)$ elements of $\mathbf{B}$ and $\mathbf{B}^{-1}$ are 0 if s is odd, using $\text{bias}\{\hat{\theta}_p^*(x;h)\} = \mathbf{e}^T\mathbf{B}^{-1}E(\mathbf{A}_n)$, some calculations give the bias expressions of $\hat{\theta}_p^*(x;h)$ stated in Theorem 4.

To calculate the asymptotic variance of $\hat{\theta}_p^*(x;h)$, we first calculate $\text{cov}(\mathbf{A}_n)$ as

$$\text{cov}(\mathbf{A}_n) = \frac{1}{n}\,E(\mathbf{G}_p^T\boldsymbol{\Delta}\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}\boldsymbol{\Sigma}\mathbf{K}_h^{1/2}\mathbf{V}^{-1}\mathbf{K}_h^{1/2}\boldsymbol{\Delta}\mathbf{G}_p) + o\{(nh)^{-1}\},$$

where $\boldsymbol{\Sigma} = \text{cov}(\mathbf{Y}_i|\mathbf{X}_i = x\mathbf{1})$ and σ_{jk} is the (j,k)th element of $\boldsymbol{\Sigma}$. The (r_1, r_2)th component of the first term is

$$\begin{aligned} n^{-1}E&\left\{\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m}\sum_{q=1}^{m}\mu_j^{(1)}\mu_l^{(1)}v_{jk}\sigma_{kl}v_{lq}K_h^{1/2}(X_j - x)\right.\\ &\times K_h^{1/2}(X_k - x)K_h^{1/2}(X_l - x)K_h^{1/2}(X_q - x)\\ &\left.\times\left(\frac{X_j - x}{h}\right)^{r_1-1}\left(\frac{X_q - x}{h}\right)^{r_2-1}\right\}\\ =\ &(nh)^{-1}\sum_{j=1}^{m}\int\{\mu_j^{(1)}\}^2\{v^{jj}\}^2\sigma_{jj}K^2(s_j)f_j(x + s_j h)\\ &\times s_j^{r_1+r_2-2}\,ds_j + o\{(nh)^{-1}\}\\ =\ &(nh)^{-1}[\mu^{(1)}\{\theta(x)\}]^2\sum_{j=1}^{m}\{v^{jj}\}^2\sigma_{jj}f_j(x)\gamma_K(r_1 + r_2 - 2)\\ &+ o\{(nh)^{-1}\}. \end{aligned}$$

Using $\text{cov}\{\hat{\theta}_p^*(x;h)\} = \mathbf{e}^T\mathbf{B}^{-1}\text{cov}(\mathbf{A}_n)\mathbf{B}^{-1}\mathbf{e}$, we have the expression of $\text{cov}\{\hat{\theta}_p^*(x;h)\}$ as that given in part (b). A direct application of the Cauchy–Schwartz inequality gives part (c).

A.5 Distribution of the Weighted Pooled Estimator Under Measurement Error

To develop the SIMEX theory, we need an asymptotic expansion for the naive estimator. In the expressions that follow, the argument $(\cdot)$ refers to $G_p^T\{(W_{ij} - w)/h\}\boldsymbol{\beta}$, the argument $(\bullet)$ refers to $\theta_N(W_{ij}, \boldsymbol{\Sigma}_{uu})$, and the argument $(\circ)$ refers to $\theta_N(w, \boldsymbol{\Sigma}_{uu})$. The first $p+1$ terms of the Taylor series expansion of $\theta_N(W_{ij}, \boldsymbol{\Sigma}_{uu})$ about $\theta_N(w, \boldsymbol{\Sigma}_{uu})$ are given by $G_p^T\{(W_{ij} - w)/h\}\boldsymbol{\beta}$. We solve $\boldsymbol{\beta}$ by

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_{ij} - \mu(\cdot)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\,\mu^{(1)}(\cdot)K_h(W_{ij} - w) \times G_p\{(W_{ij} - w)/h\} = 0.$$

It is easily seen by a first-order Taylor expansion and using (9) that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = B_n^{-1}A_n$, where

$$\begin{aligned} B_n\ =\ &n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\{\mu^{(1)}(\cdot)\}^2}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\,K_h(W_{ij} - w)\\ &\times G_p\{(W_{ij} - w)/h\}G_p^T\{(W_{ij} - w)/h\}\\ \approx\ &\mathbf{E}_p(c)[\{\mu^{(1)}(\circ)\}^2/\mathrm{V}(\circ)]\sum_{\mathrm{j}=1}^{\mathrm{m}}\mathrm{f_{j}w(w}, \boldsymbol{\Sigma}_{\mathrm{uu}})/\phi_{\mathrm{j}}(\boldsymbol{\Sigma}_{\mathrm{uu}})\\ &+ o_p(1) = B + o_p(1), \end{aligned}$$

$$A_n = A_{n1} + A_{n2},$$

$$A_{n1} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_{ij} - \mu(\bullet)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\,\mu^{(1)}(\cdot) \times K_h(W_{ij} - w)G_p\{(W_{ij} - w)/h\},$$

and

$$A_{n2} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\mu(\bullet) - \mu(\cdot)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\cdot)}\,\mu^{(1)}(\cdot) \times K_h(W_{ij} - w)G_p\{(W_{ij} - w)/h\}.$$

It is easily seen that

$$A_{n2} \approx [\{\mu^{(1)}(\circ)\}^2/V\{\mu(\circ)\}]\{h^{p+1}\theta_N^{(p+1)}(w)/(p+1)!\} \times \mathbf{E}_c(p+1)\sum_{j=1}^{m} f_{jW}(w, \boldsymbol{\Sigma}_{uu})/\phi_j(\boldsymbol{\Sigma}_{uu}),$$

and hence that

$$\hat{\theta}_N(w) - \theta_N(w) \approx h^{p+1}\theta_N^{(p+1)}(w)\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_c(p+1)/(p+1)! + \mathbf{e}^T\mathbf{B}^{-1}\mathbf{A}_{n1}. \quad \text{(A.3)}$$

Remembering that $E(Y_{ij}|W_{ij} = w) = \zeta_j(w)$ and using (9), a tedious but straightforward calculation shows that $E(\mathbf{A}_{n1}) = 0$. Hence the first term in (A.3) is the bias expansion for the naive estimate.

It is also easily seen that we can replace the argument $(\circ)$ by $(\bullet)$ in the definition of A_{n1} leading to the expression $A_{n1} = A_{n11} + A_{n12}$, where

$$A_{n11} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{Y_{ij} - \zeta_j(W_{ij}, \boldsymbol{\Sigma}_{uu})}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\bullet)}\mu^{(1)}(\bullet)K_h(W_{ij} - w) \times G_p\{(W_{ij} - w)/h\}$$

and

$$A_{n12} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\zeta_j(W_{ij}, \boldsymbol{\Sigma}_{uu}) - \mu(\bullet)}{\phi_j(\boldsymbol{\Sigma}_{uu})V(\bullet)}\mu^{(1)}(\bullet)K_h(W_{ij} - w) \times G_p\{(W_{ij} - w)/h\}.$$

Because A_{n12} is a function only of the W's, these two terms are uncorrelated.

A direct calculation shows that A_{n1} has variance asymptotically equivalent to

$$\frac{\{\mu^{(1)}(\circ)\}^2\mathbf{E}_p(\gamma)}{nhV^2(\circ)} \times \sum_{j=1}^{m} \{\mathcal{U}_j(w, \boldsymbol{\Sigma}_{uu}) + \Sigma_j(w, \boldsymbol{\Sigma}_{uu})\} f_{jW}(w, \boldsymbol{\Sigma}_{uu})/\phi_j^2(\boldsymbol{\Sigma}_{uu}).$$

We have thus shown (11), namely that the variance of $\hat{\theta}_N(w, \boldsymbol{\Sigma}_{uu})$ is asymptotically

$$\operatorname{var}\{\hat{\theta}_N(w, \boldsymbol{\Sigma}_{uu})\} \approx (nh)^{-1}Q(w, \boldsymbol{\Sigma}_{uu})\mathbf{e}^T\mathbf{E}_p^{-1}(c)\mathbf{E}_p(\gamma)\mathbf{E}_p^{-1}(c)\mathbf{e}.$$

In the case where the (Y, X, W)'s are marginally identically distributed, although not necessarily independent, simplification occurs because $\mathcal{U}_j(w, \boldsymbol{\Sigma}_{uu}) = 0, \zeta_j = \mu(\theta_N)$ and none of the terms Γ_j, ϕ_j, or f_{jW} depends on j.

We are now in a position to verify (12). The expansion (A.3), with A_{n1} replaced by $A_{n11} + A_{n12}$, can be analyzed using the same techniques as used by Carroll et al. (1999). Because the calculations are similar, although tedious, in the interest of space we have chosen not to provide them here. The key step in the proof is to show that $\operatorname{var}\{\hat{\theta}_N(x; (1+\lambda)\boldsymbol{\Sigma}_{uu})\} = O\{(nhD)^{-1}\} + O(n^{-1})$ for $\lambda > 0$, which is of smaller order than $\operatorname{var}\{\hat{\theta}_N(x; \boldsymbol{\Sigma}_{uu})\} = O\{(nh)^{-1}\}$.

A.6 Comparison of the Variances of $\hat{\theta}_{sx,\mathrm{wpe}}(x)$ and $\hat{\theta}_{sx,\mathrm{wae}}(x)$

Using the Cauchy–Schwartz inequality, we have

$$\left\{\sum_{j=1}^{m} \frac{\Gamma_j(\cdot)f_{jW}(\cdot)}{\phi_j^2}\right\}\left\{\sum_{j=1}^{m} \frac{[\mu^{(1)}\{\theta_j(\cdot)\}]^2 f_{jW}(\cdot)}{\Gamma_j(\cdot)}\right\} \geq \left\{\sum_{j=1}^{m} \frac{|\mu^{(1)}\{\theta_j(\cdot)\}| f_{jW}(\cdot)}{\phi_j}\right\}^2 \geq \left\{\sum_{j=1}^{m} \frac{\mu^{(1)}\{\theta_j(\cdot)\} f_{jW}(\cdot)}{\phi_j}\right\}^2.$$

Using equation (9) and noting $\xi_j(\cdot) = \mu\{\theta_j(\cdot)\}$, the last term is $[\mu^{(1)}\{\theta_N(\cdot)\}\sum_{j=1}^{m} f_{jW}(\cdot)/\phi_j]^2$. We have

$$\frac{\sum_{j=1}^{m} \Gamma_j(\cdot)f_{jW}(\cdot)/\phi_j^2}{\left\{\mu^{(1)}\{\theta_N(\cdot)\}\sum_{j=1}^{m} f_{jW}(\cdot)/\phi_j\right\}^2} \geq \frac{1}{\sum_{j=1}^{m} [\mu^{(1)}\{\theta_j(\cdot)\}]^2 f_{jW}(\cdot)/\Gamma_j(\cdot)}.$$

Further noting that $\mathcal{U}_j(\cdot) \geq 0$, we have $\operatorname{var}\{\hat{\theta}_{sx,\mathrm{wpe}}(x)\} \geq \operatorname{var}\{\hat{\theta}_{sx,\mathrm{wae}}(x)\}$.

[Received November 1998. Revised October 1999.]

REFERENCES

Berk, M. L., Maffeo, C., and Schur, C. L. (1993), *Research Design and Analysis Objectives*, AIDS Cost and Services Utilization Survey Report No. 1, Rockville, MD: Agency for Health Care Policy and Research.

Brumback, B. A., and Rice, J. A. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves" (with discussion), *Journal of the American Statistical Association*, 93, 961–994.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Estimation in the Presence of Measurement Error," *Biometrika*, 86, 541–554.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Carroll, R. J., Ruppert, D., and Welsh, A. (1998), "Local Estimating Equations," *Journal of the American Statistical Association*, 93, 214–227.

Cook, J. R., and Stefanski, L. A. (1994), "Simulation–Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Eckert, R. S., Carroll, R. J., and Wang, N. (1997), "Transformations to Additivity in Measurement Error Models," *Biometrics*, 53, 262–272.

Hart, J. D. (1991), "Kernel Regression Estimation With Time Series Errors," *Journal of the Royal Statistical Society*, Ser. B, 53, 173–187.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, Y. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, X., and Zhang, D. (1999), "Inference in Generalized Additive Mixed Models Using Smoothing Splines," *Journal of the Royal Statistical Society*, Ser. B, 61, 381–400.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996), "A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions," *Journal of the American Statistical Association*, 91, 1440–1449.

Ruckstuhl, A. F., Welsh, A. H., and Carroll, R. J. (2000), "Nonparametric Function Estimation of the Relationship Between Two Repeatedly Measured Variables," *Statistica Sinica*, 10, 51–71.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Silverman, B. (1984), "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898–916.

Stefanski, L. A., and Cook, J. R. (1995), "Simulation–Extrapolation: The Measurement Error Jackknife," *Journal of the American Statistical Association*, 90, 1247–1256.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995), "Modeling the Relationship of Survival to Longitudinal Data Measured With Error: Applications to Survival and CD4 Counts in Patients With AIDS," *Journal of American Statistics Association*, 90, 27–37.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999), "The Analysis of Designed Experiments and Longitudinal Data Using Smoothing Splines" (with discussion), *Applied Statistics*, 48, 269–311.

Wang, Y. (1998), "Mixed-Effects Smoothing Spline ANOVA," *Journal of the Royal Statistical Society*, Ser. B, 60, 159–174.

Wild, C. J., and Yee, T. W. (1996), "Additive Extensions to Generalized Estimating Equation Methods," *Journal of the Royal Statistical Society*, Ser. B, 58, 711–725.

Wu, C. O., Chiang, C. T., and Hoover, D. R. (1998), "Asymptotic Confidence Regions for Kernel Smoothing of a Varying Coefficient Model With Longitudinal Data," *Journal of the American Statistical Association*, 93, 1388–1402.

Wulfsohn, M. S., and Tsiatis, A. A. (1997), "A Joint Model for Survival and Longitudinal Data Measured With Error," *Biometrics*, 53, 330–339.

Zeger, S. L., and Diggle, P. J. (1994), "Semi-Parametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Semiparametric Regression for Clustered Data Using Generalized Estimating Equations

Xihong LIN and Raymond J. CARROLL

We consider estimation in a semiparametric generalized linear model for clustered data using estimating equations. Our results apply to the case where the number of observations per cluster is finite, whereas the number of clusters is large. The mean of the outcome variable μ is of the form $g(\mu) = \mathbf{X}^{t}\boldsymbol{\beta} + \theta(T)$, where $g(\cdot)$ is a link function, $\mathbf{X}$ and T are covariates, $\boldsymbol{\beta}$ is an unknown parameter vector, and $\theta(t)$ is an unknown smooth function. Kernel estimating equations proposed previously in the literature are used to estimate the infinite dimensional nonparametric function $\theta(t)$, and a profile-based estimating equation is used to estimate the finite dimensional parameter vector $\boldsymbol{\beta}$. We show that for clustered data, this conventional profile-kernel method often fails to yield a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ along with appropriate inference unless working independence is assumed or $\theta(t)$ is artificially undersmoothed, in which case asymptotic inference is possible. To gain insight into these results, we derive the semiparametric efficient score of $\boldsymbol{\beta}$, which is found to have a complicated form, and show that, unlike for independent data, the profile-kernel method does not yield a score function asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$, even when the true correlation is assumed and $\theta(t)$ is undersmoothed. We illustrate the methods with an application to infectious disease data and evaluate their finite sample performance through a simulation study.

KEY WORDS: Asymptotics; Clustered data; Consistency; Efficiency; Generalized estimating equations; Kernel method; Longitudinal data; Nonparametric regression; Partially linear model; Profile method; Sandwich estimator; Semiparametric efficient score; Semiparametric efficiency bound.

1. INTRODUCTION

Clustered data arise in many fields of biomedical research, including longitudinal studies, intervention studies, and clinical trials. Parametric regression using generalized estimating equations (GEEs) (Liang and Zeger 1986) has become a popular practice for analyzing such data. It is well understood that the GEE estimators of regression coefficients are consistent when the mean function is correctly specified even when the within-cluster correlation structure is misspecified, and that the most efficient estimator is obtained by correctly specifying the within-cluster correlation. To allow for more flexible dependence of an outcome variable on covariates, there has been substantial recent interest in modeling covariate effects nonparametrically (Lin and Carroll 2000; Hoover, Rice, Wu, and Yang 1998; Wild and Yee 1996). Lin and Carroll (2000) showed that in contrast to parametric GEEs, when standard kernel methods are used, typically the most efficient estimator of the nonparametric function is obtained by completely ignoring the within-cluster correlation; correct specification of the correlation structure generally results in an asymptotically less efficient estimator.

In many instances, a semiparametric partially generalized linear regression model is more desirable than modeling every covariate effect nonparametrically. This model assumes that the mean of the outcome variable μ depends on some covariates $\mathbf{X}$ parametrically and on some other covariate T nonparametrically in the form $g(\mu) = \mathbf{X}^{T}\boldsymbol{\beta} + \theta(T)$, where $g(\cdot)$ is a link function, $\boldsymbol{\beta}$ is an unknown parameter vector, and $\theta(\cdot)$ is an unknown smooth function. This model specification is particularly appealing when the effects of $\mathbf{X}$ (e.g., treatment) are of major interest and the effects of T (e.g., confounders) are nuisance. This is because one can make inference on the effects of $\mathbf{X}$ while making minimal assumptions on the effects of T using a fully nonparametric function.

One example is the longitudinal infectious disease study considered in Section 8. This study involved 275 preschool age children who were reexamined every 3 months for 18 months for the presence of respiratory infection (yes/no) (Diggle, Liang, and Zeger 1994). The primary interest is to study the association between respiratory infection and vitamin A deficiency (yes/no), while accounting for several confounders including age. Examination of the distribution of the vertical strokes in Figure 3 suggests that the age effect departs dramatically from linearity; the vertical strokes indicate the ages for yes (top) and no (bottom).

Because the binary exposure of vitamin A deficiency is of main interest and the age effect is nuisance, we are interested in modeling the vitamin A deficiency effect while allowing the nuisance age effect to be modeled nonparametrically.

Several authors have considered such semiparametric regression models. A key challenge of estimation in this model is that it is composed of a finite dimensional parameter vector $\boldsymbol{\beta}$ and an infinite dimensional parameter $\theta(\cdot)$. Estimation for independent nonclustered data has been considered by Carroll, Fan, Gijbels, and Wand (1997), Hastie and Tibshirani (1990), and Severini and Staniswalis (1994). These authors used the kernel method to estimate $\theta(t)$ and the profile likelihood–based method to estimate $\boldsymbol{\beta}$. They showed that the estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent and semiparametric efficient (Bickel, Klaassen, Ritov, and Wellner 1993). For longitudinal data, Zeger and Diggle (1994) considered a semiparametric model with a nonparametric time trajectory and parametric covariate effects. They estimated $\theta(t)$ using a kernel method by ignoring the within-cluster correlation, and estimated $\boldsymbol{\beta}$ using weighted least squares by accounting for the within-cluster

Xihong Lin is Associate Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *xlin@sph.umich.edu*). Her research was supported by National Cancer Institute grant CA 76404. Raymond J. Carroll is Distinguished Professor, Departments of Statistics and Biostatistics and Epidemiology, Texas A&M University, College Station TX 77843 (E-mail: *carroll@stat.tamu.edu*). His research was supported by National Cancer Institute grant CA 57030, and by the Texas A&M Center for Environmental and Rural Health via National Institute of Environmental Health Sciences grant P30 ES09106. The authors thank the editor, the associate editor, and two referees for their helpful comments and suggestions.

Journal of the American Statistical Association
September 2001, Vol. 96, No. 455, Theory and Methods

correlation. They did not study the asymptotic properties of their method. Severini and Staniswalis (1994) extended their independent data results to clustered data using profile-kernel GEEs. They claimed that the estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent for any working correlation matrix specification. Zhang, Lin Raz, and Sowers (1998) considered a semiparametric linear mixed model and estimated the nonparametric function using a smoothing spline.

In this article we consider a marginal semiparametric regression model for clustered data with $\theta(t)$ estimated using kernel estimating equations and $\boldsymbol{\beta}$ estimated using profile-based estimating equations. Our estimating equations are similar to those of Severini and Staniswalis (1994) except that different working correlation matrices are allowed in the two sets of estimating equations, and local linear regression is used instead of local average kernel regression. The main focus of this article is to investigate whether it is possible to construct a $\sqrt{n}$-consistent and efficient estimator of $\boldsymbol{\beta}$ using the profile-kernel method. This work is motivated by our observation of the diametrically opposed asymptotic properties of parametric and certain nonparametric GEEs in terms of how to obtain the most efficient estimators, the former requiring correctly specifying the correlation and the latter requiring completely ignoring the correlation. Hence we are interested in investigating whether such different asymptotic behavior affects consistency and efficiency of the estimator of $\boldsymbol{\beta}$ in the semiparametric model using the conventional profile-kernel method. In particular, does correct specification of the within-cluster correlation still yield a $\sqrt{n}$-consistent and semiparametric efficient estimator of $\boldsymbol{\beta}$?

The results that we have obtained are surprising. To obtain a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ using the conventional profile-kernel method, one generally must either artificially undersmooth $\theta(t)$ or completely ignore the within-cluster correlation by assuming working independence in the profile-kernel estimating equations. Thus, if one accounts for within-cluster correlation using the profile-kernel method, then the standard bandwidth selection methods used for estimating $\theta(t)$, such as cross-validation, fail, the sandwich covariance estimator of the estimator of $\boldsymbol{\beta}$ fails, and the conventional hypothesis tests on $\boldsymbol{\beta}$ such as the Wald and Score tests fail. With undersmoothing or working independence, asymptotically correct inference about $\boldsymbol{\beta}$ becomes possible. To gain insight into these results, we derive the semiparametric efficient score of $\boldsymbol{\beta}$, which is found to have a complicated form, and show that unlike for independent data, the profile-kernel method does not yield a score function that is asymptotically equivalent to the semiparametric efficient score for $\boldsymbol{\beta}$, even when the true correlation is assumed and $\theta(t)$ is undersmoothed. Our main conclusion is that, unlike for independent data, the conventional profile-kernel method is not semiparametric efficient and must be modified in ad hoc ways (undersmoothing) or to be made less efficient (working independence) to even be made $\sqrt{n}$ consistent.

The article is organized as follows. In Section 2 we state the semiparametric model for clustered data and in Section 3 discuss estimation of $\theta(t)$ using kernel estimating equations previously proposed in the literature and of $\boldsymbol{\beta}$ using profile estimating equations. In Section 4 we study the asymptotic properties of the profile-kernel estimators of $\boldsymbol{\beta}$ and $\theta(t)$. In Section 5 we derive the semiparametric efficient score of $\boldsymbol{\beta}$ within a likelihood framework, and show that the conventional profile-kernel estimating equations of $\boldsymbol{\beta}$ often do not yield a score equation that is asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$. In Section 6 we discuss practical implications of our results. We illustrate the methods with a simulation study in Section 7 and an application to infectious disease data in Section 8. We conclude with a discussion is Section 9.

2. A SEMIPARAMETRIC MARGINAL MODEL

In this section we present the semiparametric regression model for clustered data. Suppose that the data consist of n clusters with the ith $(i = 1, \ldots, n)$ cluster having m_i observations. Let Y_{ij} and $(\mathbf{X}_{ij}, T_{ij})$ be the response variable and the covariates of the jth $(j = 1, \ldots, m_i)$ observation in the ith cluster, where $\mathbf{X}_{ij}$ is a $p \times 1$ vector and T_{ij} is a scalar. Given the covariates $\mathbf{X}_{ij}$ and T_{ij}, the mean and the variance of the outcome variable Y_{ij} are $E(Y_{ij}) = \mu_{ij}$ and $\operatorname{var}(Y_{ij}) = \phi w_{ij}^{-1} V(\mu_{ij})$, where ϕ is a scale parameter, w_{ij} is a known weight, and $V(\cdot)$ is a known variance function. The marginal mean μ_{ij} depends on $\mathbf{X}_{ij}$ and T_{ij} through a known monotonic and differentiable link function $g(\cdot)$,

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \theta(T_{ij}), \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\theta(\cdot)$ is an unknown smooth function. We model the effects of $\mathbf{X}$ $(p \times 1)$ parametrically and the effects of T nonparametrically, and treat the within-cluster correlation parameters as nuisance parameters. In particular, it is important to note the assumption (Pepe and Couper 1997) that

$$E(Y_{ij} | \mathbf{X}_{ij}, T_{ij}) = E\{Y_{ij} | \mathbf{X}_{ij}, T_{ij}, (\mathbf{X}_{ij}, T_{ik})_{k \neq j}\}, \tag{2}$$

an assumption also made implicitly by Lin and Carroll (2000). In matrix notation, denoting by $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^T$, $g(\boldsymbol{\mu}_i) = \{g(\mu_{i1}), \ldots, g(\mu_{im_i})\}^T$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, and $\mathbf{X}_i$, and $\boldsymbol{T}_i$ similarly, we have $g(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\theta}(\boldsymbol{T}_i)$. If model (1) does not include $\theta(T_{ij})$, then it reduces to the parametric generalized linear model considered by Liang and Zeger (1986). If model (1) does not include $\mathbf{X}_{ij}^T\boldsymbol{\beta}$, then it reduces to the nonparametric model considered by Lin and Carroll (2000). Severini and Staniswalis (1994) considered a model similar to (1)–(2).

It is important to emphasize that we are considering a marginal model for the clustered data through specification of mean and variance functions. This is in the spirit of GEE-type models (Liang and Zeger 1986). Except for Gaussian data, our marginal models need not be a full semiparametric likelihood specification.

3. PROFILE-KERNEL ESTIMATING EQUATIONS

In this section we develop kernel estimating equations for $\theta(t)$ and profile estimating equations for $\boldsymbol{\beta}$. The formulation of the profile estimating equation is similar to the score equation calculated using the conventional profile likelihood approach in parametric regression. We give the motivation of these estimating equations in Section 3.1, and describe their forms in Section 3.2.

3.1 Motivation of the Profile-Kernel Estimating Equations

To motivate the profile-kernel estimating equations for $\boldsymbol{\beta}$ and $\theta(t)$ under the semiparametric model (1), we first consider the GEEs for the parametric model

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T\boldsymbol{\beta}. \qquad (3)$$

Of course, (3) is a special case of (1) when $\theta(t)=0$. Liang and Zeger (1986) proposed estimating $\boldsymbol{\beta}$ using the estimating equations

$$\sum_{i=1}^{n} \frac{\partial\mu(\mathbf{X}_i\boldsymbol{\beta})^T}{\partial\boldsymbol{\beta}}\mathbf{V}_i^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}_i)$$

$$= \sum_{i=1}^{n}\mathbf{X}_i^T\boldsymbol{\Delta}_i\mathbf{V}_i^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}_i)=0, \qquad (4)$$

where $\boldsymbol{\mu}_i = E(\mathbf{Y}_i) = \mu(\mathbf{X}_i\boldsymbol{\beta})$ with the jth component $\mu_{ij} = \mu(\mathbf{X}_{ij}^T\boldsymbol{\beta}) = g^{-1}(\mathbf{X}_{ij}^T\boldsymbol{\beta}), \boldsymbol{\Delta}_i = \text{diag}\{\mu_{ij}^{(1)}\}$, $\mu^{(1)}(\cdot)$ is the first derivative of $\mu(\cdot)$, $\mathbf{V}_i = \mathbf{S}_i^{1/2}\mathbf{R}_i(\boldsymbol{\tau})\mathbf{S}_i^{1/2}$, $\mathbf{S}_i = \text{diag}[\phi w_{ij}^{-1}V\{\mu_{ij}\}]$ contains the marginal variances of the Y_{ij}, and $\mathbf{R}_i$ is an invertible working correlation matrix, possibly depending a parameter vector $\boldsymbol{\tau}$, which can be estimated using the method of moments. Liang and Zeger (1986) showed that the GEE estimator $\hat{\boldsymbol{\beta}}$ is asymptotically consistent if the mean function μ_{ij} is correctly specified even when the working correlation matrix $\mathbf{R}_i$ is misspecified. The efficient kernal estimator of $\boldsymbol{\beta}$ is obtained by specifying $\mathbf{R}_i$ as the true correlation matrix.

Now consider kernel estimating equations for the nonparametric model

$$g(\mu_{ij}) = \theta(T_{ij}). \qquad (5)$$

Lin and Carroll (2000) considered the pth local polynomial kernel estimating equations for $\theta(t)$. We consider here the local linear kernel estimator, that is, $p=1$. Let h denote the bandwidth parameter, and let $K(\cdot)$ denote the symmetric kernel density function. Let $K_h(v) = h^{-1}K(v/h)$ and $\boldsymbol{T}_i(t)$ be an $m_i \times 2$ matrix with the jth row $\{1, (T_{ij}-t)/h\}$. Lin and Carroll (2000) considered two kernel (symmetric and asymmetric) estimating equations for $\theta(t)$ at any t,

$$\sum_{i=1}^{n}\mathbf{T}_i(t)^T\boldsymbol{\Delta}_i(t)\mathbf{K}_{ih}^{1/2}(t)\mathbf{V}_i^{-1}(t)\mathbf{K}_{ih}^{1/2}(t)\{\mathbf{Y}_i-\boldsymbol{\mu}_i(t)\}=0 \qquad (6)$$

and

$$\sum_{i=1}^{n}\mathbf{T}_i(t)^T\boldsymbol{\Delta}_i(t)\mathbf{V}_i^{-1}(t)\mathbf{K}_{ih}(t)\{\mathbf{Y}_i-\boldsymbol{\mu}_i(t)\}=0, \qquad (7)$$

where $\mathbf{K}_{ih}(t) = \text{diag}\{K_h(T_{ij}-t)\}$ and $[\boldsymbol{\mu}_i(t), \boldsymbol{\Delta}_i(t),\ \mathbf{V}_i(t), \mathbf{S}_i(t)\}$ are the same as those defined in (4) except that they are evaluated at $\mu_{ij}(t) = \mu\{\alpha_0+\alpha_1(T_{ij}-t)/h\}$, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$ is a 2×1 vector of unknown parameters. Equation (7) was also considered by Severini and Staniswalis (1994) using the local average kernel ($p=0$). Having estimated $\boldsymbol{\alpha}$ at t as $\hat{\boldsymbol{\alpha}}$, the kernel estimator of $\theta(t)$ is $\hat{\theta}(t) = \hat{\alpha}_0$. The working correlation matrix $\mathbf{R}_i$ in $\mathbf{V}_i(t)$ may again depend on a parameter vector $\boldsymbol{\tau}$, which again can be estimated using the method of moments. The kernel estimators under (6) and (7) are different except when working independence is assumed; that is, $\mathbf{R}_i = \mathbf{I}$.

Lin and Carroll (2000) showed that the two estimators under (6) and (7) have different asymptotic properties; asymptotic properties of the kernel estimator under (7) are much harder to study. The most important results of Lin and Carroll (2000) are that, unlike the parametric GEE estimator in (4), typically the asymptotically most efficient kernel estimator of the nonparametric function $\hat{\theta}(t)$ using (6) and (7) is obtained by entirely ignoring the within-cluster correlation and pretending that the observations within the same cluster were independent; that is, assuming working independence $\mathbf{R}_i = \mathbf{I}$. Correctly specifying the correlation matrix in fact typically has adverse effects and results in an asymptotically less efficient estimator of $\theta(t)$.

In view of the opposite asymptotic behaviors of parametric and nonparametric regression, we are led to ask whether using the conventional kernel method to estimate $\theta(t)$ will affect $\sqrt{n}$ consistency and efficiency of the estimation of $\boldsymbol{\beta}$. For example, is it still possible to specify an appropriate working correlation matrix in estimating equations in the semiparametric model (1) to obtain consistent and efficient estimators of $\boldsymbol{\beta}$ and $\theta(t)$? The various combinations of working independence and true correlation structure can be entertained for the separate estimating equations for $\boldsymbol{\beta}$ and $\theta(t)$. We pursue this question using profile likelihood ideas. We propose the profile-kernel estimating equations for the semiparametric model (1) in the next section, and answer these questions in Section 4 by performing asymptotic analysis.

3.2 Profile-Kernel Estimating Equations for Semiparametric Model (1)

In this section we develop estimating equations for $\boldsymbol{\beta}$ and $\theta(t)$ in the semiparametric model (1). A main feature of (1) is that $\boldsymbol{\beta}$ is a finite-dimensional parameter vector and $\theta(t)$ is an infinite-dimensional parameter. For independent data when the mean and variance functions determine a distribution, (e.g., generalized linear models), if the kernel method is used to estimate $\theta(t)$, then the profile method yields a $\sqrt{n}$-consistent and semiparametric efficient estimator of $\boldsymbol{\beta}$ (Carroll et al. 1997; Severini and Staniswalis 1994). We hence use kernel estimating equations similar to (6) and (7) to estimate $\theta(t)$, and use profile estimating equations to estimate $\boldsymbol{\beta}$ by modifying (4). We call the resulting estimating equations profile-kernel estimating equations. In the light of the discussion at the end of Section 3.1, we allow the working correlation matrices to be different in the two sets of estimating equations. In the same spirit of parametric GEEs, our primary goal is to investigate whether we can construct a $\sqrt{n}$-consistent and semiparametric efficient estimator of $\boldsymbol{\beta}$ by assuming the true correlation matrix. Our secondary goal is to investigate whether we could also construct a consistent and efficient estimator of $\theta(t)$ at the conventional nonparametric rate.

If $\boldsymbol{\beta}$ is known, then we estimate $\theta(t)$ using one of the following estimating equations:

$$\sum_{i=1}^{n}\boldsymbol{T}_i(t)^T\boldsymbol{\Delta}_i(\mathbf{X}_i, t)\mathbf{K}_{ih}^{1/2}(t)\mathbf{V}_{2i}^{-1}(\mathbf{X}_i, t)\mathbf{K}_{ih}^{1/2}(t)$$

$$\times\ \{\mathbf{Y}_i-\boldsymbol{\mu}_i(\mathbf{X}_i, t)\}=0 \qquad (8)$$

or

$$\sum_{i=1}^{n} \boldsymbol{T}_i(t)^T \boldsymbol{\Delta}_i(\mathbf{X}_i,t)\mathbf{V}_{2i}^{-1}(\mathbf{X}_i,t)\mathbf{K}_{ih}(t)\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\mathbf{X}_i,t)\} = 0, \quad (9)$$

where $\mathbf{K}_{ih}(t)$, $\boldsymbol{\mu}_i(\mathbf{X}_i, t)$, $\boldsymbol{\Delta}_i(\mathbf{X}_i, t)$, $\mathbf{V}_{2i}(\mathbf{X}_i, t) = \mathbf{S}_i^{1/2}(\mathbf{X}_i, t) \times \mathbf{R}_{2i}\mathbf{S}_i^{1/2}(\mathbf{X}_i, t)$ are the same as those in (6) and (7) except that they are evaluated at $\mu_{ij}(\mathbf{X}_{ij}, t; \boldsymbol{\beta}) = \mu\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \alpha_0 + \alpha_1(T_{ij} - t)/h\}$. Having estimated $\boldsymbol{\alpha}$ at t as $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$, the kernel estimator of $\theta(t)$ is $\hat{\theta}(t; \boldsymbol{\beta}) = \hat{\alpha}_0(\boldsymbol{\beta})$. The working correlation matrix $\mathbf{R}_{2i}$ in $\mathbf{V}_{2i}(t)$ may again depend on a parameter vector $\boldsymbol{\tau}_2$, which can be estimated using the method of moments (Liang and Zeger 1986).

Estimation of $\boldsymbol{\beta}$ proceeds by solving the profile estimating equations obtained by modifying the parametric GEEs (4) and solving

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta} + \hat{\boldsymbol{\theta}}(\boldsymbol{T}_i; \boldsymbol{\beta})\}^T}{\partial \boldsymbol{\beta}} \mathbf{V}_{1i}^{-1}(\mathbf{X}_i, \boldsymbol{T}_i) \times [\mathbf{Y}_i - \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta} + \hat{\boldsymbol{\theta}}(\boldsymbol{T}_i; \boldsymbol{\beta})\}] = 0, \quad (10)$$

where $\hat{\boldsymbol{\theta}}(\mathbf{T}_i; \boldsymbol{\beta}) = \{\hat{\theta}(T_{i1}; \boldsymbol{\beta}), \ldots, \hat{\theta}(T_{im_i}; \boldsymbol{\beta})\}^T$, $\mathbf{V}_{1i}(\mathbf{X}_i, \boldsymbol{T}_i) = \mathbf{S}_i^{1/2}(\mathbf{X}_i, \boldsymbol{T}_i)\mathbf{R}_{1i}\mathbf{S}_i^{1/2}(\mathbf{X}_i, \boldsymbol{T}_i)$, and $\mathbf{S}_i(\mathbf{X}_i, \boldsymbol{T}_i) = \operatorname{diag}\{\phi w_{ij}^{-1} V [\mu\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \hat{\theta}(T_{ij}; \boldsymbol{\beta})\}]\}$, where $\mathbf{R}_{1i}$ is a working correlation matrix depending on a parameter vector $\boldsymbol{\tau}_1$ that could be estimated using the method of moments (Liang and Zeger 1986). For example, in panel data $\mathbf{R}_{1i} \equiv \mathbf{R}$ can be estimated by $n^{-1}\sum_{i=1}^{n} \mathbf{S}_i^{-1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{S}_i^{-1/2}$, where $\mathbf{r}_i = \mathbf{Y}_i - \boldsymbol{\mu}\{\mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\theta}}(\boldsymbol{T}_i; \hat{\boldsymbol{\beta}})\}$, where $\hat{\boldsymbol{\beta}}$ is computed from working independence. The estimators $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$ jointly solving (8) or (9), and (10) are termed profile-kernel estimators.

Our asymptotics assume that $(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$ are known, but in fact it can be shown that the results apply when they are estimated. Note that we allow the working correlation matrices $\mathbf{R}_{2i}$ in (8) or (9) and $\mathbf{R}_{1i}$ in (10) to be different. The estimator of Zeger and Diggle (1994) can be viewed as a special case of our profile-kernel estimators. They considered longitudinal Gaussian data and assumed working independence when estimating $\theta(t)$; that is, $\mathbf{R}_{2i} = \mathbf{I}$ and $\mathbf{R}_{1i}$ equal to the true correlation matrix when estimating $\boldsymbol{\beta}$. Severini and Staniswalis (1994) used (8) and (10) assuming the same working correlation matrices; that is, $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{R}_i$ or, equivalently, $\mathbf{V}_{1i} = \mathbf{V}_{2i} = \mathbf{V}_i$. Note that these authors considered local average kernel estimation instead of local linear kernel estimation as in (9). We study the asymptotic properties of the general profile-kernel estimators and these special cases in Section 4.

Our results are unexpected. Specifically, the key conclusions from our asymptotic analyses are as follows:

1. If standard smoothing is used, only when $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{I}$, i.e., assuming working independence, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent.
2. For other specifications of the working correlations $\{\mathbf{R}_{1i}, \mathbf{R}_{2i}\}$, including the case when $\mathbf{R}_{1i}$ is the true correlation matrix and any specification for $\mathbf{R}_{2i}$, except for special cases, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-inconsistent unless $\theta(t)$ is undersmoothed. When $\theta(t)$ is undersmoothed and the true correlation matrix is assumed, the resulting profile-kernel estimator $\hat{\boldsymbol{\beta}}$ is not semiparametric efficient.
3. Calculation of the semiparametric efficient estimator of $\boldsymbol{\beta}$ is complicated even in the multivariate Gaussian case: construction of the semiparametric efficient score requires solving a complicated Fredholm integral equation and estimating the multivariate joint distribution of $(\mathbf{X}, \boldsymbol{T})$.

4. ASYMPTOTIC RESULTS

In this section we study the asymptotic properties of the profile-kernel estimators $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$. We focus on the symmetric local linear kernel estimating equations (8) and the profile estimating equations (10). The reason that we focus on (8) instead of (9) in our asymptotic analysis is that the asymptotic properties of the estimators under (9) are difficult to study because of the asymmetric nature of (9) (Lin and Carroll 2000). However, we show that if one uses in (9) the local *average* kernel, which includes the existing estimators (Severini and Staniswalis 1994; Zeger and Diggle 1994) as special cases, then the resulting estimators have qualitatively similar asymptotic properties to those of $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$. In what follows, let $m_i = m < \infty$, (i.e., assuming finite cluster size) and let T be a continuous observation-level covariate (e.g., a time-varying covariate in longitudinal studies).

We allow the m components of $\mathbf{X}_i$ and $\boldsymbol{T}_i$ to be correlated unless stated otherwise and assume the density of $\boldsymbol{T}_i$ to be continuous. We further assume that the $(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{T}_i)$ $(i = 1, \ldots, n)$ are iid triplets and that both $\mathbf{V}_{1i}(\boldsymbol{\mu}_i, \boldsymbol{\tau}) = \mathbf{V}_1(\boldsymbol{\mu}_i, \boldsymbol{\tau})$ and $\mathbf{V}_{2i}(\boldsymbol{\mu}_i, \boldsymbol{\tau}) = \mathbf{V}_2(\boldsymbol{\mu}_i, \boldsymbol{\tau})$ are invertible. Let $d^{(r)}(\cdot)$ denote the rth derivative of any function $d(\cdot)$, let v^{jk} denote the (j, k)th element of a matrix $\mathbf{V}^{-1}$, and let $f_j(t)$ denote the marginal density of T_{ij}. Suppose that the kernel density function $K(\cdot)$ has mean 0 and unit variance; that is, $\int sK(s)du = 0$ and $\int s^2K(s) = 1$.

We first rewrite the profile estimating equations for $\boldsymbol{\beta}$ in (10) as

$$\sum_{i=1}^{n} \widetilde{\mathbf{X}}_i^T \boldsymbol{\Delta}(\mathbf{X}_i, \boldsymbol{T}_i)\mathbf{V}_{1i}^{-1}(\mathbf{X}_i, \boldsymbol{T}_i) \times [\mathbf{Y}_i - \boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta} + \hat{\theta}(\boldsymbol{T}_i; \boldsymbol{\beta})\}] = 0, \quad (11)$$

where $\widetilde{\mathbf{X}}_i = \mathbf{X}_i + \partial\hat{\theta}(\boldsymbol{T}_i; \boldsymbol{\beta})/\partial\boldsymbol{\beta}^T$ and $\boldsymbol{\Delta}(\mathbf{X}_i, \boldsymbol{T}_i) = \operatorname{diag}[\mu^{(1)}\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \hat{\theta}(T_{ij}; \boldsymbol{\beta})\}]$. Calculations in Appendix A show that, asymptotically, $\partial\hat{\theta}(t; \boldsymbol{\beta})/\partial\boldsymbol{\beta} = -W_2^{-1}(t)\mathbf{W}_2^x(t) + op(1)$, where, suppressing the index i denoting $\mu_l = \mu\{\mathbf{X}_l^T\boldsymbol{\beta} + \theta(t)\}$ $(l = 1, \ldots, m)$,

$$W_2(t) = \sum_{l=1}^{m} E\left[\left\{\mu_l^{(1)}\right\}^2 v_2^{ll} | T_l = t\right] f_l(t)$$

and

$$\mathbf{W}_2^x(t) = \sum_{l=1}^{m} E\left[\left\{\mu_l^{(1)}\right\}^2 v_2^{ll}\mathbf{X}_l | T_l = t\right] f_l(t).$$

It follows that $\widetilde{\mathbf{X}}_i = (\widetilde{\mathbf{X}}_{i1}, \ldots, \widetilde{\mathbf{X}}_{im})^T$, where $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij} - W_2^{-1}(T_{ij})\mathbf{W}_2^x(T_{ij})$. Using these results, in Result 1 we study the asymptotic distributions of $\{\hat{\theta}(t), \hat{\boldsymbol{\beta}}\}$. A sketch of its proof is given in Appendix A.

Result 1. Let $\{\hat{\theta}(t), \hat{\boldsymbol{\beta}}\}$ denote the solution of the profile-kernel estimating equations (8) and (10), where $\hat{\theta}(t) = \hat{\theta}(t; \hat{\boldsymbol{\beta}})$. Suppose that $h \propto n^{-\alpha}$, $1/5 \le \alpha \le 1/3$ and $n \to \infty$. We then have the following:

a. If $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent, [i.e., $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(1)$], then there is an asymptotically equivalent random variable such that

$$\text{bias}\{\hat{\theta}(t)\} \approx h^2 \theta^{(2)}(t)/2 \tag{12}$$

and

$$\operatorname{var}\{\hat{\theta}(t)\} \approx \frac{\gamma}{nh} \frac{\sum_{j=1}^m E[\{\mu_j^{(1)}\}^2 \{v_2^{jj}\}^2 \sigma_{jj} | T_j = t] f_j(t)}{\{\sum_{j=1}^m E[\{\mu_j^{(1)}\}^2 v_2^{jj} | T_j = t] f_j(t)\}^2}, \tag{13}$$

where $\sigma_{jj} = \operatorname{var}(Y_j | \mathbf{X}_j, T_j) = \phi w_j^{-1} V(\mu_j)$. It follows that $\operatorname{var}\{\hat{\theta}(t)\}$ is minimized when assuming working independence $\mathbf{R}_2 = \mathbf{I}$ and is

$$\operatorname{var}\{\hat{\theta}(t)\} \approx \frac{\gamma}{nh} \left\{ \sum_{j=1}^m E\left[\left\{\mu_j^{(1)}\right\}^2 \sigma_{jj}^{-1} | T_j = t \right] f_j(t) \right\}^{-1}. \tag{14}$$

b. The estimator $\hat{\boldsymbol{\beta}}$ converges in distribution: $\sqrt{n}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - h^2 b(\boldsymbol{\beta}, \theta)/2\} \to N(0, \mathbf{V}_{\boldsymbol{\beta}})$, where, suppressing the subscript i in each term inside the expectations,

$$\mathbf{b}(\boldsymbol{\beta}, \theta) = \{E(\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \mathbf{V}_1^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}})\}^{-1} E\{\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \mathbf{V}_1^{-1} \boldsymbol{\Delta} \theta^{(2)}(\mathbf{T})\},$$
$$\mathbf{V}_{\boldsymbol{\beta}} = \{E(\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \mathbf{V}_1^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}})\}^{-1} E\{(\mathbf{Z}_1 - \mathbf{Z}_2)^T \boldsymbol{\Sigma} (\mathbf{Z}_1 - \mathbf{Z}_2)\} \times \{E(\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \mathbf{V}_1^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}})\}^{-1},$$

$\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{Y} | \mathbf{X}, \boldsymbol{T})$ and $\mathbf{Z}_1 = \mathbf{V}_1^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}}$, and the jth row of $\mathbf{Z}_2$ is

$$\mathbf{Z}_{2j} = \mu_j^{(1)} v_2^{jj} \left\{ \sum_{k=1}^m \sum_{l=1}^m E\left[\widetilde{\mathbf{X}}_k \mu_k^{(1)} v_1^{kl} \mu_l^{(1)} | T_l = T_j) \right] \right\} \times W_2^{-1}(T_j) f_j(T_j).$$

c. If these two conditions—working independence is assumed in both (8) and (10), (i.e., $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{I}$) and $(\mathbf{X}_{ij}, T_{ij})$ have the same marginal density, [i.e., $f_j(\mathbf{X}_{ij}, T_{ij}) = f(\mathbf{X}_{ij}, T_{ij})$]—are satisfied, then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent; that is, the bias term $\mathbf{b}(\boldsymbol{\beta}, \boldsymbol{\theta}) = 0$ and $\sqrt{n}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\} \to N(0, \widetilde{\mathbf{V}}_{\boldsymbol{\beta}})$ in distribution, where, suppressing the subscript i in each term inside the expectations,

$$\widetilde{\mathbf{V}}_{\boldsymbol{\beta}} = \{E(\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}})\}^{-1} E\{\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}}\} \times \{E(\widetilde{\mathbf{X}}^T \boldsymbol{\Delta} \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\Delta} \widetilde{\mathbf{X}})\}^{-1},$$

and $\boldsymbol{\Sigma}_d$ is a diagonal matrix with the diagonal elements of $\boldsymbol{\Sigma}$, (i.e., σ_{jj}) on the diagonal.

d. For other specifications of the working correlation matrices $\mathbf{R}_{1i}$ and $\mathbf{R}_{2i}$, including the true correlation matrix, $\hat{\boldsymbol{\beta}}$ is often $\sqrt{n}$ inconsistent; that is, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \infty$ in distribution. However, if one assumes that $nh^4 \to 0$ [i.e., undersmooths $\theta(t)$], then for any specification of the working correlation matrices $\mathbf{R}_{1i}$ and $\mathbf{R}_{2i}$, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \mathbf{V}_{\boldsymbol{\beta}})$ in distribution.

In general, $\mathbf{V}_{\boldsymbol{\beta}}$ can be estimated by replacing terms in its expression by estimates of those terms. We conjecture that the bootstrap can also be used. The results in part a of Result 1 are similar to those of Lin and Carroll (2000) when the covariate $\mathbf{X}$ is absent in model (1), except that the variance of $\hat{\theta}(t)$ now involves conditional expectations of $\mathbf{X}_j$ given T_j. These results suggest that if the profile estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent, then $\hat{\theta}(t)$ is consistent and asymptotically normal at the regular nonparametric rate. The most efficient estimator of $\hat{\theta}(t)$ is obtained by completely ignoring the within-cluster correlation.

To see why the bias term $\boldsymbol{b}(\boldsymbol{\beta}, \theta) \neq 0$ for non-identity working correlation matrices, consider linear models for multivariate normal $\mathbf{Y}_i$. Suppose that the marginal density of $\{\mathbf{X}_{ij}, T_{ij}\}$ $(j = 1, \ldots, m)$ is the same. Then the jth component of $\widetilde{\mathbf{X}}$ is $\widetilde{\mathbf{X}}_j = \mathbf{X}_j - E(\mathbf{X}_j | T_j)$. It follows that the second term of $\boldsymbol{b}(\boldsymbol{\beta}, \theta)$ is $E\{\widetilde{\mathbf{X}}^T \mathbf{V}_1^{-1} \theta^{(2)}(\boldsymbol{T})\} = \sum_{j=1}^m \sum_{k=1}^m E\{C_{jk}(T_k) v_1^{jk} \theta(T_k)\}$, where $C_{jk}(T_k) = E(\mathbf{X}_j | T_k) - E\{E(\mathbf{X}_j | T_j) | T_k)\}$ is generally not equal to 0 except when $j = k$. This means that the bias term $\boldsymbol{b}(\boldsymbol{\beta}, \theta) \neq 0$ unless we assume working independence, (i.e., $\mathbf{R}_1 = \mathbf{I}$), or $E(\mathbf{X}_j | T_j, T_k) = E(\mathbf{X}_j | T_j)$ for any j, k (e.g., when $\mathbf{X}$ and $\boldsymbol{T}$ are independent).

Simple calculations show that for multivariate normal $\mathbf{Y}$, if $\mathbf{X}$ and $\boldsymbol{T}$ are independent, then $\hat{\boldsymbol{\beta}}$ in fact is $\sqrt{n}$ consistent for any arbitrary working correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$. Furthermore, as shown in Section 5, if one assumes $\mathbf{R}_{1i}$ equal to the true correlation matrix in (10) and working independence $\mathbf{R}_{2i} = \mathbf{I}$ in (8), then $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and semiparametric efficient, and $\hat{\theta}(t)$ is efficient as well. The foregoing independence assumption of $\mathbf{X}$ and $\boldsymbol{T}$ is strong and difficult to satisfy in practice if both covariates $\mathbf{X}$ and T are time-varying covariates. But if $\mathbf{X}$ contains only one-time covariates and T is time in longitudinal studies, then this condition is satisfied. Note that the outcome needs to be normally distributed for the foregoing results to hold. For non-Gaussian data, if the true correlation matrix is used, even when $\mathbf{X}$ and T are independent, then $\hat{\boldsymbol{\beta}}$ is still $\sqrt{n}$ inconsistent.

Result 1 assumes that $\theta(t)$ is estimated using the symmetric local linear kernel estimating equation (8). Severini and Staniswalis (1994) and Zeger and Diggle (1994) proposed slightly different estimators. They estimated $\theta(t)$ by replacing the *symmetric local linear* kernel estimating equation (8) with the *asymmetric local average* kernel estimating equation, which is obtained by letting $\mu(\mathbf{X}_{ij}, t) = \mu(\mathbf{X}_{ij}^T \boldsymbol{\beta} + \alpha_0)$ and replacing $\boldsymbol{T}_i(t)$ by $\mathbf{1}_i$ in (9). We denote these estimators by $\{\hat{\boldsymbol{\beta}}_*, \hat{\theta}_*(t)\}$. Specifically, Severini and Staniswalis (1994) assumed the same working correlation matrix in both $\theta(t)$ and $\boldsymbol{\beta}$ estimating equations, that is, $\mathbf{R}_{1i} = \mathbf{R}_{2i} = \mathbf{R}_i$. Zeger and Diggle (1994) considered Gaussian data and assumed $\mathbf{R}_{1i}$ equal to the true correlation and $\mathbf{R}_{2i} = \mathbf{I}_i$ (working independence). It can be shown that the asymptotic properties of

$\{\hat{\boldsymbol{\beta}}_*, \hat{\theta}_*(t)\}$ are similar to those of $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$ in Result 1, and that the conclusions are the same.

Computation. A Fisher–Sivring algorithm for computation for the working indendence estimation is given in Appenix C.

5. SEMIPARAMETRIC EFFICIENT SCORE

It is of substantial interest to understand why the profile-kernel estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ inconsistent when the true correlation matrix is used unless $\theta(t)$ is undersmoothed. One way to address this question is to define a likelihood function for $\mathbf{Y}_i$ and compare how the profile-kernel estimating equation (10) differs from the semiparametric efficient score for $\boldsymbol{\beta}$ (Bickel et al., 1993).

The motivation of this investigation is as follows. For independent data, (i.e., the cluster size $m = 1$), suppose that the distribution of the outcome Y belongs to the linear exponential family. If $\theta(t)$ is smoothed using standard kernel methods (e.g., cross-validation), then the profile-kernel estimating equation of $\boldsymbol{\beta}$ is asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$ (Carroll et al. 1997; Severini and Staniswalis 1994). The resulting profile estimator $\hat{\boldsymbol{\beta}}$ hence is $\sqrt{n}$ consistent and semiparametric efficient. If one uses an estimating equation for $\boldsymbol{\beta}$ asymptotically different from the semiparametric efficient score [e.g., by simply replacing $\tilde{\mathbf{X}}_i$ in (11) (simplified for $m = 1$) by $\mathbf{X}_i$], then the resulting estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ inconsistent unless $\theta(t)$ is undersmoothed (Rice 1986).

Our key findings in this section are as follows. First, the semiparametric efficient score of $\boldsymbol{\beta}$ for multivariate Gaussian data is complicated and requires solving the Fredholm integral equation of the second kind and estimating the joint distribution of $\mathbf{X}_i$ and $\boldsymbol{T}_i$. Second, if regular smoothing is used for estimating $\theta(t)$, then the profile-kernel score of $\boldsymbol{\beta}$ estimates the semiparametric efficient score with a nonzero bias. This explains why the profile-kernel estimator $\hat{\boldsymbol{\beta}}$ is often $\sqrt{n}$ inconsistent. Finally, when $\hat{\theta}(t)$ is undersmoothed, the profile-kernel estimator of $\boldsymbol{\beta}$ is $\sqrt{n}$ consistent but is still not semiparametric efficient, except for special cases.

We first derive the semiparametric efficient score of $\boldsymbol{\beta}$. We assume a constant cluster size $1 < m < \infty$ and suppress the index i. To understand the fundamental issues involved, we consider $\mathbf{Y}$ to be multivariate normal $N\{\mathbf{X}\boldsymbol{\beta} + \theta(\boldsymbol{T}), \mathbf{V}\}$, where $\theta(\boldsymbol{T}) = \{\theta(T_1), \ldots, \theta(T_m)\}^T$ and $\mathbf{V}$ is assumed known.

In Appendix B we show that the semiparametric efficient score of β is

$$\{\mathbf{X} - \boldsymbol{\varphi}_*(\boldsymbol{T})\}^T \mathbf{V}^{-1} \{\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \theta(\boldsymbol{T})\}, \tag{15}$$

where $\boldsymbol{\varphi}_*(\boldsymbol{T}) = \{\boldsymbol{\varphi}_*(T_1), \ldots, \boldsymbol{\varphi}_*(T_m)\}^T$, $\boldsymbol{\varphi}_*(T_j) = \{\varphi_{*1}(T_j), \ldots, \varphi_{*p}(T_j)\}^T$, and p is the dimension of $\boldsymbol{\beta}$. The semiparametric efficiency bound of $\boldsymbol{\beta}$ is $E\{[\mathbf{X} - \boldsymbol{\varphi}_*(\boldsymbol{T})]^T \mathbf{V}^{-1} [\mathbf{X} - \boldsymbol{\varphi}_*(\boldsymbol{T})]\}$. The function $\boldsymbol{\varphi}_*(t)$ solves

$$\sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} E\{[\mathbf{X}_j - \boldsymbol{\varphi}_*(T_j)] | T_k = t\} f_k(t) = 0, \tag{16}$$

where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_m)^T$, v^{jk} is the (j, k)th element of $\mathbf{V}^{-1}$, and $f_k(t)$ is the density of T_k. Simple calculations show that (16) can be written as the Fredholm integral equation of the second kind (Bronshtein and Semendyayev 1985, sec. 8.4)

$$\boldsymbol{\varphi}_*(t) + \int H(t, s) \boldsymbol{\varphi}_*(s) ds = \boldsymbol{q}(t), \tag{17}$$

where $H(t, s)$ and $\boldsymbol{q}(t)$ are defined as

$$H(t, s) = \frac{\sum \sum_{j \neq k} v^{jk} f(T_j = s, T_k = t)}{\sum_{j=1}^{m} v^{jj} f(T_j = t)}$$

and

$$q(t) = \frac{\sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} E(\mathbf{X}_j | T_k = t) f(T_k = t)}{\sum_{j=1}^{m} v^{jj} f(T_j = t)},$$

where $f(\cdot)$ denotes a density function.

If $H(t, s)$ is square-integrable, then (17) has only one solution, except when the eigenvalues of (17) contain -1 and its solution can be written as $\boldsymbol{\varphi}_*(t) = -\int \Gamma(t, s) \boldsymbol{q}(s) ds + \boldsymbol{q}(t)$, where $\Gamma(t, s)$ is called the resolvent kernel and can be written as the Fredholm series, $\Gamma(t, s) = \sum_{k=0}^{\infty} H_k(t, s) / \sum_{k=0}^{\infty} \delta_k$, with $\delta_0 = 0, H_0(t, s) = H(t, s)$, $\delta_k = k^{-1} \int H_{k-1}(t, t) dt$, and $H_k(t, s) = H_{k-1}(t, s)\delta_k - \int H(t, u) H_{k-1}(u, s) du$ (Bronshtein and Semendyayev, 1985, sec. 8.4.7). An alternative expression of $\Gamma(t, s)$ is given by the Neumann series (Bronshtein and Semendyayev 1985, sec. 8.4.6). The foregoing Fredholm series always converges but is of little use when numerically calculating $\boldsymbol{\varphi}_*(t)$, because in most cases the approximation is inadequate for small values of k. More useful is the Nyström method (Bronshtein and Semendyayev 1985, sec. 8.4.8).

The foregoing discussion suggests that construction of the semiparametric efficient score of $\boldsymbol{\beta}$ is complicated even in the multivariate normal case. One needs to solve the complicated integral equation (17), which requires estimating the pairwise joint densities of (T_j, T_k) and the pairwise conditional expectations $E(\mathbf{X}_j | T_k)$ when calculating $H(t, s)$ and $\boldsymbol{q}(t)$. However, in the special case when the marginal density of $(\mathbf{X}_j, T_j)$ is the same and $E(\mathbf{X}_j | T_j, T_k) = E(\mathbf{X}_j | T_k)$ (e.g., when $\mathbf{X}$ and T are independent), simple calculations show that the solution of (16) has the closed form $\boldsymbol{\varphi}_*(t) = E(\mathbf{X}_j | T_j = t)$.

We now study for multivariate Gaussian data how the semiparametric efficient score (15) asymptotically differs from the profile-kernel estimating equation of $\boldsymbol{\beta}$ in (10) when the working correlation matrix $\mathbf{R}$ is the true correlation matrix. Using the results in Appendix A, we can easily show that the profile estimating equation for $\boldsymbol{\beta}$ in (11) is asymptotically equivalent to

$$(\tilde{\mathbf{X}}^T \mathbf{V}^{-1} - \mathbf{Z}_2^T)\{\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \theta(\boldsymbol{T})\} + \tilde{\mathbf{X}}^T \mathbf{V}^{-1} \boldsymbol{\theta}^{(2)}(T) h^2 / 2, \tag{18}$$

where the jth component of $\tilde{\mathbf{X}}$ is $\tilde{\mathbf{X}}_j = \mathbf{X}_j - E(\mathbf{X}_j | T_j)$ and $\mathbf{Z}_2$ is defined in Result 1. A comparison between (15) and (18) suggests that they are often different and that (18) is often subject to a nonzero bias. Even when $\theta(t)$ is undersmoothed [i.e., the second bias term in (18) is 0], some calculations show that the first term in (18) is still generally different from (15). In other words, the profile-kernel score (10) is often asymptotically different from the semiparametric efficient score (15). But when $\mathbf{X}$ and T are independent, they are the same asymptotically, and the profile-kernel estimator of $\hat{\boldsymbol{\beta}}$ hence is $\sqrt{n}$ consistent and semiparametric efficient. Some calculations show that the same conclusion holds for the profile-kernel estimator $\hat{\boldsymbol{\beta}}_*$ when $\hat{\theta}_*(t)$ is the average kernel estimator obtained using the asymmetric kernel estimating equation (9); see Section 4.

It is difficult to construct the semiparametric efficient score directly using the complicated form of $\varphi_*(t)$ in (15), because this involves theoretical density functions and expectations. This raises an open question on how to construct a practical semiparametric efficient estimator of $\boldsymbol{\beta}$. It is a reasonable conjecture that if such a construction is pushed through, then undersmoothing will not be required.

6. PRACTICAL IMPLICATIONS OF THE THEORETICAL RESULTS AND COMPUTATION OF THE ESTIMATES

Cross-Validation. Conventional bandwidth selection techniques, such as cross-validation by deleting one cluster data at a time, fail unless working independence is assumed. Because the bandwidth h chosen by cross-validation satisfies $h = O(n^{-1/5})$, $\hat{\boldsymbol{\beta}}$ will be $\sqrt{n}$ inconsistent unless working independence is assumed (Result 1). Unfortunately, there is no generally accepted data-driven way to choose h to undersmooth $\theta(t)$, although ad hoc methods have been proposed (Brockmann, Gasser, and Herrmann 1993). In our experience, we have found that multiplying the bandwidth by $n^{-2/15}$, which makes $h \propto n^{-1/3}$, often works quite well in practice. Presumably, other methods (e.g., higher-order kernels, twicing) can be used to eliminate the bias.

Sandwich Method. The sandwich method, which is commonly used in calculating the covariance estimator of $\hat{\boldsymbol{\beta}}$ in estimating equations (Liang and Zeger 1986), will give an inconsistent estimator of cov$(\hat{\boldsymbol{\beta}})$ unless working independence is assumed. This is because it ignores the extra $\mathbf{Z}_2$ term in $\mathbf{V}_{\beta}$ in part b of Result 1. This is true even when one undersmoothes $\theta(t)$. We conjecture that the bootstrap can be used.

Hypothesis Testing. One is often interested in testing H_0 : $\boldsymbol{\beta} = 0$ or part of $\boldsymbol{\beta}$ is 0. If conventional smoothing techniques such as cross-validation are used, then the Wald test and the score test for H_0 will be inconsistent unless working independence is assumed or $\theta(t)$ is undersmoothed. For example, when the Wald test is used, $\hat{\boldsymbol{\beta}}$ in fact estimates the true $\boldsymbol{\beta}$ plus the bias term $b(\boldsymbol{\beta}, \theta)h^2/2$.

Functional Data Analysis. The simplest functional regression model (Ramsay and Dalzell 1991) is $Y_i(t) = \theta(t) + \epsilon_i(t)$, where i indexes the ith subject, t indexes time t, and $\epsilon_i(t)$ is an error whose distribution is a Gaussian process with mean 0 and cov$\{e(t), e(s)\} = \sigma(t, s)$. Rice and Silverman (1991) considered estimating $\theta(t)$ using a smoothing spline. The results of Lin and Carroll (2000) suggest that the most efficient estimator of $\theta(t)$ when the kernel method is used is obtained by entirely ignoring the correlation of the repeated measures of $Y_i(t)$ over time. In the presence of covariates $\mathbf{X}_i(t) = \{X_{i1}(t), \ldots, X_{ip}(t)\}^T$, a semiparametric functional regression model could be considered,

$$Y_i(t) = \mathbf{X}_i(t)^T\boldsymbol{\beta} + \theta(t) + \epsilon_i(t). \tag{19}$$

The semiparametric model (1) is a discrete version of (19).

Suppose that the profile-kernel method is used to estimate $\{\boldsymbol{\beta}, \theta(t)\}$. Our results suggest that (a) if $\mathbf{X}_i(t)$ is a vector of one-time subject-level covariates (i.e., $\mathbf{X}_i(t) = \mathbf{X}_i$ free of t), by specifying $\mathbf{R}_1$ as the true correlation matrix and $\mathbf{R}_2 = \mathbf{I}$, $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and semiparametric efficient and $\hat{\theta}(t)$ is asymptotically efficient as well, and (b) if $\mathbf{X}_i(t)$ contains time-varying covariates (i.e., $\mathbf{X}$ and T are not independent), then one must assume working independence ($\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$) or undersmooth $\hat{\theta}(t)$ to obtain a $\sqrt{n}$ consistent (but inefficient) estimator of $\hat{\boldsymbol{\beta}}$.

It is important to emphasize that our results assume that the number of observations per subject m is finite, as is common in longitudinal studies. With T being time, our asymptotic analysis thus assumes that observations from different subjects may be observed at different time points asymptotically, but the number of observations per subject remains bounded.

Computation. A Fisher-Sivring algorithm for computation for the working indendence estimator is given in Appendix C.

7. SIMULATION STUDY

We conducted a simulation study to evaluate the finite-sample performance of the profile-kernel method. Each dataset comprised $n = 100$ subjects and $m_i = 3$ observations per subject over time. The covariate vector $\mathbf{X}_{ij}$ was set at $\mathbf{X}_{ij} = (X_{1ij}, X_{2i})^T$, where X_{1ij} a time-varying covariate and X_{2i} is a subject level covariate that takes value 1 for half of the subjects and 0 for the other half and mimics a binary treatment indicator. We generated X_{1ij} and T_{ij} according to the model $X_{1ij} = b_i + e_{ij}$ and $T_{ij} = b_i + e'_{ij}$, where $b_i \sim$ uniform$(-1, 1)$ and e_{ij} and e'_{ij} are independent and follow uniform$(-1, 1)$. This setup allows the X_{1ij} and the T_{ij} to be correlated with each other and over time between their repeated measures with exchangeable correlation .5. Conditional on $\mathbf{X}_{ij}$ and T_{ij}, we generated the outcome Y_{ij} from multivariate normal with mean $\mu_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2i} + \theta(T_{ij})$, where $\beta_1 = \beta_2 = 1.0$ and $\theta(t) = \sin(2t)$, and Y_{ij} has variance 1 and exchangeable correlation .5.

We generated 200 datasets with $N = 300$ observations each and analyzed them using the profile-kernel methods. For each simulated dataset, we first assumed working independence when we calculated the profile-kernel estimate of $\boldsymbol{\beta}$ and $\theta(t)$ and estimated the bandwidth parameter h needed for the kernel estimate of $\theta(t)$ using cross-validation by deleting one subject data at a time. We next calculated the profile-kernel estimate of $\boldsymbol{\beta}$ and $\theta(t)$ by accounting for the within-subject correlation. Specifically, we estimated the true covariance of $\mathbf{Y}_i$ using the method of moments and calculated the bandwidth parameter h by multiplying the cross-validation bandwidth estimate by $n^{-2/15}$. This undersmooths $\theta(t)$ and eliminates the bias term (Sec. 6), at least theoretically.

Table 1 gives the averaged estimated regression coefficients of β_1 and β_2, along with their empirical and estimated standard errors (SEs) when working independence is assumed and when the true covariance of $\mathbf{Y}_i$ is estimated. When assuming working independence, we estimated the SEs of $\hat{\boldsymbol{\beta}}$ using the sandwich estimate given in Appendix C. When assuming that the true covariance is estimated, we estimated the SEs of $\hat{\boldsymbol{\beta}}$ using a finite-sample estimate of $\mathbf{V}_{\beta}$ given in part b of

Table 1. Means and Standard Errors of Regression Coefficient Estimates Over 200 Replications

	Working independence			True covariance		
Parameter	Mean	Empirical SE	Estimated SE	Mean	Empirical SE	Estimated SE
β_1	1.005	.088	.084	1.002	.075	.070
β_2	1.020	.160	.160	1.022	.161	.158

NOTE: True values are $\beta_1 = 1.0$ and $\beta_1 = 1.0$.

Result 1. Table 1 reports the averages of the estimated standard errors over 200 replications. The results in the table show that the profile-kernel method performs well in finite samples and that the biases in the profile-kernel estimates of $\boldsymbol{\beta}$ are minimal under both covariance assumptions. The estimate of $\boldsymbol{\beta}_1$, the coefficient of the time-varying covariate X_1, is more efficient when the true covariance is estimated than when working independence is assumed. However, no gain in efficiency is realized in $\hat{\boldsymbol{\beta}}_2$ by estimating the true covariance of $\mathbf{Y}_i$. This is because X_2 is a subject-level covariate and is independent of T_{ij} and the design is balanced with respect to X_2. The simulation results are consistent with the theory. The estimated SEs of $\hat{\boldsymbol{\beta}}$ also agree well with the simulated SEs.

Figure 1 compares the true nonparametric function $\theta(t)$ to the kernel estimates of $\theta(t)$ when assuming working independence and when the true covariance is estimated. Both kernel estimates of $\theta(t)$ are close to the true $\theta(t)$. Figure 2 compares the SEs of these two kernel estimates. It suggests that assuming working independence gives a more efficient kernel estimate of $\theta(t)$ than that achieved when assuming the true covariance. These results agree well with the theory.

8. APPLICATION TO THE INFECTIOUS DISEASE DATA

In this section we apply the semiparametric model (1) to analyzing the longitudinal infectious disease data introduced in Section 1. A total of 1,200 binary indicators for the presence of respiratory infection (0 = no, 1 = yes) were collected on 275 preschool-age children examined every quarter for up to six consecutive quarters. The primary interest was to study the association between respiratory infection and the exposure variable vitamin A deficiency, which was manifested by xerophthalmia status (0 = no; 1 = yes), while adjusting for several key confounders. These confounders include age in years, sex (0 = male, 1 = female), height for age, and stunting status (0 = no, 1 = yes). (For a detailed description of the covariates, see Zeger and Karim 1991.)

Examination of the distribution of the vertical strokes in Figure 3 suggests that the age effect departs dramatically from linearity. To avoid possible confounding of misspecification of the age effects on estimation of the effect of the key exposure xerophthalmia, we consider a semiparametric logistic model for the jth observation of the ith subject as

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \theta(\text{age}_{ij}), \qquad (20)$$

where $\mathbf{X}_{ij}$ comprises xerophthalmia status, seasonal cosine and sine, sex, height for age, and stunting, and $\theta(\text{age}_{ij})$ is a smooth function of age. Examination of the data suggested that the height for age effect was linear, and hence we included it in $\mathbf{X}_{ij}$.

We used the profile-kernel method assuming working independence using the algorithm in Appendix C and calculated the SEs using the sandwich method. We chose the bandwidth parameter h using the empirical bias bandwidth selection (EBBS) method (Ruppert 1997). Figure 3 shows the estimated nonparametric function of age and its 95% confidence interval. The risk of respiratory infection increased slightly during the first 2 years of life and decreased thereafter. Table 2 gives the estimated regression coefficients $\boldsymbol{\beta}$. The data provide no evidence for vitamin A deficiency on respiratory infection, but strong evidence for the association between respiratory infection and sex and season.

To examine whether a simple parametric model can fit the data equally well as the semiparametric model, we fit a parametric GEE model with $\theta(\text{age})$ to be quadratic assuming

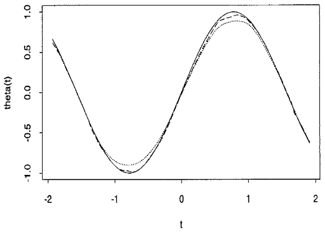


Figure 1. True and Estimated Nonparametric Functions $\hat{\theta}(t)$ Based on 200 Replications: (—— True; - - - - assuming working independence; – – – assuming that the true covariance is estimated).

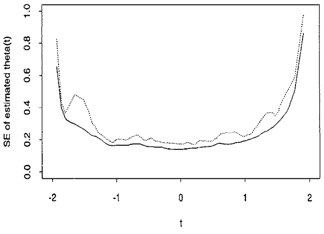


Figure 2. Empirical Pointwise SEs of the Estimated Nonparametric Functions $\hat{\theta}(t)$ Based on 200 Replications: (—— assuming working independence; - - - - assuming that the true covariance is estimated).

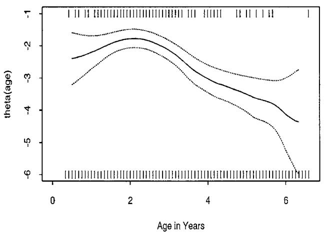


Figure 3. Estimated Kernel Estimate $\hat{\theta}$(age) When Fitting the Semiparametric Model (20) to the Infectious Disease Data Assuming Working Independence and Its 95% Pointwise Confidence Intervals (—— $\hat{\theta}$(age); - - - - 95% confidence interval). The vertical strokes at 0 and − 6 indicate the occurrence of 1 and 0 in the response.

working independence. Figure 4 compares the semiparametric kernel estimate of $\theta(t)$ to its quadratic counterpart (Diggle et al. 1994, p. 161). The semiparametric kernel estimate suggests that some excess nonlinearity may be undetected by the quadratic age model, a conjecture confirmed by the fact that a cubic age model fit using GEE had a statistically significant cubic age term (p value .02). Table 2 compares the regression coefficients $\boldsymbol{\beta}$ estimated using the semiparametric model and the parametric quadratic age model. The coefficient estimates of stunting were considerably different using the two methods, although the other coefficient estimates are similar. This difference was due mainly to misspecification of the quadratic age effect.

9. DISCUSSION

We have considered a marginal semiparametric partially linear generalized linear model for clustered data, where the effects of some covariates $\mathbf{X}$ are modeled parametrically as $\mathbf{X}\boldsymbol{\beta}$ and the effect of some other covariate T is modeled nonparametrically as $\theta(t)$. Our results apply to the case where the number of observations per cluster is finite and the number of clusters is large. The profile-kernel estimating equations in the literature are used for estimation. The results are unexpected.

We show that for clustered data, this conventional profile-kernel method fails to yield a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ unless working independence is assumed or $\theta(t)$ is artificially undersmoothed. Under working independence, one may need to greatly sacrifice efficiency to achieve $\sqrt{n}$ consistency of $\boldsymbol{\beta}$. When $\theta(t)$ is artificially undersmoothed, the profile-kernel estimator of $\boldsymbol{\beta}$ is still not semiparametric efficient, except for special cases.

Table 2. Regression Coefficient Estimates in Analysis of the Infectious Disease Data Using the Semiparametric Model and the Quadratic Age Model

	Semiparametric model		*Quadratic age model*	
	Estimate	*SE*	*Estimate*	*SE*
Vitamin A	.611	.529	.629	.413
Seasonal cosine	−.587	.210	−.590	.172
Seasonal sine	−.161	.183	−.170	.148
Sex	−.508	.295	−.485	.240
Height	−.026	.035	−.030	.029
Stunting	.463	.525	.272	.417

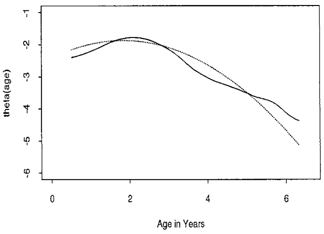


Figure 4. Comparison of the Kernel Estimate $\hat{\theta}$(age) (——) and the Quadratic Estimate of Age (- - - -).

To explain why the profile-kernel method fails in clustered data, we have derived the semiparametric efficient score of $\boldsymbol{\beta}$ for multivariate normal semiparametric models. We show that unlike in the independent data case, the profile-kernel method fails to provide an estimated score equation that is asymptotically equivalent to the semiparametric efficient score of $\boldsymbol{\beta}$. Even in this simple multivariate normal case, the semiparametric efficient score of $\boldsymbol{\beta}$ is complicated and requires solving the Fredholm integral equation and estimating the pairwise joint distributions of all observations $(\mathbf{X}_j, \mathbf{X}_k, T_j, T_k)$ in the same cluster. Direct estimation of such densities is complicated and could well be infeasible or cumbersome, especially when cluster sizes vary from one cluster to another. For example, in longitudinal data, different subjects could have different numbers of observations, and these different observations might be observed at different time points. Estimation of the joint distribution of $\mathbf{X}$ and $\boldsymbol{T}$ is hence difficult. One strategy is to assume a parametric model for $\mathbf{X}$ and $\boldsymbol{T}$ to estimate the joint distribution of $\mathbf{X}$ and $\boldsymbol{T}$. But this could lead to an inconsistent estimator of $\boldsymbol{\beta}$ if such a parametric model for $\mathbf{X}$ and $\boldsymbol{T}$ is misspecified. This leaves an open question on how to construct a semiparametric efficient estimator of $\boldsymbol{\beta}$ in practice for clustered data. Further research is needed.

We should note that the results in this article assume that T_{ij} varies within each cluster. If T_{ij} is a cluster-level covariate (i.e., $T_{ij} = T_i$), then, in contrast to the results reported in this paper, Lin and Carroll (2001) showed that the profile-kernel method works as usual and yields a $\sqrt{n}$ consistent and semiparametric efficient estimate of $\boldsymbol{\beta}$ if the true covariance is assumed and regular smoothing is used.

APPENDIX A: PROOF OF RESULT 1

A Note on Technical Conditions

It is possible to write down detailed technical conditions that would allow rigorous proofs of the results that follow for panel data. We have chosen not to do so, both in the interest of space and also because similar details have been written down by other authors in similar situations, without any real impact on statistical practice. These authors include Carroll et al. (1997), Carroll, Knickerbocker, and Wang (1995), Carroll and Wand (1991), Severini and Staniswalis (1994), and Severini and Wong (1992).

However, there is one situation for which it is easy to write down technical conditions leading to precise proofs—namely, the Gaussian linear case with constant true and working covariance matrices independent of $\boldsymbol{\beta}$. Happily, this is the problem of most interest, because all of our global conclusions have been made using this problem as an illustration.

To do this, one must first assume that, as in Carroll et al. (1995) and Severini and Staniswalis (1994), the $(T_{ij})_i$ have common compact support over j and their marginal and joint densities are bounded away from 0 on this support. We assume that $h \propto n^{-\alpha}$, where $1/5 \leq \alpha \leq 1/3$. Then, using the techniques of Mack and Silverman (1982) or Marron and Härdle (1986), one can show that (A.2) holds *uniformly* in t. In some cases, (as in Carroll et al. 1995), it is easier to prove this by restricting attention to $(T_{ij})_j$ that fall within a proper compact subset of the common support, in which case statements of results must be modified appropriately. In either case, the Gaussian linear problem means that nonparametric regressions are standard ones and do not involve solving nonlinear equations.

We now note the other key features of the Gaussian case. For the Gaussian case, (A.3)–(A.4) are *exact*, with $\widetilde{\mathbf{X}}_i$ defined just after (A.2) being *independent* of $\boldsymbol{\beta}$. In particular, the term $o_p(1)$ in (A.3) equals 0. With the uniformity of (A.2), the calculations following (A.3)–(A.4) are then routine.

Sketch of the Proof

To prove part a, we first assume that $\boldsymbol{\beta}$ is known and show that the asymptotic bias and variance of $\hat{\theta}(t;\boldsymbol{\beta})$ are given in (12) and (13). The proof is similar to appendix A.4 of Lin and Carroll (2000) and is hence omitted. Following that work, simple application of the Cauchy–Schwartz inequality shows that $\operatorname{var}\{\hat{\theta}(t;\boldsymbol{\beta})\}$ is minimized when $\mathbf{R}_2 = \mathbf{I}$ and is given in (14). We next study the distribution of $\hat{\theta}(t;\hat{\boldsymbol{\beta}})$ when $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent; that is, $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = O_p(1)$. We write

$$\begin{aligned}
&\sqrt{nh}\{\hat{\theta}(t;\hat{\boldsymbol{\beta}})-\theta(t)\}\\
&\quad= \sqrt{nh}\{\hat{\theta}(t;\hat{\boldsymbol{\beta}})-\hat{\theta}(t;\boldsymbol{\beta})\}+\sqrt{nh}\{\hat{\theta}(t;\boldsymbol{\beta})-\theta(t)\}\\
&\quad= \sqrt{h}\left\{\frac{\hat{\theta}(t;\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right\}\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}\\
&\qquad+\sqrt{nh}\{\hat{\theta}(t;\boldsymbol{\beta})-\theta(t)\}+o_p(1), \qquad (A.1)
\end{aligned}$$

where $\hat{\theta}(t;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T = -W_2^{-1}(t)\mathbf{W}_2^x(t)+o_p(1) = O_p(1)$, where $W_2(t)$ and $\mathbf{W}_2^x(t)$ are defined in Section 4. Because $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = O_p(1)$, the first term in (A.1) is $o_p(1)$. Hence the asymptotic distribution of $\hat{\theta}(t;\hat{\boldsymbol{\beta}})$ is the same as that of $\hat{\theta}(t;\boldsymbol{\beta})$.

We now study the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. First, using part a of Result 1 and following Lin and Carroll (2000), we have

$$\begin{aligned}
\hat{\theta}(t;\boldsymbol{\beta})-\theta(t) = W_2^{-1}(t)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{ij}^{(1)}v_{2i}^{jj}K_h(T_{ij}-t)(Y_{ij}-\mu_{ij})&\\
+\frac{\theta^{(2)}(t)h^2}{2}+o_p\{n^{-1/2}\}.& \qquad (A.2)
\end{aligned}$$

Define $\widetilde{\mathbf{X}}_i$ as $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij}+\partial\hat{\theta}(T_{ij};\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T = \mathbf{X}_{ij}-W_2^{-1}(T_{ij})\mathbf{W}_2^x(T_{ij})$. A linear Taylor expansion of (10) gives

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\} = \boldsymbol{D}_n^{-1}\{\sqrt{n}\mathbf{C}_n\}+o_p(1), \qquad (A.3)$$

where

$$\boldsymbol{D}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}\boldsymbol{\Delta}_i\widetilde{\mathbf{X}}_i$$

and

$$\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}[\mathbf{Y}_i-\boldsymbol{\mu}\{\mathbf{X}_i\boldsymbol{\beta}+\hat{\boldsymbol{\theta}}(\mathbf{T}_i;\boldsymbol{\beta})\}]. \qquad (A.4)$$

Denote $\boldsymbol{D} = \lim_{n\to\infty}\boldsymbol{D}_n = E(\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\widetilde{\mathbf{X}})$. Simple calculations show that $\boldsymbol{C}_n$ can be expanded as $\boldsymbol{C}_n = \boldsymbol{C}_{1n}-\boldsymbol{C}_{2n}+o_p(1)$, where, denoting $\boldsymbol{\mu}_i = \mu\{\mathbf{X}_i\boldsymbol{\beta}+\theta(\boldsymbol{T}_i)\}$ and $\mathbf{Z}_{1i}^T = \widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}$,

$$\boldsymbol{C}_{1n} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{1i}^T(\mathbf{Y}_i-\boldsymbol{\mu}_i)$$

and

$$\boldsymbol{C}_{2n} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{X}}_i^T\boldsymbol{\Delta}_i\mathbf{V}_{1i}^{-1}\boldsymbol{\Delta}_i\{\hat{\boldsymbol{\theta}}(\boldsymbol{T}_i;\boldsymbol{\beta})-\boldsymbol{\theta}(\boldsymbol{T}_i)\}.$$

Obtaining asymptotic distribution of $\sqrt{n}\boldsymbol{C}_{1n}$ is simple. Now examine the distribution of $\sqrt{n}\boldsymbol{C}_{2n}$. Using the Taylor expansion (A.2), we have

$$\begin{aligned}
\boldsymbol{C}_{2n} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}\{\hat{\theta}(T_{ik};\boldsymbol{\beta})-\theta(T_{ik})\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}\Bigg\{\Bigg[W_2^{-1}(T_{ik})\frac{1}{n}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}v_{2i'}^{j'j'}\\
&\qquad\times K_h(T_{i'j'}-T_{ik})(Y_{i'j'}-\mu_{i'j'})\Bigg]+\frac{h^2}{2}\theta^{(2)}(T_{ik})\Bigg\}+o_p(1)\\
&= \frac{1}{n}\sum_{i'=1}^{n}\sum_{j'=1}^{m}\mu_{i'j'}^{(1)}v_{2i'}^{j'j'}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}W_2^{-1}(T_{ik})\\
&\qquad\qquad\times K_h(T_{ik}-T_{i'j'})\Bigg\}(Y_{i'j'}-\mu_{i'j'})\\
&\quad+\frac{h^2}{2}\Bigg\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\widetilde{\mathbf{X}}_{ij}\mu_{ij}^{(1)}v_{1i}^{jk}\mu_{ik}^{(1)}\theta^{(2)}(T_{ik})\Bigg\}+o_p(1)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{Z}_{2ij}(Y_{ij}-\mu_{ij})+\frac{h^2}{2}\sum_{j=1}^{m}\sum_{k=1}^{m}\\
&\quad\times E\Big\{\widetilde{\mathbf{X}}_j\mu_j^{(1)}v_1^{jk}\mu_k^{(1)}\theta^{(2)}(T_k)\Big\}+o_p(1)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_{2i}^T(\mathbf{Y}_i-\boldsymbol{\mu}_i)+\frac{h^2}{2}E\{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\theta^{(2)}(\boldsymbol{T})\}+o_p(1),
\end{aligned}$$

where $\mathbf{Z}_{2i} = \{\mathbf{Z}_{2i1},\ldots,\mathbf{Z}_{2im}\}^T$ and

$$\mathbf{Z}_{2ij} = \mu_{ij}^{(1)}v_{2i}^{jj}\Bigg\{\sum_{k=1}^{m}\sum_{l=1}^{m}E\Big(\widetilde{\mathbf{X}}_k\mu_k^{(1)}v_1^{kl}\mu_l^{(1)}\Big|T_l=T_{ij}\Big)\Bigg\}W_2^{-1}(T_{ij})f_j(T_{ij}).$$

It follows that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \boldsymbol{D}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{Z}_{1i}-\mathbf{Z}_{2i})(\mathbf{Y}_i-\boldsymbol{\mu}_i)&\\
+\sqrt{nh^4}\boldsymbol{b}(\boldsymbol{\beta},\theta)/2+o_p(1),& \qquad (A.5)
\end{aligned}$$

where the bias term $\boldsymbol{b}(\boldsymbol{\beta},\theta) = \boldsymbol{D}^{-1}E\{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}\mathbf{V}_1^{-1}\boldsymbol{\Delta}\theta^{(2)}(\boldsymbol{T})\}$. Equivalently,

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}-h^2b(\boldsymbol{\beta},\theta)/2\} \to N(0,\mathbf{V}_{\boldsymbol{\beta}}),$$

where $\mathbf{V}_{\boldsymbol{\beta}} = \boldsymbol{D}^{-1}E\{(\mathbf{Z}_1-\mathbf{Z}_2)^T\boldsymbol{\Sigma}(\mathbf{Z}_1-\mathbf{Z}_2)\}\boldsymbol{D}^{-1}$ with $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{Y}|\mathbf{X},\boldsymbol{T})$.

One can see easily that the bias term $\boldsymbol{b}(\boldsymbol{\beta},\theta)$ in (A.5) is generally nonzero. Under conventional asymptotics, $n\to\infty$, $h\to 0$, and $nh\to\infty$, to obtain a $\sqrt{n}$ consistent estimate of $\boldsymbol{\beta}$, one must identify working correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ to make the bias term

$b(\boldsymbol{\beta}, \theta) = 0$. Simple calculations show that this requires assuming working independence $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$, and the same marginal joint density of $(\mathbf{X}_j, T_j)$; that is, $f_j(\mathbf{X}_j, T_j) = f(\mathbf{X}_j, T_j)$. Under these two assumptions,

$$\boldsymbol{b}(\boldsymbol{\beta}, \theta) = \boldsymbol{D}^{-1} E\left\{\sum_{j=1}^{m} E\left[\widetilde{\mathbf{X}}_j \left\{\mu_j^{(1)}\right\}^2 \sigma_{jj}^{-1} | T_j\right] \theta^{(2)}(T_j)\right\},$$

where $\widetilde{\mathbf{X}}_j = \mathbf{X}_j - E[\{\mu_j^{(1)}\}^2 \sigma_{jj}^{-1} \mathbf{X}_j | T_j])^{-1} E[\{\mu_j^{(1)}\}^2 \sigma_{jj}^{-1} | T_j])^{-1}$. One can see easily that $E[\widetilde{\mathbf{X}}_j \{\mu_j^{(1)}\}^2 \sigma_{jj}^{-1} | T_j] = 0$. It follows that $\boldsymbol{b}(\boldsymbol{\beta}, \theta) = 0$. Similar calculations show that $\mathbf{Z}_{2i} = 0$. This implies $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow N(0, \widetilde{\mathbf{V}}_{\boldsymbol{\beta}})$, where $\widetilde{\mathbf{V}}_{\boldsymbol{\beta}}$ is given in part c of Result 1 and can be estimated using a sandwich estimator.

For any nonidentity working correlation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, even when $\mathbf{R}_1$ and $\mathbf{R}_2$ are the true correlation matrices, under the foregoing conventional asymptotics [e.g., with h chosen using cross-validation; i.e., $h = O(n^{-1/5})$], the bias term $\sqrt{nh^4}\boldsymbol{b}(\boldsymbol{\beta}, \theta) \rightarrow \infty$. This means that $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ inconsistent and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow \infty$. Furthermore, $\mathbf{Z}_{2i} \neq 0$, implying that the standard sandwich estimator will be an inconsistent estimator of $\mathbf{V}_{\boldsymbol{\beta}}$, because it estimates $\boldsymbol{D}^{-1} E\{\mathbf{Z}_1 \boldsymbol{\Sigma} \boldsymbol{Z}_1^T\} \boldsymbol{D}^{-1}$ and ignores the nonzero term $\boldsymbol{Z}_2$. If one undersmooths $\theta(t)$ by letting $nh^4 \rightarrow 0$, then the bias term $\sqrt{nh^4}\boldsymbol{b}(\boldsymbol{\beta}, \theta) \rightarrow 0$, and $\hat{\boldsymbol{\beta}}$ will be $\sqrt{n}$ consistent for arbitrary working correlation matrices $(\mathbf{R}_1, \mathbf{R}_2)$.

APPENDIX B: SEMIPARAMETRIC EFFICIENT SCORE

We focus on the case where $\mathbf{X}_j$ and $\boldsymbol{\beta}$ are scalars (i.e., $p = 1$) and briefly discuss how to extend this result to the case where $\mathbf{X}_j$ and $\boldsymbol{\beta}$ are vectors. Let $f(\beta, \theta)$ denote the multivariate normal density of $\mathbf{Y} \sim N\{\mathbf{X}\beta + \theta(\boldsymbol{T}), \mathbf{V}\}$, where $\theta(\boldsymbol{T}) = \{\theta(T_1), \ldots, \theta(T_m)\}^T$. Following Begun, Hall, Huang, and Wellner (1983), we first calculate the Hellinger derivative with respect to $\theta(\cdot)$. Suppose that the sequence $\{\theta_n(t)\}$ satisfies $\sqrt{n}\{\theta_n(t) - \theta(t)\} - \varphi(t) \rightarrow 0$ as $n \rightarrow \infty$ for any given continuous function $\varphi(t)$. The Hellinger derivative $A\varphi(\cdot)$ with respect to $\theta(\cdot)$ is defined as

$$2n^{1/2}\left\{\frac{f^{1/2}(\beta, \theta_n) - f^{1/2}(\beta, \theta)}{f^{1/2}(\beta, \theta)}\right\} - \frac{2A\varphi}{f^{1/2}(\beta, \theta)} \rightarrow 0, \qquad \text{as} \quad n \rightarrow \infty,$$

where A denotes a linear operator. Denote $f_n = f\{\beta, \boldsymbol{\theta}_n\}$ and $f = f\{\beta, \boldsymbol{\theta}\}$, where $\boldsymbol{\theta}_n = \{\theta_n(T_1), \ldots, \theta_n(T_m)\}^T$ and $\boldsymbol{\theta} = \{\theta(T_1), \ldots, \theta(T_m)\}^T$. Let $\ell_n = \log f_n$ and $\ell = \log f$. A simple Taylor expansion shows that

$$\begin{aligned} 2\sqrt{n}\left\{\frac{\sqrt{f_n} - \sqrt{f}}{\sqrt{f}}\right\} &= \sqrt{n}\left\{\frac{f_n - f}{f}\right\} + o_p(1) \\ &= \sqrt{n}\{\ell_n - \ell\} + o_p(1) \\ &= \frac{\partial \ell}{\partial \boldsymbol{\theta}^T}\{\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})\} + o_p(1) \\ &= \varphi(T)^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\} + o_p(1). \end{aligned}$$

It follows that $2A\varphi(\boldsymbol{T})/f^{1/2} = \varphi(\boldsymbol{T})^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\}$. Let $\dot{\ell}_\beta = \partial \ell / \partial \beta = \mathbf{X}^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\}$. Then the semiparametric efficient score $\dot{\ell}_\beta^{\sharp}$ of β is

$$\begin{aligned} \dot{\ell}_\beta^{\sharp} &= \dot{\ell}_\beta - 2A\varphi_*(\boldsymbol{T})/f^{1/2}(\beta, \theta) \\ &= \{\mathbf{X} - \varphi_*(\boldsymbol{T})\}^T \mathbf{V}^{-1}\{\mathbf{Y} - \mathbf{X}\beta - \theta(\boldsymbol{T})\}, \end{aligned}$$

which is (15), where $\varphi_*(t)$ satisfies

$$E\{\dot{\ell}_\beta^{\sharp}[A\varphi(\boldsymbol{T})/f^{1/2}]\} = E\{[\mathbf{X} - \varphi_*(\boldsymbol{T})]^T \mathbf{V}^{-1} \varphi(\boldsymbol{T})\} = 0 \qquad \text{(B.1)}$$

for all functions $\varphi(\boldsymbol{T}) = \{\varphi(T_1), \ldots, \varphi(T_m)\}^T$, where $\varphi_*(\boldsymbol{T}) = \{\varphi_*(T_1), \ldots, \varphi_*(T_m)\}^T$. The semiparametric efficiency bound of $\boldsymbol{\beta}$ is $E\{[\dot{\ell}_\beta^{\sharp}]^2\}$. Equation (B.1) can be written as

$$\begin{aligned} &\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E\{[X_j - \varphi_*(T_j)]\varphi(T_k)\} \\ &\qquad = \sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E[E\{E[X_j - \varphi_*(T_j)|T_k]\}\varphi(T_k)] = 0. \end{aligned}$$

Simple calculations show that this equation can be written as

$$\int\left[\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk}\{E[X_j - \varphi_*(T_j)|T_k = t]\} f_k(t)\right]\varphi(t)\,dt = 0$$

for any $\varphi(t)$. It follows that $\varphi_*(t)$ must solve $\sum_{j=1}^m \sum_{k=1}^m \times v^{jk}\{E[X_j - \varphi_*(T_j)|T_k = t]\} f_k(t) = 0$, which is (16).

To extend the results to the case where $\mathbf{X}_j$ and $\boldsymbol{\beta}$ are vectors, we need to find $\varphi_*(t)$ for each component of $\mathbf{X}_j$ using (16) (Begun et al. 1983). Specifically, we calculate $\boldsymbol{\varphi}_*(t) = \{\varphi_{*1}(t), \ldots, \varphi_{*p}(t)\}^T$, where, letting X_{jr} denote the rth component of $\mathbf{X}_j$, $\varphi_{*r}(t)$ solves

$$\sum_{j=1}^{m}\sum_{k=1}^{m} v^{jk} E[\{X_{jr} - \varphi_{*r}(T_j)\}|T_k = t] f_k(t) = 0.$$

Hence semiparametric efficient score of $\boldsymbol{\beta}$ is given by (15) and (16).

APPENDIX C: COMPUTATION ASSUMING WORKING INDEPENDENCE

In this section we assume working independence ($\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$) in the profile-kernel estimating equations (8) and (10), and discuss the use of the Fisher scoring algorithm to solve for $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}(t)$, where $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent. Specifically, under working independence ($\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$), (8) and (10) are solved in the following steps:

1. Assume a parametric function for $\theta(t)$, [e.g., $\theta(t) = \alpha_0 + \alpha_1 t$], and fit a parametric generalized linear model, $g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \alpha_0 + \alpha_1 T_{ij}$, to obtain an initial value of $\boldsymbol{\beta}$.
2. Given the value of $\boldsymbol{\beta}$, use the Fisher scoring algorithm to solve (8) (with $\mathbf{R}_2 = \mathbf{I}$) for $t = T_{11}, \ldots, T_{nm_n}$. This gives $\{\hat{\theta}(T_{11}; \boldsymbol{\beta}), \ldots, \hat{\theta}(T_{nm_n}; \boldsymbol{\beta}\})$.
3. Update $\boldsymbol{\beta}$ using the one-step Fisher scoring algorithm to solve (10) (with $\mathbf{R}_1 = \mathbf{I}$) given the $\hat{\theta}(T_{ij}; \boldsymbol{\beta})$ in step 2.
4. Iterate between steps 2 and 3 until convergence.

In step 2, it can be easily shown that the Fisher scoring algorithm updates $\hat{\boldsymbol{\alpha}}$ by

$$\begin{aligned} &\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{T}_{ij}(t) W_{ij}(t) K_h(T_{ij} - t) \boldsymbol{T}_{ij}(t)^T\right\}\hat{\boldsymbol{\alpha}} \\ &\qquad = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{T}_{ij}(t) W_{ij}(t) K_h(T_{ij} - t) y_{ij}(t), \end{aligned}$$

where $W_{ij}(t) = \{\mu_{ij}^{(1)}(t)\}^2 V_{ij}^{-1}(t)$ is the generalized linear model working weight and $y_{ij}(t) = \boldsymbol{T}_{ij}(t)^T \boldsymbol{\alpha} + \mu_{ij}^{(1)}(t)\{Y_{ij} - \mu_{ij}(t)\}$ is the generalized linear model working vector.

In step 3, the one-step Fisher scoring algorithm updates $\boldsymbol{\beta}$ using the weighted least squares,

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} \widetilde{\mathbf{X}}_{ij} W_{ij} \widetilde{\mathbf{X}}_{ij}^T\right\}\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \widetilde{\mathbf{X}}_{ij} W_{ij} y_{ij},$$

where $W_{ij} = \{\mu_{ij}^{(1)}\}^2 V^{-1}(\mu_{ij})$ is the working weight, $y_{ij} = \widetilde{\mathbf{X}}_{ij}^T \boldsymbol{\beta} + \mu_{ij}^{(1)}(Y_{ij} - \mu_{ij})$ is the working vector, $\mu_{ij} = \mu\{\mathbf{X}_{ij}^T\boldsymbol{\beta} + \hat{\theta}(T_{ij};\boldsymbol{\beta})\}$, and $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij} + \partial\hat{\theta}(T_{ij};\boldsymbol{\beta}))\partial\boldsymbol{\beta}$. To calculate $\widetilde{\mathbf{X}}_{ij}$, we need to construct a consistent estimate of $\partial\hat{\theta}(t;\boldsymbol{\beta}))\partial\boldsymbol{\beta}$. Using the results in Section 4, we can easily see that a consistent estimator of $\partial\hat{\theta}(t;\boldsymbol{\beta}))\partial\boldsymbol{\beta}$ is

$$-\frac{\sum_{i=1}^n \sum_{j=1}^{m_i} W_{ij}(t) K_h(T_{ij}-t)\mathbf{X}_{ij}}{\sum_{i=1}^n \sum_{j=1}^{m_i} W_{ij}(t) K_h(T_{ij}-t)}.$$

The covariance estimators of $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}(t)$ at convergence are sandwich estimators given by

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}) = &\left\{\sum_{i=1}^n \widetilde{\mathbf{X}}_i^T \mathbf{W}_i \widetilde{\mathbf{X}}_i\right\}^{-1} \left\{\sum_{i=1}^n \widetilde{\mathbf{X}}_i^T \boldsymbol{\Delta}_i \boldsymbol{\Sigma}_{id}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)\right.\\ &\left.\times (\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_{id}^{-1} \boldsymbol{\Delta}_i \widetilde{\mathbf{X}}_i\right\} \left\{\sum_{i=1}^n \widetilde{\mathbf{X}}_i^T \mathbf{W}_i \widetilde{\mathbf{X}}_i\right\}^{-1}\end{aligned}$$

and

$$\operatorname{cov}\{\hat{\theta}(t)\} = \boldsymbol{e}_1^T \boldsymbol{\Omega}_1^{-1}(t) \boldsymbol{\Omega}_2(t) \boldsymbol{\Omega}_1^{-1}(t) \boldsymbol{e}_1,$$

where $\boldsymbol{e}_1 = (1, 0)^T$, $\boldsymbol{\Delta}_i = \operatorname{diag}\{\mu_{ij}^{(1)}\}$, and $\boldsymbol{\Sigma}_{id} = \operatorname{diag}\{V_{ij}\}$ and all are evaluated at $\{\hat{\boldsymbol{\beta}}, \hat{\theta}(t)\}$ and

$$\boldsymbol{\Omega}_1(t) = \sum_{i=1}^n \mathbf{T}_i^T(t) \mathbf{W}_i(t) \mathbf{K}_{ih}(t) T_i(t)$$

and

$$\begin{aligned}\boldsymbol{\Omega}_2(t) = &\sum_{i=1}^n \boldsymbol{T}_i^T(t) \boldsymbol{\Delta}_i(t) \boldsymbol{\Sigma}_{id}^{-1} \mathbf{K}_{ih}(t) [\mathbf{Y}_i - \boldsymbol{\mu}_i(t)]\\ &\times [\mathbf{Y}_i - \boldsymbol{\mu}_i(t)]^T \mathbf{K}_{ih}(t) \boldsymbol{\Sigma}_{id}^{-1} \boldsymbol{\Delta}_i(t) \boldsymbol{T}_i(t).\end{aligned}$$

Estimation of $\theta(t)$ requires choosing the bandwidth parameter h. One approach is to use cross-validation by deleting one cluster at a time. Another approach is to extend Ruppert's (1997) empirical bias bandwidth selection (EBBS) method to clustered data. We use the EBBS method to choose h for given $\boldsymbol{\beta}$. (For details, see Lin and Carroll 2000.)

[Received February 2000. Revised January 2001.]

REFERENCES

Begun, J. M., Hall, W. J., Huang, W., and Wellner, J. A. (1983), "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *The Annals of Statistics*, 11, 432–452.

Bickel, P. J., Klaassen, C. J., Ritov, Y., and Wellner, J. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Brockmann, M., Gasser, T., and Herrmann, E. (1993), "Locally Adaptive Bandwidth Choice for Kernel Regression Estimators," *Journal of the American Statistical Association*, 88, 1302–1309.

Bronshtein, J. N., and Semendyayev, K. A. (1985), *Handbook of Mathematics*, New York: Van Nostrand Reinhold.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Carroll, R. J., Knickerbocker, R. K., and Wang, C. Y. (1995), "Dimension Reduction in Semiparametric Measurement Error Models," *The Annals of Statistics*, 23, 161–181.

Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error," *Journal of the Royal Statistical Society*, Ser. B, 53, 573–585.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, Y. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520–534.

——— (2001), "Semiparametric Regression For Clustered Data," *Biometrika*, in press.

Mack, Y., and Silverman, B. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 60, 405–415.

Marron, J. S., and Härdle, W. (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91–113.

Pepe, M. S., and Couper, D. (1997), "Modeling Partly Conditional Means With Longitudinal Data," *Journal of the American Statistical Association*, 92, 991–998.

Ramsay, J. O., and Dalzell, C. J. (1991), "Some Tools for Functional Data Analysis" (with discussions), *Journal of the Royal Statistical Society*, Ser. B, 53, 539–572.

Rice, J. (1986), "Convergence Rates for Partially Splined Models," *Statistical and Probability Letters*, 4, 203–208.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves," *Journal of the Royal Statistical Society*, Ser. B 53, 233–243.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

Wild, C. J., and Yee, T. W. (1996), "Additive Extensions to Generalized Estimating Equation Methods," *Journal of the Royal Statistical Society*, Ser. B, 58, 711–725.

Zeger, S. L., and Diggle, P. J. (1994), "Semi-Parametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semi-Parametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.

J. R. Statist. Soc. B (2006)
68, *Part* 1, *pp.* 69–88

Semiparametric estimation in general repeated measures problems

Xihong Lin

Harvard School of Public Health, Boston, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA

[Received January 2005. Revised September 2005]

Summary. The paper considers a wide class of semiparametric problems with a parametric part for some covariate effects and repeated evaluations of a nonparametric function. Special cases in our approach include marginal models for longitudinal or clustered data, conditional logistic regression for matched case–control studies, multivariate measurement error models, generalized linear mixed models with a semiparametric component, and many others. We propose profile kernel and backfitting estimation methods for these problems, derive their asymptotic distributions and show that in likelihood problems the methods are semiparametric efficient. Although generally not true, it transpires that with our methods profiling and backfitting are asymptotically equivalent. We also consider pseudolikelihood methods where some nuisance parameters are estimated from a different algorithm. The methods proposed are evaluated by using simulation studies and applied to the Kenya haemoglobin data.

Keywords: Clustered and longitudinal data; Generalized estimating equations; Generalized linear mixed models; Kernel method; Marginal models; Measurement error; Nonparametric regression; Partially linear model; Profile method; Semiparametric efficient score; Semiparametric information bound; Time-dependent covariate

1. Introduction

This paper considers a wide class of semiparametric problems with some covariates modelled parametrically and repeated evaluations of a nonparametric function of a covariate. We propose profile kernel and backfitting estimation methods for these problems, derive their asymptotic distributions and show that in likelihood problems the methods are semiparametric efficient.

To obtain some sense of the generality of our approach, consider the following examples, all of which can be solved by using our approach. The first four are new, in the sense that neither the semiparametric efficient score function nor a constructive method of estimation and inference that achieve efficiency is known. In contrast, the fifth example has a large literature.

1.1. Example 1

One of the most common designs in epidemiology is the matched case–control study, which is a design that is attracting considerable interest in genetic epidemiology; see for example Schaid (1999). Matched case–control studies consist of groups that have discordant responses. Thus, in

Address for correspondence: Xihong Lin, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA.
E-mail: xlin@hsph.harvard.edu

1369–7412/06/68069

the 1–1 matched study, we consider matched pairs of subjects, with disease responses (Y_{i1}, Y_{i2}) that are constrained to be discordant, so that $Y_{i1}+Y_{i2}=1$. The underlying prospective semiparametric logistic regression model is that

$$\mathrm{pr}(Y_{ij}=1|X_{ij}, Z_{ij}) = H\{b_i + X_{ij}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij})\},$$

where $H(v)=1/\{1+\exp(-v)\}$ is the logistic distribution function, b_i is a nuisance parameter depending on the matched set, X_{ij} is a covariate vector whose effect is modelled parametrically and Z_{ij} is a scalar covariate whose effect is modelled by using a nonparametric smooth function $\theta_0(\cdot)$. Let $\tilde{X}_i=(X_{i1}, X_{i2})$ and $\tilde{Z}_i=(Z_{i1}, Z_{i2})$. Because the data are constrained to be discordant, and we do not want to model the stratum effects b_i, inference is based on the conditional likelihood function

$$\mathrm{pr}(Y_{i1}=1, Y_{i2}=0|\tilde{X}_i, \tilde{Z}_i, Y_{i1}+Y_{i2}=1) = H\{(X_{i1}-X_{i2})^{\mathrm{T}}\beta_0 + \theta_0(Z_{i1}) - \theta_0(Z_{i2})\}. \qquad (1)$$

Note that in equation (1) the stratum effects have been eliminated, and that in the likelihood $\theta_0(\cdot)$ is evaluated twice at different values of Z. In more complex matched studies, $\theta_0(\cdot)$ is evaluated more than twice, e.g. the $1-M$ matched design.

1.2. Example 2

Hafner (1998) and Carroll *et al.* (2002) studied

$$Y_i = \sum_{j=1}^{m} \beta_0^{j-1}\theta_0(Z_{ij}) + \varepsilon_i,$$

a model that arises in finance. The algorithm that was proposed by Carroll *et al.* (2002) for this case is extremely unwieldy and difficult to implement, because it is based on an integration estimator (Linton and Nielson, 1995). Our methodology in this case is far easier to implement and has the advantage of being semiparametric efficient in the Gaussian case.

1.3. Example 3

Generalized linear mixed models (Breslow and Clayton, 1993) have become popular as a means of quantifying and understanding variability. The simplest such model for binary data is the random-intercept model

$$\mathrm{pr}(Y_{ij}=1|X_{ij}, Z_{ij}, b_i) = \mu\{X_{ij}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij}) + b_i\},$$

where $\mu(\cdot)$ is the inverse of a link function and $b_i = \mathrm{normal}(0, \sigma_0^2)$. Here the variance component σ_0^2 may be of interest in itself and may in some cases depend on components of X such as gender; see Heagerty and Kurland (2001) for an example.

1.4. Example 4

As discussed in a data example in Section 5.1.2, consider problems in which family i has m children, each of whom have a base-line measure Z_{ij} for $j=1,\ldots,m$, but for whom there are repeated measures Y_{ijk} over time for $k=1,\ldots,K$ and a possible repeated time-varying covariate X_{ijk}. A reasonable marginal model for the Y_{ijk} is that their means are $\mu\{X_{ijk}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij})\}$ for a known inverse link function $\mu(\cdot)$, and a covariance matrix Σ reflecting the structure of the problem. In this case, note that the function $\theta_0(\cdot)$ is evaluated m times for different children per family.

1.5. Example 5

Consider a repeated measures Gaussian partially linear problem where for the ith subject responses $\tilde{Y}_i = (Y_{i1}, \ldots, Y_{im})^{\mathrm{T}}$ and predictors $\tilde{X}_i = (X_{i1}, \ldots, X_{im})^{\mathrm{T}}$ and $\tilde{Z}_i = (Z_{i1}, \ldots, Z_{im})^{\mathrm{T}}$ are observed, with Z_{ij} scalar. The basic model is that, for a known function $\mu(\cdot)$ and a true but unknown function $\theta_0(z)$,

$$Y_{ij} = \mu\{X_{ij}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij})\} + \varepsilon_{ij}, \tag{2}$$

where, given $(\tilde{X}_i, \tilde{Z}_i)$, $\tilde{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^{\mathrm{T}}$ has mean 0 and covariance matrix $\Sigma(\tau_0)$ for a parameter τ_0. Note that the function $\theta_0(\cdot)$ is evaluated repeatedly, and thus this problem is very much different from the standard partially linear model (Severini and Staniswalis, 1994). This problem has a large literature, with many kernel-based methods (Zeger and Diggle (1994), Hoover *et al.* (1998), Lin and Ying (2001), Wu and Zhang (2002) and many others), all of them estimating $\theta_0(\cdot)$ while ignoring the correlation structure. Lin and Carroll (2000, 2001) and Fan and Li (2004) made an effort to incorporate the correlation structure in the estimation procedure within the traditional kernel framework. However, Lin and Carroll (2000) showed that the optimal estimator of $\theta_0(\cdot)$ within the standard kernel framework requires ignoring the correlation. There is also an extensive spline-based literature (Wild and Yee, 1996; Zhang *et al.*, 1998; Wang, 1998; Rice and Wu, 2001). Fixing $\Sigma(\tau_0)$ and pretending normality, Wang *et al.* (2004) developed kernel-based consistent and asymptotically normal estimators for β_0: these are semiparametric efficient when $\tilde{\varepsilon}_i$ is actually Gaussian.

These examples can be placed into a common framework. There is a *criterion function* $\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}, \mathcal{B})$, where $\tilde{\eta}$ has m components representing $\theta(Z_1), \ldots, \theta(Z_m)$ and $\mathcal{B}$ is a vector of parameters. For true values $\tilde{\eta}_0$ and $\mathcal{B}_0$, the criterion function satisfies

$$0 = E[\{\partial\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}_0, \mathcal{B}_0)/\partial(\tilde{\eta}_0, \mathcal{B}_0)\}|\tilde{X}, \tilde{Z}]. \tag{3}$$

For example, consider the model that is given in equation (2). Here, $\mathcal{B}_0 = (\beta_0, \tau_0)$ and the criterion function is the Gaussian log-likelihood

$$-\tfrac{1}{2}\log[\det\{\Sigma(\tau_0)\}] - \tfrac{1}{2}(\tilde{Y} - \tilde{X}\beta_0 - \tilde{\eta}_0)^{\mathrm{T}}\Sigma^{-1}(\tau_0)(\tilde{Y} - \tilde{X}\beta_0 - \tilde{\eta}_0).$$

The criterion function in example 1 is given in equation (1), and examples 2–4 also have explicit forms.

In this paper, we show how to compute efficient estimators of the nonparametric component $\theta_0(\cdot)$ for problems with and without the parametric component $\mathcal{B}_0$. The method that is defined in Section 2 is based on a likelihood-type generalization of the basic kernel method of Wang (2003) to the general problem (3). The methods are applicable to likelihood and non-likelihood problems, the only constraint being that condition (3) holds.

In Section 3 we take up estimation of the parameter $\mathcal{B}$. In this context, we derive two general methods, one incorporating profile likelihood ideas and the other based on the often easier to compute backfitting algorithm. Lin and Carroll (2001) and Wang *et al.* (2005) proposed estimating-equation-based profile kernel methods for a marginal generalized semiparametric model that was similar to the normal model (2) for clustered data. Hu *et al.* (2004) proposed a backfitting method under the normal model (2). Our profile likelihood method and the backfitting algorithm are likelihood-type extensions of the methods of Wang *et al.* (2005) and Hu *et al.* (2004) to the general setting in expression (3). We show that in our case, using the smoother of Section 2, profiling and backfitting have identical limit distributions. The folklore of course is that backfitting and profiling are in general asymptotically equivalent, independent of the method of smoothing, but in general this is not so (Hu *et al.*, 2004). However, our use of an efficient smoother allows us to show that backfitting and profiling are asymptotically equivalent. It should be noted that undersmoothing of the nonparametric function is required by backfitting but not required by profiling. In this section, we also describe the semiparametric efficient score function when $\mathcal{L}(\cdot)$ is a likelihood function, and we show in our case that our method achieves the semiparametric information bound.

In many problems, there are nuisance parameters that can be estimated relatively conveniently by alternative means. In the example that was considered by Wang *et al.* (2005), the covariance matrix $\Sigma(\tau_0)$ depends on a parameter τ_0. The parameter τ_0 is conveniently estimated by the simple device of ignoring the correlation of the data, forming residuals from the fit and then using the method of moments. This is a pseudolikelihood approach. In Section 4, we derive the limiting distribution of the pseudolikelihood estimator in the general case.

Section 5 first describes example 4 in detail. We illustrate example 4 by using the Kenya haemoglobin data and a simulation study. The second case that is considered in Section 5 is a multivariate measurement error problem. The formulation of the measurement error model is new even in the parametric measurement error model literature. Sketches of the technical arguments are given in the appendices and detailed proofs can be found at `http://www.bepress.com/harvardbiostat` and also at `http://www.stat.tamu.edu/~carroll/papers.php`.

2. The nonparametric case

Before describing methods for the general semiparametric problem, we describe methods when there is no parametric component, which is a problem of interest in its own right. In the nonparametric case, the criterion function is $\mathcal{L}(\tilde{Y},\tilde{X},\tilde{\eta})=\mathcal{L}\{\tilde{Y},\tilde{X},\theta(Z_1),\ldots,\theta(Z_m)\}$ where $\eta_j=\theta(Z_j)$ $(j=1,\ldots,m)$. Define $\mathcal{L}_{j\theta}(\cdot)=\partial\mathcal{L}(\tilde{Y},\tilde{X},\eta_1,\ldots,\eta_m)/\partial\eta_j$ and $\mathcal{L}_{jk\theta}(\cdot)=\partial^2\mathcal{L}(\tilde{Y},\tilde{X},\eta_1,\ldots,\eta_m)/\partial\eta_j\,\partial\eta_k$ $(j,k=1,\ldots,m)$. We assume that $0=E[\mathcal{L}_{j\theta}\{\tilde{Y},\tilde{X},\theta(Z_1),\ldots,\theta(Z_m)\}|\tilde{X},\tilde{Z}]$. Let $K(\cdot)$ be a symmetric density function with variance 1.0, and define $G_{ij}(z,h)=\{1,(Z_{ij}-z)/h\}$. Let $f_j(z)$ be the marginal density of Z_{ij}.

We propose to estimate $\theta(\cdot)$ by solving the kernel estimating equation

$$0=\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(Z_{ij}-z)\,G_{ij}(z,h)\,\mathcal{L}_{j\theta}\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}(Z_{i1}),\ldots,\hat{\theta}(z)+\hat{\theta}^{(1)}(z)(Z_{ij}-z),\ldots,\hat{\theta}(Z_{im})\},\tag{4}$$

where $\hat{\theta}^{(1)}(z)$ denotes the first derivative of $\hat{\theta}(z)$. Following Wang (2003), we propose to solve the kernel estimating equation (4) for $\hat{\theta}(z)$ in the following iterative fashion. Suppose that the current estimate of $\theta(\cdot)$ at the $(l-1)$th step is $\hat{\theta}_{[l-1]}(\cdot)$. Then $\hat{\theta}_{[l]}(z)=\hat{\alpha}_0$, where $(\hat{\alpha}_0,\hat{\alpha}_1)$ solve

$$0=\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(Z_{ij}-z)\,G_{ij}(z,h)\,\mathcal{L}_{j\theta}\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}_{[l-1]}(Z_{i1}),\ldots,\alpha_0+\alpha_1(Z_{ij}-z)/h,\ldots,\hat{\theta}_{[l-1]}(Z_{im})\}.\tag{5}$$

At convergence, $\hat{\theta}(z)$ solves the kernel estimating equation (4). In Gaussian cases such as in examples 1 and 2, iteration is actually not needed, with explicit solutions being available; see Lin *et al.* (2004) for example 1, and see also Section 5.1 for another example. Define $\mathcal{L}(\cdot)=\mathcal{L}\{\tilde{Y},\tilde{X},\theta(Z_1),\ldots,\theta(Z_m)\}$, and similarly for its derivatives. Make the definitions

$$\Omega(z)=\sum_{j=1}^{m}f_j(z)\,E\left\{\mathcal{L}_{jj\theta}(\cdot)|Z_j=z\right\}$$

and

$$\mathcal{A}(B,z_1,z_2)=\sum_{j=1}^{m}\sum_{k\neq j}^{m} f_j(z_1)\,E\{\mathcal{L}_{jk\theta}(\cdot)\,B(Z_k,z_2)/\Omega(Z_k)|Z_j=z_1\},$$

$$\mathcal{Q}(z_1,z_2)=\sum_{j=1}^{m}\sum_{k\neq j}^{m} f_{jk}(z_1,z_2)\,E\{\mathcal{L}_{jk\theta}(\cdot)|Z_j=z_1,Z_k=z_2\}/\Omega(z_2),$$

$$\Lambda(g,z)=\sum_{j=1}^{m}\sum_{k\neq j}^{m} f_j(z)\,E\{\mathcal{L}_{jk\theta}(\cdot)g(Z_k)|Z_j=z\}/\Omega(z),$$

where $f_j(z)$ is the density of Z_j and $f_{jk}(z_1,z_2)$ is the bivariate density of (Z_j,Z_k). Let $\mathcal{G}(z_1,z_2)$ and $b(z)$ be the solutions to

$$\mathcal{G}(z_1,z_2)=\mathcal{Q}(z_1,z_2)-\mathcal{A}(\mathcal{G},z_1,z_2), \tag{6}$$

$$b(z)=\theta^{(2)}(z)-\Lambda(b,z). \tag{7}$$

2.1. Result 1: expansion for the nonparametric part

Suppose that the Z_{ij} have support on a compact set and that their joint and marginal densities are bounded away from zero on that set. Assume that the algorithm converges to a unique solution and that equations (6) and (7) have unique solutions. Let the bandwidth sequence satisfy $nh^2\to\infty$ and $nh^6\to 0$. Let $\phi=\int z^2\,K(z)\,\mathrm{d}z$. Denote by $\theta_0(z)$ the true function. Then, at convergence,

$$\hat{\theta}(z)-\theta_0(z)=(h^2/2)\phi\, b(z)-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(Z_{ij}-z)\varepsilon_{ij}/\Omega(z)$$
$$+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}\,\mathcal{G}(z,Z_{ij})/\Omega(z)+o_p(n^{-1/2}), \tag{8}$$

where $\varepsilon_{ij}=\mathcal{L}_{j\theta}\{\tilde{Y}_i,\tilde{X}_i,\theta_0(Z_{i1}),\ldots,\theta_0(Z_{im})\}$. Thus, the asymptotic bias and variance of $\hat{\theta}(z)$ are

$$E\{\hat{\theta}(z)\}-\theta_0(z)=(h^2/2)\phi\, b(z)+o(h^2), \tag{9}$$

$$\mathrm{var}\{\hat{\theta}(z)\}=\frac{1}{nh}\frac{\psi}{\Omega^2(z)}\sum_{j=1}^{m}E(D_{jj}|Z_j=z)\,f_j(z)+o\{(nh)^{-1}\}, \tag{10}$$

where $\psi=\int K^2(s)\,\mathrm{d}s$ and D_{jj} is the jth diagonal element of $\mathrm{cov}(\tilde{\varepsilon}_i|\tilde{X}_i,\tilde{Z}_i)$, where $\tilde{\varepsilon}_i=(\varepsilon_{i1},\ldots,\varepsilon_{im})^{\mathrm{T}}$.

Remark 1. Equations (8)–(10) agree with the results of Wang (2003) in the special cases that were considered by her. In equation (8), since the first two terms are of order $O_p\{h^2+(nh)^{-1/2}\}$ whereas the third is of order $O_p(n^{-1/2})$, the first two terms dominate. The proof of result (8) is similar to that of Wang (2003) and is given in the technical report that was mentioned at the end of Section 1.

Remark 2. Note that equation (9) has design-density-dependent bias. It is possible to remove this. Suppose that the algorithm is run with an undersmoothing bandwidth $h_1=o(n^{-1/4})$, thus obtaining $\hat{\theta}(z,h_1)$ at convergence. Let $\hat{\theta}_{\mathrm{os}}(z,h)$ be the estimator that is defined by doing one step of the iteration from $\hat{\theta}(z,h_1)$, but now with bandwidth h, where $h/h_1\to 0$ as $n\to\infty$. Then result (8) still holds except that the bias term $(h^2/2)\phi b(z)$ is replaced by $(h^2/2)\phi\theta^{(2)}(z)$. The proof of this argument is a routine application of lemma 1 and equation (24) in Appendix A.1 starting from expansion (8).

3. The semiparametric case: methods and results

In this section, we formulate the profile kernel and backfitting estimation methods for $\mathcal{B}_0$ in the semiparametric model $\mathcal{L}(\tilde{Y}, \tilde{X}, \tilde{\eta}_0, \mathcal{B}_0)$, state their asymptotic distributions and show that, when the criterion function $\mathcal{L}(\cdot)$ is a log-likelihood function conditional on $(\tilde{Z}, \tilde{X})$, our method achieves the semiparametric information bound.

3.1. Estimation: profile kernel and backfitting methods

To estimate $\mathcal{B}$, we propose profile kernel and backfitting methods. For any $\mathcal{B}$, we first obtain the modified kernel estimate of $\hat{\theta}(z,\mathcal{B})$ and its first derivative $\hat{\theta}^{(1)}(z,\mathcal{B})$ with respect to z by solving

$$0 = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} K_h(Z_{ij}-z)\, G_{ij}(z,h)\, \mathcal{L}_{j\theta}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1},\mathcal{B}),\ldots,\hat{\theta}(z,\mathcal{B}) + h\,\hat{\theta}^{(1)}(z,\mathcal{B})(Z_{ij}-z)/h,\ldots,\hat{\theta}(Z_{im},\mathcal{B}),\mathcal{B}\}. \tag{11}$$

To solve equation (11) we suggest the following iterative algorithm. Suppose that the current estimate in the iteration is $\hat{\theta}_{[l-1]}(z,\mathcal{B})$. Then we update to $\hat{\theta}_{[l]}(z,\mathcal{B})$ by solving (α_0,α_1) in the equation

$$0 = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} K_h(Z_{ij}-z)\, G_{ij}(z,h)\, \mathcal{L}_{j\theta}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[l-1]}(Z_{i1},\mathcal{B}),\ldots,\alpha_0 + \alpha_1(Z_{ij}-z)/h,\ldots,\hat{\theta}_{[l-1]}(Z_{im},\mathcal{B}),\mathcal{B}\}.$$

Set $\hat{\theta}_{[l]}(z,\mathcal{B}) = \alpha_0$. At convergence, for any fixed $\mathcal{B}$, we have the kernel estimator $\hat{\theta}(z,\mathcal{B})$.

We now define two methods for estimating $\mathcal{B}_0$. The *profile kernel estimator* $\hat{\mathcal{B}}_{\mathrm{p}}$ maximizes

$$\sum_{i=1}^{n} \mathcal{L}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1},\mathcal{B}),\ldots,\hat{\theta}(Z_{im},\mathcal{B}),\mathcal{B}\}.$$

Maximization of the profile likelihood requires calculating the derivative $\hat{\theta}_{\mathcal{B}}(z,\mathcal{B}) = \partial\hat{\theta}(z,\mathcal{B})/\partial\mathcal{B}$. This can be computed by numerical differentiation: in addition, in Appendix A.6, we show how to use an algorithm that is very similar to equation (5) to compute $\hat{\theta}_{\mathcal{B}}(z,\mathcal{B})$ by solving a kernel estimating equation.

In some cases, the profile kernel method may be difficult to implement numerically owing to the additional required computation of $\hat{\theta}_{\mathcal{B}}(z,\mathcal{B})$. Instead, a *backfitting* algorithm can be used. In the iterative backfitting algorithm, suppose that the current estimate is $\mathcal{B}_*$. The updated backfitting estimate then maximizes $\mathcal{B}$ in the function

$$\sum_{i=1}^{n} \mathcal{L}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1},\mathcal{B}_*),\ldots,\hat{\theta}(Z_{im},\mathcal{B}_*),\mathcal{B}\}.$$

The fully iterated solution to this algorithm is denoted by $\hat{\mathcal{B}}_{\mathrm{b}}$. It is somewhat more general to write the updated backfitting estimate as the solution in $\mathcal{B}$ to

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Psi_i(\mathcal{B}_*,\mathcal{B}) \\ &= \sum_{i=1}^{n} \mathcal{L}_{\mathcal{B}}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1},\mathcal{B}_*),\ldots,\hat{\theta}(Z_{im},\mathcal{B}_*),\mathcal{B}\}, \end{aligned} \tag{12}$$

where

$$\mathcal{L}_{\mathcal{B}}\{\tilde{Y}_i, \tilde{X}_i, \theta(Z_1),\ldots,\theta(Z_m),\mathcal{B}\} = \partial\mathcal{L}\{\tilde{Y}_i, \tilde{X}_i, \theta(Z_1),\ldots,\theta(Z_m),\mathcal{B}\}/\partial\mathcal{B}.$$

In general problems of this type, Hu *et al.* (2004) have shown that backfitting and profiling lead to different asymptotic distributions. However, Hu *et al.* (2004) also showed that in example 5 and equation (2) the use of the smoother that is defined in equation (5) leads to profiling and backfitting being asymptotically equivalent. Thus we would conjecture that the same equivalence holds in our general problem, a conjecture which is verified in Section 3.3. It should be noted that, as shown in Section 3.3, to obtain a $\sqrt{n}$-consistent estimator of $\mathcal{B}$, undersmoothing of the nonparametric function $\theta(z)$ is required by the backfitting method: no such undersmoothing is needed when the profile kernel method is used.

3.2. Optimal semiparametric score

To study the asymptotic properties of the profile kernel and backfitting estimators of $\mathcal{B}$, we first derive the semiparametric efficiency bound and efficient semiparametric score function in the case that $\mathcal{L}(\cdot)$ is a likelihood function.

3.2.1. *Result 2: semiparametric efficiency bound*

Assume that $(\tilde{Y}_i, \tilde{X}_i, \tilde{Z}_i)$ are independent and identically distributed, and that $\mathcal{L}(\cdot)$ is a likelihood function conditional on $(\tilde{X}, \tilde{Z})$. Then the optimal semiparametric score function is

$$\mathcal{L}_{\mathcal{B}}(\cdot) + \sum_{j=1}^{m} \mathcal{L}_{j\theta}(\cdot)\, \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0), \tag{13}$$

where the argument is $\{\tilde{Y}, \tilde{X}, \theta_0(Z_1), \ldots, \theta_0(Z_m), \mathcal{B}_0\}$, and $\theta_{\mathcal{B}}(Z_j, \mathcal{B}_0)$ is the asymptotic limit of $\hat{\theta}_{\mathcal{B}}(Z_j, \mathcal{B}_0)$ and $\mathcal{B}_0$ is the true value of $\mathcal{B}$. In addition, the asymptotic covariance matrix of the optimal semiparametric estimator is $n^{-1}\mathcal{V}^{-1}$, where

$$\mathcal{V} = \operatorname{cov}\{\mathcal{L}_{\mathcal{B}}(\theta_0, \mathcal{B}_0) + \sum_{j=1}^{m} \mathcal{L}_{j\theta}(\theta_0, \mathcal{B}_0)\, \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0)\}. \tag{14}$$

The proof of result (13) is given in Appendix A.2.

3.3. Asymptotic distribution theory

We study in this section the asymptotic properties of the profile kernel estimator $\hat{\mathcal{B}}_{\mathrm{p}}$ and the backfitting estimator $\hat{\mathcal{B}}_{\mathrm{b}}$ under a general criterion function $\mathcal{L}(\cdot)$. To study the asymptotic properties of the profile kernel estimator $\hat{\mathcal{B}}_{\mathrm{p}}$, we first provide the asymptotic properties of the kernel estimator of the derivative $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$. Define $\mathcal{L}_{j\theta\mathcal{B}}(\cdot) = \partial \mathcal{L}_{j\theta}(\tilde{Y}, \tilde{X}, \eta_1, \ldots, \eta_m, \mathcal{B})/\partial \mathcal{B}$, and

$$\begin{aligned} \varepsilon_{ij}^{\#}(\theta, \mathcal{B}) = {} & \mathcal{L}_{j\theta\mathcal{B}}\{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}), \ldots, \theta(Z_{im}), \mathcal{B}\} \\ & + \sum_{k=1}^{m} \mathcal{L}_{jk\theta}\{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}), \ldots, \theta(Z_{im}), \mathcal{B}\}\, \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}). \end{aligned}$$

As we show in Appendix A.4, $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B}_0) = \theta_{\mathcal{B}}(z, \mathcal{B}_0) + o_p(1)$, where $\theta_{\mathcal{B}}(z, \mathcal{B}_0)$ satisfies

$$0 = \sum_{j=1}^{m} f_j(z)\, E\{\varepsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0) | Z_j = z\}. \tag{15}$$

Define

$$\mathcal{F} = E\{\mathcal{L}_{\mathcal{B}\mathcal{B}} + \sum_{j=1}^{m} \mathcal{L}_{j\theta\mathcal{B}}(\cdot)\, \theta_{\mathcal{B}}^{\mathrm{T}}(Z_j, \mathcal{B}_0)\},$$

where $\mathcal{L}_{\mathcal{B}\mathcal{B}}(\cdot) = \partial^2 \mathcal{L}(\cdot)/\partial \mathcal{B}^2$.

3.3.1. *Result 3: profile kernel method*

Assume that $(\tilde{Y}_i, \tilde{X}_i, \tilde{Z}_i)$ are independent and identically distributed, and that $0 = E\{\mathcal{L}_{\mathcal{B}}(\cdot)|\tilde{Z}\} = E\{\mathcal{L}_{j\theta}(\cdot)|\tilde{Z}\}$. Suppose further that the bandwidth $h \propto n^{-c}$ with $\frac{1}{5} \leqslant c \leqslant \frac{1}{3}$. Then

$$n^{1/2}(\hat{\mathcal{B}}_p - \mathcal{B}_0) = -\mathcal{F}^{-1} n^{-1/2} \sum_{i=1}^{n} \{\mathcal{L}_{i\mathcal{B}} + \sum_{j=1}^{m} \varepsilon_{ij}\, \theta_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0)\} + o_p(1) \qquad (16)$$

$$\rightarrow \text{normal}(0, \mathcal{F}^{-1}\mathcal{V}\mathcal{F}^{-1}),$$

where $\varepsilon_{ij} = \mathcal{L}_{ij\theta}(\cdot)$ and $\mathcal{V}$ is defined in equation (14). In the case that $\mathcal{L}(\cdot)$ is a log-likelihood conditioned on $(\tilde{X}, \tilde{Z})$, $\mathcal{F} = -\mathcal{V}$, the resulting asymptotic variance is $\mathcal{V}^{-1}$, and the profile estimator is semiparametric efficient. The proof of result (16) is given in Appendix A.4.

3.3.2. *Result 4: backfitting method*

Make the same assumptions as in result 3, except that $nh^4 \rightarrow 0$ is required, i.e. undersmoothing is required. Then the backfitting estimator $\hat{\mathcal{B}}_{\text{b}}$ has the same asymptotic distribution as does the profile estimator $\hat{\mathcal{B}}_{\text{p}}$. The proof is given in Appendix A.5.

3.3.3. *Result 5: covariance matrix estimation*

Consistent estimates of $\mathcal{F}$ and $\mathcal{V}$ can be constructed as follows. Let $\hat{\mathcal{L}}_{i\mathcal{B}}$, $\hat{\mathcal{L}}_{ij\theta}$, $\hat{\mathcal{L}}_{i\mathcal{B}\mathcal{B}}$ and $\hat{\mathcal{L}}_{ij\theta\mathcal{B}}$ be the estimated versions of the quantities indicated. Let $\hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B})$ be the solution of the kernel estimating equation (36) in Appendix A.6. Then a consistent estimator of $\mathcal{V}$ is the sample covariance matrix of the terms

$$\hat{\mathcal{L}}_{i\mathcal{B}} + \sum_{j=1}^{m} \hat{\mathcal{L}}_{ij\theta}\, \hat{\theta}_{\mathcal{B}}(Z_{ij}, \hat{\mathcal{B}}).$$

Further, a consistent estimator of $\mathcal{F}$ is

$$\hat{\mathcal{F}} = n^{-1} \sum_{i=1}^{n} \{\hat{\mathcal{L}}_{i\mathcal{B}\mathcal{B}} + \hat{\mathcal{L}}_{ij\theta\mathcal{B}}\, \hat{\theta}^{\text{T}}_{\mathcal{B}}(Z_{ij}, \hat{\mathcal{B}})\}.$$

4. Pseudolikelihood with nuisance parameters

In many problems, it is convenient to estimate a subset of parameters by alternative algorithms. For example, in the partially linear model problem of Wang *et al.* (2004), the mean functions are $X_{ij}^{\text{T}}\beta_0 + \theta_0(Z_{ij})$ and the covariance matrix is $\Sigma_{\varepsilon 0}$. In our notation, $\mathcal{B}_0 = \{\beta_0^{\text{T}}, \text{vec}^{\text{T}}(\Sigma_{\varepsilon 0})\}^{\text{T}}$. Wang *et al.* (2004) provided an initial estimate $\hat{\Sigma}_{\varepsilon p}$ of $\Sigma_{\varepsilon 0}$, and then applied our algorithm only to β while pretending that $\Sigma_{\varepsilon 0}$ is known and equal to $\hat{\Sigma}_{\varepsilon p}$.

Problems such as this are easily handled in our context as follows. Suppose that $\mathcal{B}^{\text{T}} = (\kappa^{\text{T}}, \gamma^{\text{T}})$ and that we have a preliminary estimate $\hat{\gamma}_{\text{prelim}}$ with the property that it has the asymptotic expansion

$$n^{1/2}(\hat{\gamma}_{\text{prelim}} - \gamma_0) = n^{-1/2} \sum_{i=1}^{n} \mathcal{U}_i + o_p(1),$$

where $E(\mathcal{U}) = 0$. Let $e_1 = (I, 0)$ so that $\kappa = e_1\mathcal{B}$ and write $(\mathcal{F}_{11}, \mathcal{F}_{12}) = e_1\mathcal{F}$. Then, in Appendix A.4 at equation (31), we show that, for either profiling or backfitting,

$$n^{1/2}(\hat{\kappa} - \kappa_0) = -\mathcal{F}_{11}^{-1}[n^{-1/2} \sum_{i=1}^{n} \{\mathcal{L}_{i\kappa} + \sum_{j=1}^{m} \mathcal{L}_{ij\theta}\, \theta_{\kappa}(Z_{ij}, \mathcal{B}_0)\} + \mathcal{F}_{12} n^{1/2}(\hat{\gamma}_{\text{prelim}} - \gamma_0)] + o_p(1)$$

$$= -\mathcal{F}_{11}^{-1} n^{-1/2} \sum_{i=1}^{n} \{\mathcal{L}_{i\kappa} + \sum_{j=1}^{m} \mathcal{L}_{ij\theta}\theta_{\kappa}(Z_{ij}, \mathcal{B}_0) + \mathcal{F}_{12}\mathcal{U}_i\} + o_p(1),$$

from which the covariance of the asymptotic distribution of $n^{1/2}(\hat{\kappa}-\kappa_0)$ follows. In some cases, such as that investigated by Wang *et al.* (2004), $\mathcal{F}_{12}=0$, in which case the asymptotic covariance matrix becomes $\mathcal{F}_{11}^{-1}\mathcal{V}_{11}\mathcal{F}_{11}^{-1}$. In either case, a consistent estimator of the asymptotic covariance matrix is easily constructed.

5. Examples

In this section, we provide two examples to illustrate applications of our methods in the general likelihood-type framework that was described in Section 1. Our first example concerns multilevel hierarchical data where inference is based on a likelihood, whereas the second example is on longitudinal data with covariates that are measured with error where the likelihood inference is difficult and a non-likelihood criterion function is used.

5.1. Data with common Z-values

In some situations, the Z_{ij} have sets of common values in a way that the first m_1 observations have common value Z_{i1}^*, the next m_2 have common value Z_{i2}^*, etc. For example, consider problems in which there are n families, family i ($i=1,\ldots,n$) has L_i children, the jth child ($j=1,\ldots,L_i$) has a base-line measure Z_{ij}^* and repeated measures Y_{ijk} over time for $k=1,\ldots,m_{ij}$ and a possible repeated time-varying covariate X_{ijk}. Consider a three-level hierarchical model

$$Y_{ijk}=X_{ijk}^{\mathrm{T}}\beta_0+\theta_0(Z_{ij}^*)+\varepsilon_{ijk}, \tag{17}$$

where $i=1,\ldots,n$ (e.g. the ith family), $j=1,\ldots,L_i$ (e.g. the jth member in the ith family), $k=1,\ldots,m_{ij}$ (e.g. the kth time point). Equation (17) models the effect of the base-line subject level covariate Z_{ij}^* nonparametrically and other covariates X_{ijk} parametrically. Denote the covariance matrix of ε_i by Σ_i, which is a $\Sigma_{j=1}^{L_i}m_{ij}\times\Sigma_{j=1}^{L_i}m_{ij}$ matrix. Assuming that Σ_i is known, the criterion function is

$$(\tilde{Y}_i-\tilde{X}_i\beta-(\theta(Z_{i1}^*)e_{i1}^{\mathrm{T}},\ldots,\theta(Z_{iL_i}^*)e_{iL_i}^{\mathrm{T}})^{\mathrm{T}})^{\mathrm{T}}\Sigma_i^{-1}(\tilde{Y}_i-\tilde{X}_i\beta-(\theta(Z_{i1}^*)e_{i1}^{\mathrm{T}},\ldots,\theta(Z_{iL_i}^*)e_{iL_i}^{\mathrm{T}})^{\mathrm{T}}), \tag{18}$$

where e_{ij} is an $m_{ij}\times 1$ vector of 1s. Let $\varepsilon_{ij}=(\varepsilon_{ij1},\ldots,\varepsilon_{ijm_{ij}})^{\mathrm{T}}$, $\varepsilon_i=(\varepsilon_{i1}^{\mathrm{T}},\ldots,\varepsilon_{iL_i}^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\varepsilon}=(\varepsilon_1^{\mathrm{T}},\ldots,\varepsilon_n^{\mathrm{T}})^{\mathrm{T}}$. Now partition Σ_i as follows: the (jk)th block $\Sigma_{i,jk}=\mathrm{cov}(\varepsilon_{ij},\varepsilon_{ik})$ and the dimension of $\Sigma_{i,jk}$ is $m_{ij}\times m_{ik}$. Denote $\Sigma_i^{-1}=\{\Sigma_i^{jk}\}$, where the partition of Σ_i^{-1} is the same as Σ_i. Chen and Jin (2005) considered a problem that was similar to our setting without the parametric component and proposed to apply Wang's (2003) smoothing algorithm, pretending that the repeated base-line values of Z_{ij}^* from the same subject were distinct over time. Estimation based on our criterion function (18) effectively accounts for the nature that the data have common Z-values and would yield a more efficient estimator.

Specifically, for any given β, define $\mathcal{Y}_{ijk}=\mathcal{Y}_{ijk}(\beta)=Y_{ijk}-X_{ijk}^{\mathrm{T}}\beta$, and define $\mathcal{Y}_{ij}$, $\mathcal{Y}_i$ and $\tilde{\mathcal{Y}}$ in the same fashion as ε_{ij}, ε_i and $\tilde{\varepsilon}$. Define $Z_i^*=(Z_{i1}^*,\ldots,Z_{iL_i}^*)^{\mathrm{T}}$ and $\tilde{Z}^*=(Z_{11}^*,\ldots,Z_{m,L_n}^*)^{\mathrm{T}}$ and define $\tilde{X}=(X_{11},\ldots)^{\mathrm{T}}$. Then, the linear kernel estimating equation at the lth iteration is

$$\sum_{i=1}^{n}\sum_{j=1}^{L_i}K_h(Z_{ij}^*-z)\,G_{ij}(z)(0,\ldots,0,e_{ij}^{\mathrm{T}},0,\ldots,0)\Sigma_i^{-1}\{\mathcal{Y}_i-\mu_i(Z_i^*,z_0)\}=0, \tag{19}$$

where $G_{ij}(z)$ is defined in Section 2 and

$$\mu_i(Z_i^*,z_0)=(\hat{\theta}_{[l-1]}(Z_{i1}^*)e_{i1}^{\mathrm{T}},\ldots,\{\hat{\alpha}_0+\hat{\alpha}_1(Z_{ij}^*-z)\}e_{ij}^{\mathrm{T}},\ldots,\hat{\theta}_{[l-1]}(Z_{iL_i}^*)e_{iL_i}^{\mathrm{T}})^{\mathrm{T}}.$$

In Appendix A.7, we give an explicit closed form solution to equation (19): no iteration is necessary, and equation (19) is only a descriptive device. Indeed, we derive an explicit form of a smoother matrix $\mathcal{S}$ such that $\hat{\theta}(\tilde{Z}^*, \beta) = \mathcal{S}\tilde{\mathcal{Y}}(\beta) = \mathcal{S}\tilde{Y} - \mathcal{S}\tilde{X}\beta$, where $\mathcal{S}$ is given in equation (38). This means that the profile kernel estimator of β is also explicit, i.e. non-iterative, since it is the generalized least squares estimator in the model with responses $(I - \mathcal{S}_*)\tilde{Y}$ and predictors $(I - \mathcal{S}_*)\tilde{X}$, where $\mathcal{S}_*$ is the expanded version of $\mathcal{S}$ that is appropriate for the smoothing of all the responses by accounting for the common Z_{ij} within the same subject, i.e. $\mathcal{S}_* = E\mathcal{S}$, where $E = \text{diag}(e_{11}, \ldots, e_{nL_n})$ is an $N \times \Sigma_{i=1}^n L_i$ matrix and

$$N = \sum_{i=1}^{n} \sum_{j=1}^{L_i} m_{ij}$$

is the total sample size. The profile kernel estimator is

$$\hat{\beta} = \{\tilde{X}^{\mathrm{T}}(I - \mathcal{S}_*)^{\mathrm{T}}\tilde{\Sigma}^{-1}(I - \mathcal{S}_*)\tilde{X}\}^{-1}\tilde{X}^{\mathrm{T}}(I - \mathcal{S}_*)^{\mathrm{T}}\tilde{\Sigma}^{-1}(I - \mathcal{S}_*)\tilde{Y}, \tag{20}$$

where $\tilde{\Sigma} = \text{diag}(\Sigma_1, \ldots, \Sigma_n)$.

5.1.1. *Simulation study*

We applied our method to the case of $n = 100$ clusters with six observations per cluster, with $Z_{i1} = Z_{i2} = Z_{i3}$ and $Z_{i4} = Z_{i5} = Z_{i6}$, i.e. we fit the hierarchical model (17) with $n = 100$ families, $L = 2$ subjects per family and $m = 3$ repeated measures over time per subject. We assume that the correlation structure is autoregressive with correlation 0.60 between repeated measures over time and common between-subject (within-family) correlation 0.20: let Σ denote the resulting covariance matrix. The true function was $\theta_0(z) = \sin(8z - 2)$. The Z-values were generated as independent uniform distributions, whereas the X-values were bivariate independent uniform distributions minus the corresponding value of Z. The true value was $\beta_0 = (1, 1)^{\mathrm{T}}$.

The Epanechnikov kernel was used. Working independence was based on bandwidths that were selected by using the method of Ruppert *et al.* (1995). The covariance matrix $\hat{\Sigma}$ of the ε_{ij} was estimated as the sample covariance matrix of the residuals formed by a preliminary working independence regression spline fit. We used pseudolikelihood, with the estimated covariance matrix fixed as above. Both the method that ignored the fact that there were common values of Z and our method were applied with bandwidth selected via the following simple device. For a given β we formed $Y_{ij} - X_{ij}^{\mathrm{T}}\beta$ and then calculated $\hat{\theta}(\cdot)$ by using the closed form expression (38). With $\mathcal{S}$ as the smoother matrix, $\text{cov}\{\hat{\theta}(\cdot)\}$ is estimated as $\mathcal{S}\,\text{diag}(\hat{\Sigma})\mathcal{S}^{\mathrm{T}}$, and the estimated average variance of the fit follows directly. Bias was estimated as in Wang (2003). We then minimized the estimated mean-squared error as a function of the bandwidth. The estimator of the profile kernel estimator of β was calculated by using the closed form formula (20).

In 1000 simulated data sets, both weighted methods achieved over 70% greater mean-squared error efficiency for estimating β_0 than the working independence estimator. For estimating $\theta_0(z)$, the method that ignored the common Z-values was 35% more efficient in mean-squared error than working independence, but our method was 65% more efficient.

5.1.2. *Analysis of the Kenya haemoglobin data*

We applied our method to analyse a subset of the Kenya haemoglobin data to study the changes in haemoglobin level over time in the first year since birth and the risk factors of haemoglobin among Kenyan children. This subset contained $n = 68$ families with $L = 2$ children per family and $m = 4$ repeated measures per child over time in the first year since birth. Haemoglobin level was measured at each visit and visit times varied from child to child. The risk factors of interest include the mother's age at child birth, child sex and placental parasitemia density PDEN, a marker for malaria, which could affect haemoglobin. Log-transformation was applied to PDEN to make the normality assumption plausible. A preliminary analysis showed that the effect of mother's age was non-linear. We considered the semiparametric model (17) and modelled the mother's age effect nonparametrically, and sex, PDEN and time effects parametrically. Specifically, we set Z_{ij} to be the mother's age at birth, $X_{ijk} = \{\text{sex, logpden, month, (month}-4)_+\}$, where sex = 1 if female and sex = 0 if male, logpden = log(PDEN+1), the function $f_+ = f$ if $f > 0$ and $f_+ = 0$ if $f \leqslant 0$. Note that the terms {months, $(\text{month}-4)_+$} model the time effect as a piecewise linear function with a knot at 4 months. This trend is observed by preliminary analysis of the data.

In our analysis, we used pseudolikelihood, with the following modifications from the simulation. We started with an estimate of Σ as obtained from a preliminary regression spline fit and then estimated the bandwidth by using leaving one mother out cross-validation, and thus obtained estimates of $\theta_0(\cdot)$ and β_0. From this, we formed residuals $Y_{ij} - X_{ij}^{\mathrm{T}}\hat{\beta} - \hat{\theta}(Z_{ij}, \hat{\beta})$, re-estimated the covariance matrix, re-estimated the bandwidth, etc., repeating this process 10 times.

For numerical stability, we standardized the haemoglobin level. We obtained an estimated residual variance of 0.66, an estimated autocorrelation of 0.20 and an estimated between-child (within-mother) correlation of 0.13. The estimated cross-validation bandwidth was 0.23. The correlation was low or moderate in this example. In Fig. I, we compared the estimated non-

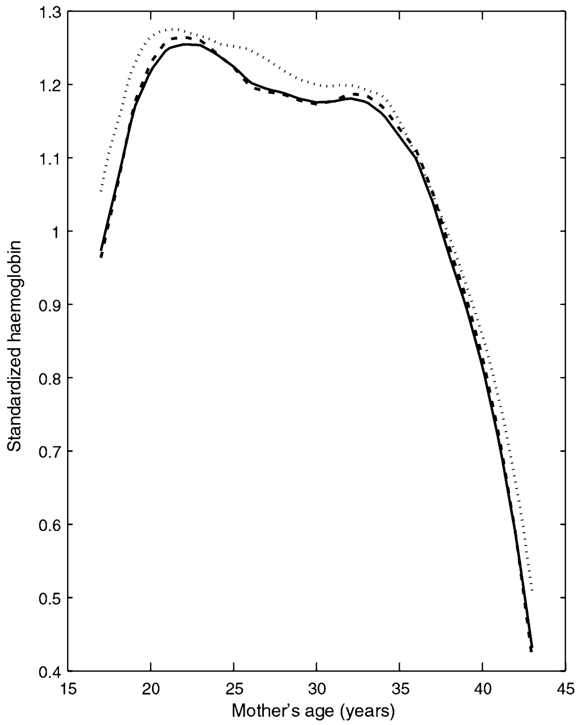


Fig. 1. Estimated nonparametric curve of the effect of mother's age at birth on child haemoglobin by fitting the semiparametric model (17) to the Kenya haemoglobin data: ———, efficient estimate when common Z-values are ignored; -------, method proposed; ·······, working independence fit

Table 1. Profile kernel estimates regression coefficients of the semiparametric model (17) applied to the Kenya haemoglobin data

	Coefficients from the following models:		
	Working independence	*Structured covariance (ignoring ties)*	*Structured covariance (accounting for ties)*
Month	−0.418 (0.0378†)	−0.397 (0.039‡) (0.043§)	−0.397 (0.039‡) (0.043§)
(Month−4)$_+$	0.147 (0.028)	0.129 (0.028) (0.028)	0.129 (0.028) (0.028)
Sex	−0.122 (0.072)	−0.122 (0.080) (0.087)	−0.122 (0.080) (0.087)
LNPDEN	−0.010 (0.013)	−0.009 (0.015) (0.017)	−0.009 (0.014) (0.016)

†Naïve standard error ignoring correlation.
‡Model-based standard error.
§Sandwich standard error.

parametric curve estimates of the effects of mother's age at birth, using the working independence kernel estimator and our proposed likelihood-based kernel estimator (with or without accounting for ties in mother's age). The estimated curves were similar. Children's haemoglobin increased with mother's age at birth for mothers who were younger than 22 years old, then decreased slightly with mother's age until at age early 30 years and then started decreasing quickly with mother's age, indicating that children are likely to have much lower haemoglobin levels if mothers give birth after early 30 years of age, i.e. giving birth after early 30 years is likely to increase children's risk of anaemia (low haemoglobin) considerably.

As expected, since the correlation was not high, the estimates of the regression coefficients β were roughly the same for the working independence kernel fit with bandwidths selected by using the method of Ruppert *et al.* (1995), the method of Wang *et al.* (2004) ignoring the common Z-values and our method accounting for the common Z-values. Estimated standard errors were computed ignoring the correlation for the working independence methods, and using the sandwich method for our likelihood-based methods. These standard errors were roughly the same in all cases. The results are given in Table 1. The haemoglobin level drops quickly after birth and decreases at a slower rate after month 4. Neither sex nor placental parasitemia density affects the haemoglobin level significantly.

5.2. Measurement error models

Here we consider the multivariate partially linear measurement error model

$$Y_{ij} = C_{ij}^{\mathrm{T}}\beta_0 + \theta_0(Z_{ij}) + \varepsilon_{ij}, \qquad (21)$$

where $\tilde{\varepsilon}_i$ has covariance matrix $\Sigma_{\varepsilon 0}$. Instead of observing C_{ij} we observe $W_{ij} = C_{ij} + U_{ij}$. Define $\tilde{U}_i = (U_{i1}, \ldots, U_{im})^{\mathrm{T}}$. These measurement errors have mean 0 and the property that $\mathrm{cov}\{\mathrm{vec}(\tilde{U}_i)\} = \Sigma_{u0}$, which is assumed here to be known. There is to date no literature on this problem other than Lin and Carroll (2000), which came to unsatisfactory conclusions such as that in panel data it was better to ignore the correlation structure in the responses.

Define $G(\Sigma_\varepsilon, \Sigma_{u0}) = E(\tilde{U}^{\mathrm{T}}\Sigma_\varepsilon^{-1}\tilde{U})$ and define $\mathcal{K}(\Sigma_{u0}, \beta) = E(\tilde{U}\beta\beta^{\mathrm{T}}\tilde{U}^{\mathrm{T}})$. Note that

$$\beta^{\mathrm{T}}G(\Sigma_\varepsilon, \Sigma_{u0})\beta = \mathrm{tr}\{\Sigma_\varepsilon^{-1}E(\tilde{U}\beta\beta^{\mathrm{T}}\tilde{U}^{\mathrm{T}})\} = \mathrm{tr}\{\Sigma_\varepsilon^{-1}\mathcal{K}(\Sigma_{u0}, \beta)\}.$$

In equation (21), $\mathcal{B} = (\beta, \tau, \Sigma_\varepsilon)$ and the criterion function is

$$\tfrac{1}{2}\log\{\det(\Sigma_\varepsilon^{-1})\} + \tfrac{1}{2}\beta^{\mathrm{T}} G(\Sigma_\varepsilon, \Sigma_{u0})\beta - \tfrac{1}{2}(\tilde{Y} - \tilde{W}\beta - \theta(\tilde{Z}))^{\mathrm{T}}\Sigma_\varepsilon^{-1}(\tilde{Y} - \tilde{W}\beta - \theta(\tilde{Z})). \qquad (22)$$

Equation (22) is new even in the *parametric* measurement error literature.

For symmetric matrices Σ, $\partial\{\log(|\Sigma|)\}/\partial\Sigma = 2\Sigma^{-1} - \mathrm{diag}(\Sigma^{-1})$ and $\partial\{\mathrm{tr}(\Sigma A)\}/\partial\Sigma = 2A - \mathrm{diag}(A)$. It is readily seen that the derivative of expression (22) with respect to β, Σ_ε and θ evaluated at the true parameters has expectation 0, and thus expression (22) satisfies the essential condition (3).

In this problem, the backfitting algorithm is computationally convenient. Of course, for given $\mathcal{B} = (\beta, \Sigma_\varepsilon)$, forming the estimate $\hat{\theta}(z, \mathcal{B})$ is easy since it is simply the estimate of Wang (2003) applied to the terms $Y_{ij} - W_{ij}^{\mathrm{T}}\beta$. Indeed, define $\mathcal{Y} = (Y_{11}, \ldots, Y_{nm})^{\mathrm{T}}$, $\mathcal{Z} = (Z_{11}, \ldots, Z_{nm})^{\mathrm{T}}$ and $\mathcal{W} = (W_{11}, \ldots, W_{nm})^{\mathrm{T}}$. Then as Lin *et al.* (2004) showed, there is a smoother matrix $\mathcal{S} = \mathcal{S}(\Sigma_\varepsilon)$ such that $\hat{\theta}(\mathcal{Z}, \mathcal{B}) = \mathcal{S}(\mathcal{Y} - \mathcal{W}\beta)$. If $\hat{\beta}_{\mathrm{c}}$, $\hat{\mathcal{B}}_{\mathrm{c}}$ and $\hat{\Sigma}_{\varepsilon,\mathrm{c}}$ are the current estimates, the updated estimates are

$$\hat{\beta}_{\mathrm{new}} = \{n^{-1} \sum_{i=1}^{n} \tilde{W}_i^{\mathrm{T}} \hat{\Sigma}_{\varepsilon,\mathrm{c}}^{-1} \tilde{W}_i - G(\hat{\Sigma}_{\varepsilon,\mathrm{c}}, \Sigma_{u0})\}^{-1} n^{-1} \sum_{i=1}^{n} \tilde{W}_i^{\mathrm{T}} \hat{\Sigma}_{\varepsilon,\mathrm{c}}^{-1} \{\tilde{Y}_i - \hat{\theta}(\tilde{Z}_i, \hat{\mathcal{B}}_c)\},$$

$$\hat{\Sigma}_{\varepsilon,\mathrm{new}} = n^{-1} \sum_{i=1}^{n} \{\tilde{Y}_i - \tilde{W}_i\hat{\beta}_{\mathrm{c}} - \hat{\theta}(\tilde{Z}_i, \hat{\mathcal{B}}_{\mathrm{c}})\}\{\tilde{Y}_i - \tilde{W}_i\hat{\beta}_{\mathrm{c}} - \hat{\theta}(\tilde{Z}_i, \hat{\mathcal{B}}_{\mathrm{c}})\}^{\mathrm{T}} - \mathcal{K}(\Sigma_{u0}, \hat{\beta}_c). \qquad (23)$$

Profile pseudolikelihood estimates are also easily constructed. Let $\tilde{\Sigma}_\varepsilon = I_n \otimes \Sigma_\varepsilon$. Let $\mathcal{W}_* = (I - \mathcal{S})\mathcal{W}$ and $\mathcal{Y}_* = (I - \mathcal{S})\mathcal{Y}$. Then, for given Σ_ε, the profile estimate of β is given by

$$\{\mathcal{W}_*^{\mathrm{T}} \tilde{\Sigma}_\varepsilon^{-1} \mathcal{W}_* - n\, G(\Sigma_\varepsilon, \Sigma_{u0})\}^{-1} \mathcal{W}_*^{\mathrm{T}} \tilde{\Sigma}_\varepsilon^{-1} \mathcal{Y}_*.$$

A simple estimate of Σ_ε is to form the working independence estimate of β and to apply equation (23).

6. Discussion

This paper has described nonparametric and semiparametric methods in cases where the nonparametric function is evaluated repeatedly within a sampling unit. Examples discussed included old and new versions of marginal longitudinal and clustered data, matched case–control studies, generalized linear mixed models, common additive models that are linked by a parameter and multivariate measurement error models. The methodology is motivated by the use of a criterion function that would be used if the problem were a parametric problem: if the criterion function is a likelihood, then our methods are semiparametric efficient. We showed that backfitting and profiling gave asymptotically the same results, although undersmoothing is needed for backfitting. We also showed how to use pseudolikelihood methods within our context when some of the parameters are more conveniently estimated by alternative algorithms. In a very different problem, namely nonparametric regression of additive models, Mammen *et al.* (1999) proposed a ‘smooth backfitting’ algorithm that does not require undersmoothing. It is of future research interest to extend this method to our setting.

Although we have motivated the methodology by basing it on criterion functions, the approach is considerably more general. Our approach really only requires the following. First, we need a set of unbiased estimating functions $\mathcal{L}_{j\theta}\{\tilde{Y}, \tilde{X}, \theta_0(Z_1), \ldots, \theta_0(Z_m), \mathcal{B}_0\}$ that satisfy condition (3). Second, we need an estimating function $\Psi_{\mathcal{B}}\{\tilde{Y}, \tilde{X}, \theta_0(Z_1), \ldots, \theta_0(Z_m), \mathcal{B}_0, \mathcal{B}_0\}$ taking the place of equation (12) and also satisfying condition (3): the double argument in $\mathcal{B}_0$ is meant to allow for the possibility of using backfitting. It is useful to use the symbols $\mathcal{L}$ and Ψ to emphasize that the derivative of the former with respect to $\mathcal{B}$ need not be the same as the derivative of the latter with respect to the jth component of θ. It can be shown that result 1 and equation (8) still hold with the same notation, as does the fundamental identity (15). The basic backfitting expansion (30) in Appendix A.3, as well as the definition of $\mathcal{F}$ in result 3, also holds with $\mathcal{L}$ replaced by Ψ. It then becomes straightforward to derive the asymptotic distribution of the estimate of $\mathcal{B}_0$: note here, however, that $\mathcal{T}_1 + \mathcal{T}_2$ need no longer be symmetric. The asymptotic covariance matrix of the resulting estimator $\hat{\mathcal{B}}$ is more complicated than that given in equation (16), because it involves the implicitly defined function $\mathcal{G}$ in equation (6). However, the bootstrap method that bootstraps clusters can be used to estimate the covariance of $\hat{\mathcal{B}}$ (Chen *et al.*, 2003).

Acknowledgements

Lin's research was supported by a grant from the National Cancer Institute (CA-76404). Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We thank Naisyin Wang for many helpful suggestions.

Appendix A: Sketch of technical arguments

Detailed proofs are given in the technical report at `http://www.bepress.com/harvardbiostat/` and also at `http://www.stat.tamu.edu/~carroll/papers.php`.

A.1. A key technical lemma

Lemma 1. Let $\hat{\theta}_{[l]}(\cdot)$ be the estimate at the lth stage of the iteration. Then

$$\hat{\theta}_{[l]}(z) = \theta_0(z) - \frac{h^2}{2} b_{[0]}(z) - n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{K_h(Z_{ij}-z)\varepsilon_{ij}}{\Omega(z)} - n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k \neq j}^{m} \frac{K_h(Z_{ij}-z)}{\Omega(z)} \times \mathcal{L}_{ijk\theta}(\cdot)\{\hat{\theta}_{[l-1]}(Z_{ik}) - \theta_0(Z_{ik})\} + o_p(n^{-1/2}), \tag{24}$$

where $b_{[0]}(z) = \theta^{(2)}(z)$, and the argument is $\{\tilde{Y}_i, \tilde{X}_i, \theta(z_{i1}), \ldots, \alpha_0 + \alpha_1(z_{ij} - z)/R, \ldots, \theta(z_{im})\}$. Here is a brief sketch of equation (24). By Taylor series expansion, we have

$$0 = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\, G_{ij}(z,h)\, \mathcal{L}_{ij\theta}(\cdot) + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\, G_{ij}(z,h)\, G_{ij}^{\mathrm{T}}(z,h) \times \mathcal{L}_{ijj\theta}(\cdot) \begin{pmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\alpha}_1 - \alpha_1 \end{pmatrix} + o_p(n^{-1/2}),$$

where the argument is $\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[l-1]}(Z_{i1}), \ldots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \ldots, \hat{\theta}_{[l-1]}(Z_{im})\}$. It is easily seen that the sum in the second argument converges at the appropriate rate to $\Omega(z)I_2$, where I_2 is the 2×2 identity matrix (again, this is because K has variance 1.0). Hence,

$$\begin{aligned} -\Omega(z)(\hat{\alpha}_0 - \alpha_0) &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\, \mathcal{L}_{ij\theta}(\cdot) + o_p(n^{-1/2}) \\ &= A_{1n} + A_{2n} + o_p(n^{-1/2}), \end{aligned}$$

$$A_{1n} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\, \mathcal{L}_{j\theta}\{\tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \ldots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \ldots, \theta_0(Z_{im})\},$$

$$\begin{aligned} A_{2n} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)[\mathcal{L}_{j\theta}\{&\tilde{Y}_i, \tilde{X}_i, \hat{\theta}_{[l-1]}(Z_{i1}), \ldots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \ldots, \hat{\theta}_{[l-1]}(Z_{im})\} \\ &- \mathcal{L}_{j\theta}\{\tilde{Y}_i, \tilde{X}_i, \theta_0(Z_{i1}), \ldots, \alpha_0 + \alpha_1(Z_{ij} - z)/h, \ldots, \theta_0(Z_{im})\}]. \end{aligned}$$

Some calculation shows that

$$A_{1n} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\varepsilon_{ij} + (h^2/2)\, b_{[0]}(z)\, \Omega(z) + o_p(n^{-1/2}),$$

and A_{2n} is equal to the third term in equation (24).

A.2. Proof of result 2: semiparametric efficient score

We use Begun *et al.* (1983). In their set-up, their 'f' is our $\exp(\mathcal{L})$, their 'θ' is our $\mathcal{B}$ and their 'g' is our θ. It is easily derived that their '$2\rho_\theta / f^{1/2}$' is our $\mathcal{L}_\mathcal{B}$. Similarly, for an arbitrary function $\gamma(\cdot)$, their '$2A\beta/f^{1/2}$' is $\Sigma_{j=1}^{m} \mathcal{L}_{j\theta}(\cdot)\gamma(Z_j)$. This means that their equation (3.1) is the following. The semiparametric optimal score is of the form

$$\mathcal{L}_\mathcal{B}(\cdot) - \sum_{j=1}^{m} \mathcal{L}_{j\theta}(\cdot)\, \gamma_*(Z_j),$$

where $\gamma_*(\cdot)$ is such that, for all $\gamma(\cdot)$,

$$0 = E[\{\mathcal{L}_\mathcal{B}(\cdot) - \sum_{j=1}^{m} \mathcal{L}_{j\theta}(\cdot)\, \gamma_*(\cdot)\} \sum_{k=1}^{m} \mathcal{L}_{k\theta}(\cdot)\, \gamma(Z_k)]. \tag{25}$$

We now show that $\gamma_*(\cdot) = -\theta_\mathcal{B}(\cdot)$ satisfies condition (25). To see this, interchange the indices j and k and note that condition (25) means that we must show that for arbitrary $\gamma(\cdot)$

$$0 = E\{\sum_{j=1}^{m} \mathcal{L}_\mathcal{B}(\cdot)\, \mathcal{L}_{j\theta}(\cdot)\gamma(Z_j) + \sum_{j=1}^{m} \sum_{k=1}^{m} \mathcal{L}_{j\theta}(\cdot)\, \mathcal{L}_{k\theta}(\cdot)\, \theta_\mathcal{B}(Z_k)\, \gamma(Z_j)\}.$$

Condition on $(\tilde{X}, \tilde{Z})$ and note that, because $\mathcal{L}(\cdot)$ is a likelihood function given $(\tilde{X}, \tilde{Z})$,

$$E\{\mathcal{L}_\mathcal{B}(\cdot)\, \mathcal{L}_{j\theta}(\cdot) | \tilde{X}, \tilde{Z}\} = -E\{\mathcal{L}_{j\theta\mathcal{B}}(\cdot) | \tilde{X}, \tilde{Z}\},$$
$$E\{\mathcal{L}_{j\theta}(\cdot)\, \mathcal{L}_{k\theta}(\cdot) | \tilde{X}, \tilde{Z}\} = -E\{\mathcal{L}_{jk\theta}(\cdot) | \tilde{X}, \tilde{Z}\}.$$

Thus we must show that, for arbitrary $\gamma(\cdot)$,

$$\begin{aligned} 0 &= \sum_{j=1}^{m} E[\gamma(Z_j)\{\mathcal{L}_{j\theta\mathcal{B}}(\cdot) + \sum_{k=1}^{m} \mathcal{L}_{jk\theta}(\cdot)\, \theta_\mathcal{B}(Z_k)\}] \\ &= \sum_{j=1}^{m} E\{\gamma(Z_j)\, \varepsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0)\}, \end{aligned} \tag{26}$$

where $\varepsilon_{ij}^{\#}(\theta_0, \mathcal{B}_0)$ is defined in Section 3.3. This last step follows by conditioning the expectation in equation (26) on Z_j and then applying equation (29) below.

A.3. Sketch proof of equation (15): fundamental identity

Since

$$n^{-1} \sum_{i=1}^{n} \{(Z_i - z)/h\}\, K_h(Z_i - z) = o_p(1)$$

one can show that, for any $\mathcal{B}$,

$$0 = \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\, \mathcal{L}_{j\theta}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \ldots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B}\}. \tag{27}$$

Differentiating equation (27) with respect to $\mathcal{B}$, we obtain

$$0 = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\{\mathcal{L}_{j\theta\mathcal{B}}(\cdot) + \sum_{k=1}^{m} \mathcal{L}_{ijk\theta}(\cdot)\, \hat{\theta}_\mathcal{B}(Z_{ik}, \mathcal{B})\},$$

with argument $\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}), \ldots, \hat{\theta}(Z_{im}, \mathcal{B}), \mathcal{B}\}$. Taking limits and evaluating at $\mathcal{B}_0$ yields equation (15).

Recall the definition of $\varepsilon^{\#}_{ij}(\theta, \mathcal{B})$ that is given in Section 3.3. Define

$$H_j(z) = E\{\varepsilon^{\#}_{ij}(\theta_0, \mathcal{B}_0)|Z_j = z\}. \tag{28}$$

It follows from equation (15) that $0 = \Sigma_{j=1}^{m} f_j(z)\, H_j(z)$, and hence that, for any function $B(\cdot)$,

$$0 = E\Big\{\sum_{j=1}^{m} B(Z_j)\, H_j(Z_j)\Big\}. \tag{29}$$

We shall use this equality repeatedly.

A.4. Sketch proof of result 3: asymptotic distribution for profiling

Recall that $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$, where $\mathcal{F}_1 = E(\mathcal{L}_{\mathcal{B}\mathcal{B}})$ and $\mathcal{F}_2 = E\{\Sigma_{j=1}^{m} \mathcal{L}_{j\theta\mathcal{B}}(\cdot)\, \theta^{\mathrm{T}}_{\mathcal{B}}(Z_j, \mathcal{B}_0)\}$. Also, define

$$\mathcal{F}_3 = E\Big\{\sum_{j=1}^{m}\sum_{k=1}^{m} \mathcal{L}_{jk\theta}(\cdot)\, \theta_{\mathcal{B}}(Z_j, \mathcal{B}_0)\, \theta^{\mathrm{T}}_{\mathcal{B}}(Z_k, \mathcal{B}_0)\Big\}.$$

It is an easy consequence of equation (29) that $\mathcal{F}_2 + \mathcal{F}_3 = 0$, so that $\mathcal{F} = \mathcal{F}_1 + 2\mathcal{F}_2 + \mathcal{F}_3$.

Let $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B}) = \partial\hat{\theta}(z, \mathcal{B})/\partial\mathcal{B}$, and let its limit as $n \to \infty$ be $\theta_{\mathcal{B}}(z, \mathcal{B})$. Then the profile estimator solves the equation $0 = A_1(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta}) + A_2(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta})$, where

$$A_1(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta}) = n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_{i\mathcal{B}_{\mathrm{p}}}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \hat{\mathcal{B}}_{\mathrm{p}}), \ldots, \hat{\theta}(Z_{im}, \hat{\mathcal{B}}_{\mathrm{p}}), \hat{\mathcal{B}}_{\mathrm{p}}\},$$

$$A_2(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta}) = n^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta}\{\tilde{Y}_i, \tilde{X}_i, \hat{\theta}(Z_{i1}, \hat{\mathcal{B}}_{\mathrm{p}}), \ldots, \hat{\theta}(Z_{im}, \hat{\mathcal{B}}_{\mathrm{p}}), \hat{\mathcal{B}}_{\mathrm{p}}\}\, \hat{\theta}_{\mathcal{B}}(Z_{ij}, \hat{\mathcal{B}}_{\mathrm{p}}).$$

A Taylor series expansion shows that

$$A_1(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta}) = n^{-1/2} \sum_{i=1}^{n} \mathcal{L}_{i\mathcal{B}}(\cdot) + n^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta\mathcal{B}}(\cdot)\{\hat{\theta}(Z_{ij}, \mathcal{B}_0) - \theta(Z_{ij})\} + (\mathcal{F}_1 + \mathcal{F}_2)n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0) + o_p(1), \tag{30}$$

where the symbol '·' here means evaluated at θ and $\mathcal{B}_0$. Similarly, we have that

$$\begin{aligned} A_2(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta}) = {}& n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta\mathcal{B}}(\cdot)\theta^{\mathrm{T}}_{\mathcal{B}}(Z_{ij})n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0) + n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot)\, \theta_{\mathcal{B}\mathcal{B}}(Z_{ij})n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0) \\ & + n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} \mathcal{L}_{ijk\theta}(\cdot)\, \theta_{\mathcal{B}}(Z_{ij})\, \theta^{\mathrm{T}}_{\mathcal{B}}(Z_{ik})n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0) \\ & + n^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta}\{\tilde{Y}_1, \tilde{X}_i, \hat{\theta}(Z_{i1}, \mathcal{B}_0), \ldots, \hat{\theta}(Z_{im}, \mathcal{B}_0), \mathcal{B}_0\}\hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) + o_p(1). \end{aligned}$$

The first and third terms sum to $(\mathcal{F}_2 + \mathcal{F}_3)n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0) + o_p(1)$. Because $E\{\mathcal{L}_{ij\theta}(\cdot)|\tilde{Z}_i\} = 0$, the second term is $o_p(1)$. The last term can be decomposed, so that

$$\begin{aligned} A_2(\hat{\mathcal{B}}_{\mathrm{p}}, \hat{\theta}) = {}& (\mathcal{F}_2 + \mathcal{F}_3)n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0) + n^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot)\, \theta_{\mathcal{B}}(Z_{ij}) \\ & + n^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} \mathcal{L}_{ijk\theta}(\cdot)\theta_{\mathcal{B}}(Z_{ij})\{\hat{\theta}(Z_{ik}, \mathcal{B}_0) - \theta(Z_{ik})\} \\ & + n^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot)\{\hat{\theta}_{\mathcal{B}}(Z_{ij}, \mathcal{B}_0) - \theta_{\mathcal{B}}(Z_{ij})\} + o_p(1). \end{aligned}$$

Recall that $\varepsilon_{ij} = \mathcal{L}_{ij\theta}(\cdot)$ and $H_j(z)$ as defined in equation (28). If

$$P_{ij} = \mathcal{L}_{ij\theta\mathcal{B}}(\cdot) + \sum_{k=1}^{m} \mathcal{L}_{ijk\theta}(\cdot)\, \theta_{\mathcal{B}}(Z_{ik}),$$

we have

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}}-\mathcal{B}_0)=n^{-1/2}\sum_{i=1}^{n}\{\mathcal{L}_{i\mathcal{B}}+\sum_{j=1}^{m}\varepsilon_{ij}\,\theta_{\mathcal{B}}(Z_j,\mathcal{B}_0)\}+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}H_j(Z_{ij})\{\hat{\theta}(Z_{ij},\mathcal{B}_0)-\theta(Z_{ij})\}$$
$$+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\{P_{ij}-H_j(Z_{ij})\}\{\hat{\theta}(Z_{ij},\mathcal{B}_0)-\theta(Z_{ij})\}$$
$$+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathcal{L}_{ij\theta}(\cdot)\{\hat{\theta}_{\mathcal{B}}(Z_{ij},\mathcal{B}_0)-\theta_{\mathcal{B}}(Z_{ij})\}+o_p(1). \quad (31)$$

We can show that the last three terms of equation (31) are all $o_p(1)$. The proof of the last term uses the asymptotic expansion

$$\hat{\theta}_{\mathcal{B}}(z,\mathcal{B}_0)-\theta_{\mathcal{B}}(z,\mathcal{B}_0)=(h^2/2)\{b_1(z)+b_2(z)\}-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(Z_{ij}-z)\varepsilon_{ij}\,\Omega_1(z)$$
$$-n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}K_h(Z_{ij}-z)\,\varepsilon_{ij}^{\#}(\theta_0,\mathcal{B}_0)\,\Omega_2(z)+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}\mathcal{G}_1(z,Z_{ij})$$
$$+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{ij}^{\#}(\theta_0,\mathcal{B}_0)\,\mathcal{G}_2(z,Z_{ij})+o_p(n^{-1/2}), \quad (32)$$

for some functions $b_j(\cdot)$, $\Omega_j(\cdot)$ and $\mathcal{G}_j(\cdot)$ $(j=1,2)$. The detailed proofs are given in the technical report that is mentioned at the end of Section 1.

A.5. Sketch proof of result 4: asymptotic distribution for backfitting

Using the notation of Appendix A.4, for backfitting we are solving the equation $0=A_1(\hat{\mathcal{B}}_{\mathrm{b}},\hat{\theta})$. Using the results in Appendix A.4, we have

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_{\mathrm{b}}-\mathcal{B}_0)=n^{-1/2}\sum_{i=1}^{n}\mathcal{L}_{i\mathcal{B}}(\cdot)-n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\mathcal{L}_{ijk\theta}(\cdot)\,\theta_{\mathcal{B}}(Z_{ik},\mathcal{B}_0)\{\hat{\theta}(Z_{ij},\mathcal{B}_0)-\theta_0(Z_{ij})\}. \quad (33)$$

Since the profile estimator satisfies

$$-\mathcal{F}n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}}-\mathcal{B}_0)=n^{-1/2}\sum_{i=1}^{n}\{\mathcal{L}_{i\mathcal{B}}(\cdot)+\sum_{j=1}^{m}\mathcal{L}_{ij\theta}(\cdot)\,\theta_{\mathcal{B}}(Z_{ij},\mathcal{B}_0)\}+o_p(1), \quad (34)$$

we see that we must show that the second terms in equations (33) and (34) are asymptotically equivalent. Make the definitions

$$\Omega(z_0)=\sum_{j=1}^{m}f_j(z_0)\,E\{\mathcal{L}_{jj\theta}(\cdot)|Z_j=z_0\},$$

$$P_1(z_0)=\sum_{j=1}^{m}f_j(z_0)\,E\{\mathcal{L}_{j\theta\mathcal{B}}(\cdot)|Z_j=z_0\}/\Omega(z_0),$$

$$P_2(z_0)=E\Big[\sum_{j=1}^{m}E\{\mathcal{L}_{j\theta\mathcal{B}}(\cdot)|Z_j\}\,\mathcal{G}(Z_j,z_0)/\Omega(Z_j)\Big],$$

$$P_3(z_0)=\sum_{j=1}^{m}\sum_{k\neq j}^{m}\{f_j(z_0)/\Omega(z_0)\}\,E\{\mathcal{L}_{jk\theta}(\cdot)\,\theta_{\mathcal{B}}(Z_k,\mathcal{B}_0)|Z_j=z_0\}.$$

Recalling equation (15), we see that

$$P_1(z_0)=-\sum_{j=1}^{m}\sum_{k=1}^{m}\{f_j(z_0)/\Omega(z_0)\}\,E\{\mathcal{L}_{jk\theta}(\cdot)\,\theta_{\mathcal{B}}(Z_k,\mathcal{B}_0)|Z_j=z_0\}$$
$$=-\theta_{\mathcal{B}}(z_0,\mathcal{B}_0)-P_3(z_0),$$

$$P_2(z_0) = \int P_1(z)\, \mathcal{G}(z, z_0)\, \mathrm{d}z$$
$$= -\int \theta_{\mathcal{B}}(z, \mathcal{B}_0)\, \mathcal{G}(z, z_0)\, \mathrm{d}z - \int P_3(z)\, \mathcal{G}(z, z_0)\, \mathrm{d}z.$$

We now plug in result (8) into the second term of equation (33). Noting the assumption that $nh^4 \to 0$, some calculation shows that this second term is asymptotically equivalent to $C_{n1} + C_{n2}$, where

$$C_{n1} = -n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot)\, P_1(Z_{ij}) + o_p(1),$$

$$C_{n2} = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot) P_2(Z_{ij}) + o_p(1).$$

Collecting the expressions for $P_1(z)$ and $P_2(z)$, it thus follows that

$$-\mathcal{F}\{n^{1/2}(\hat{\mathcal{B}}_{\mathrm{b}} - \mathcal{B}_0) - n^{1/2}(\hat{\mathcal{B}}_{\mathrm{p}} - \mathcal{B}_0)\} = S_n + o_p(1)$$

where

$$S_n = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot)\{P_3(Z_{ij}) - \int P_3(z)\, \mathcal{G}(z, Z_{ij})\, \mathrm{d}z + \int \theta_{\mathcal{B}}(z, \mathcal{B}_0)\, \mathcal{G}(z, Z_{ij})\, \mathrm{d}z.$$

Now, using the definition of $Q(z_1, z_2)$ above equation (6) and the definition of $\mathcal{A}(\cdot)$ just above equation (6), one can show that

$$\int \theta_{\mathcal{B}}(z, \mathcal{B}_0)\, \mathcal{G}(z, z_0)\, \mathrm{d}z = P_3(z_0) - \int \theta_{\mathcal{B}}(z, \mathcal{B}_0)\, \mathcal{A}(\mathcal{G}, z, z_0)\, \mathrm{d}z.$$

Hence

$$S_n = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_{ij\theta}(\cdot) \int \{\theta_{\mathcal{B}}(z, \mathcal{B}_0)\, \mathcal{A}(\mathcal{G}, z, Z_{ij}) - P_3(z)\, \mathcal{G}(z, Z_{ij})\}\, \mathrm{d}z.$$

We thus need to show that, for all z_0,

$$0 = \int \{\theta_{\mathcal{B}}(z, \mathcal{B}_0)\, \mathcal{A}(\mathcal{G}, z, z_0) - P_3(z)\, \mathcal{G}(z, z_0)\}\, \mathrm{d}z.$$

Its proof is given in the technical report that was mentioned at the end of Section 1.

A.6. Computation of $\hat{\theta}_{\mathcal{B}}(z, \mathcal{B})$

We first derive the first-degree polynomial kernel estimating equation for $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$. Differentiating equation (11) with respect to $\mathcal{B}$ gives the linear kernel estimating equation for $\theta_{\mathcal{B}}(z; \mathcal{B})$. Let $\theta(z, \mathcal{B})$ be the asymptotic limit of $\hat{\theta}(z, \mathcal{B})$. Let $\Theta_i(\tilde{Z}_i, \mathcal{B}) = (\theta(Z_{i1}, \mathcal{B}), \ldots, \theta(Z_{im}, \mathcal{B}))^{\mathrm{T}}$ and $\Theta_{i\mathcal{B}}(\tilde{Z}_i, \mathcal{B}) = (\theta_{\mathcal{B}}(Z_{i1}, \mathcal{B}), \ldots, \theta_{\mathcal{B}}(Z_{im}, \mathcal{B}))^{\mathrm{T}}$. Denote the estimating function

$$e_{ij}(\tilde{Y}_i, \tilde{X}_i, \Theta_i, \Theta_{i\mathcal{B}}) = \mathcal{L}_{ij\theta\mathcal{B}}(\cdot) + \sum_{k=1}^{m} \mathcal{L}_{ijk\theta}(\cdot)\, \theta_{\mathcal{B}}(Z_{ik}, \mathcal{B}), \tag{35}$$

where $\cdot = \{\tilde{Y}_i, \tilde{X}_i, \theta(Z_{i1}, \mathcal{B}), \ldots, \theta(Z_{im}, \mathcal{B})\}$. Equation (35) is the same as $\varepsilon^{\#}_{ij}(\theta, \mathcal{B})$ that was defined in Section 3.3, but as shown below a slightly different notation is needed in our arguments. Then

$$\sum_{j=1}^{m} E\{e_{ij}(\cdot) | Z_{ij} = z\}\, f_j(z) = 0;$$

see equation (29). The kernel estimating equation for $\hat{\theta}_{\mathcal{B}}(z; \mathcal{B})$ can be written as

$$R_n = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} K_h(Z_{ij} - z)\, G_{ij}(z, h)\, e_{ij}\{\tilde{Y}_i, \tilde{X}_i, \hat{\Theta}_{ij}(z, \tilde{Z}_i, \mathcal{B}), \hat{\Theta}_{ij\mathcal{B}}(z, \tilde{Z}_i, \mathcal{B})\} = 0, \tag{36}$$

where

$$\hat{\Theta}_{ij}(z,\tilde{Z}_i,\mathcal{B})=(\hat{\theta}(Z_{i1},\mathcal{B}),\ldots,\hat{\theta}(z,\mathcal{B})+h\,\hat{\theta}^{(1)}(z,\mathcal{B})(Z_{ij}-z)/h,\ldots,\hat{\theta}(Z_{im},\mathcal{B}))^{\mathrm{T}},$$

$$\hat{\Theta}_{ij\mathcal{B}}(z,\tilde{Z}_i,\mathcal{B})=(\hat{\theta}_{\mathcal{B}}(Z_{i1},\mathcal{B}),\ldots,\hat{\theta}_{\mathcal{B}}(z,\mathcal{B})+h\,\hat{\theta}_{\mathcal{B}}^{(1)}(z,\mathcal{B})(Z_{ij}-z)/h,\ldots,\hat{\theta}_{\mathcal{B}}(Z_{im},\mathcal{B}))^{\mathrm{T}}.$$

Equation (36) can be used to show that $\hat{\theta}_{\mathcal{B}}(z,\mathcal{B})$ has the asymptotic expansion (32), and it can be computed by a similar algorithm to that which was used to compute $\hat{\theta}(z,\mathcal{B})$. If we refer to equation (2) of Lin *et al.* (2004), we can make the following substitutions. First replace their $B_{ij}^{\mathrm{T}}(t)V^{-1}Y_i$ by $G_{ij}(z,h)\,\mathcal{L}_{ij\theta\mathcal{B}}\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}_{ij}(z,\tilde{Z}_i,\mathcal{B}),\mathcal{B}\}$. Then replace $B_{ij}^{\mathrm{T}}(t)V^{-1}\mu_{i(j)}(t)$ by

$$G_{ij}(z,h)\sum_{k=1}^{m}\mathcal{L}_{ijk\theta\mathcal{B}}\{\tilde{Y}_i,\tilde{X}_i,\hat{\theta}_{ij}(z,\tilde{Z}_i,\mathcal{B}),\mathcal{B}\}\,\hat{\theta}_{ij\mathcal{B}}(z,\tilde{Z}_i,\mathcal{B}).$$

Although this is a vector form rather than the scalar form in Lin *et al.* (2004), their same method can be used to find an explicit, closed form solution for $\hat{\theta}_{\mathcal{B}}(z,\mathcal{B})$.

A.7. Explicit algorithm for method in Section 5.1

Equation (19) can be rewritten as

$$\sum_{i=1}^{n}\sum_{j=1}^{L_i}K_h(Z_{ij}^{*}-z_0)\,G_{ij}(z_0)[e_{ij}^{\mathrm{T}}\Sigma_i^{jj}\{\mathcal{Y}_{ij}-G_{ij}(z_0)^{\mathrm{T}}\alpha e_{ij}\}+e_{ij}^{\mathrm{T}}\sum_{k\neq j}^{L_i}\Sigma_i^{jk}\{\mathcal{Y}_{ik}-\hat{\theta}_{[l-1]}(Z_{ik}^{*})e_{ik}\}],$$

where $\mathcal{Y}_{ij}=(\mathcal{Y}_{ij1},\ldots,\mathcal{Y}_{ijm_{ij}})^{\mathrm{T}}$ is an $m_{ij}\times 1$ vector and $\mathcal{Y}_i=(\mathcal{Y}_{i1}^{\mathrm{T}},\ldots,\mathcal{Y}_{iL_i}^{\mathrm{T}})^{\mathrm{T}}$. It follows that

$$\left\{\sum_{i=1}^{n}\sum_{j=1}^{L_i}K_h(Z_{ij}^{*}-z_0)\,G_{ij}(z_0)e_{ij}^{\mathrm{T}}\Sigma_i^{jj}e_{ij}\,G_{ij}^{\mathrm{T}}(z_0)\right\}\hat{\alpha}$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{L_i}K_h(Z_{ij}^{*}-z_0)\,G_{ij}(z_0)\left[e_{ij}^{\mathrm{T}}\Sigma_i^{jj}e_{ij}\hat{\theta}_{[l-1]}(Z_{ij}^{*})+e_{ij}^{\mathrm{T}}\sum_{k=1}^{L_i}\Sigma_i^{jk}\{\mathcal{Y}_{ik}-\hat{\theta}_{[l-1]}(Z_{ik}^{*})e_{ik}\}\right]. \tag{37}$$

Denote by $M=\Sigma_{i=1}^{n}\Sigma_{j=1}^{L_i}m_{ij}$ the total sample size and $L=\Sigma_{i=1}^{n}L_i$ the total number of family members, i.e. the number of levels of the second hierarchical level. Let $\tilde{G}(z_0)=(G_{11}(z_0),\ldots,G_{nL_n}(z_0))^{\mathrm{T}}$, which is an $L\times p$ design matrix, $\tilde{Z}=(Z_{11}^{*},\ldots,Z_{nL_n}^{*})^{\mathrm{T}}$ be an $L\times 1$ vector containing distinct observed values of Zs, $K_{dh}(z_0)=\mathrm{diag}\{K_h(Z_{11}^{*}-z_0),\ldots,K_h(Z_{nL_n}^{*}-z_0)\}$, which is an $L\times L$ matrix, $E=\mathrm{diag}(e_{11},\ldots,e_{nL_n})$, which is an $M\times L$ matrix, $\tilde{\Sigma}^{d}=\mathrm{diag}(\Sigma_1^{d},\ldots,\Sigma_n^{d})$, $\Sigma_i^{d}=\mathrm{diag}(\Sigma_i^{11},\ldots,\Sigma_i^{L_iL_i})$ and $\tilde{\Sigma}=\mathrm{diag}(\Sigma_1,\ldots,\Sigma_n)$, $\mathcal{Y}=(\mathcal{Y}_1^{\mathrm{T}},\ldots,\mathcal{Y}_n^{\mathrm{T}})^{\mathrm{T}}$. Note that $\hat{\theta}^{(l+1)}(z_0)=\hat{\alpha}_0$. Writing equation (37) in a matrix form, simple calculation shows that

$$\hat{\theta}^{(l+1)}(z_0)=\delta^{\mathrm{T}}\{\tilde{G}(z_0)^{\mathrm{T}}K_{dh}(z_0)E^{\mathrm{T}}\tilde{\Sigma}^{d}E\,\tilde{G}(z_0)\}^{-1}\,\tilde{G}(z_0)^{\mathrm{T}}K_{dh}(z_0)\{E^{\mathrm{T}}\tilde{\Sigma}^{-1}\mathcal{Y}+E^{\mathrm{T}}(\tilde{\Sigma}^{d}-\tilde{\Sigma}^{-1})E\hat{\theta}_{[l-1]}(\tilde{Z}^{*})\},$$

where $\delta=(1,0,\ldots,0)^{\mathrm{T}}$. Let

$$K_{wh}^{\mathrm{T}}(z_0)=\delta^{\mathrm{T}}\{\tilde{G}(z_0)^{\mathrm{T}}K_{dh}(z_0)E^{\mathrm{T}}\tilde{\Sigma}^{d}E\tilde{G}(z_0)\}^{-1}\,\tilde{G}(z_0)^{\mathrm{T}}K_{dh}(z_0),$$

and $K_w=(K_{wh}(Z_{11}^{*}),\ldots,K_{wh}(Z_{nL_n}^{*}))^{\mathrm{T}}$, which is an $L\times L$ matrix. Then we have

$$\hat{\theta}^{(l+1)}(\tilde{Z}^{*})=K_w\{E^{\mathrm{T}}\tilde{\Sigma}^{-1}\mathcal{Y}+E^{\mathrm{T}}(\tilde{\Sigma}^{d}-\tilde{\Sigma}^{-1})E\hat{\theta}_{[l-1]}(\tilde{Z}^{*})\}.$$

Write $\hat{\theta}_{[l]}(\tilde{Z}^{*})=\mathcal{S}_{[l]}E^{\mathrm{T}}\tilde{\Sigma}^{-1}\mathcal{Y}$. Note that $\mathcal{S}_{[l]}$ is an $L\times L$ square matrix. At convergence $\mathcal{S}_{[l]}\rightarrow\mathcal{S}$, where $\mathcal{S}$ satisfies

$$\mathcal{S}=K_w\{I+E^{\mathrm{T}}(\tilde{\Sigma}^{d}-\tilde{\Sigma}^{-1})\}E\mathcal{S}.$$

It follows that

$$\mathcal{S}=\{I+K_wE^{\mathrm{T}}(\tilde{\Sigma}^{-1}-\tilde{\Sigma}^{d})E\}^{-1}K_w.$$

Hence at convergence

$$\hat{\theta}(\tilde{Z}^{*})=\{I+K_wE^{\mathrm{T}}(\tilde{\Sigma}^{-1}-\tilde{\Sigma}^{d})E\}^{-1}K_wE^{\mathrm{T}}\tilde{\Sigma}^{-1}\mathcal{Y}. \tag{38}$$

If $m_{ij}\equiv 1$ then $E=I$. The results then reduce to those in Lin *et al.* (2004).

Note that E, $\tilde{\Sigma}^{-1}$ and $\tilde{\Sigma}^{d}$ are all block diagonal matrices. The above matrix calculations can then be greatly simplified. Specifically, partition K_w as an $n\times n$ block matrix with the (i,i')th block denoted by $K_{w,ii'}$ which is an $L_i \times L_{i'}$ matrix. Write $E = \text{diag}(E_1, \ldots, E_n)$ and $K_{dh} = \text{diag}(K_{dh,1}, \ldots, K_{dh,n})$, where $E_i = \text{diag}(e_{i1}, \ldots, e_{in_i})$ and $K_{dh,i}(z_0) = \text{diag}\{K_h(Z^*_{i1} - z_0), \ldots, K_h(Z^*_{iL_i} - z_0)\}$. Write $\tilde{G}(z_0) = (\tilde{G}_i(z_0)^{\mathrm{T}}, \ldots, \tilde{G}_n(z_0)^{\mathrm{T}})^{\mathrm{T}}$. Then

$$K^{\mathrm{T}}_{wh}(z_0) = \delta^{\mathrm{T}}\left\{\sum_{i=1}^{n} \tilde{G}_i(z_0)^{\mathrm{T}} K_{dh,i}(z_0) E_i^{\mathrm{T}} \tilde{\Sigma}_i^d E_i \tilde{G}_i(z_0)\right\}^{-1} \{\tilde{G}_1(z_0)^{\mathrm{T}} K_{dh,1}(z_0), \ldots, \tilde{G}_n(z_0)^{\mathrm{T}} K_{dh,n}(z_0)\}.$$

For equation (38), partition the matrix $K_w E^{\mathrm{T}}(\tilde{\Sigma}^{-1} - \tilde{\Sigma}^d)E$ in the same fashion as K_w into an $n \times n$ block matrix and the computation can be simplified in a similiar way.

References

Begun, J. H., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983) Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.*, **11**, 432–452.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Carroll, R. J., Härdle, W. and Mammen, E. (2002) Estimation in an additive model when components are linked parametrically. *Econometr. Theory*, **18**, 886–912.

Chen, K. and Jin, Z. (2005) Local polynomial regression analysis of clustered data. *Biometrika*, **92**, 59–74.

Chen, X., Linton, O. and Van Keilegom, I. (2003) Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591–1608.

Fan, J. and Li, R. (2004) New estimation and model selection procedures for semiparametric modeling in longitudinal data. *J. Am. Statist. Ass.*, **99**, 710–723.

Hafner, C. M. (1998) *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility.* Heidelberg: Physica.

Heagerty, P. J. and Kurland, B. F. (2001) Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika*, **88**, 973–985.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, Y. (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Hu, Z., Wang, N. and Carroll, R. J. (2004) Profile-kernel versus backfitting in the partially linear model for longitudinal/clustered data. *Biometrika*, **91**, 251–262.

Lin, D. Y. and Ying, Z. (2001) Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *J. Am. Statist. Ass.*, **96**, 103–126.

Lin, X. and Carroll, R. J. (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Ass.*, **95**, 520–534.

Lin, X. and Carroll, R. J. (2001) Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Ass.*, **96**, 1045–1056.

Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2004) Equivalent kernels of smoothing splines in nonparametric regression for longitudinal/clustered data. *Biometrika*, **91**, 177–194.

Linton, O. B. and Nielson, J. P. (1995) A kernel method for estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–101.

Mammen, E., Linton, O. and Nielsen, J. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.

Rice, J. A. and Wu, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257–1270; correction, **91** (1996), 1380.

Schaid, D. J. (1999) Case-parents design for gene-environment interaction. *Genet. Epidem.*, **16**, 261–273.

Severini, T. A. and Staniswalis, J. G. (1994) Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Ass.*, **89**, 501–511.

Wang, N. (2003) Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, **90**, 43–52.

Wang, N., Carroll, R. J. and Lin, X. (2005) Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Statist. Ass.*, **100**, 147–157.

Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc.* B, **60**, 159–174.

Wild, C. J. and Yee, T. W. (1996) Additive extensions to generalized estimating equation methods. *J. R. Statist. Soc.* B, **58**, 711–725.

Wu, H. and Zhang, J. Y. (2002) Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Ass.*, **97**, 883–897.

Zeger, S. L. and Diggle, P. J. (1994) Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998) Semiparametric stochastic mixed models for longitudinal data. *J. Am. Statist. Ass.*, **93**, 710–719.

J. R. Statist. Soc. B (2006)
68, *Part* 2, *pp.* 179–199

Wavelet-based functional mixed models

Jeffrey S. Morris

University of Texas MD Anderson Cancer Center, Houston, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA

[Received April 2004. Revised October 2005]

Summary. Increasingly, scientific studies yield functional data, in which the ideal units of observation are curves and the observed data consist of sets of curves that are sampled on a fine grid. We present new methodology that generalizes the linear mixed model to the functional mixed model framework, with model fitting done by using a Bayesian wavelet-based approach. This method is flexible, allowing functions of arbitrary form and the full range of fixed effects structures and between-curve covariance structures that are available in the mixed model framework. It yields nonparametric estimates of the fixed and random-effects functions as well as the various between-curve and within-curve covariance matrices. The functional fixed effects are adaptively regularized as a result of the non-linear shrinkage prior that is imposed on the fixed effects' wavelet coefficients, and the random-effect functions experience a form of adaptive regularization because of the separately estimated variance components for each wavelet coefficient. Because we have posterior samples for all model quantities, we can perform pointwise or joint Bayesian inference or prediction on the quantities of the model. The adaptiveness of the method makes it especially appropriate for modelling irregular functional data that are characterized by numerous local features like peaks.

Keywords: Bayesian methods; Functional data analysis; Mixed models; Model averaging; Nonparametric regression; Proteomics; Wavelets

1. Introduction

Technological innovations in science and medicine have resulted in a growing number of scientific studies that yield functional data. Here, we consider data to be functional if

(a) the ideal units of observation are curves and
(b) the observed data consist of sets of curves sampled on a fine grid.

Ramsay and Silverman (1997) coined 'functional data analysis' as an inclusive term for the analysis of data for which the ideal units are curves. They stated that the common thread uniting these methods is that they must deal with both replication, or combining information across N curves, and regularity, or exploiting the smoothness to borrow strength between the measurements within a curve. The key challenge in functional data analysis is to find effective ways to deal with both of these issues simultaneously.

Much of the existing functional data analysis literature deals with exploratory analyses, and more work developing methodology to perform inference is needed. The complexity and high dimensionality of these data make them challenging to model, since it is difficult to construct

Address for correspondence: Jeffrey S. Morris, University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Unit 447, Houston, TX 77030-4009, USA.
E-mail: jeffmo@mdanderson.org

1369–7412/06/68179

models that are reasonably flexible, yet feasible to fit. When the observed functions are well represented by simple parametric forms, parametric mixed models (Laird and Ware, 1982) can be used to model the functions (see Verbeke and Molenberghs (2000)). When simple parametric forms are insufficient, however, nonparametric approaches allowing arbitrary functional forms must be considered. There are numerous papers in the recent literature applying kernels or fixed knot splines to this problem of modelling replicated functional data (e.g. Rice and Silverman (1991), Shi *et al.* (1996), Zhang *et al.* (1998), Wang (1998), Staniswallis and Lee (1998) Brumback and Rice (1998), Rice and Wu (2001), Wu and Zhang (2002), Guo (2002), Liang *et al.* (2003) and Wu and Liang (2004)). Some of these models are very flexible, with many allowing different fixed effect functions of arbitrary form and some also allowing random-effect functions to be of arbitrary form. Among the most flexible of these is that of Guo (2002), who introduced a functional mixed model allowing functional fixed and random-effect functions of arbitrary form, with the modelling done by using smoothing splines. All of these approaches are based on smoothing methods using global bandwidths and penalties, so they are not well suited for modelling irregular functional data that are characterized by spatial heterogeneity and local features like peaks.

This type of functional data is frequently encountered in scientific research, e.g. in biomarker assessments on a spatial axis on colonic crypts (Grambsch *et al.*, 1995; Morris *et al.*, 2003a), in measurements of activity levels by using accelerometers (Gortmaker *et al.*, 1999) and mass spectrometry proteomics (Morris *et al.*, 2005). Our main focus in this paper is modelling functions of this type. In existing literature, data like these are successfully modelled in the single-function setting by using kernels with local bandwidths or splines with free knots or adaptive penalties. However, it is not straightforward to generalize these approaches to the multiple-function setting, since the positions of the local features may differ across curves. It is possible for the mean functions to be spiky but the curve-to-curve deviations smooth, the mean functions to be smooth but the curve-to-curve deviations spiky, or for both the mean functions and the curve-to-curve deviations to be spiky. This requires flexible and adaptive modelling of both the mean and the covariance structure of the data.

Wavelet regression is an alternative method that can effectively model spatially heterogeneous data in the single-function setting (e.g. Donoho and Johnstone (1995)). Morris *et al.* (2003a) extended these ideas to a specific multiple-function setting—hierarchical functional data—which consists of functions observed in a strictly nested design. The fully Bayesian modelling approach yielded adaptively regularized estimates of the mean functions in the model, estimates of random-effect functions and posterior samples which could be used for Bayesian inference. However, the method that was presented in Morris *et al.* (2003a) has limitations that prevent its more general use. It can model only nested designs and hence cannot be used to model functional effects for continuous covariates, functional main and interaction effects for crossed factors, and cannot jointly model the effects of multiple covariates. Also, it cannot handle other between-curve correlation structures, such as serial correlation that might occur in functions that are sampled sequentially over time. Further, Morris *et al.* (2003a) made restrictive assumptions on the curve-to-curve variation that do not accommodate non-stationarities that are commonly encountered in these types of functional data, such as different variances and different degrees of smoothness at different locations in the curve-to-curve deviations (see Fig. 1 in Section 4.2). Finally, Morris *et al.* (2003a) did not provide general use code that could be used to analyse other data sets.

In this paper, we develop a unified Bayesian wavelet-based approach for the much more general functional mixed models framework. This framework accommodates any number of fixed and random-effect functions of arbitrary form, so it can be used for the broad range of mean and between-curve correlation structures that are available in the mixed model setting. The random-effect distributions are allowed to vary over strata, allowing different groups of curves to differ with respect to both their mean functions and covariance surfaces. We also make much less restrictive assumptions on the form of the curve-to-curve variability that accommodate important types of non-stationarity and result in more adaptively regularized representations of the random-effect functions. As in Morris *et al.* (2003a), we obtain posterior samples of all model quantities, which can be used to perform any desired Bayesian inference. We also present a completely data-based method for selecting the regularization parameters of the method, which allows the procedure to be applied without any subjective prior elicitation, if desired, and these regularization parameters are allowed to differ across fixed effect functions. The additional flexibilities that we have built into the method that is presented in this paper has led to increased computational challenges, but we have tackled these and developed general use code for implementing the method that is sufficiently efficient to handle extremely large data sets. We make this code freely available on the Web (http://biostatistics.mdanderson.org/Morris/papers.html), so researchers need not write their own code to implement our method.

The remainder of the paper is organized as follows. In Section 2, we introduce wavelets and wavelet regression. In Section 3, we describe our functional mixed model framework. In Section 4, we describe the wavelet-based functional mixed models methodology, presenting the wavelet space model, describing the covariance assumptions that we make and specifying prior distributions. In Section 5, we describe the Markov chain Monte Carlo (MCMC) procedure that we use to obtain posterior samples of our model quantities and explain how we use these for inference. In Section 6, we apply the method to an example functional data set and, in Section 7, we present a discussion of the method. Technical details and derivations are in Appendix A.

2. Wavelets and wavelet regression

Wavelets are families of orthonormal basis functions that can be used to represent other functions parsimoniously. For example, in $L^2(\Re)$, an orthogonal wavelet basis is obtained by dilating and translating a *mother wavelet* ψ as

$$\psi_{jk}(t) = 2^{j/2}\,\psi(2^j t - k)$$

with j and k integers. A function g can then be represented by the wavelet series

$$g(t) = \sum_{j,k\in\Im} d_{jk}\,\psi_{jk}(t),$$

with wavelet coefficients

$$d_{jk} = \int g(t)\,\psi_{jk}(t)\,\mathrm{d}t$$

describing features of the function g at the spatial locations indexed by k and frequencies indexed by j. In this way, the wavelet decomposition provides a location and scale decomposition of the function.

Let $\mathbf{y} = (y_1, \ldots, y_T)$ be a row vector containing values of a function that is taken at T equally spaced points. A fast algorithm, the *discrete wavelet transform* (DWT), exists for decomposing $\mathbf{y}$ into a set of T wavelet and scaling coefficients (Mallat, 1989). This transform requires only $O(T)$ operations when T is a power of 2. The DWT can also be represented as matrix multiplication by an orthogonal matrix $W' = (W_1', W_2', \ldots, W_J', V_J')$ where J is the coarsest level of the transform. A DWT applied to the vector $\mathbf{y}$ of observations $\mathbf{d} = \mathbf{y}W'$ decomposes the data into sets of wavelet and scaling coefficients $\mathbf{d} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_J, \mathbf{c}_J)$, where $\mathbf{d}_j = \mathbf{y}W_j'$ are the wavelet coefficients at level or scale j and $\mathbf{c}_J = \mathbf{y}V_J'$ are the scaling coefficients. For simplicity, we refer to the entire set of wavelet and scaling coefficients $\mathbf{d}$ as simply the wavelet coefficients. Each wavelet level j contains K_j coefficients. A similar algorithm for the inverse reconstruction, the inverse discrete wavelet transform (IDWT), also exists.

Wavelet regression is a nonparametric regression technique that is useful for modelling functional data that are spiky or otherwise characterized by local features. Suppose that we observe a response vector $\mathbf{y}$, represented by a row vector of length T on an equally spaced grid $\mathbf{t}$ and assumed to be some unspecified function of t plus white noise, i.e. $\mathbf{y} = g(\mathbf{t}) + \varepsilon$, with $\varepsilon \sim \text{MVN}(0, \sigma_e^2 I_T)$. Wavelet regression follows three steps. First, the data are projected into the wavelet space by using the DWT. The corresponding wavelet space model is $\mathbf{d} = \boldsymbol{\theta} + \varepsilon^*$, where $\mathbf{d} = \mathbf{y}W'$ are the empirical wavelet coefficients, $\boldsymbol{\theta} = g(\mathbf{t})W'$ are the true function's wavelet coefficients and $\varepsilon^* = \varepsilon W' \sim \text{MVN}(0, \sigma_e^2 I_T)$ is the noise in the wavelet space.

Since the wavelet transform tends to distribute white noise equally among all wavelet coefficients but concentrates the signal on a small subset, most wavelet coefficients will tend to be small and to consist almost entirely of noise, with the remaining few wavelet coefficients being large in magnitude and containing primarily signal. Thus, we can denoise the signal and regularize the observed function by taking the smallest wavelet coefficients and thresholding them or shrinking them strongly towards zero. This is done either by using thresholding rules (e.g. Donoho and Johnstone (1995)) or by placing a mean 0 shrinkage prior on the true wavelet coefficients (e.g. Abramovich *et al.* (1998)). An effective prior in this context should give rise to a non-linear shrinkage profile, so that smaller coefficients are strongly shrunken whereas larger ones are left largely unaffected. This thresholding or shrinkage of the wavelet coefficients constitutes the second step of wavelet regression. Third, the thresholded or shrunken estimators of the true wavelet coefficients $\boldsymbol{\theta}$ are transformed back to the data space by using the IDWT, yielding a nonparametric estimator of the function. This procedure accomplishes *adaptive regularization*, meaning that the functional estimates are denoised or regularized in a way that tends to retain dominant local features in the function. With the exception of Morris *et al.* (2003a), previous literature on wavelet regression for functional responses has focused on the single-function setting.

3. Functional mixed model

Here we introduce the functional mixed model framework on which we base our methodology. This framework represents an extension of Laird and Ware (1982) to functional data, where the forms of the fixed and random-effect functions are left completely unspecified. Other researchers (e.g. Shi *et al.* (1996), Brumback and Rice (1998), Rice and Wu (2001), Wu and Zhang (2002), Guo (2002) and Wu and Liang (2004)) have worked with similar models, although none have made the same modelling assumptions that we describe here.

Suppose that we observe a sample of N curves $Y_i(t)$, $i = 1, \ldots, N$, on a compact set $\mathcal{T}$, which is assumed without a loss of generality to be [0,1]. Our functional mixed model is given by

$$\mathbf{Y}(t) = X\,\mathbf{B}(t) + Z\,\mathbf{U}(t) + \mathbf{E}(t), \qquad (1)$$

where $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_N(t))'$ is a vector of observed functions, 'stacked' as rows. Here, $\mathbf{B}(t) = (B_1(t), \ldots, B_p(t))'$ is a vector of fixed effect functions with corresponding $N \times p$ design matrix X, $U(t) = (U_1(t), \ldots, U_m(t))'$ is a vector of random-effect functions with corresponding $N \times m$ design matrix Z and $\mathbf{E}(t) = (E_1(t), \ldots, E_N(t))'$ is a vector of functions representing the residual error processes.

Definition 1. A set of N stacked functions, $\mathbf{A}(t)$, all defined on the same compact set $\mathcal{T}$, is a realization from a *multivariate Gaussian process* with $N \times N$ between-row covariance matrix Λ and within-function covariance surface $\Sigma \in \mathcal{T} \times \mathcal{T}$, denoted $\mathbf{A}(t) \sim \mathcal{MGP}(\Lambda, \Sigma)$, if the rows of $\Lambda^{-1/2}\mathbf{A}(t)$ are independent mean 0 Gaussian processes with covariance surface $\Sigma(t_1, t_2)$, where $\Lambda^{-1/2}$ is the inverse matrix square root of Λ. This assumption implies that the covariance between $A_i(t_1)$ and $A_{i'}(t_2)$ is given by $\Lambda_{ii'}\,\Sigma(t_1, t_2)$. This distribution is the functional generalization of the matrix normal distribution (see Dawid (1981)). Note that a scalar identifiability condition must be set on either Λ or Σ, since letting $\Lambda = \Lambda/c$ and $\Sigma = \Sigma * c$ for some constant $c > 0$ yields the same likelihood. For example, we can set $\Lambda_{11} = 1$.

The set of random-effect functions $\mathbf{U}(t)$ is assumed to be a realization from a multivariate Gaussian process with $m \times m$ between-function covariance matrix P and within-function covariance surface $Q(t_1, t_2)$, denoted by $\mathbf{U}(t) \sim \mathcal{MGP}(P, Q)$. The residual errors are assumed to follow $\mathbf{E}(t) \sim \mathcal{MGP}(R, S)$, which is independent of $\mathbf{U}(t)$.

This model is very general and includes many other models that are commonly used for functional data as special cases. For example, it reduces to a simple linear mixed model when the functional effects are represented by parametric linear functions. When $N = 1$, the model simplifies to a form in which traditional smoothing spline and wavelet regression models for single functions can be represented. If we omit the random effects and assume a factorial structure on the fixed effects, we obtain functional analysis-of-variance models. Model (1) also includes the hierarchical functional model that was presented by Morris *et al.* (2003a) as a special case.

This proposed model is very flexible. The fixed effects can be mean functions, functional main effects, functional interactions, functional linear coefficients for continuous covariates, interactions of functional coefficients with other effects or any combination of these. The design matrix Z and between-curve correlation matrices P and R can be chosen to accommodate a myriad of different covariance structures between curves that may be suggested by the experimental design. These include simple random-effects, in which case $P = I$, as well as structures for functional data from nested designs, split-plot designs, subsampling designs and designs involving repeated functions over time. The random-effect portion of the model may be partitioned into

$$Z\,\mathbf{U}(t) = \sum_{h=1}^{H} Z_h\,\mathbf{U}_h(t)$$

with $\mathbf{U}_h(t) \sim \mathcal{MGP}(P_h, Q_h)$, e.g. to allow multiple hierarchical levels of random effects or to allow different random-effects distributions for different strata.

This model is similar to the functional mixed model in Guo (2002), with a couple of key differences. Guo (2002) assumed independent random-effect functions ($P = R = I$ in our framework), whereas our model, by introducing P and R, can accommodate correlation across the functions. Also, Guo (2002) assumed a structure on Q that is different from what we do here. For each level of random effects h, Guo assumed that $Q_h = L_h + \Sigma/\lambda_h$, where $L_h = \sigma_h^2 M'DM$ is the covariance that is induced by random intercept and linear terms whose design matrix is M, D is a structured 2×2 covariance matrix (which was assumed diagonal in Guo's example) and σ_h^2 is a variance component that is estimated from the data. The parameter λ_h is a scalar smoothing parameter that is estimated from the data, and the correlation matrix Σ is fixed on the basis of the reproducing kernel for the chosen spline basis. Our assumptions on Q are described later in Section 4.2.

Of course, we cannot directly fit model (1), since in practice we observe only samples of the continuous curves on some discrete grid. A discrete version of this model is given below, assuming that all observed functions are sampled on a common equally spaced grid $\mathbf{t} = (t_1 \ldots t_T)'$.

Recall that, by our definition of functional data (sampled on a very fine grid), this assumption is not especially restrictive, since, if the grid is sufficiently fine, interpolation can be used to obtain a common grid without substantively changing the observed data. The model is

$$Y = XB + ZU + E, \tag{2}$$

where Y is an $N \times T$ matrix of observed curves on the grid $\mathbf{t}$, B is a $p \times T$ matrix of fixed effects, U is an $m \times T$ matrix of random effects and E is an $N \times T$ matrix of residual errors. As defined above, X is an $N \times p$ matrix and Z is an $N \times m$ matrix, and the two are the design matrices for the fixed and random-effect functions respectively. Following the notation of Dawid (1981), U follows a matrix normal distribution with $m \times m$ between-row covariance matrix P and $T \times T$ between-column covariance matrix Q, which we denote by $U \sim \mathcal{MN}(P, Q)$. Another way to represent this structure is to say that $\text{vec}(U') \sim \text{MVN}(0, P \otimes Q)$, where $\text{vec}(A)$ is the vectorized version of a matrix A obtained by stacking the columns and '$\otimes$' is the Kronecker product, both defined as in Harville (1997). This assumption implies that the covariance between U_{ij} and $U_{i'j'}$ is $P_{ii'}Q_{jj'}$. The residual error matrix E is assumed to be $\mathcal{MN}(R, S)$. The within-random-effect curve covariance surface Q and residual error covariance surface S are $T \times T$ covariance matrices that are discrete approximations of the corresponding covariance surfaces in $\mathcal{T} \times \mathcal{T}$.

4. Wavelet-based functional mixed model

Having presented a conceptual functional mixed model for correlated functional data, we now describe our nonparametric wavelet-based approach to fit it. Our approach consists of three basic steps.

(a) Compute the empirical wavelet coefficients for each observed curve, which we think of as projecting the observed curves from the data space to the wavelet space.
(b) Use Markov chain Monte Carlo methods to obtain posterior samples for quantities in the wavelet space version of the functional mixed model. Projecting to the wavelet space allows modelling to be done in a more parsimonious and computationally efficient manner and causes regularization to be performed as a natural consequence of the modelling through shrinkage priors placed on the fixed effects portion of the model.
(c) Transform the wavelet space quantities back to the data space, yielding posterior samples of all quantities in the data space model, which can be used to perform Bayesian estimation, inference and prediction.

The first step involves decomposing each observed function, sampled on an equally spaced grid of size T, into a set of T wavelet coefficients. This projection from the data space into the wavelet space is done by applying the DWT to each row of Y and can be conceptualized as the right matrix multiplication $D = YW'$, where W is the orthogonal DWT matrix. The $N \times T$ matrix D contains the empirical wavelet coefficients for all observed curves, with row i containing wavelet and scaling coefficients for curve i and the columns double indexed by the scale j and location k, with $j = 1, \dots, J$ and $k = 1, \dots, K_j$.

4.1. Wavelet space model

Right matrix multiplication of both sides of model (2) by the DWT matrix W' yields a wavelet space version of the model:

$$D = XB^* + ZU^* + E^*, \tag{3}$$

where X and Z are the design matrices as in model (2), $B^* = BW'$ is a $p \times T$ matrix whose rows contain the wavelet coefficients for the p fixed effect functions on the grid, $U^* = UW'$ is an $m \times T$ matrix whose rows contain the wavelet coefficients for the m random-effect functions and $E^* = EW'$ is an $N \times T$ matrix consisting of the residual errors in the wavelet space. Like D, the columns of B^*, U^* and E^* are all double indexed by the wavelet coefficients' scale j and location k. The linearity of the DWT makes it easy to compute the induced distributional assumptions of the random matrices in the wavelet space, $U^* \sim \mathcal{MN}(P, Q^*)$ and $E^* \sim \mathcal{MN}(R, S^*)$, where $Q^* = WQW'$ and $S^* = WSW'$. Note that the between-row covariance structure is retained when projecting into the wavelet space; only the column covariance changes.

4.2. Covariance assumptions

Before we fit model (3), it is necessary to specify some structure on the various covariance matrices since their large dimensions make it infeasible to estimate them in a completely unstructured fashion. We model P and R by using parametrically structured covariance matrices as in linear mixed models, which can be chosen on the basis of either the experimental design or an empirical investigation of the data. The vectors of the covariance parameters indexing matrices P and R are denoted by $\mathbf{\Omega}_P$ and $\mathbf{\Omega}_R$ respectively.

For Q and S, we propose a parsimonious structure in the wavelet space that yields a flexible class of covariance surfaces in the data space. As is frequently done in wavelet regression, we assume that the wavelet coefficients within a given curve are independent across j and k, making Q^* and S^* diagonal. The heuristic justification that is frequently given for this assumption is the whitening property of the wavelet transform, which is discussed in Johnstone and Silverman (1997). The diagonal elements are allowed to vary across both wavelet scales j and locations k, yielding $Q^* = \text{diag}(q^*_{jk})$ and $S^* = \text{diag}(s^*_{jk})$. For convenience, we denote these sets of variance components by $\mathbf{\Omega}_Q$ and $\mathbf{\Omega}_S$ respectively.

This structure requires only T parameters instead of the $T(T+1)/2$ parameters that would be required to estimate each of these matrices in an unstructured fashion, yet it is sufficiently flexible to emulate a wide range of covariance structures that are commonly encountered in functional data. For example, when $T = 256$, only 256 parameters are required instead of the 32896 for the unstructured representation. Independence in the wavelet space does not imply independence in the data space unless the variance components are identical across all wavelet scales j and locations k, since heterogeneity in variances across wavelet coefficients at different levels induces serial dependences in the data. In general, larger variances at low frequency scales correspond to stronger serial correlations, and thus smoother functions.

Further, since the variance components are free to vary across both scale j and location k, this structure accommodates non-stationarity, e.g. allowing the curve-to-curve variances and the smoothness in the curve-to-curve deviations both to vary over t. These types of non-stationarities are frequently encountered in complex functional data but cannot be accommodated when the variance components are allowed to vary only over j (see Fig. 1). It is typical in existing wavelet regression literature for the wavelet space variance components to vary over j, but not k (e.g. Abramovich *et al.* (1998), Morris *et al.* (2003a), Abramovich and Angelini (2003) and Antoniadis and Sapatinas (2004)). This may be a necessary practical restriction in the single-function case, but not in the multiple-function case, since the replicate functions allow the variance components to be estimable even when they also vary by k. To our knowledge, this is the first paper allowing these variance components to depend on both j and k.

To illustrate the flexibility of these assumptions, we randomly generated 200 realizations from a Gaussian process with mean $\mu(t)$ and covariance $S(t_1, t_2)$ on an equally spaced grid of length 256 on $(0, 1)$. From top to bottom, Fig. 1(a) contains the true mean function $\mu(t)$, the true

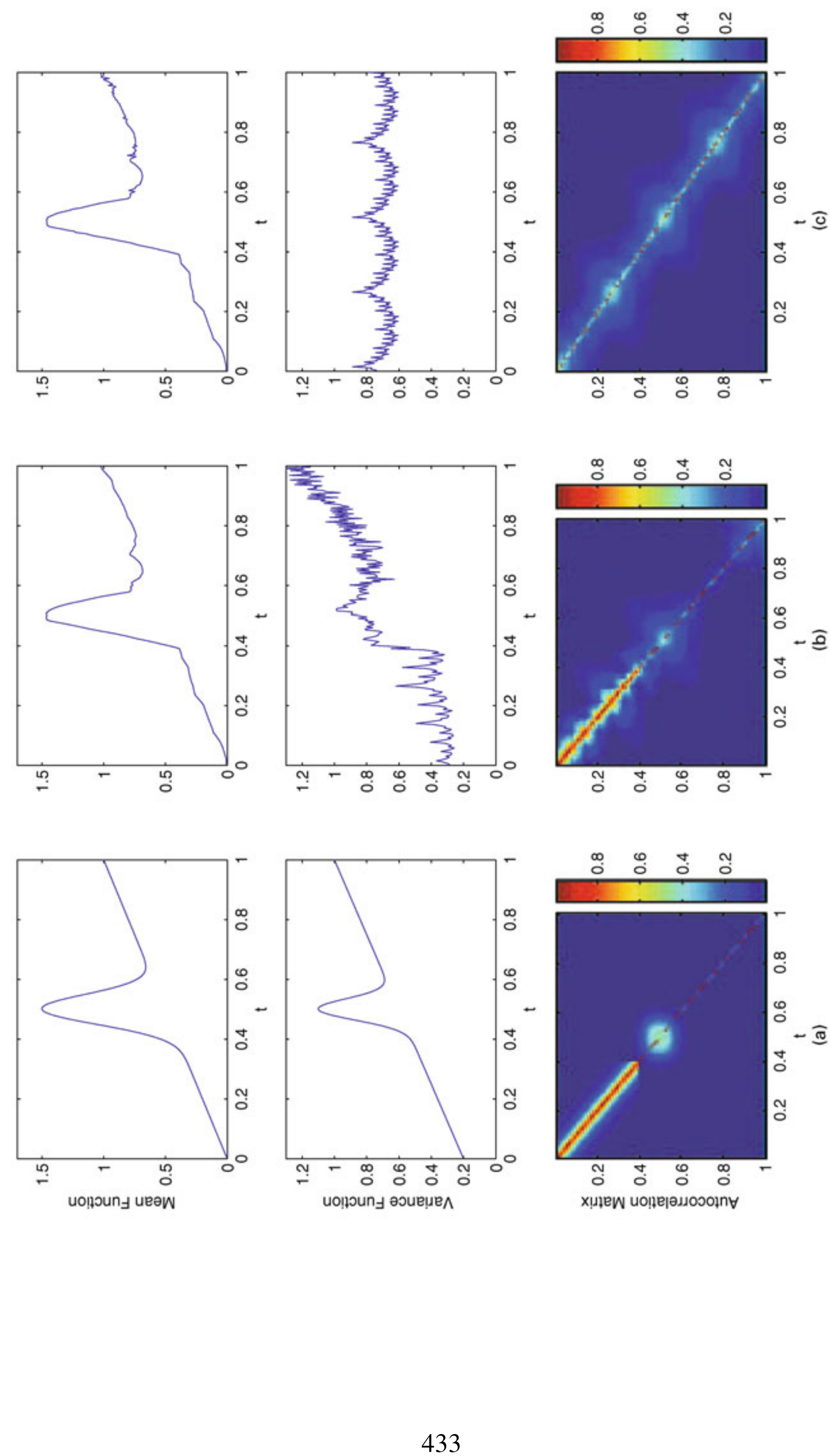


Fig. 1. Simulated data (see the discussion in Section 4.2 for details): (a) truth; (b) estimated, wavelet space variance components indexed by scale j and location k; (c) estimated, wavelet space variance components indexed by scale j only

variance function $v(t) = \text{diag}(S)$ and the true autocorrelation surface $\rho_S(t_1, t_2) = v^{-1/2} S v^{-1/2}$. Figs 1(b) and 1(c) contain the posterior mean estimates of these quantities by using wavelet-based methods. Both assume independence across wavelet coefficients, but Fig. 1(b) allows the wavelet space variance components to vary across scale j and location k, and Fig. 1(c) allows them to vary across j only, as assumed in Morris *et al.* (2003a) and other work involving wavelet regression. The framework that is used in Fig. 1(b) is sufficiently flexible to pick up on the non-stationary features of S, whereas Fig. 1(c) is not. Specifically, it can model the increasing variance in t, the extra variance near the peak at 0.5, the different degrees of smoothness in the region (0,0.4) and (0.6,1) and the extra autocorrelation from the peak at 0.5. Also note that it appears to have done a marginally better job of denoising the estimate of the mean function. These same principles apply to the covariance across random-effect functions.

Another advantage of this independence assumption is that it allows us to fit the wavelet space model (3) one column (wavelet coefficient) at a time. This greatly simplifies the computational procedure and allows much larger data sets to be fitted by using this method.

4.3. Adaptive regularization using a multiple-shrinkage prior

To obtain adaptively regularized representations of the fixed effect functions $B_i(t)$, as is standard in Bayesian implementations of wavelet regression, we place a mixture prior on B^*_{ijk}, the wavelet coefficient at scale j and location k for fixed effect i:

$$B^*_{ijk} = \gamma^*_{ijk} \mathcal{N}(0, \tau_{ijk}) + (1 - \gamma^*_{ijk}) I_0, \qquad (4)$$

$$\gamma^*_{ijk} = \text{Bernoulli}(\pi_{ij}),$$

where I_0 is a point mass at zero and γ^*_{ijk} is an indicator of whether wavelet coefficient (j, k) is 'important' for representing the signal for fixed effect function i. The hyperparameter π_{ij} is the prior probability that a wavelet coefficient at wavelet scale j is important for representing the fixed effect function i, and τ_{ijk} is the prior variance of any important wavelet coefficient at location k and level j for fixed effect i.

The quantities π_{ij} and τ_{ijk} are regularization parameters. For example, smaller π_{ij} will result in more attenuation in the features of fixed effect function i occurring at a frequency indicated by scale j. By indexing these parameters by i and j, we allow different degrees of regularization for different fixed effect functions and at different frequencies. See Morris *et al.* (2003a) for a discussion of the intuition behind how this prior leads to adaptive regularization. It is possible to elicit values for these regularization parameters, taking into account some of the considerations that were discussed in Morris *et al.* (2003a) or Abramovich *et al.* (1998), or to estimate them from the data by using an empirical Bayes procedure. Section 4.4 describes one such procedure.

In this modelling framework, the random-effect functions $U_i(t)$ are also regularized as a result of the mean 0 Gaussian distribution on their wavelet coefficients. Morris *et al.* (2003b) described how the regularization of the random-effect functions in their wavelet-based hierarchical functional model was governed by the relative sizes of corresponding variance components and residual errors. The same principles also apply here, although here our regularization is more adaptive than in Morris *et al.* (2003a) since we allow the wavelet space variance components for both the random effects and the residual errors to depend on scale j and location k. To explain, wavelet coefficients that are indexed by (j, k) that tend to be important for representing even a small number of random-effect functions will have relatively large subject level variance components q_{jk}. These large variances will lead to less shrinkage of these coefficients, and thus the features that are represented by these coefficients will tend to be preserved in the regularized random-effect function estimates. Wavelet coefficients that are unimportant for representing the random-effect functions will be close to 0, leading to small variance components, strong shrinkage and regularization of the features corresponding to these coefficients.

This regularization is sufficiently adaptive to model very spiky random-effect functions, as demonstrated in supplementary material that is available at `http://biostatistics.mdanderson.org/Morris/papers.html`. A major advantage of our approach is that the random-effect functions' regularization parameters are simply the variance components of the model, which are directly estimated from the data, and thus need not be arbitrarily chosen. Further, in our Bayesian approach, the uncertainty of their estimation is automatically propagated throughout any inference that is done.

It may be possible to obtain even more adaptively randomized random-effect functions by assuming a mixture prior like equation (4) on the wavelet coefficients for the random-effect functions. However, by doing so, we would lose some of the coherency that is evident in models (1)–(3), since the random-effect functions would no longer be Gaussian in the data space. Further, we would not be able to marginalize over the random-effect functions in our model fitting (see Section 5), which would increase the computational burden for implementing the method. Since we are satisfied with the degree of adaptiveness that is afforded by our Gaussian assumptions with variances depending on j and k, we do not further pursue this idea in this paper.

4.4. *Empirical Bayes method for selecting shrinkage hyperparameters*

Here we present a data-based procedure for determining the shrinkage hyperparameters for the fixed effect functions in the wavelet-based functional mixed model. We estimate these hyperparameters by using maximum likelihood while conditioning on consistent estimates of the variance components in the model. This method is an extension of the work of Clyde and George (2000), which they later adapted to the hierarchical functional framework (Clyde and George, 2003).

First we introduce some notation. Consider the quantities

$$\hat{B}^*_{ijk,\mathrm{MLE}} = \{X_i'(\Sigma_{jk})^{-1}X_i\}^{-1}X_i'(\Sigma_{jk})^{-1}(\mathbf{d}_{jk} - \mathbf{X}_{(-i)}\hat{B}^*_{(-i)jk,\mathrm{MLE}}), \tag{5}$$

$$\begin{aligned} V_{ijk} &= \mathrm{var}(\hat{B}^*_{ijk,\mathrm{MLE}}) \\ &= \{X_i'(\Sigma_{jk})^{-1}X_i\}^{-1}, \end{aligned} \tag{6}$$

where X_i is the ith column of the design matrix and $X_{(-i)}$ is the design matrix with column i omitted, and

$$\Sigma_{jk} = ZP(\Omega_P)Z' * q^*_{jk} + R(\Omega_R) * s^*_{jk} \tag{7}$$

is the marginal variance of $\mathbf{d}_{jk}$. Note that $\hat{B}^*_{ijk,\mathrm{MLE}}$ is the maximum likelihood estimator (MLE) of B^*_{ijk} conditional on the covariance parameters and the other fixed effects and $\surd V_{ijk}$ is the standard error of the MLE. Taking their ratio yields

$$\zeta_{ijk} = \hat{B}^*_{ijk,\mathrm{MLE}} / \surd V_{ijk}, \tag{8}$$

which can be thought of as a standardized score for the wavelet coefficient at scale j and location k from fixed effect function i.

We assume that $\tau_{ijk} = V_{ijk}\Upsilon_{ij}$ for some parameters Υ_{ij}, allowing full flexibility in these regularization parameters across different scales, but making the ratio of regularization parameters within a given scale proportional to the size of the variance of the MLE for that coefficient. This allows us to estimate Υ_{ij} from the data. Assuming knowledge of V_{ijk}, it can be shown that the likelihood for Υ_{ij} and π_{ij} can be represented by

$$l(\Upsilon_{ij},\pi_{ij}) \propto (1+\Upsilon_{ij})^{-\Sigma_{k=1}^{K_j}\gamma^*_{ijk}/2} \exp\left\{-\frac{1}{2}\sum_{k=1}^{K_j}\zeta^2_{ijk}\gamma^*_{ijk}/(1+\Upsilon_{ij})\right\}$$
$$\times (\pi_{ij})^{\Sigma_{k=1}^{K_j}\gamma^*_{ijk}}(1-\pi_{ij})^{K_j-\Sigma_{k=1}^{K_j}\gamma^*_{ijk}}. \tag{9}$$

On the basis of this likelihood, local maximum likelihood estimates of π_{ij} and Υ_{ij} can be obtained by iterating through the following steps until convergence is achieved:

$$\hat{\gamma}^*_{ijk} = \frac{\hat{O}_{ijk}}{1+\hat{O}_{ijk}};$$
$$\hat{O}_{ijk} = \frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}}(1+\hat{\Upsilon}_{ij})^{-1/2}\exp\left(\frac{1}{2}\zeta^2_{ijk}\frac{\hat{\Upsilon}_{ij}}{1+\hat{\Upsilon}_{ij}}\right);$$
$$\hat{\Upsilon}_{ij} = \max\left(0, \sum_{k=1}^{K_j}\hat{\gamma}^*_{ijk}\zeta^2_{ijk}\Big/\sum_{k=1}^{K_j}\hat{\gamma}^*_{ijk} - 1\right);$$
$$\hat{\pi}_{ij} = \sum_{k=1}^{K_j}\frac{\hat{\gamma}^*_{ijk}}{K_j}.$$

This procedure can be applied while conditioning on consistent estimators of the variance components, e.g. method-of-moment estimators or MLEs, giving $\hat{V}_{ijk}$ of V_{ijk}. Then the empirical Bayes estimates of π_{ij} and τ_{ijk} are given by $\hat{\pi}_{ij}$ and $\hat{V}_{ijk} * \hat{\Upsilon}_{ij}$ respectively.

5. Posterior sampling by using Markov chain Monte Carlo methods

After specifying diffuse proper priors for the variance components, we are left with a fully specified Bayesian model for the functional data. Since the posterior distributions of parameters are not available in closed form, we use MCMC sampling to obtain posterior samples for all the parameters in model (3). We work with the marginalized likelihood where the random effects have been integrated out, which improves the mixing properties of the sampler over a naïve Gibbs sampler. We alternate between sampling the fixed effects B^* and the covariance parameters $\mathbf{\Omega}$; then we later sample the random-effects U^* whenever they are of interest. Following are the details of the sampling procedure that we use.

(a) For each wavelet coefficient (j,k), sample fixed effect i from $f(B^*_{ijk}|\mathbf{d}_{jk}, B^*_{(-i)jk}, \mathbf{\Omega})$, where $B^*_{(-i)jk}$ is the set of all fixed effects coefficients at scale j and location k except the ith. As shown in Appendix A, this distribution is a mixture of a point mass at 0 and a normal distribution, with the normal mixture proportion α_{ijk} and the mean and variances of the normal μ_{ijk} and v_{ijk} respectively given by

$$\begin{aligned}\alpha_{ijk} &= \Pr(\gamma_{ijk}=1|\mathbf{d}_{jk}, B^*_{(-i)jk}, \mathbf{\Omega}) \\ &= O_{ijk}/(O_{ijk}+1),\end{aligned} \tag{10}$$

$$O_{ijk} = \pi_{ij}/(1-\pi_{ij})\mathrm{BF}_{ijk},$$

$$\mathrm{BF}_{ijk} = (1+\tau_{ijk}/V_{ijk})^{-1/2}\exp\{\tfrac{1}{2}\zeta^2_{ijk}(1+V_{ijk}/\tau_{ijk})\}, \tag{11}$$

$$\mu_{ijk} = \hat{B}_{ijk,\mathrm{MLE}}(1+V_{ijk}/\tau_{ijk})^{-1}, \tag{12}$$

$$v_{ijk} = V_{ijk}(1+V_{ijk}/\tau_{ijk})^{-1}, \tag{13}$$

where $\hat{B}^*_{ijk,\mathrm{MLE}}$, V_{ijk}, Σ_{jk} and ζ_{ijk} are defined as in equations (5)–(8) above. O_{ijk} and BF_{ijk} have an interesting interpretation. They are the posterior odds and Bayes factor respectively for deciding whether wavelet coefficient (j,k) is important for representing function i, conditional on the covariance parameters Ω and other fixed effects. The posterior means of the B_{ijk} will be Bayesian model-averaged estimators that have averaged over models where B_{ijk} is either 0 or not. Alternatively, a soft thresholding approach could be used whereby $\hat{B}_{ijk}=0$ if the estimated posterior probability that $|B_{ijk}|>0$ (i.e. $\gamma_{ijk}=1$) from the MCMC algorithm is less than some threshold.

(b) For each wavelet coefficient (j,k), sample the elements q^*_{jk} and s^*_{jk} of $\mathbf{\Omega}_Q$ and $\mathbf{\Omega}_S$ by using a random-walk Metropolis–Hastings step. The objective function is

$$f(q^*_{jk},s^*_{jk}|\mathbf{d}_{jk},B^*_{jk},\mathbf{\Omega}_P,\mathbf{\Omega}_R)\propto|\Sigma_{jk}|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{d}_{jk}-XB^*_{jk})'\Sigma^{-1}_{jk}(\mathbf{d}_{jk}-XB^*_{jk})\}\,f(q^*_{jk},s^*_{jk}).$$

We use an independent Gaussian density, truncated at zero and centred at the previous parameter values, as the proposal for each parameter. We automatically estimate the proposal variance from the data by using estimates of the variance of the maximum likelihood estimates. Wolfinger *et al.* (1994) provided details of how to compute maximum likelihood estimates and their standard errors in linear mixed models. The details of the Metropolis–Hastings procedure are available at `http://biostatistics.mdanderson.org/Morris/papers.html`.

(c) Sample the between-curve covariance parameters $\mathbf{\Omega}_P$ and $\mathbf{\Omega}_R$ by using a single random-walk Metropolis–Hastings step. If the random-effects and residual errors are assumed to be independent and homoscedastic across samples ($P=I$ and $R=I$), then there are no parameters to update in this step. The assumption of independence between the wavelet coefficients allows the Metropolis–Hastings objective function to factor into the product of independent pieces for each wavelet coefficient:

$$f(\mathbf{\Omega}_P,\mathbf{\Omega}_R|D,B^*,\mathbf{\Omega}_Q,\mathbf{\Omega}_S)\propto\prod_{j,k}|\Sigma_{jk}|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{d}_{jk}-XB^*_{jk})'\Sigma^{-1}_{jk}(\mathbf{d}_{jk}-XB^*_{jk})\}\,f(\mathbf{\Omega}_P,\mathbf{\Omega}_R),$$

where Σ_{jk} is given by equation (7) above. The details of implementation are similar to those for the previous step. Again, we use an independent truncated Gaussian distribution with mean at the previous parameter values for the proposal distribution, with the proposal variance automatically determined from the data.

(d) Sample the random effects $\mathbf{u}^*_{jk}$ for each (j,k) from their full conditional $f(\mathbf{u}^*_{jk}|\mathbf{d}_{jk},B^*_{jk},\mathbf{\Omega})$, which is easily seen to be Gaussian distributed with mean $\{\Psi^{-1}_{jk}+(q^*_{jk}*P)^{-1}\}^{-1}\Psi^{-1}_{jk}\hat{\mathbf{u}}_{NS,jk}$ and variance $\{\Psi^{-1}_{jk}+(q^*_{jk}*P)^{-1}\}^{-1}$, where $\Psi_{jk}=\{Z'(s^*_{jk}*R)^{-1}Z\}^{-1}$ and

$$\hat{\mathbf{u}}_{NS,jk}=\{Z'(s^*_{jk}*R)^{-1}Z\}^{-1}Z'(s^*_{jk}*R)^{-1}(\mathbf{d}_{jk}-XB^*_{jk}).$$

If the random effects are not desired, we can omit this step and thus speed up the MCMC algorithm, since the previous steps work with the marginalized likelihood.

Code for applying this method is available at `http://biostatistics.mdanderson.org/Morris/papers.html`.

5.1. Bayesian inference and prediction

The MCMC algorithm that was described above yields posterior samples for all quantities in the wavelet space mixed model (3). These posterior samples can then be projected back into the data space by using the IDWT, yielding posterior samples of the quantities in model (2). Specifically, posterior samples for each fixed effect function $B_i(t)$ on the grid $\mathbf{t}$ are obtained by applying the IDWT to each posterior sample of the corresponding vector of wavelet coefficients $B_i^* = (B_{i11}^*, \ldots, B_{iJK_J}^*)$, and similarly for the random-effect functions. Further, posterior samples of the covariance matrices Q and S are obtained by applying the two-dimensional IDWT to the posterior samples of the diagonal matrices Q^* and S^*, following Vannucci and Corradi (1999).

Given the posterior samples, we can then construct any Bayesian estimators and perform any desired Bayesian inference. See Gelman *et al.* (2004) for an overview of Bayesian analysis and inference, and a description of the types of inference that are possible given posterior samples. For example, we can construct pointwise credible intervals for fixed effect functions or compute posterior probabilities for any hypotheses of interest. These can involve any transformation or combination of the parameters in the model. Since we have posterior samples for entire functions, marginal inference can be done for single locations on the function or joint inference can be done over regions of the function. It is also straightforward to compute posterior predictive distributions $f(Y^*|Y)$ for a future observed curve Y^* given data Y, since

$$f(Y^*|Y) = \int f(Y^*|B, U, \Omega)\, f(B, U, \Omega|Y)\, \mathrm{d}B\, \mathrm{d}U\, \mathrm{d}\Omega,$$

which can be estimated via Monte Carlo integration using the posterior samples as $G^{-1} \times \Sigma_g f(Y^*|B^{(g)}, U^{(g)}, \Omega^{(g)})$, where the superscript (g) indicates the posterior sample from iteration g of the MCMC algorithm. This inference and prediction appropriately account for all sources of variation in the model. For example, they do not condition on estimates of the variance components as if they were known but automatically propagate the uncertainty of their estimation throughout inference. This is one of the advantages of using a unified Bayesian modelling approach.

6. Example

Nutrition researchers at Texas A&M University conducted a rat carcinogenesis experiment to investigate whether the type of dietary fat (fish-oil or corn oil) plays a role in modulating important colon cancer biomarkers during the initiation stage of carcinogenesis, the first hours after exposure to a carcinogen. In this study, they fed 30 rats one of the two diets for 14 days, exposed them to a carcinogen and then sacrificed them at one of five times after exposure to the carcinogen (0, 3, 6, 9 or 12 h). They removed and dissected each rat's colon and then used immunohistochemical staining to obtain measurements of various cancer biomarkers, including the deoxyribonucleic acid (DNA) adduct level, a measurement of the amount of DNA damage occurring from the exposure to the carcinogen, O^6-methylguanine-DNA methyltransferase (MGMT), a DNA repair enzyme that repairs this carcinogen-induced damage, and apoptosis, the selective elimination of damaged cells.

They quantified each biomarker for a separate set of roughly 25 crypts in the distal region of each rat's colon. Crypts are finger-like structures extending into the colon wall in which all colon cells reside. A cell's relative depth within its crypt is related to its age and stage in the cell cycle, so it is an important factor to consider when assessing biomarker modulation. Using image analysis software, they quantified the MGMT levels on a fine grid along the side of each selected crypt, resulting in an observed curve for each crypt containing the biomarker quantifications as a function of relative depth within the crypt. The relative depth in the crypt was coded such that an observation at the base of the crypt was relative cell position 0, whereas an observation at the lumenal surface was relative cell position 1. Fig. 2 contains the observed curves from two crypts from two rats. Note that these functions appear very irregular, with

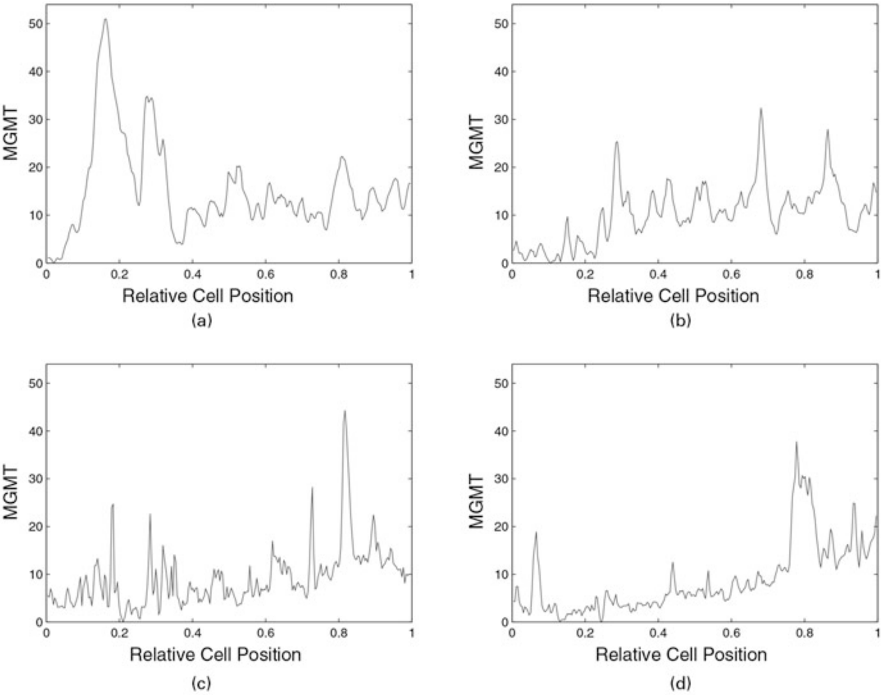


Fig. 2. Sample curves of MGMT intensity levels as a function of relative depth within the crypts: (a) fish-oil diet 12 h, rat 1, crypt 1; (b) fish-oil diet 12 h, rat 1, crypt 2; (c) corn oil diet 12 h, rat 1, crypt 1; (d) corn oil diet 12 h, rat 1, crypt 2

many spikes presumably corresponding to local areas in the crypt with high biomarker levels (Morris *et al*., 2003a), e.g. the nuclei of the cells. The full data set consists of 738 such observed curves, each sampled on an equally spaced 256-unit grid.

The MGMT data were analysed by Morris *et al*. (2003a), and it was found that corn-oil-fed rats had lower MGMT expression near the lumenal surface at 12 h after exposure to the carcinogen than did fish-oil-fed rats. Our goal here is to relate the levels of the other biomarkers to the MGMT expression levels, and to see whether this 12 h-effect remains after adjusting for these other biomarkers as covariates. For each rat, we obtained measurements of the continuous covariates mean DNA adduct level and apoptotic index (the percentage of cells undergoing apoptosis) across its crypts in the upper third compartment, i.e. the compartment that is closest to the lumenal surface. We would like to assess whether there is a relationship between the amount of DNA damage and/or the amount of apoptosis near the lumenal surface of the crypts and the levels of MGMT, and whether these relationships depend on relative cell position and/or diet. These covariates were not considered in Morris *et al*. (2003a) and could not be accommodated by their hierarchical functional model.

Our design matrix X had $p=14$ columns, with the first 10 indicating the rat's diet by time group. Columns 11 and 12 contained the mean DNA adduct level in the upper third of the crypt for rats fed the fish- and corn oil diets respectively. These columns were standardized to have mean 0 and standard deviation 1. Columns 13 and 14 contained the apoptotic index in the upper third of the crypt for rats fed the fish- and corn oil diets respectively. To model the correlation between crypts from the same rat, we included random-effect functions for each rat. The residual errors represented the sum of the crypt-to-crypt variability and any within-function noise. We assumed that rats and crypts within rats were independent and identically distributed, so we let $P=R=I$. We used the Daubechies wavelet with eight vanishing moments (Daubechies, 1992) at $J=8$ levels. Other wavelet bases yielded substantively equivalent results. After a burn-in of 1000, we ran the MCMC algorithm for 20000 iterations, keeping every 10. The Metropolis–Hastings acceptance probabilities for the variance components were all between 0.12 and 0.39. Trace plots of the model parameters are available at `http://biostatistics.mdanderson.org/Morris/papers.html` and reveal that the MCMC algorithm converged and mixed very well.

Fig. 3 contains the posterior mean functional coefficients corresponding to the DNA adduct level and apoptotic index covariates for fish- and corn-oil-fed rats. The estimate for the DNA adduct level top coefficient was negative near the lumenal surface for rats that were fed fish-oil

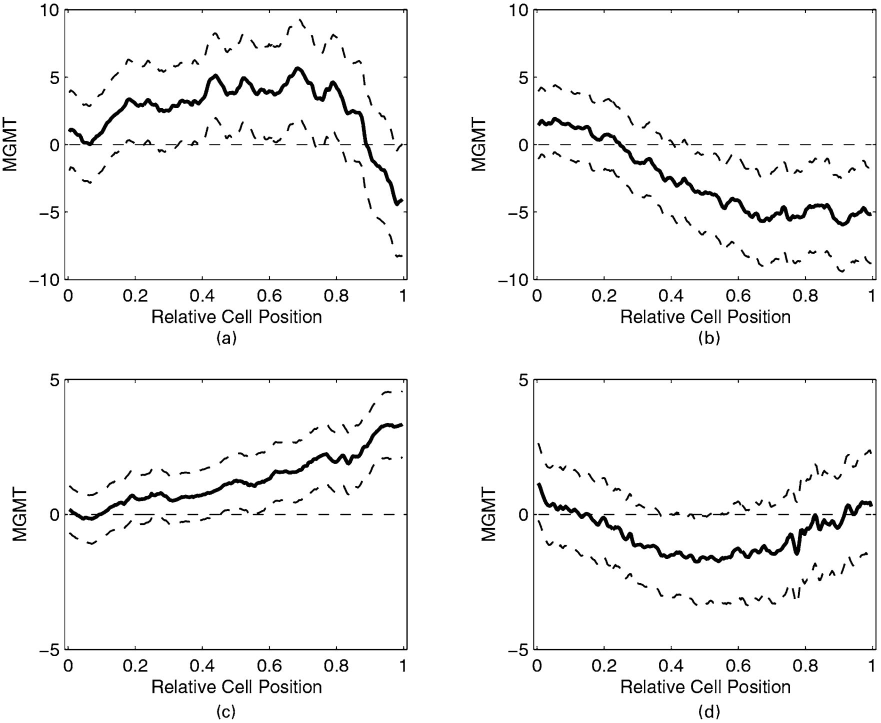


Fig. 3. MGMT results: posterior mean and 95% pointwise posterior credible intervals for functional linear coefficients (for the corresponding continuous covariates in a functional mixed model that also includes categorical effects for the 10 diet–time combinations and random-effect functions for each rat): (a) DNA adduct level, top third of the crypt, fish-oil diet; (b) DNA adduct level, top third of the crypt, corn oil diet; (c) apoptotic index, top third of the crypt, fish-oil diet; (d) apoptotic index, top third of the crypt, corn oil diet

or corn oil, meaning that animals with high levels of DNA damage near the lumenal surface tended also to have lower levels of MGMT near the lumenal surface. The posterior probabilities that the coefficient at the top of the crypt was less than 0 were 0.947 and 0.989 for fish- and corn oil diets respectively. This negative relationship extended to the middle of the crypts for corn-oil-fed rats, but not for fish-oil-fed rats, for whom the estimate was positive. The posterior probability that the fish-oil coefficient at the middle of the crypt (relative cell position 0.5) was greater than that for the corn oil coefficient was 0.9965.

For fish-oil-fed rats, the apoptotic index top coefficient was positive throughout nearly the entire crypt, with the coefficient increasing in a roughly linear fashion moving up the crypt. The posterior probability that this coefficient was greater than 0 at the lumenal surface for fish- and corn-oil-fed rats was greater than 0.9995 and 0.612 respectively, and the posterior probability that the coefficient for fish-oil-fed rats was greater than that for corn-oil-fed rats was 0.9815. The interpretation of these results is that the fish-oil-fed animals who had a large amount of apoptosis near their lumenal surface also had high levels of the DNA repair enzyme MGMT near their lumenal surface, meaning that the two major mechanisms for dealing with DNA damage were correlated. This relationship was not so strong for corn-oil-fed animals.

With DNA adduct level and apoptotic index and their interactions with diet included in the model, the difference between the fish-oil and corn oil diets at 12 h near the lumenal surface that was found in Morris *et al.* (2003a) was no longer evident (the posterior probability that the effect for fish-oil was greater than that for corn oil was only 0.674, whereas it was greater than 0.9995 without covariates in the model). One interpretation of this result is that the differences in MGMT between diets at the lumenal surface may be explained by the previously observed DNA adduct level and apoptosis effects (Hong *et al.*, 2000), whereby rats on fish-oil diets had lower DNA adduct levels and higher apoptotic rates at the lumen surface than rats fed corn oil diets.

7. Discussion

Functional data are increasingly encountered in scientific studies, and there is a need for systematic methods for analysing these complex and large data sets and extracting the meaningful information that is contained inside them. In this paper, we have introduced a unified Bayesian wavelet-based modelling approach for functional data that is a vast extension over the hierarchical functional method that was introduced by Morris *et al.* (2003a). Although applied to just one example here, our approach is sufficiently flexible to be applied to a very broad range of functional data sets and to address a large number of potential research questions. If we substitute higher dimensional wavelet transforms for the one-dimensional transforms that are described here, our methodology is immediately extendable to higher dimensional functional data, e.g. image data.

The underlying functional mixed models framework is very flexible, allowing the same wide range of mean and covariance structures as in mixed effects models, while allowing functional fixed and random effects of unspecified form. We perform our modelling in the wavelet space, which provides a natural mechanism for adaptive regularization using mixture prior distributions, and also allows us to model the high dimensional covariance matrices Q and S describing the form of the curve-to-curve deviations in a parsimonious manner. As in much work in wavelet regression, we assume independence in the wavelet space, but unlike existing work in wavelet regression we allow the wavelet space variance components to vary across both scale j and location k. This provides a large amount of flexibility, accommodating various types of non-stationarity that is commonly encountered in functional data, including heteroscedasticity and varying degrees of smoothness at different locations in the curve-to-curve deviations; see Fig. 1.

This flexibility allows us to model many different types of functional data and also results in more adaptive regularization in the representations of the fixed and random-effect functions. This approach can effectively accommodate spiky fixed effect functions and/or spiky random-effect functions. In our example, the fixed effect and rat level random-effect functions were smooth, but the crypt level deviations were spiky.

After running an MCMC algorithm, we obtain posterior samples of the fixed and random-effect functions and various covariance matrices in the model, which can be used to perform any desired Bayesian estimation, inference or prediction. Credible intervals can be constructed and posterior probabilities of hypotheses can be computed for any transformation or function of the model parameters, e.g. averaging over different intervals or looking at specific locations of interest. Also, predictive densities for future curves can be estimated. Although our method is Bayesian, the only informative priors that we use in our analyses involve the shrinkage hyperparameters, which can be estimated from the data by using the empirical Bayes method that we describe, if desired. Another advantage of the Bayesian approach is that there is a natural mechanism for handling measurement error or missingness, both in covariates and in the functional responses, since the missing or error prone data can simply be treated as parameters that are updated from their complete conditional distributions as part of the MCMC algorithm. Also, the structure of our framework makes it possible to consider functional hypothesis testing using Bayes factors or mixture priors with positive probabilities placed on zero functions. These ideas require further development, however, so are beyond the scope of this paper and are topics of future investigation.

There is some recent and on-going related work on functional analysis of variance using wavelets. Unlike here, the major focus in these papers is on developing frequentist functional hypothesis tests. Fan and Lin (1998) presented methods for functional testing using wavelets, although their framework did not include random effects. Abramovich and Angelini (2003) allowed functional random effects but only dealt with one-way analysis-of-variance mean structures. Antoniadis and Sapatinas (2004) also allowed functional random effects, and they described a functional mixed modelling framework that is similar to model (1), but they did not accommodate correlated random-effect functions.

There are other important differences between our modelling framework and those which were used in Fan and Lin (1998), Abramovich and Angelini (2003) and Antoniadis and Sapatinas (2004). Whereas we let the wavelet space variance components depend on scale j and location k, they only allowed them to depend on j, which places strong restrictions on functional forms of the between-curve deviations (see **Fig. 1**), which we expect should affect any subsequent inference. Also, since we specify diffuse proper priors for the wavelet space variance components for the random effects and update them within the MCMC algorithm, we estimate these parameters from the data and propagate the uncertainty of their estimation throughout subsequent inference. These variance components both model the curve-to-curve variability and serve as regularization parameters for the random-effect functions. In Antoniadis and Sapatinas (2004), the user simply fixes the relative sizes of these variance components across different wavelet scales j and then only estimates a single scalar variance component from the data. Abramovich and Angelini (2003) described a data-based method for estimating them, but they condition on these estimates as though they were known, and thus the inference that they describe does not account for their estimation error.

Antoniadis and Sapatinas (2004) and Abramovich and Angelini (2003) focused on functional hypothesis testing for fixed effect functions and, in Antoniadis and Sapatinas (2004), random-effect functions. This is clearly of interest in many contexts but is not the only relevant question with functional data. For example, the primary interest in many applications is not simply testing whether the function is identically 0, but rather identifying specific regions or features of the curves that differ from zero. No inferential procedures for these questions are described by them. One example is mass spectrometry proteomics, where the functions are characterized by many peaks corresponding to different proteins in the sample. The primary goal is not simply to decide whether there are any systematic differences in the mean curves for different groups of patients, but rather to identify which regions of the curves demonstrate differences. These specific regions can subsequently be mapped to individual proteins that may serve as useful biomarkers in medical applications.

We have developed easy-to-use code for implementing our method that we make freely available via `http://biostatistics.mdanderson.org/Morris/papers.html`. The minimum information that a user needs to supply includes a matrix of observed functions Y, fixed and random-design matrices X and Z, and a specification of the desired covariance structures and wavelet bases to use. Method-of-moments and generalized least squares starting values, vague proper priors on the variance components and empirical Bayes values for the hyperparameters are all automatically computed by the program and can be used, if desired. The program also contains an automatic, data-based method for determining the proposal variances that are necessary for the Metropolis–Hastings steps that are used to sample the large number of covariance parameters in the model. This method appears to work very well with none of the fine tuning that is normally required when implementing random-walk Metropolis–Hastings algorithms. This feature is key in making our method practically implementable for high dimensional functional data.

Acknowledgements

We thank Phil Brown, Marina Vannucci, Louise Ryan, Kevin Coombes, Keith Baggerly, Peter Mueller and Yuan Ji for useful discussions regarding this work. We also thank Joanne Lupton, Rob Chapkin, Nancy Turner and Meeyoung Hong for the colon carcinogenesis data and Dick Herrick for his assistance in helping to deal with various computational issues that arose in coding the method. We also thank the Associate Editor and referees, whose questions and insightful comments have led to a much improved paper. Morris's effort was supported by the National Cancer Institute (grant CA-107304). Carroll's research was supported by the National Cancer Institute (grant CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (grant P30-ES09106).

Appendix A: Conditional distribution for fixed effects

Here we show that the conditional distribution $(B^*_{ijk}|\mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)$ is a mixture of a point mass at zero and a normal distribution, with normal mixing proportion α_{ijk} given by equation (10) and the mean and variances of the normal μ_{ijk} and v_{ijk} given by equations (12) and (13) respectively.

Recall that, after integrating the random effects out of model (3), we have $\mathbf{d}_{jk} \sim \mathrm{MVN}(X\mathbf{B}^*_{jk}, \Sigma_{jk})$ where

$$\Sigma_{jk} = ZP(\boldsymbol{\Omega}_P)Z' * q^*_{jk} + R(\boldsymbol{\Omega}_R) * s^*_{jk}$$

as defined in equation (7). The prior for B^*_{ijk} is given by equation (4), which is a mixture of an $N(0, \tau_{ijk})$ distribution and a point mass at 0, with γ^*_{ijk} the indicator for the normal component of the mixture, which itself has a Bernoulli(π_{ij}) prior distribution.

We can write

$$f(B^*_{ijk}|\mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega) = \int f(B^*_{ijk}|\gamma^*_{ijk}, \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)\, f(\gamma^*_{ijk}|\mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)\, \mathrm{d}\gamma^*_{ijk}$$

$$= f(B^*_{ijk} | \gamma^*_{ijk} = 1, \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)\, \Pr(\gamma^*_{ijk} = 1 | \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega) \tag{14}$$

$$+ f(B^*_{ijk} | \gamma^*_{ijk} = 0, \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)\, \Pr(\gamma^*_{ijk} = 0 | \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega). \tag{15}$$

We shall first show that $f(B^*_{ijk} | \gamma^*_{ijk} = 1, \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)$ in expression (14) is normal with mean μ_{ijk} and variance v_{ijk}. Second, we shall show that $\Pr(\gamma^*_{ijk} = 1 | \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)$ in expression (14) is equal to α_{ijk}. It is trivial to show that in expression (15) $f(B^*_{ijk} | \gamma^*_{ijk} = 0, \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega) = I_0$ and $\Pr(\gamma^*_{ijk} = 0 | \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega) = 1 - \alpha_{ijk}$. First note that

$$f(B^*_{ijk} | \gamma^*_{ijk} = 1, \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega) \propto f(\mathbf{d}_{jk} | B^*_{ijk}, B^*_{(-i)jk}, \Omega)\, f(B^*_{ijk} | \gamma^*_{ijk} = 1)$$

$$\propto \exp\{-\tfrac{1}{2}(\mathbf{d}^*_{jk} - X_i B^*_{ijk})' \Sigma^{-1}_{jk} (\mathbf{d}^*_{jk} - X_i B^*_{ijk})\} \tag{16}$$

$$\times \exp(-\tfrac{1}{2} B^{*2}_{ijk} / \tau_{ijk}), \tag{17}$$

where $\mathbf{d}^*_{jk} = (\mathbf{d}_{jk} - X_{(-i)} B^*_{(-i)jk})$ are the 'residuals' after conditioning on the other fixed effect parameters. Multiplying expression (16) by the constant term

$$\exp[-\tfrac{1}{2} \operatorname{tr}\{(X_i' \Sigma^{-1}_{jk} X_i)(X_i' \Sigma^{-1}_{jk} X_i)^{-1} (X_i' \Sigma^{-1}_{jk} X_i)(X_i' \Sigma^{-1}_{jk} X_i)^{-1}\}],$$

reorganizing the terms within the trace and simplifying yields

$$\exp[-\tfrac{1}{2}(B^*_{ijk} - \hat{B}^*_{ijk,\mathrm{MLE}})' V^{-1}_{ijk} (B^*_{ijk} - \hat{B}^*_{ijk,\mathrm{MLE}})\}, \tag{18}$$

where

$$\hat{B}^*_{ijk,\mathrm{MLE}} = (X_i' \Sigma^{-1}_{jk} X_i)^{-1} X_i' \Sigma^{-1}_{jk} \mathbf{d}^*_{jk}$$

and $V_{ijk} = (X_i' \Sigma^{-1}_{jk} X_i)^{-1}$, as defined in equations (5) and (6). Combining the terms in expressions (18) and (17) and completing the square leaves us with $\exp\{-\tfrac{1}{2}(B^*_{ijk} - \mu_{ijk})^2 / v_{ijk}\}$, which is the kernel of an $N(\mu_{ijk}, v_{ijk})$ distribution, thus proving the first part.

For the second part, note that $\Pr(\gamma^*_{ijk} = 1 | \mathbf{d}_{jk}, B^*_{(-i)jk}, \Omega)$ can be written as $O_{ijk}/(O_{ijk} + 1)$, where O_{ijk} is the conditional odds of $\gamma^*_{ijk} = 1$ *versus* $\gamma^*_{ijk} = 0$, which can be written as a product of the prior odds $\pi_{ij}/(1 - \pi_{ij})$ and the conditional Bayes factor

$$\mathrm{BF}_{ijk} = \frac{f(\mathbf{d}_{jk} | \gamma^*_{ijk} = 1, B^*_{(-i)jk}, \Omega)}{f(\mathbf{d}_{jk} | \gamma^*_{ijk} = 0, B^*_{(-i)jk}, \Omega)}. \tag{19}$$

All that needs to be done is to show that BF_{ijk} simplifies into expression (11).

Consider the numerator of equation (19), which is

$$f(\mathbf{d}_{jk} | \gamma^*_{ijk} = 1, B^*_{(-i)jk}, \Omega) = \int f(\mathbf{d}_{jk} | B^*_{ijk}, B^*_{(-i)jk}, \Omega)\, f(B^*_{ijk} | \gamma^*_{ijk} = 1)\, \mathrm{d}B^*_{ijk}.$$

Given that

$$(\mathbf{d}_{jk} | B^*_{ijk}, B^*_{(-i)jk}, \Omega) \sim \mathrm{MVN}(X_i B^*_{ijk} + X_{(-i)} B^*_{(-i)jk}, \Sigma_{jk})$$

and $(B^*_{ijk} | \gamma^*_{ijk} = 1) \sim N(0, \tau_{ijk})$, some algebraic rearrangements and simplifications followed by the integration with respect to B^*_{ijk} reveal that

$$(\mathbf{d}_{jk} | \gamma^*_{ijk} = 1, B^*_{(-i)jk}, \Omega) \sim \mathrm{MVN}(X_{(-i)} B^*_{(-i)jk}, \Sigma_{jk} + X_i X_i' \tau_{ijk}),$$

or equivalently

$$(\mathbf{d}^*_{jk} | \gamma^*_{ijk} = 1, B^*_{(-i)jk}, \Omega) \sim \mathrm{MVN}(0, \Sigma_{jk} + X_i X_i' \tau_{ijk}).$$

It is trivial to show that $f(\mathbf{d}_{jk} | \gamma^*_{ijk} = 0, B^*_{(-i)jk}, \Omega)$ in the denominator of equation (19) is an $\mathrm{MVN}(X_{(-i)} B^*_{(-i)jk}, \Sigma_{jk})$ density. Thus, we can write the conditional Bayes factor BF_{ijk} as

$$\mathrm{BF}_{ijk} = \frac{|\Sigma_{jk} + X_i X_i' \tau_{ijk}|^{-1/2}}{|\Sigma_{jk}|^{-1/2}} \exp[-\tfrac{1}{2}(\mathbf{d}^*_{jk})' \{(\Sigma_{jk} + X_i X_i' \tau_{ijk})^{-1} - \Sigma^{-1}_{jk}\} \mathbf{d}^*_{jk}]. \tag{20}$$

Consider the first part of equation (20). Multiplying the numerator and denominator by $|\Sigma^{-1}_{jk}|^{-1/2}$, this simplifies to $|I_N + \tau_{ijk} X_i X_i' \Sigma^{-1}_{jk}|^{-1/2}$, where I_N is an $N \times N$ identity matrix, and recall that N is the number of observed functions. By the properties of determinants, we can rewrite this as the scalar quantity $(1 + \tau_{ijk} X_i' \Sigma^{-1}_{jk} X_i)^{-1/2}$, which is the first part of equation (11).

Now consider the exponent in equation (20). Using the well-known identity

$$\Sigma_1^{-1} - \Sigma_0^{-1} = -\Sigma_0^{-1} \boldsymbol{u}\boldsymbol{v}' \Sigma_0^{-1} / (1 + \boldsymbol{v}' \Sigma_0^{-1} \boldsymbol{u})$$

that holds whenever $\Sigma_1 = \Sigma_0 + \boldsymbol{u}\boldsymbol{v}'$, we can rewrite this expression and perform a series of simplifications

$$\begin{aligned}
&= \exp[(\tau_{ijk}/2)(1 + \tau_{ijk} X_i' \Sigma_{jk}^{-1} X_i)^{-1} \{(\mathbf{d}_{jk}^*)' (\Sigma_{jk}^{-1} X_i X_i' \Sigma_{jk}^{-1}) \mathbf{d}_{jk}^*\}] \\
&= \exp((\tau_{ijk}/2)[(X_i' \Sigma_{jk}^{-1} X_i)^{-1} \{(X_i' \Sigma_{jk}^{-1} X_i)' + \tau_{ijk}\}^{-1} (\mathbf{d}_{jk}^*)' \Sigma_{jk}^{-1} X_i X_i' \Sigma_{jk}^{-1} \mathbf{d}_{jk}^*]) \\
&= \exp\left[\frac{1}{2} \frac{(\mathbf{d}_{jk}^*)' \Sigma_{jk}^{-1} X_i (X_i' \Sigma_{jk}^{-1} X_i)^{-1} (X_i' \Sigma_{jk}^{-1} X_i)^{-1} X_i' \Sigma_{jk}^{-1} \mathbf{d}_{jk}^* \tau_{ijk}}{(X_i' \Sigma_{jk}^{-1} X_i)^{-1} \{\tau_{ijk} + (X_i' \Sigma_{jk}^{-1} X_i)\}}\right] \\
&= \exp\{\tfrac{1}{2} (\hat{B}_{ijk,\mathrm{MLE}}^{*2} / V_{ijk})(1 + V_{ijk}/\tau_{ijk})^{-1}\},
\end{aligned}$$

which, by letting $\zeta_{ijk} = \hat{B}_{ijk,\mathrm{MLE}}^{*} / \sqrt{V_{ijk}}$, gives us the second part of equation (11).

References

Abramovich, F. and Angelini, C. (2003) Testing in mixed-effects FANOVA models. *Technical Report RP SOR-03-03*. Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv.

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc.* B, **60**, 725–749.

Antoniadis, A. and Sapatinas, T. (2004) Estimation and inference in functional mixed-effects models. *Technical Report TR-15-2004*. Department of Mathematics and Statistics, University of Cyprus, Nicosia.

Brumback, B. A. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Ass.*, **93**, 961–976.

Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B, **62**, 681–698.

Clyde, M. and George, E. I. (2003) Discussion on 'Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis' (by J. S. Morris, M. Vannucci, P. J. Brown and R. J. Carroll). *J. Am. Statist. Ass.*, **98**, 584–585.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Dawid, A. P. (1981) Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.

Donoho, D. and Johnstone, I. M. (1995) Adapting to unknown smoothness by wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

Fan, J. and Lin, S. K. (1998) Tests of significance when data are curves. *J. Am. Statist. Ass.*, **93**, 1007–1021.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. New York: Chapman and Hall.

Gortmaker, S., Peterson, K., Wiecha, J., Sobol, A., Dixit, S., Fox, M. and Laird, N. (1999) Reducing obesity via a school-based interdisciplinary intervention among youth: planet health. *Arch. Ped. Adolesc. Med.*, **153**, 409–418.

Grambsch, P. M., Randall, B. L., Bostick, R. M., Potter, J. D. and Louis, T. A. (1995) Modeling the labeling index distribution: an application of functional data analysis. *J. Am. Statist. Ass.*, **90**, 813–821.

Guo, W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121–128.

Harville, D. (1997) *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Hong, M. Y., Lupton, J. R., Morris, J. S., Wang, N., Carroll, R. J., Davidson, L. A., Elder, R. and Chapkin, R. S. (2000) Dietary fish oil reduces O^6-methylguanine DNA adduct levels in the rat colon in part by increasing apoptosis during tumor initiation. *Cancer Epidem. Biomark. Prevn*, **9**, 819–826.

Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc.* B, **59**, 319–351.

Laird, N. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Liang, H., Wu, H. and Carroll, R. J. (2003) The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, **4**, 297–312.

Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A. and Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. *Bioinformatics*, **21**, 1764–1775.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003a) Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *J. Am. Statist. Ass.*, **98**, 573–583.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003b) Discussion on 'Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis' (by J. S. Morris, M. Vannucci, P. J. Brown and R. J. Carroll). *J. Am. Statist. Ass.*, **98**, 591–597.

Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. New York: Springer.

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.

Rice, J. A. and Wu, C. O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Appl. Statist.*, **45**, 151–163.

Staniswalis, J. G. and Lee, J. J. (1998) Nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass.*, **93**, 1403–1418.

Vannucci, M. and Corradi, F. (1999) Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. R. Statist. Soc.* B, **61**, 971–986.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wang, Y. (1998) Mixed effects smoothing spline analysis of variance. *J. R. Statist. Soc.* B, **60**, 159–174.

Wolfinger, R., Tobias, R. and Sall, J. (1994) Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM J. Scient. Comput.*, **15**, 1294–1310.

Wu, H. and Liang, H. (2004) Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scand. J. Statist.*, **31**, 3–20.

Wu, H. and Zhang, J. T. (2002) Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Ass.*, **97**, 883–897.

Zhang, D., Lin, X., Raz, J. and Sowers, M. F. (1998) Semiparametric stochastic mixed models for longitudinal data. *J. Am. Statist. Ass.*, **93**, 710–719.

Chapter 6
Robustness

By Roger Koenker and Douglas Simpson

About the Authors. Roger Koenker is William B. McKinley Professor of Economics and Professor of Statistics at the University of Illinois at Urbana-Champaign. He is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the Econometric Society. In 2010 he was the recipient of the Emanuel and Carol Parzen Prize for Statistical Innovation. He first met Ray on a seminar visit to the University of North Carolina at Chapel Hill in 1982, by which time he was already a devoted admirer of Ray's work, and he is profoundly grateful for the impetus that it had provided for his own work.

Douglas Simpson is Professor and Chair of Statistics at the University of Illinois at Urbana-Champaign. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics. He received his PhD in Statistics from the University of North Carolina at Chapel Hill in 1985 under the direction of Ray Carroll and David Ruppert. He first met Ray as a graduate student at UNC in the early 1980s and found him to an extraordinarily insightful person and effective mentor. He is deeply grateful for Ray's exceptional mentorship and research collaborations over the years.

Selected Papers on Robustness

[ROB-1]-[303] Ruppert D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 77, 828–838.

[ROB-2]-[155] Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics*, 10, 429–441.

[ROB-3]-[83] Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models, with applications to logistic regression. *Biometrika*, 73, 413–425.

[ROB-4]-[78] Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84, 460–466.

[ROB-5]-[104] Simpson, D. G., Ruppert, D., and Carroll, R. J. (1992). One-step GM-estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, 87, 439–450.

We are delighted to have the opportunity to introduce Ray's papers on robustness in statistics. These papers include important breakthrough developments such as the first rigorous asymptotic analysis of trimmed least squares, the first robust heteroscedastic regression estimators, the first efficient bounded-influence estimators for generalized linear regression, and the first bounded-influence regression estimators with high breakdown points. Like so many of Ray's publications, each of these widely cited papers is a marvel of clarity, innovation, and deep analysis.

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_6,

Trimmed Least Squares

Ray's long and extremely fruitful collaboration with David Ruppert began with a pair of 1978 technical reports on "trimmed least squares" estimation of the linear model that were eventually merged into a single 1980 *Journal of the American Statistical Association* paper (Ruppert and Carroll, 1980 [ROB-1]). Their papers consider two distinct approaches to constructing analogues of the trimmed mean for linear regression. The first is a proposal made by Koenker and Bassett (1978) in their "regression quantile" paper that had appeared earlier that year in *Econometrica*. It is remarkable that this paper appeared on the Chapel Hill radar screen at all, much less that it provoked such an immediate and profound response. Why? Perhaps it was only an early portent of the extraordinary energy and insight of the two collaborators, but the real explanation seems to have more to do with the appeal of *disproving* the conjecture that the Koenker–Bassett proposal behaves like a trimmed mean. Koenker and Bassett (1978) claim that in the classical linear model with symmetric i.i.d. errors, the weighted least squares estimator

$$\tilde{\beta}_\alpha = (X^T W X)^{-1} X^T W y$$

with diagonal weighting matrix, $W = \mathrm{diag}[I\{x_i^T \hat{\beta}(\alpha) < y_i < x_i^T \hat{\beta}(1-\alpha)\}]$, and $\hat{\beta}(\alpha) = \mathrm{argmin} \sum_i \rho_\alpha(y_i - x_i^T b)$ with $\rho_\alpha(u) = u\{\alpha - I(u < 0)\}$, satisfies,

$$n^{1/2}(\tilde{\beta}_\alpha - \beta) \rightsquigarrow \mathscr{N}(0, \sigma^2(\alpha, F)(X^T X)^{-1}),$$

where $\sigma^2(\alpha, F)$ is the α-Winsorized variance of the error distribution, F. It is entirely plausible that simpler estimators based on residuals from a preliminary fit of the model by some form of M-estimator could achieve the same objective. Contrary to this contra-conjecture, Ruppert and Carroll demonstrate that residual-based trimming has far less attractive asymptotic behavior than that of $\tilde{\beta}_\alpha$. Indeed, for asymmetric error distributions, they establish an even more general link with the limiting behavior of the trimmed mean. In the process, they provide a much more elegant approach to the large-sample theory, including a Bahadur representation, for the regression quantiles, $\hat{\beta}(\alpha)$, that helped to stimulate an extensive body of subsequent research. It wasn't until almost a decade later that Alan Welsh (Welsh, 1987) showed that residual-based trimming, if replaced by Winsorization of the residuals, can also achieve similar asymptotic objectives, thereby vindicating the original intuition that motivated their paper.

Robust Heteroscedastic Regression

Ray's next several papers with David Ruppert consider estimation and inference in heteroscedastic linear models. The role of heterogeneous scale in M-estimation was often neglected in the early robustness literature; Ray's papers confront the issue head-on, developing robust tests and proposing weighted M-estimator methods to achieve adaptive efficiency of the location component of the regression parameter. Their 1982 *Annals of Statistics* paper (Carroll and Ruppert 1982 [ROB-2]) foreshadows several later papers on optimal weighting and may be seen as a precursor of subsequent work on adaptive estimation of the linear model and the more general analysis of transformation models. While acknowledging the possible advantages of joint estimation of the location and scale components, some cautionary remarks about risks of misspecification inherent in such "feedback" methods are also made. Although "robustness" was an active concern in the early 1980s, much of the research had focused on location scale and had only just begun to make headway on regression problems. [ROB-2] is the first of a number of Ray's articles pushing the robustness literature beyond the location-scale families that dominated much of the 1970s and early 1980s.

Optimal Robust Score Functions

Continuing the push to expand robust methods into important practical areas, two key papers with Len Stefanski, David Ruppert, and Hans Künsch provide remarkable extensions of Hampel's constrained optimality theory for M-estimators to the class of generalized linear models.

In his Berkeley PhD thesis, Hampel develops a univariate constrained optimality result for censored maximum likelihood score functions (Hampel, 1968; Hampel et al., 1986), constructing a Fisher consistent robust estimating equation by censoring a shifted version of the likelihood score function and establishing that the resulting estimator had minimum asymptotic variance among M-estimators with the same bound on the influence function, a measure of local sensitivity to outliers. For symmetric error models, Krasker and Welsch (1982) develop an extension of Hampel's optimality result to multiple linear regression in a widely cited paper. The considerable technical challenges entail addressing the potential for unbounded influence of observations that are extreme in the design space as well as those that are outliers in the response space while maintaining equivariance to affine transformations of the regression variables.

In a tour de force generalization of both the Hampel result and the Krasker–Welsch result, Stefanski, Carroll, and Ruppert (1986 [ROB-3]) consider robust estimation for generalized linear regression. They not only extend the constrained optimality theory to the more general class of models, which necessitates treatment of an implicitly defined location functional for the censored score function, but they provide a simplified proof that encompasses the earlier result as well. This paper also introduces a convenient one-step method to reduce the considerable computational burden. In a related work around the same time, Simpson, Carroll, and Ruppert (1987) solve a conjecture of Hampel's regarding the asymptotic normality of the optimal M-estimator under a discrete models when the censoring points have positive mass.

In a remarkable follow-up paper, Künsch, Stefanski, and Carroll (1989 [ROB-4]) resolve many of the technical and conceptual limitations of the earlier work by introducing the concept of conditionally unbiased M-estimation, where the Fisher consistency is required to hold conditionally over all possible designs X. This strong concept implies the marginal Fisher consistency (averaged over the design space) of the earlier paper. This results in a very clean general constrained optimality result for bounded influence estimating equations for generalized linear models. The theory provides justification for Mallow's type leverage weights, which Ray expanded upon in a number of different papers, including Carroll and Welsh (1989), on the ability of generalized M-estimators (GM) to handle asymmetric errors, and Carroll and Pederson (1993) on robust logistic regression with bounded influence. In typical fashion for Ray's publications, the latter paper demonstrates impressively the performance of the proposed bounded influence estimator on a real and nontrivial dataset, obtaining new insights into the data that are not obvious using other methods.

Bounded Influence Regression with High Breakdown Point

The breakdown point is an appealing worst case measure of the stability of an statistical function: determine the fraction of data replacement that could force the statistical function to arbitrary values. The minimum such fraction is the breakdown point. In the univariate setting the sample median is an example of a statistical estimators with breakdown point $\approx 1/2$, whereas the sample mean has a breakdown point of $1/n$, indicating its sensitivity to a single outlier.

In the late 1970s, it was discovered that a wide class of regression equivariant M-estimators, which included the known bounded influence estimators, could not achieve a breakdown point higher than $1/(p+1)$ asymptotically, where p is the number of regression parameters (Maronna, Bustos, and Yohai, 1979). Following this discovery, the search was on to find regression equivariant estimators as well as multivariate location and scatter estimates that could achieve higher, dimension-free breakdown points. The first practical implementation of a such a high-breakdown regression estimator is the least median of squares (LMS) estimator of Rousseeuw (1984), in which the elemental set sampling algorithm for computing is also introduced. While this regression equivariant estimator has an impressive breakdown point $\approx 1/2$, its rate of convergence is $n^{-1/3}$, leading to inefficient parametric inferences and an unbounded local influence function. Nonetheless, this proposal demonstrates that a dimension-free lower bound on the breakdown point is possible in the regression setting and provides a means for outlier resistent exploratory data analysis. In the multivariate setting, Rousseeuw and van Zomeren (1990) establish that the minimum volume ellipsoid containing half the data (MVE) can be used to define location vector and scatter matrices that combine equivariance with breakdown points $\approx 1/2$ in the multivariate setting.

Several further developments led to estimators with high breakdown points and better root-n rates of convergence. However, by the late 1980s, it was still unknown whether or not a regression equivariant estimator could have both a high breakdown point and a bounded influence function. The one-step GM estimators of Simpson, Ruppert, and Carroll (1992 [ROB-5]) are the first such estimators. The approach is to employ the Bickel (1975) concept of a one-step Newton–Raphson adjustment, which can improve the rate of convergence of an initial estimator, with a high breakdown starting estimator such as LMS. Using a one-step Mallows-type estimator with design leverage downweighting, the asymptotics and breakdown point are established as well as the heuristic bounded influence function. In order to insure high breakdown regression equivariant design weights, the proposal is to downweight the regression score function inversely proportional to a power of the robust Mahalanobis distance of x from the robust center of the design distribution based on a high breakdown scatter matrix such as the MVE. Ever concerned about practical inferences, Ray also introduced the standard error breakdown concept, which led to a recommendation for stronger leverage downweighting of extreme observations than the standard Mallows weights in the literature. The resulting family of estimators is relatively straightforward to implement on top of existing functions for LMS regression and MVE location and scatter.

Simpson and Yohai (1998) follow up on this work, rigorously establishing the influence function and breakdown point of the corresponding statistical functionals. Following the publication of Simpson et al. (1992 [ROB-5]), the question remained open for several years whether a fully iterated estimator could retain the high breakdown point while achieving a bounded influence function. This question was finally answered in He, Simpson, and Wang (2000), which establishes that certain fully iterated heteroscedastic t regression estimators can have high breakdown points as well as bounded influence functions, and further, will be fully asymptotically efficient under the corresponding heteroscedastic t error model. The key idea of highly robust Mallows weights from [ROB-5] paper is central in this paper as well.

References

Other publications by Ray Carroll cited in this chapter.

Carroll, R. J., and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, 55, 693–706.

Carroll, R. J., and Welsh, A. H., (1989). A note on asymmetry and robustness in linear regression. *The American Statistician*, 42, 285–287.

Simpson, D. G., Carroll, R. J., and Ruppert, D. (1987). M-estimation for discrete data: asymptotic distribution theory and implications. *Annals of Statistics*, 15, 657–669.

Publications by other authors cited in this chapter.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70, 428–434.

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. Thesis. University of California, Berkeley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P.J., and Stahel, W. (1986) *Robust Statistics*, Wiley: New York.

He, X., Simpson, D. G., and Wang, G. (2000). Breakdown points of t-type regression estimators. *Biometrika*, 87, 675–687.

Koenker, R. and Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation using alternative definitions of sensitivity. *Journal of the American Statistical Association*, 77, 595–605.

Maronna, R. A., Bustos, O. H., and Yohai, V. J. (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, Springer-Verlag: New York.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633–639.

Simpson, D. G. and Yohai, V. J., (1998). Functional stability of one-step GM-estimators in approximately linear regression. *Annals of Statistics*, 26, 1147–1169.

Welsh, A. H. (1987). The trimmed mean in the linear model. *Annals of Statistics*, 15, 20–36.

Trimmed Least Squares Estimation in the Linear Model

DAVID RUPPERT and RAYMOND J. CARROLL*

We consider two methods of defining a regression analog to a trimmed mean. The first was suggested by Koenker and Bassett and uses their concept of regression quantiles. Its asymptotic behavior is completely analogous to that of a trimmed mean. The second method uses residuals from a preliminary estimator. Its asymptotic behavior depends heavily on the preliminary estimate; it behaves, in general, quite differently than the estimator proposed by Koenker and Bassett, and it can be inefficient at the normal model even if the percentage of trimming is small. However, if the preliminary estimator is the average of the two regression quantiles used with Koenker and Bassett's estimator, then the first and second methods are asymptotically equivalent for symmetric error distributions.

KEY WORDS: Regression analog; Trimmed mean; Regression quantile; Preliminary estimator; Linear model; Trimmed least squares.

1. INTRODUCTION

This article is concerned with the linear model

$$y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z} \ , \tag{1.1}$$

where $y' = (y_1, \ldots, y_n)$, $\mathbf{X}$ is a $n \times p$ matrix of known constants whose ith row is x'_i, $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ is a vector of unknown parameters, and $\mathbf{Z}' = (Z_1, \ldots, Z_n)$ is a vector of independent, identically distributed random variables with unknown distribution function F. Despite the advantages, including efficiency when F is normal, of the least squares (LS) estimator of $\boldsymbol{\beta}$, this estimator is inefficient when F has heavier tails than the Gaussian distribution, and the estimator possesses high sensitivity to spurious observations. This inefficiency to heavy-tailed F has been amply demonstrated for the location submodel by a Monte Carlo study (Andrews et al. 1972) and by asymptotics (e.g., Table 1 of this article). The presence of spurious data can be modeled by letting F be a mixture of the distribution function of the "good" data, for example, standard normal, and that of the "bad" data, for example, normal with variance exceeding 1. Such an F will have heavier tails than a normal distribution, and inefficiency with heavy-tailed F appears to be closely related to sensitivity to outliers. Huber (1977, p. 3) stated that "for most practical purposes, 'distributionally robust' and 'outlier resistant' are interchangeable." For the location model, three classes of estimators have been proposed as alternatives to the sample mean, and M, L, and R estimators (see Huber 1977 for an introduction). Among the L estimates, the trimmed mean is particularly attractive because it is easy to compute and is efficient under a variety of circumstances.

As with M estimates, trimmed means can be used to form confidence intervals (Huber 1970). The Monte Carlo study by Gross (1976) indicates that the validity (agreement of nominal and actual significance levels) of such confidence intervals will not be wholly satisfactory if n is small (Gross studied $n = 10$ and $n = 20$), but with F nonnormal, they appear to be as valid as the standard confidence interval based on the sample mean and standard deviation and the t distribution.

Hogg (1974) favored trimmed means for the previous reasons and because they can serve as a basis for adaptive estimators. Stigler (1977) applied robust estimators to data from 18th- and 19th-century experiments designed to measure basic physical constants. He concluded that "the 10% trimmed mean (the smallest nonzero trimming percentage included in the study) emerges as the recommended estimator (p. 1070)."

One might argue, of course, that although L estimates have desirable properties, they really offer no advantages over other estimators. After all, Jaeckel (1971) has shown that if F is symmetric, then for each L estimator of location, there are asymptotically equivalent M and R estimators. However, without knowledge of F it is not possible to match up an L estimator with its corresponding M and R estimators.

We do not believe that trimmed means are always preferable to M estimates but rather that they are worthwhile alternatives to M estimates, particularly to Huber's M estimate.

For the linear model, Bickel (1973) has proposed a class of one-step L estimators depending on a preliminary estimate of $\boldsymbol{\beta}$, but although these have good asymptotic efficiencies, they are computationally complex and are apparently not invariant to reparameterization.

In this article we consider two other methods of defining a regression analog to the trimmed mean. In the location problem, both estimates reduce to the trimmed mean. The first, which we call $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\alpha)$ for $0 < \alpha < \frac{1}{2}$, requires a preliminary estimate, which is denoted by $\hat{\boldsymbol{\beta}}_0$. Suppose that the residuals from $\hat{\boldsymbol{\beta}}_0$ are calculated and that those observations corresponding to the $[n\alpha]$ smallest and $[n\alpha]$ largest residuals are removed. Then $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(= \hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\alpha))$

* David Ruppert is Assistant Professor and Raymond J. Carroll is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. This research was supported by NSF Grant MCS78-01240 and by the Air Force Office of Scientific Research under Contracts AFOSR-75-2796 and AFOSR-80-0080. The present article is a synthesis of two earlier manuscripts. Many helpful suggestions were provided by two editors, two associate editors, and several referees. We are especially grateful for the suggestion to combine the two manuscripts.

© Journal of the American Statistical Association
December 1980, Volume 75, Number 372
Theory and Methods Section

is defined as the LS estimator calculated from the remaining observations.

The definition of $\hat{\beta}_{\mathrm{PE}}$ was motivated by the applied statisticians' practice of examining the residuals from a LS fit, removing the points with large (absolute) residuals and recalculating the LS solution with the remaining observations. Generally, there is no formal rule for deciding which points to remove, but $\hat{\beta}_{\mathrm{PE}}$ is at least similar to this practice.

The second method of defining an analog to the trimmed mean was proposed by Koenker and Bassett (1978), who extended the concept of quantiles to the linear model. Let $0 < \theta < 1$. Define

$$\psi_\theta(x) = \theta - \mathrm{I}(x < 0) \ , \qquad (1.2)$$

there $I(x < a)$ is the indicator of the set $\{x\colon x < a\}$, and

$$\rho_\theta(x) = x\psi_\theta(x) \ .$$

Then they called $\hat{\beta}(\theta)$ any value of b that solves

$$\sum_{i=1}^{n} \rho_\theta(y_i - x_i b) = \min! \ , \qquad (1.3)$$

a θth regression quantile. (Recall that x'_i is the ith row of $\mathbf{X}$.) Koenker and Bassett's Theorem 4.2 states that regression quantiles have asymptotic behavior similar to that of sample quantiles in the location problem. (Since $\hat{\beta}(\theta)$ is an M estimate, its large-sample behavior can also be deduced from standard M estimate theory, as we show later.) Therefore, L estimates consisting of linear combinations of a fixed number of order statistics—for example, the median, trimean, and Gastwirth's estimator—are easily extended to the linear model and have the same asymptotic efficiencies as in the location model. As Koenker and Bassett pointed out, regression quantiles can be computed by standard linear programming techniques. They also suggested the following trimmed LS estimators ($\hat{\beta}_{\mathrm{KB}}$): Remove from the sample any observations whose residual from $\hat{\beta}(\alpha)$ is negative or whose residual from $\hat{\beta}(1 - \alpha)$ is positive, and calculate the LS estimator using the remaining observations. They conjectured that if $\lim_{n\to\infty} n^{-1}(\mathbf{X}'\mathbf{X}) = Q$ (positive definite), then the asymptotic covariance of $\hat{\beta}_{\mathrm{KB}}(\alpha)$ is $n^{-1}\sigma^2(\alpha, F)Q^{-1}$, where $n^{-1}\sigma^2(\alpha, F)$ is the variance of an α-trimmed mean from a population with distribution F.

In this article we develop asymptotic expansions for $\hat{\beta}(\theta)$ and $\hat{\beta}_{\mathrm{KB}}(\alpha)$ that provide simple proofs of Koenker and Bassett's Theorem 4.2 and their conjecture about the asymptotic covariance of $\hat{\beta}_{\mathrm{KB}}(\alpha)$.

The close analogy betwen the asymptotic distributions of trimmed means and the trimmed LS estimator $\hat{\beta}_{\mathrm{KB}}(\alpha)$ is remarkable. Perhaps even more surprising is that the distribution of the estimator $\hat{\beta}_{\mathrm{PE}}$ depends heavily on that of the preliminary estimator $\hat{\beta}_0$. In particular, if the preliminary estimate is either the LS or least-absolute-deviation (LAD) estimator, then $\hat{\beta}_{\mathrm{PE}}$ is inefficient at the normal model (for LS this was surprising) and is not a regression analog to the trimmed mean.

Our results are such that we are able to find a choice of $\hat{\beta}_0$ so $\hat{\beta}_{\mathrm{PE}}$ corresponds to a trimmed mean when the error distribution F is symmetric. The "right choice" for $\hat{\beta}_0$ is the average of the αth and $(1 - \alpha)$th regression quantiles (i.e., $\hat{\beta}_0 = \frac{1}{2}(\hat{\beta}(\alpha) + \hat{\beta}(1 - \alpha))$.

Hogg (1974) mentioned that adaptive estimators can be constructed from estimators similar or identical to $\hat{\beta}_{\mathrm{PE}}(\alpha)$ with α a function of the residuals from $\hat{\beta}_0$. The advantage of this class of adaptive estimators, he felt, was that they "would correspond more to the trimmed means for which we can find an error structure" (p. 917). However, from the previously described results, we can conclude that even if α is nonstochastic, estimators of the type suggested by Hogg will not necessarily have error structures that correspond to the trimmed mean.

Besides its nice asymptotic covariance, $\hat{\beta}_{\mathrm{KB}}$ has another desirable property. In the location model, if F is asymmetric then there is no natural parameter to estimate. In the linear model, if the design matrix is centered so one column, for example, the first, consists entirely of ones and the remaining columns each sum to zero, then our expansions show that for each $0 < \alpha < \frac{1}{2}$,

$$n^{\frac{1}{2}}(\hat{\beta}_{\mathrm{KB}}(\alpha) - \beta - \delta(\alpha)) \xrightarrow{\mathcal{L}} N(0, Q^{-1}\sigma^2(\alpha, F)) \ ,$$

where $\delta(\alpha)$ is a vector whose components are all zero except for the first. Therefore, the ambiguity about the parameter being estimated involves only the intercept and none of the slope parameters. However, this is also true for M estimates (see, e.g., Carroll and Ruppert 1979 or Carroll 1979).

We present a large-sample theory of confidence ellipsoids and general linear hypothesis testing that is similar to that of LS estimation. The same theory holds for $\hat{\beta}_{\mathrm{PE}}$ when $\hat{\beta}_0 = (\hat{\beta}(\alpha) + \hat{\beta}(1 - \alpha))/2$, but only if F is symmetric.

The methods in this article can be applied to other estimators. For example, let $\hat{\beta}_{\mathrm{A}}(\alpha) = \hat{\beta}_{\mathrm{A}}$ be the LS estimate, after the points with the $[2\alpha N]$ largest absolute residuals from $\hat{\beta}_0$ are removed. In Section 6 we state results for $\hat{\beta}_{\mathrm{A}}$. Their proofs are omitted but are similar to the proofs of analogous results for $\hat{\beta}_{\mathrm{PE}}$.

In Section 2 we give notation and assumptions. In Section 3, asymptotic representations of $\hat{\beta}_{\mathrm{PE}}$ are developed, and their significance is discussed in Section 4. Section 5 contains asymptotic results for $\hat{\beta}_{\mathrm{KB}}$, and Section 6 discusses conditions under which $\hat{\beta}_{\mathrm{KB}}$ and $\hat{\beta}_{\mathrm{PE}}$ are asymptotically equivalent. In Section 7, we compare the asymptotic behavior of $\hat{\beta}_{\mathrm{PE}}$ for several choices of $\hat{\beta}_0$. Large-sample inference is the subject of Section 8. Several examples with real data are considered in Section 9. All proofs are found in the Appendix.

2. NOTATION AND ASSUMPTIONS

Although y, $\mathbf{X}$, and $\mathbf{Z}$ in (1.1) depend on n, this is not made explicit in the notation. Let $\mathbf{e}' = (1, 0, \ldots, 0)$ $(1 \times p)$, and let $\mathbf{I}_p$ be the $p \times p$ identity matrix. Whenever r is a scalar, $\mathbf{r} = r\mathbf{e}$. For $0 < p < 1$, define $\eta(p) = F^{-1}(p)$. Suppose $0 < \alpha_1 < \frac{1}{2} < \alpha_2 < 1$, and define

$\xi_1 = \eta(\alpha_1)$ and $\xi_2 = \eta(\alpha_2)$. Let $N_p(\mu, \mathbf{\Sigma})$ denote the p-variate Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$. We make the following assumptions throughout.

1. F has a continuous density f that is positive on the support of F.

2. Letting $(x_{il}, \ldots, x_{ip}) = x'_i$ be the ith row of $\mathbf{X}$, $x_{i1} = 1$ for $i = 1, \ldots, n$ and

$$\sum_{i=1}^{n} x_{ij} = 0 \ , \quad \text{for} \quad j = 2, \ldots, p \ .$$

3. $\lim_{n\to\infty} \left(\max_{j \leq p \text{ and } i \leq n} (n^{-\frac{1}{2}} |x_{ij}|) \right) = 0 \ .$

4. There exists positive definite $\mathbf{Q}$ such that

$$\lim_{n\to\infty} n^{-1}(\mathbf{X'X}) = \mathbf{Q} \ .$$

5. $(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} - \theta\mathbf{e}) = O_p(n^{-\frac{1}{2}})$ for some constant θ.

We assume that $\xi_{\frac{1}{2}} = 0$. By Assumption 2, this involves no loss in generality. Assumption 5 is satisfied by many estimators, including the LAD or median regression (see Corollary 2) and, if $Ee_1^2 < \infty$, the LS estimators.

The residuals from the preliminary estimate $\hat{\boldsymbol{\beta}}_0$ are

$$r_i = y_i - x'_i\hat{\boldsymbol{\beta}}_0 = Z_i - x'_i(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \ .$$

Let r_{1n} and r_{2n} be the $[n\alpha]$th and $[n(1 - \alpha)]$th ordered residuals, respectively. Then the estimate $\hat{\boldsymbol{\beta}}_{\text{PE}}$ is a LS estimate that is calculated after all observations are removed that satisfy

$$r_i \leq r_{1n} \quad \text{or} \quad r_i \geq r_{2n} \ . \tag{2.1}$$

Because of Assumption 1, asymptotic results are unaffected by requiring strict inequalities in (2.1). Let $a_i = 0$ or 1 according to whether i satisfies (2.1) or not, and let $\mathbf{A}$ be the $n \times n$ diagonal matrix, with $\mathbf{A}_{ii} = a_i$. The matrix $\mathbf{A}$ indicates which observations are not trimmed. Thus

$$\hat{\boldsymbol{\beta}}_{\text{PE}}(\alpha) = (\mathbf{X'AX})^{-}\mathbf{X'A}y \ ,$$

where $(\mathbf{X'AX})^{-}$ is a generalized inverse for $\mathbf{X'AX}$. (Later we show that

$$n^{-1}(\mathbf{X'AX}) \xrightarrow{P} (1 - 2\alpha)Q \ ;$$

thus $\Pr((\mathbf{X'AX}) \text{ is invertible}) \to 1$.)

Since $\hat{\boldsymbol{\beta}}_{\text{KB}}$ behaves similarly to a trimmed mean, even for asymmetric F and for asymmetric trimming, we do not restrict ourselves to symmetric trimming when defining $\hat{\boldsymbol{\beta}}_{\text{KB}}$.

Let $\alpha = (\alpha_1, \alpha_2)$ and define $\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha)$ to be a LS estimator calculated after all observations are removed that satisfy

$$y_i - x'_i\hat{\boldsymbol{\beta}}(\alpha_2) \geq 0 \quad \text{or} \quad y_i - x'_i\hat{\boldsymbol{\beta}}(\alpha_1) \leq 0 \ . \tag{2.2}$$

(Again asymptotic results are unaffected by requiring strict inequalities in (2.2), which is Koenker and Bassett's suggestion.) Let $b_i = 0$ or 1 according to whether i satisfies (2.2) or not, and let $\mathbf{B}$ be the $n \times n$ diagonal matrix with $\mathbf{B}_{ii} = b_i$. Then

$$\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) = (\mathbf{X'BX})^{-} \ (\mathbf{X'B}y) \ ,$$

where $(\mathbf{X'BX})^{-}$ is a generalized inverse of $(\mathbf{X'BX})$. (Again, for n sufficiently large, $\mathbf{X'BX}$ will be invertible.) Let

$$\begin{aligned} \phi(x) &= \xi_1/(\alpha_2 - \alpha_1) \quad \text{if} \quad x < \xi_1 \ , \\ &= x/(\alpha_2 - \alpha_1) \quad \text{if} \quad \xi_1 \leq x \leq \xi_2 \ , \\ &= \xi_2/(\alpha_2 - \alpha_1) \quad \text{if} \quad \xi_2 < x \ . \end{aligned} \tag{2.3}$$

Define

$$\delta(\alpha) = (\alpha_2 - \alpha_1)^{-1} \int_{\xi 1}^{\xi_2} x dF(x) \ ,$$

and letting $\kappa_j = (\xi_j - \boldsymbol{\delta}(\alpha))$, define

$$\sigma^2(\alpha, F) = (\alpha_2 - \alpha_1)^{-2} \left(\int_{\xi 1}^{\xi_2} (x - \delta(\alpha))^2 \, dF(x) \right. \\ \left. + \alpha_1\kappa_1^2 + (1 - \alpha_2)\kappa_2^2 - ((1 - \alpha_2)\kappa_2 + \alpha_1\kappa_1)^2 \right) \ .$$

By, for example, deWet and Venter (1974, Equation (6)), $\sigma^2(\alpha, F)/n$ is the asymptotic variance of a trimmed mean, with trimming proportions α_1 and $1 - \alpha_2$ from a population with distribution F.

3. MAIN RESULTS FOR $\hat{\boldsymbol{\beta}}_{\text{PE}}$

First we will find relationships of the form

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{PE}} - \boldsymbol{\beta}) \approx n^{-\frac{1}{2}} \sum_{i=1}^{n} G(x_i, \mathbf{Z}_i) + n^{\frac{1}{2}}H(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \ , \tag{3.1}$$

where G and H are given functions. We then show that in many special cases (including LS and LAD) the latter term in (3.1) can be further expanded so that

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{PE}} - \boldsymbol{\beta}) \approx n^{-\frac{1}{2}} \sum_{i=1}^{n} G(x_i, \mathbf{Z}_i) + n^{-\frac{1}{2}} \sum_{i=1}^{n} H^*(x_i, \mathbf{Z}_i) \tag{3.2}$$

for some function H^*. It is then a simple matter to obtain the limit distribution of $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{PE}} - \boldsymbol{\beta})$ from (3.2).

In this section we only consider symmetric trimming, so we assume $\alpha_1 = 1 - \alpha_2 = \alpha$. Now define $a = \xi_2 f(\xi_2) - \xi_1 f(\xi_1)$ and $c_i = (\mathbf{I} - \mathbf{ee}')x_i = (0, x_{i2}, \ldots, x_{ip})'$. The specific relationship of form (3.1) is the following:

Theorem 1: As $n \to \infty$,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{PE}} - \boldsymbol{\beta}) = (1 - 2\alpha)^{-1} n^{-\frac{1}{2}} \sum_{i=1}^{n} Q^{-1} c_i \mathbf{Z}_i \mathbf{I}(\xi_1 \leq \mathbf{Z}_i \leq \xi_2) \\ + (1 - 2\alpha)^{-1} a n^{\frac{1}{2}} (\mathbf{I} - \mathbf{ee}')(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) \\ + n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{e}\boldsymbol{\phi}(\mathbf{Z}_i) + o_p(1) \ . \tag{3.3}$$

We call the first entry of $\boldsymbol{\beta}$ the intercept, and the remaining entries are called the slopes. Since premultiplication of a vector by $(\mathbf{I} - \mathbf{ee}')$ simply replaces the first coordinate by 0, the first two terms on the right-hand side (rhs) of (3.3) represent the slope estimates. Note the similarity (and the difference) between the first term and a representation of the (untrimmed) LS estimate, $\hat{\boldsymbol{\beta}}$; since

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Z}\ ,$$

and

$$(n^{-1}\,\mathbf{X}'\mathbf{X}) \to \mathbf{Q}\ ,$$

it follows that

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} Q^{-1}x_i\mathbf{Z}_i + o_p(1)\ .$$

The second term indicates the contribution of the preliminary estimate to the trimmed LS estimate; this contribution is only to the slope estimates. Since only the first coordinate of $\mathbf{e}$ is nonzero, the third term on the rhs of (3.3) is a representation of the intercept estimate and is identical to a representation of the trimmed mean in the location model (cf. Corollary 1).

To specify the relationship of form (3.2), we make the following assumption:

6. For some function g,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} Q^{-1}x_i g(\mathbf{Z}_i) + o_p(1)\ .$$

As indicated, Assumption 6 holds with $g(x) = x$ if $\hat{\boldsymbol{\beta}}_0$ is the LS estimate. By Theorem 5.3, Assumption 6 holds with $g(x) = (f(0))^{-1}(\frac{1}{2} - \mathbf{I}(x < 0))$ if $\hat{\boldsymbol{\beta}}_0$ is the LAD estimate. As an immediate consequence of Theorem 1, we have our main result.

Theorem 2:

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}} - \boldsymbol{\beta})$$
$$= (1-2\alpha)^{-1}n^{-\frac{1}{2}} \sum_{i=1}^{n} Q^{-1}c_i\{\mathbf{Z}_i I(\xi_1 \le \mathbf{Z}_i \le \xi_2) + ag(\mathbf{Z}_i)\}$$
$$+ n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{e}\phi(\mathbf{Z}_i) + o_p(1)\ . \tag{3.4}$$

In the next section, limit distributions are obtained from (3.4) for various special cases. Both (3.3) and (3.4) show how the preliminary estimate influences the asymptotic behavior of $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$.

As a special case of Theorem 2, we obtain

Corollary 1: In the location model ($p = 1$ and $x_i = 1$ for all i)

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}} - \boldsymbol{\beta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi(\mathbf{Z}_i) + o_p(1)\ .$$

The key technical step in the proofs is an asymptotic linearity result for ordered residuals, which generalizes work of Bahadur (1966) and Ghosh (1971) for the location model.

Lemma 1: For $0 < \theta < 1$, let $r_{\theta n}$ be the $[n\theta]$th ordered residual from $\hat{\boldsymbol{\beta}}_0$. Then

$$n^{\frac{1}{2}}(r_{\theta n} - \eta(\theta)) = f(\eta(\theta))^{-1}n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_\theta(\mathbf{Z}_i - \eta(\theta))$$
$$- \mathbf{e}'n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) + o_p(1)\ .$$

(Recall that $\psi_\theta(x) = \theta - I(x < 0)$.)

4. ASYMPTOTIC BEHAVIOR OF $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$

In this section we show that Theorem 2 leads to these basic conclusions about $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$:

1. The intercept estimate is asymptotically unbiased if F is symmetric; 2. The slope estimates are asymptotically unbiased even if F is asymmetric; 3. The asymptotic variance of the intercept, which does not depend on the choice of $\hat{\boldsymbol{\beta}}_0$, is that of the trimmed mean in the location model; and 4. The asymptotic covariance matrix of the slopes depends on $\hat{\boldsymbol{\beta}}_0$ and in general will be difficult to estimate.

Let $\mathbf{0}$ be a $(p - 1) \times 1$ vector of zeros. By Assumption 2, there is a $\tilde{\mathbf{Q}}$ such that

$$\mathbf{Q} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{Q}} \end{bmatrix} \quad\text{and}\quad \mathbf{Q}^{-1} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{Q}}^{-1} \end{bmatrix}.$$

Moreover,

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} c_i c'_i = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \tilde{\mathbf{Q}} \end{bmatrix},$$

and

$$\mathbf{Q} \sum_{i=1}^{n} c_i = \begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}. \tag{4.1}$$

If we estimate $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$, then the asymptotic bias of the intercept is

$$E\phi(\mathbf{Z}_1) = (1 - 2\alpha)^{-1}\int_{\xi_1}^{\xi_2} x\,dF(x)\ ,$$

which is zero if F is symmetric about zero. By (3.4) and (4.1), the slope estimates are asymptotically unbiased, even if F is asymmetric. The asymptotic variance of the normalized intercept is

$$\sigma^2(\alpha, F) = \operatorname{var}\phi(\mathbf{Z}_1)\ ,$$

the asymptotic variance of the normalized α-trimmed mean in the location model. The intercept is asymptotically uncorrelated with the slopes, and the asymptotic covariance matrix of the normalized slopes is $\tilde{Q}^{-1}\sigma^2(\alpha, g, F)$, where

$$\sigma^2(\alpha, g, F)$$
$$= (1 - 2\alpha)^{-2}\operatorname{var}(\mathbf{Z}_1 I(\xi_1 \le \mathbf{Z}_1 \le \xi_2) + ag(\mathbf{Z}_1))\ .$$

We see that the asymptotic distribution of the intercept estimate does not depend on the choice of $\hat{\boldsymbol{\beta}}_0$, provided $(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) = O_p(n^{-\frac{1}{2}})$. On the other hand, we see from (3.4) that the slope estimates depend upon $\hat{\boldsymbol{\beta}}_0$, since the unusual situation in which $a = 0$ is ruled out by Assumption 1. Using the Lindeberg central limit theorem and Theorem 2, it is easy to show that under Assumptions 3 and 4, $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}} - \boldsymbol{\beta} - \mathbf{e}E\phi(\mathbf{Z}_1))$ converges in distribution to a normal law.

In general, large-sample statistical inference based on $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ will be a challenging problem because of the difficulties of estimating $a = (\xi_2 f(\xi_2) - \xi_1 f(\xi_1))$. Obtaining reasonably good estimates of the density f might take very large sample sizes.

5. MAIN RESULTS FOR $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$

In Section 1, we defined a θth regression quantile to be any value of b that solves (1.5). There may be multiple solutions, though in our few examples we found that the solution was always unique. However, the asymptotic results we present do not depend on the rule used to select one. We suppose, then, that a definite rule has been used, and we denote this solution by $\hat{\boldsymbol{\beta}}(\theta)$.

For $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ we obtain an asymptotic representation that is similar to those for $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ but perhaps simpler.

Theorem 3: The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ satisfies

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{KB}}(\alpha) - \boldsymbol{\beta}) = Q^{-1}n^{-\frac{1}{2}} \cdot \{\sum_{i=1}^{N} x'_i(\boldsymbol{\phi}(\mathbf{Z}_i) - E\boldsymbol{\phi}(\mathbf{Z}_i)) + \boldsymbol{\delta}(\alpha)\} + o_p(1) \ , \quad (5.1)$$

and therefore

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{KB}}(\alpha) - \boldsymbol{\beta} - \boldsymbol{\delta}(\alpha)) \xrightarrow{\mathcal{L}} N_p(\mathbf{0}, \sigma^2(\alpha, F)Q^{-1}) \ . \quad (5.2)$$

Expression (5.1) is similar to a result of deWet and Venter (1974, Equation (5)). Note that (5.2) verifies Koenker and Bassett's (1978) hypothesis on the covariance of $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$. Moreover, the bias of $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ for $\boldsymbol{\beta}$ involves only the intercept β_1 and not the slopes. Also, $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ is asymptotically unbiased if F is symmetric.

Theorem 3 requires the next result on regression quantiles. Define $\boldsymbol{\beta}(\theta) = \boldsymbol{\beta} + \eta(\theta)$. The next theorem, which is a special case of a general result for M estimators, shows that $(\hat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta))$ is essentially a sum of independent random variables.

Theorem 4: The following representation holds:

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)) = n^{-\frac{1}{2}}(f(\eta(\theta)))^{-1}\mathbf{Q}^{-1} \sum_{i=1}^{n} x_i\psi_\theta(\mathbf{Z}_i - \eta(\theta)) + o_p(1) \ .$$

Theorem 4 and the Lindeberg central limit theorem provide an easy proof of Koenker and Bassett's (1978) Theorem 4.2, which we state as a corollary.

Corollary 2: Let $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\theta_1, \ldots, \theta_m; F)$ be the symmetric $m \times m$ matrix defined by

$$\Omega_{ij} = \frac{\theta_i(1 - \theta_j)}{f(\eta(\theta_i))f(\eta(\theta_j))} , \quad 1 \leq i \leq j \leq m \ .$$

Then

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\theta_1) - \boldsymbol{\beta}(\theta_1), \ldots, \hat{\boldsymbol{\beta}}(\theta_m) - \boldsymbol{\beta}(\theta_m)) \xrightarrow{\mathcal{L}} N_{mp}(\mathbf{0}, \boldsymbol{\Omega} \otimes \mathbf{Q}^{-1}) \ .$$

6. A CHOICE OF $\hat{\boldsymbol{\beta}}_0$ FOR WHICH $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ AND $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ ARE ASYMPTOTICALLY EQUIVALENT

We have seen that $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ is a close analog to the trimmed mean, but the behavior of $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ depends on $\hat{\boldsymbol{\beta}}_0$ and is in general not similar to that of a trimmed mean. One might ask whether $\hat{\boldsymbol{\beta}}_0$ can be chosen so that $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ has the same asymptotic distribution as $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$. The answer is yes, provided that F is symmetric and that we allow only symmetric trimming.

Let $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{RQ}, \alpha)(= \hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{RQ}))$ be $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ when $\hat{\boldsymbol{\beta}}_0$ is the average of the αth and $(1 - \alpha)$th regression quantiles (i.e., $\hat{\boldsymbol{\beta}}_0 = (\hat{\boldsymbol{\beta}}(\alpha) + \boldsymbol{\beta}(1 - \alpha))/2$). Then, by Theorem 4, this $\hat{\boldsymbol{\beta}}_0$ satisfies Assumption 6 with

$$g(x) = (2f(\xi_1))^{-1}\psi_\alpha(x - \xi_1) + (2f(\xi_2))^{-1}\psi_{1-\alpha}(x - \xi_2) \ .$$

If F is symmetric, then $\xi_1 = - \xi_2$, $f(\xi_1) = f(\xi_2)$, and therefore

$$ag(x) = \xi_1 I(x \leq \xi_1) + \xi_2 I(x \geq \xi) \ . \quad (6.1)$$

By (3.4) and (6.1),

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(RQ) - \boldsymbol{\beta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} Q^{-1}x_i\boldsymbol{\phi}(\mathbf{Z}_i) + o_p(1) \ ,$$

and therefore, since $\boldsymbol{\delta}(\alpha) = 0$, (5.1) implies

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{KB}} - \boldsymbol{\beta}_{\mathrm{PE}}(RQ)) \xrightarrow{P} \mathbf{0} \ , \quad (6.2)$$

so that asymptotically there is no difference between trimming with this preliminary estimate and using Koenker and Bassett's (1978) proposal. (However, (6.2) does not necessarily hold if F is asymmetric.)

Notice that (5.1) and (6.2) imply that

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(RQ) - \hat{\boldsymbol{\beta}}) \xrightarrow{\mathcal{L}} N(0, \mathbf{Q}^{-1}\sigma^2(\alpha, F)) \ .$$

7. COMPARISONS OF SEVERAL CHOICES OF $\hat{\boldsymbol{\beta}}_0$

The choice of $\hat{\boldsymbol{\beta}}_0$ should be based on the efficiency of the resulting $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$, not on its similarity to $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$. In this section we find further support for using $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{RQ})$ by comparing $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{RQ})$ with two other estimators, $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{LS})$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{LAD})$, which are $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ with $\hat{\boldsymbol{\beta}}_0$ equal to the LS estimator and $\hat{\boldsymbol{\beta}}(.5)$, respectively. Comparisons are made within the family of contaminated normal distributions, which has long been used to study the behavior of statistical procedures under heavy-tailed distributions. (Stigler 1973 represents an interesting account of its early use.) These distributions have the form

$$F(x) = (1 - \epsilon)\Phi(x) + \epsilon\Phi(x/b) \ ,$$

where $0 < \epsilon < 1$ and Φ is standard normal distribution. Typically, $b > 1$ and $\Phi(x/b)$ is the distribution of the "bad" data, whereas ϵ is the proportion of "bad" observations. Recall that the asymptotic variance of the intercept does not depend on $\hat{\boldsymbol{\beta}}_0$ and that the asymptotic covariance matrix of the slopes is $\tilde{Q}^{-1}\sigma^2(\alpha, g, F)$, where $\tilde{Q}^{-1}$ depends only on the sequence of design matrices. Therefore, we can compare the estimators by using only $\sigma^2(\alpha, g, F)$. Table 1 displays $\sigma^2(\alpha, g, F)$ for several choices of α, ϵ, and b, and for g corresponding to $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{LS})$, $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{LAD})$, and $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}(\mathrm{RQ})$. For comparison, we include the standardized asymptotic variance (i.e., σ^2 where $\sigma^2 Q^{-1}$ is the asymptotic covariance matrix) for the LS estimate and two M estimates, a Huber and a Hampel. Both of the

1. *Variances of the Asymptotic Distribution of Slope Estimators of Contaminated Normal Distributions*

ϵ[a]	b[b]	*Estimator*											
					Trimmed Least Squares								
					$\hat{\beta}_{PE}$(LS) *(Least Squares As Preliminary Estimate)*			$\hat{\beta}_{PE}$(LAD) *(Least Absolute Deviation As Preliminary Estimate)*			$\hat{\beta}_{PE}$(RQ) *(Average of* α*th and 1-*α*th Regression Quantiles As Preliminary Estimate)*		
		Least Squares	*Huber Proposal 2*	*Hampel One Step*	$\alpha = .05$	$\alpha = .10$	$\alpha = .25$	$\alpha = .05$	$\alpha = .10$	$\alpha = .25$	$\alpha = .05$	$\alpha = .10$	$\alpha = .25$
Normal		1.00	1.04	1.04	1.30	1.36	1.26	1.54	1.83	2.14	1.03	1.06	1.19
.05	3.0	1.40	1.16	1.17	1.38	1.51	1.58	1.54	1.88	2.26	1.16	1.17	1.29
.05	5.0	2.20	1.20	1.23	1.43	1.71	2.15	1.51	1.87	2.28	1.20	1.20	1.31
.05	10.0	5.95	1.23	1.28	1.68	2.66	4.81	1.46	1.85	2.30	1.25	1.23	1.33
.10	3.0	1.80	1.30	1.32	1.44	1.64	1.88	1.56	1.93	2.39	1.32	1.30	1.39
.10	5.0	3.40	1.40	1.47	1.45	1.96	2.99	1.46	1.90	2.44	1.46	1.38	1.45
.10	10.0	10.90	1.49	1.61	1.48	3.32	8.09	1.34	1.85	2.47	1.65	1.45	1.49
.25	3.0	3.00	1.90	1.94	1.79	1.97	2.74	1.82	2.12	2.87	2.14	1.85	1.80
.25	5.0	7.0	2.46	2.68	2.49	2.09	5.13	2.37	1.92	2.99	4.11	2.39	2.01
.25	10.0	25.75	3.20	4.26	6.50	1.88	15.66	5.51	1.65	3.06	13.65	3.69	2.19

[a] Proportion of contamination.

[b] Standard deviation of contamination.

NOTE: The asymptotic covariance matrix is $\hat{Q}^{-1}$ multiplied by the displayed quantity.

M estimates use Huber's Proposal 2 to obtain scale equivariance. The Huber uses

$$\psi(x) = \min(2, |x|)\,\mathrm{sign}(x) \ ,$$

and the Hampel uses

$$\begin{aligned} \psi(x) &= x && \text{if } 0 \le x \le 1.5 \ , \\ &= 1.5 && \text{if } 1.5 \le x \le 3.5 \ , \\ &= (8 - x)/3 && \text{if } 3.5 \le x \le 8 \ , \\ &= 0 && \text{if } 8 \le x \ , \end{aligned}$$

and $\psi(-\,x) = -\,\psi(x)$. (For discussion of Huber's Proposal 2, see Carroll and Ruppert 1979.) Several conclusions can be drawn from Table 1.

1. $\hat{\beta}_{PE}$(LS) and $\hat{\beta}_{PE}$(LAD) are inefficient at the normal distribution.
2. $\hat{\beta}_{PE}$(RQ) is quite efficient at the normal model.
3. Under heavy contamination (b large or ϵ large) $\hat{\beta}_{PE}$(LS), $\hat{\beta}_{PE}$(LAD), and $\hat{\beta}_{PE}$(RQ) are relatively efficient, compared with LS. Also, $\hat{\beta}_{PE}$(RQ) and $\hat{\beta}_{PE}$(LAD) compare well with the M estimates, but $\hat{\beta}_{PE}$(LS) does poorly compared with the M estimates, if $\epsilon = .25$, $b = 10$, and $\alpha = .25$. (Intuitively, one can expect that when $\alpha = .25$, $\hat{\beta}_{PE}$(LS) will be heavily influenced by its preliminary estimate, which estimates β poorly for these b and ϵ.)
4. For a fixed distribution, the asymptotic variance of $\hat{\beta}_{PE}$(LS) is not necessarily a monotone function of α, $0 < \alpha < \frac{1}{2}$.

Because of Conclusions (1) and (3), the practice of fitting by LS or LAD, removing points corresponding to extreme residuals, and computing the LS estimate from the trimmed sample is not an adequate substitute for robust methods of estimation.

Conclusion (4) seems surprising at first but has an intuitive explanation. If $\alpha = 0$, then $\hat{\beta}_{PE}$ is the LS estimate, and as $\alpha \to \frac{1}{2}$, $\hat{\beta}_{PE}$ converges to the preliminary estimate. Thus $\hat{\beta}_{PE}$(LS) should be virtually equal to the LS estimate for α close to 0 or $\frac{1}{2}$. Consequently, letting $\sigma^2(\alpha, \text{LS}, F)$ equal $\sigma^2(\alpha, g, F)$, where g corresponds to $\hat{\beta}_0 = \text{LS}$, we might expect that $\sigma^2(\alpha, \text{LS}, \Phi)$ will decrease

2. *Finite and Asymptotic Variances of* $n^{1/2}\,\hat{\beta}_A$(LS) *in the Location Model* ($\alpha = .10$)

ϵ[a]	b[b]	$n = 50$ $NI = 1{,}000$	$n = 100$ $NI = 1{,}000$	$n = 200$ $NI = 500$	$n = 300$ $NI = 500$	$n = 400$ $NI = 850$	*Asymptotic*
Normal		1.31	1.36	1.37	1.32	1.35	1.36
.05	3	1.47	1.49	1.50	1.47	1.48	1.51
.05	5	1.57	1.65	1.70	1.66	1.65	1.71
.05	10	2.10	2.36	2.54	2.51	2.40	2.66
.10	3	1.58	1.58	1.65	1.63	1.60	1.64
.10	5	1.74	1.83	1.97	1.90	1.90	1.96
.10	10	2.24	2.51	2.92	2.99	3.03	3.32
.25	3	2.01	1.93	1.94	1.96	1.96	1.97
.25	5	2.12	2.05	2.08	2.11	2.07	2.09
.25	10	2.98	2.42	2.14	2.13	2.11	1.88

[a] Proportion of contamination.

[b] Standard deviation of contamination.

NOTE: n = sample size; NI = no. of Monte Carlo simulations.

as $\alpha \downarrow 0$ or $\alpha \uparrow \frac{1}{2}$. For F (a heavy-tailed distribution), we might expect that $\sigma^2(\alpha, \text{LS}, F)$ will increase as $\alpha \downarrow 0$ and as $\alpha \uparrow \frac{1}{2}$.

If, instead of removing those observations with the $[n\alpha]$ smallest and $[n\alpha]$ largest residuals from $\hat{\boldsymbol{\beta}}_0$, we remove those observations with the $[2n\alpha]$ largest absolute residuals, then the asymptotic variance of the intercept is the same as that of the slopes. Specifically, let $\hat{\boldsymbol{\beta}}_{\text{A}}(\alpha)$ $(= \hat{\boldsymbol{\beta}}_{\text{A}})$ be the estimate formed in this manner. Then if F is symmetric,

$$(1-2\alpha)n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{A}}-\boldsymbol{\beta})$$

$$=n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{Q}^{-1}x_i\{\mathbf{Z}_i\mathbf{I}(\xi_1 \leq \mathbf{Z}_i \leq \xi_2)+a(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta})\}+o_p(1) \ ,$$

and if Assumption 6 holds, then

$$(1-2\alpha)n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{A}}-\boldsymbol{\beta})$$

$$=n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{Q}^{-1}x_i\{\mathbf{Z}_i\mathbf{I}(\xi_1 \leq \mathbf{Z}_i \leq \xi_2)+ag(\mathbf{Z}_i)\}+o_p(1) \ ,$$

which in the location case reduces to

$$(1 - 2\alpha)n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{\text{A}} - \boldsymbol{\beta})$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^{n} \{\mathbf{Z}_i\mathbf{I}(\xi_1 \leq \mathbf{Z}_i \leq \xi_2) + ag(\mathbf{Z}_i)\} + o_p(1) \ .$$

The proofs are similar to those of Theorems 1 and 2 and are omitted.

Since $\hat{\boldsymbol{\beta}}_{\text{A}}$ is particularly easy to compute in the location model, it is very suitable for Monte Carlo studies. It is hoped that such studies will indicate the degree of agreement between the asymptotic and finite sample variances of $\hat{\boldsymbol{\beta}}_{\text{PE}}$ as well as $\hat{\boldsymbol{\beta}}_{\text{A}}$. Table 2 displays the variance of $\hat{\boldsymbol{\beta}}_{\text{A}}(\text{LS})$ (i.e., $\hat{\boldsymbol{\beta}}_{\text{A}}$ with $\hat{\boldsymbol{\beta}}_0$ the LS estimate, for $n = 50$, 100, 200, 300, and 400). Throughout $\alpha = .10$. The Monte Carlo swindle (Gross 1973) was used as a variance reduction technique. One sees from Table 2 that convergence of the variance to its asymptotic value can be extremely slow for some distributions (e.g., $b = 10$ and $\epsilon = .10$ or .25).

8. LARGE-SAMPLE INFERENCE

Here we sketch a large-sample methodology of confidence ellipsoids and hypothesis testing based on $\hat{\boldsymbol{\beta}}_{\text{KB}}$. For symmetric trimming and symmetric F, the theory is applicable to $\hat{\boldsymbol{\beta}}_{\text{PE}}(\text{RQ})$ as well. The asymptotic covariance matrix $\sigma^2(\alpha, F)Q^{-1}$ can be consistently estimated, since $n^{-1}\ \mathbf{X}'\mathbf{X} \rightarrow \mathbf{Q}$ and a consistent estimate of $\sigma^2(\alpha, F)$ is provided by Theorem 5.

Theorem 5: Let S be the sum of squares for residuals calculated from the trimmed sample, that is,

$$S = y'\mathbf{B}(\mathbf{I}_p - \mathbf{X}(\mathbf{X}'\mathbf{B}\mathbf{X})^{-}\mathbf{X}')\mathbf{B}y \ .$$

Let $c_j = \mathbf{e}'[\hat{\boldsymbol{\beta}}(\alpha_j) - \hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha)]$ for $j = 1$, 2, and

$$s^2(\alpha, F) = (\alpha_2 - \alpha_1)^{-2}((n - p)^{-1}\ S$$
$$+ \alpha_1 c_1{}^2 + (1 - \alpha_2)c_2{}^2 - (\alpha_1 c_1 + (1 - \alpha_2)c_2)^2) \ .$$

Then

$$s^2(\alpha, \text{F}) \xrightarrow{p} \sigma^2(\alpha, F) \ .$$

Theorem 6: Suppose m is the number of observations that have been removed by trimming. For $0 < \epsilon < 1$, let $F(n_1, n_2, \epsilon)$ denote the $(1 - \epsilon)$ quantile of the F distribution, with n_1 and n_2 degrees of freedom, and let

$$d(n_1, n_2, \epsilon) = (\alpha_2 - \alpha_1)^{-1}\ S^2(\alpha, F)n_1F(n_1, n_2, \epsilon) \ .$$

Suppose for some integer ℓ, $\mathbf{K}$ and $\mathbf{c}$ are matrices of sizes $\ell \times p$ and $\ell \times 1$, respectively, and that $\mathbf{K}$ has rank ℓ. If $\mathbf{K}'(\boldsymbol{\beta} + \boldsymbol{\delta}(\alpha)) = \mathbf{c}$, then

$$\lim_{n\to\infty} P\{(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) - \mathbf{c})'[\mathbf{K}'(\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{K}]^{-1}$$
$$\cdot(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) - c) \geq d(\ell, n - m - p, \epsilon)\} = \epsilon \ .$$

Letting $\mathbf{K} = \mathbf{I}_p$ and $\mathbf{c} = \boldsymbol{\beta} - \boldsymbol{\delta}(\alpha)$, the confidence ellipsoid

$$(\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) - \boldsymbol{\beta} - \boldsymbol{\delta}(\alpha))'(\mathbf{X}'\mathbf{A}\mathbf{X})(\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) - \boldsymbol{\beta} - \boldsymbol{\delta}(\alpha))$$
$$\leq d(\ell, n - m - p, \epsilon) \quad (8.1)$$

for $\boldsymbol{\beta} + \boldsymbol{\delta}(\alpha)$ has an asymptotic confidence coefficient of $(1 - \epsilon)$. Moreover, if we test

$$\text{H}_0 \colon \mathbf{K}'(\boldsymbol{\beta} + \boldsymbol{\delta}(\alpha)) = \mathbf{c}$$

versus

$$\text{H}_1 \colon \mathbf{K}'(\boldsymbol{\beta} + \boldsymbol{\delta}(\alpha)) = \mathbf{c}$$

by rejecting H_0 whenever

$$(\mathbf{K}'\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) - \mathbf{c})'[\mathbf{K}'(\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'(\hat{\boldsymbol{\beta}}_{\text{KB}}(\alpha) - \mathbf{c})$$
$$\geq d(\ell, n - m - p, \epsilon), \quad (8.2)$$

then the asymptotic size of our test is ϵ.

Of course, in the special cases in which $\alpha_1 = 0$, $\alpha_2 = 1$ (so $m = 0$ and $\mathbf{A} = \mathbf{I}$), and F is Gaussian, (8.1) is an exact $1 - \epsilon$ confidence ellipsoid and (8.2) is an exact size ϵ test.

9. EXAMPLES

In this section we contrast the results obtained for different estimates when applied to two data sets: (a) the *stackloss* data set given by Brownlee (1965) and further analyzed using M estimates by Andrews (1974) and (b) a set of measurements of *water salinity* and river discharge taken in North Carolina's Pamlico Sound (see Table 3). In the first case, stackloss was regressed against air flow, temperature, and acid; salinity was regressed against salinity lagged two weeks, river discharge, and a linear time trend using the second data set. The estimates we consider are listed in Table 4. Both Huber and Andrews are M estimates and are calculated by the iterative solution to

$$\sum_{i=1}^{N} \psi((y_i - x'_i\boldsymbol{\beta})/s)x_i = 0 \ ,$$

where $s = \text{MAD}/C$, C is a constant, and MAD is the median of the absolute values of the residuals. For Huber, $C = .6745$ and $\psi(\text{z}) = \max(-1.25, \min(\text{z}, 1.25))$. This

3. The Water Salinity Data Set

OBS	SALINITY	SALLAG	TREND	H2OFLOW	YEAR
1	7.6	8.2	4	23.005	72
2	7.7	7.6	5	23.873	
3	4.3	4.6	0	26.417	73
4	5.9	4.3	1	24.868	
5	5.0	5.9	2	29.895	
6	6.5	5.0	3	24.200	
7	8.3	6.5	4	23.215	
8	8.2	8.3	5	21.862	
9	13.2	10.1	0	22.274	74
10	12.6	13.2	1	23.830	
11	10.4	12.6	2	25.144	
12	10.8	10.4	3	22.430	
13	13.1	10.8	4	21.785	
14	12.3	13.1	5	22.380	
15	10.4	13.3	0	23.927	75
16	10.5	10.4	1	33.443	
17	7.7	10.5	2	24.859	
18	9.5	7.7	3	22.686	
19	12.0	10.0	0	21.789	76
20	12.6	12.0	1	22.041	
21	13.6	12.1	4	21.033	
22	14.1	13.6	5	21.005	
23	13.5	15.0	0	25.865	77
24	11.5	13.5	1	26.290	
25	12.0	11.5	2	22.932	
26	13.0	12.0	3	21.313	
27	14.1	13.0	4	20.769	
28	15.1	14.1	5	21.393	

NOTE: The values are biweekly averages of SALINITY at time period i. SALLAG = salinity lagged 2 weeks, i = TREND = one of the six biweekly periods in March–May, and H2OFLOW = river discharge in time i. (Since only spring data are used, SALLAG is not always the previous value of SALINITY.)

choice of ψ should give results for normal data similar to those for the regression analogues of a 10% trimmed mean. The Andrews estimate uses $C = 1$ and $\psi(z) = \text{sine}\,(Z)I(|Z| \leq \pi)$.

We defined $\hat{\beta}_{KB}$ a bit differently than in Section 2. Both data sets have four independent variables, and each regression quantile hyperplane passes through four observations. Therefore, if one defines $\hat{\beta}_{KB}$ as in Section 2, at least eight observations are trimmed. Instead, we defined $\hat{\beta}_{KB}$ by requiring strict inequality in (2.2). If $\alpha = (.1, .9)$, this leads to no trimming for the stackloss data and only two observations trimmed for the salinity data, so we use $\alpha = (.15, .85)$. Then observations 4, 9, and 21 are trimmed in the stackloss data, and observations 1, 13, 15, and 17 are trimmed in the salinity data.

An important advantage of $\hat{\beta}_{PE}$(RQ) over $\hat{\beta}_{KB}$ is that residuals from a preliminary estimate are rarely tied (at least in these data sets), and with $\hat{\beta}_{PE}$(RQ) one can have the actual percentage of trimming close to any specified α. The observations deleted when calculating $\hat{\beta}_{PE}$(RQ, .10) are 1, 3, 9, and 21 for the stackloss data and 1, 11, 13, 15, 16, and 17 for the salinity data.

Since both data sets have outliers, asymptotic theory and Monte Carlo studies for the location problem (Andrews et al. 1972) lead us to expect that LS estimates will be worst, that Andrews will do very well, and that Huber, $\hat{\beta}_{PE}$(RQ), and $\hat{\beta}_{KB}$ will have roughly comparable performances. Of course, with these data the true parameters are unknown, and we can only measure performance by closeness of fit to the bulk of the observations, for example, with MAD or interquartile range of the residuals (IQR). Using either MAD or IQR as criteria, our study does seem to agree with our expectations. The redescending M estimator (Andrews) appears to be best overall.

Also, we have included $\hat{\beta}(.5)$, the LAD estimate. Its performance was quite good here, but of course it is known to have poor efficiency at the normal model.

In Table 4, we list the regression coefficients, MAD, and IQR for each estimator. The figure shows box plots of the residuals and was obtained from the SAS package.

LS computations were performed on SAS. Regression quantiles were computed using MPS/360, a linear programming package, and LPMPS, a preprocessor for MPS/360 (McKeown and Rubin 1977).

10. SUMMARY

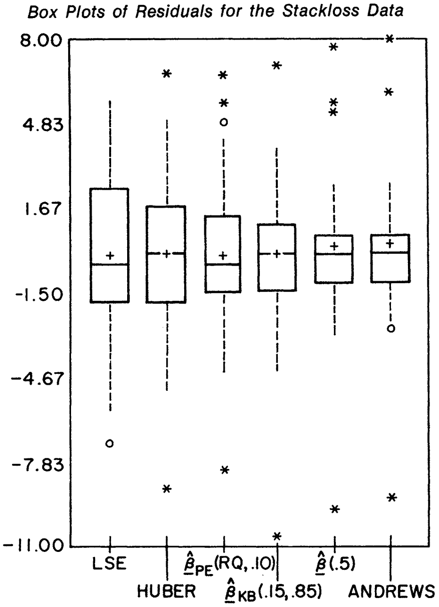
We have considered two methods of defining a trimmed LS estimator: $\hat{\beta}_{KB}$, which uses Koenker and Bas-

4. Regression Coefficients, MAD, and IQR for the Stackloss Data

Code	Intercept	Air Flow	Temperature	Acid	MAD	IQR
LSE	39.92	−.72	−1.30	.15	1.92	3.12
$\hat{\beta}(.50)$	39.69	−.83	−.57	.06	1.18	1.71
$\hat{\beta}_{KB}(.15)$	42.83	−.93	−.63	.10	1.60	2.49
$\hat{\beta}_{PE}$(RQ,.10)	40.37	−.72	−.96	.07	1.37	2.59
Huber	41.00	−.83	−.91	.13	1.63	3.07
Andrews	37.20	−.82	−.52	.07	.99	1.50

Water Salinity Data

Code	Intercept	SALLAG	TREND	H2OFLOW	MAD	IQR
LSE	9.59	.777	−.026	−.295	.72	1.38
$\hat{\beta}(.50)$	14.21	.740	−.111	−.458	.50	.98
$\hat{\beta}_{KB}(.15)$	9.69	.800	−.128	−.290	.67	1.36
$\hat{\beta}_{PE}$(RQ,.10)	14.49	.774	−.160	−.488	.60	1.05
Huber	13.36	.756	−.094	−.439	.56	1.02
Andrews	17.22	.733	−.196	−.578	.47	.83



sett's (1978) regression quantiles, and $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$, which uses a preliminary estimate.

Despite its intuitive appeal, $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$(LS) may not be very satisfactory when based on an arbitrary preliminary estimate. Its behavior will depend heavily on the choice of the preliminary estimate. Some choices (e.g., median regression) result in very inefficient trimmed estimates at the normal distribution, even if the trimming proportion is small. Other choices (e.g., LS) can lead to low efficiency for heavy-tailed distributions, especially if the trimming proportion is high. Moreover, the contribution of the preliminary estimate to the variance of $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ depends on the density of the error distribution and might be difficult to estimate in practical situations.

The estimate $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ behaves analogously to a trimmed mean. Also, for a particular choice of preliminary estimate, namely, the average of two regression quantiles, $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$ (which for this preliminary estimate we call $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$(RQ)), is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$, provided the error distribution is symmetric.

For moderately sized data sets, $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$(RQ) has one major advantage over $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$; with $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$(RQ) the proportion of observations rejected can be made quite close to any specified α. Since the number of observations lying on a regression quantile hyperplane is typically equal to the number of independent variables, $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ does not share this property with $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$(RQ).

The trimmed estimates $\hat{\boldsymbol{\beta}}_{\mathrm{KB}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PE}}$(RQ) seem to be worthwhile alternatives to M estimates based on Huber's ψ, but they are not necessarily adequate substitutes for redescending M estimates.

APPENDIX

Lemma A.1: With probability one there exists no vector $\mathbf{b}$ and $p+1$ rows of $\mathbf{X}$, $x_{i(1)}, \ldots, x_{i(p+1)}$, such that $y_i = x'_{i(j)}\mathbf{b}$ for $j = 1, \ldots, p+1$.

Proof: Routine; use the continuity of F.

Lemma A.2: Let $r_1, \ldots, r_n$ be the residuals from $\hat{\boldsymbol{\beta}}_0$, suppose $0 < \theta < 1$, and let μ_n be a sequence of solutions to

$$\sum_{i=1}^{n} \rho_\theta(r_i - \mu_n) = \min \ .$$

Then

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_\theta(r_i - \mu_n) \rightarrow 0 \text{ almost surely} \ . \qquad \text{(A.1)}$$

In addition, the sequence of solutions $\hat{\boldsymbol{\beta}}(\theta)$ of (1.5) satisfies

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i\psi_\theta(y_i - x'_i\hat{\boldsymbol{\beta}}(\theta)) \rightarrow \mathbf{0} \text{ almost surely} \ . \qquad \text{(A.2)}$$

Proof: We prove only (A.2) because (A.1) can be demonstrated in a similar manner.

Let $\{\mathbf{e}_j\}_{j=1}^{p}$ be the standard basis of R^p. Define

$$G_j(a) = \sum_{i=1}^{n} \rho_\theta(y_i - x'_i(\hat{\boldsymbol{\beta}}(\theta) + \mathrm{a}\mathbf{e}_j)) \ ,$$

and let $H_j(t)$ be the derivative from the right of G_j, so that

$$H_j(a) = \sum_{i=1}^{n} x_{ij}\psi_\theta(y_i - x'_i(\hat{\boldsymbol{\beta}}(\theta) + a\mathbf{e}_j)) \ .$$

Notice that $H_j(a)$ is nondecreasing. Therefore, for $\epsilon > 0$

$$H_j(-\epsilon) \leq H_j(0) \leq H_j(\epsilon) \ ,$$

and because $G_j(a)$ achieves its minimum at $a = 0$,

$$H_j(-\epsilon) \leq 0 \quad \text{and} \quad H_j(\epsilon) \geq 0 \ .$$

Consequently,

$$|H_j(0)| \leq H_j(\epsilon) - H_j(-\epsilon) \ . \qquad \text{(A.3)}$$

Letting $\epsilon \rightarrow 0$ in (A.3), we see that

$$|H_j(0)| \leq \sum_{i=1}^{n} |x_{ij}|\mathbf{I}(y_i - x'_i\hat{\boldsymbol{\beta}}(\theta) = 0) \ .$$

Now (A.2) follows from Lemma A.1.

Lemma A.3: For $\boldsymbol{\Delta} \in R^p$, define

$$\mathbf{M}(\boldsymbol{\Delta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i\psi_\theta(\mathbf{Z}_i - x_i\boldsymbol{\Delta}n^{-\frac{1}{2}} - \eta(\theta)) \ .$$

Then for all $L > 0$

$$\sup_{0 \leq ||\boldsymbol{\Delta}|| \leq L} \|\mathbf{M}(\boldsymbol{\Delta}) - \mathbf{M}(0) + f(\eta(\theta))\mathbf{Q}\boldsymbol{\Delta}\| = o_p(1) \ . \qquad \text{(A.4)}$$

Proof: The result follows from Lemma 4.1 of Bickel (1975) because

$$E(\mathbf{M}(\boldsymbol{\Delta}) - \mathbf{M}(\mathbf{0})) \to -f(\eta(\theta))\mathbf{e}'\boldsymbol{\Delta} \ .$$

Remark: Equation (A.4) is a special case of the conclusion of Jurecková's (1977) Theorem 4.1, which she proves under conditions different from ours. Her C_{ji} is our $x_{ij}n^{-\frac{1}{2}}$.

Proof of Lemma 1: Since $\mu = r_{\theta n}$ is a solution to

$$\sum_{i=1}^{n} \rho_\theta(r_i - \mu) = \min! \ ,$$

(A.1) implies that

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_\theta(\mathbf{Z}_i - \eta(\theta) - x'_i((\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) + \mathbf{e}(r_{\theta n} - \eta(\theta)))) \to 0 \text{ almost surely} \ . \quad \text{(A.5)}$$

Define $V(\boldsymbol{\Delta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_\theta(\mathbf{Z}_i - x'_i\boldsymbol{\Delta}n^{-\frac{1}{2}} - \eta(\theta))$. Using the method of Jurecková (1977, proof of Lemma 5.2) and (A.4), we can show that for all $\epsilon > 0$ there exists η, K, and n_0 such that

$$\Pr(\inf_{|e'\Delta|>K} |V(\boldsymbol{\Delta})| < \eta) < \epsilon \quad \text{for} \quad n \geq n_0 \ . \quad \text{(A.6)}$$

Next, (A.5) and (A.6) allows us to conclude that

$$n^{\frac{1}{2}}(\mathbf{e}'(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) + r_{\theta n} - \eta(\theta)) = O_p(1) \ . \quad \text{(A.7)}$$

By (A.7) and Assumption 5 we may substitute $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + \mathbf{e}(r_{\theta n} - \eta(\theta)))$ for $\boldsymbol{\Delta}$ in (A.4) and complete the proof by examining first coordinates, using Assumption 2.

Lemma A.4: Let $\mathbf{D}_{in}(= \mathbf{D}_i)$ be a $r \times c$ matrix. Suppose

$$\sup_n \, (n^{-1} \sum_{i=1}^{n} \|\mathbf{D}_i\|^2) < \infty \ ,$$

where $\|\mathbf{D}_i\|^2 = \mathrm{Tr}\mathbf{D}'_i\mathbf{D}_i$ is the Euclidean norm of $\mathbf{D}_i$. Let $\mathbf{I}$ be an open interval containing ξ_1 and ξ_2 and let the function $g(x)$ be defined for all x and Lipschitz continuous on $\mathbf{I}$. For $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, and $\boldsymbol{\Delta}_3$ in R^p and $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$, define

$$\mathbf{T}(\boldsymbol{\Delta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{D}_i g(\mathbf{Z}_i + \boldsymbol{\Delta}'_3 x_i n^{-\frac{1}{2}}) \cdot \mathbf{I}\{\xi_1 + x'_i\boldsymbol{\Delta}_1 n^{-\frac{1}{2}} < \mathbf{Z}_i < \xi_2 + x'_i\boldsymbol{\Delta}_2 n^{-\frac{1}{2}}\} \ .$$

Then, for all $M > 0$,

$$\sup_{\|\Delta\| \leq M} |\mathbf{T}(\boldsymbol{\Delta}) - \mathbf{T}(\mathbf{0}) - E(\mathbf{T}(\boldsymbol{\Delta}) - \mathbf{T}(\mathbf{0}))| = o_p(1) \ .$$

Proof: The proof is very similar to that of Bickel's (1975) Lemma 4.1 and is omitted here, but it can be found in Ruppert and Carroll (1978).

Proof of Theorem 1: For $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ in R^p and $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2)$, define

$$\mathbf{U}(\boldsymbol{\Delta}) = n^{-1} \sum_{i=1}^{n} x_i x'_i \mathbf{I}(\xi_1 + x'_i\boldsymbol{\Delta}_1 n^{-\frac{1}{2}} \leq \mathbf{Z}_i \leq \xi_2 + x'_i\boldsymbol{\Delta}_2 n^{-\frac{1}{2}}\}$$

and

$$\mathbf{W}(\boldsymbol{\Delta}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i z_i \mathbf{I}\{\xi_1 + x'_i\boldsymbol{\Delta}_1 n^{-\frac{1}{2}} \leq \mathbf{Z}_i \leq \xi_2 + x'_i\boldsymbol{\Delta}_2 n^{-\frac{1}{2}}\} \ .$$

Using Lemma A.4, it is easy to show that for all $M > 0$,

$$\sup_{0 \leq \|\boldsymbol{\Delta}\| \leq M} |\mathbf{U}(\boldsymbol{\Delta}) - (1 - 2\alpha)\mathbf{Q}| = o_p(1) \quad \text{(A.8)}$$

and

$$\sup_{0 \leq \|\boldsymbol{\Delta}\| \leq M} |\mathbf{W}(\boldsymbol{\Delta}) - \mathbf{W}(\mathbf{0}) - \mathbf{Q}(\boldsymbol{\Delta}_2\xi_2 f(\xi_2) - \boldsymbol{\Delta}_1\xi_1 f(\xi_1))| = o_p(1) \ . \quad \text{(A.9)}$$

Then, using the fact that $x'_i\mathbf{e} = 1$, we have

$$I\{r_{1n} \leq r_i \leq r_{2n}\} = I\{\xi_1 + x'_i((\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) + \mathbf{e}(r_{1n} - \xi_1)) \leq \mathbf{Z}_i \leq \xi_2 + x'_i((\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) + \mathbf{e}(r_{2n} - \xi_2))\} \ ,$$

and so replacing $\boldsymbol{\Delta}_\ell$ by $n^{\frac{1}{2}}((\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}) + \mathbf{e}(r_{\ell n} - \xi_\ell))$ for $\ell = 1, 2$ in (A.8) and (A.9), we have

$$n^{-1}(\mathbf{X}'\mathbf{A}\mathbf{X}) = (1 \times 2\alpha)Q + o_p(1) \quad \text{(A.10)}$$

and

$$n^{-\frac{1}{2}}\mathbf{X}'\mathbf{A}(y - \mathbf{A}\mathbf{X}\boldsymbol{\beta}) = W(\mathbf{0}) + \mathbf{Q}\{\xi_2 f(\xi_2) n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + \mathbf{e}(r_{2n} - \xi_2)) - \xi_1 f(\xi_1) n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta} + \mathbf{e}(r_{1n} - \xi_1))\} + o_p(1) \ . \quad \text{(A.11)}$$

By (A.10),

$$n^{\frac{1}{2}}(\mathbf{X}'\mathbf{A}(y - \mathbf{A}\mathbf{X}\boldsymbol{\beta})) = (1 - 2\alpha)n^{\frac{1}{2}}\mathbf{Q}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}} - \boldsymbol{\beta}) + o_p(1) \ . \quad \text{(A.12)}$$

By (A.11), (A.12), and Lemma 1,

$$(1 - 2\alpha)n^{\frac{1}{2}}\mathbf{Q}(\hat{\boldsymbol{\beta}}_{\mathrm{PE}} - \boldsymbol{\beta}) = \mathbf{W}(\mathbf{0}) + \mathbf{Q}\{\xi_2\mathbf{e}n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_{1-\alpha}(\mathbf{Z}_i - \xi_2) - \xi_1\mathbf{e}n^{-\frac{1}{2}} \sum_{i=1}^{n} \psi_\alpha(\mathbf{Z}_i - \xi_1) + n^{\frac{1}{2}}a(\mathbf{I} - \mathbf{e}\mathbf{e}')(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta})\} + o_p(1) \ .$$

Then (3.3) follows from the definition of $W(\mathbf{0})$.

Proof of Theorem 4: Using (A.4) and the method of Jurecková (1977, proof of Lemma 5.2), we can show that

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\boldsymbol{\theta})) = O_p(1) \ .$$

Therefore, we can substitute $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{\beta}(\boldsymbol{\theta}))$ for $\boldsymbol{\Delta}$ in (A.4) and use (A.2) to obtain

$$\boldsymbol{\Delta}(0) = f(\eta(\theta))\mathbf{Q}n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}(\theta)) + o_p(1) \ ,$$

and Theorem 4 follows easily.

Proof of Theorem 3: The proof is quite similar to the proof of Theorem 1 and can be found in Ruppert and Carroll (1978).

Proof of Theorem 5: For $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$, $\boldsymbol{\Delta}_3$ in R^p, define $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_3)$ and

$$V(\boldsymbol{\Delta}) = n^{-1} \sum_{i=1}^{n} (\mathbf{Z}_i - x'_i\boldsymbol{\Delta}_1 n^{-\frac{1}{2}} - \delta(\alpha))^2 \cdot \mathbf{I}(\xi_1 + x'_i\boldsymbol{\Delta}_2 n^{-\frac{1}{2}} \leq \mathbf{Z}_i \leq \xi_2 + x'_i\boldsymbol{\Delta}_3 n^{-\frac{1}{2}}) \ .$$

We see that

$$S = nV(\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{KB}}(\alpha) - (\boldsymbol{\beta} + \delta(\alpha)),\ \sqrt{n}(\hat{\boldsymbol{\beta}}(\alpha_1) - \boldsymbol{\beta}(\alpha_1)),\ \sqrt{n}(\hat{\boldsymbol{\beta}}(\alpha_2) - \boldsymbol{\beta}(\alpha_2)))\ .$$

Applying Lemma A.4 with $g(x) = x^2$ and $D_i = 1$, we have for $M > 0$

$$\sup_{|\boldsymbol{\Delta}| \le M} |V(\boldsymbol{\Delta}) - V(\mathbf{0}) - E(V(\boldsymbol{\Delta}) - V(\mathbf{0}))| = o_p(1)\ .$$

By a Taylor expansion of F and additional simple calculations,

$$E(V(\boldsymbol{\Delta}) - V(\mathbf{0})) \rightarrow 0\ ,$$

whence

$$\sup_{|\boldsymbol{\Delta}| \le M} |V(\boldsymbol{\Delta}) - V(\mathbf{0})| = o_p(1)\ .$$

Therefore, by Corollary 2 and (5.2), we have

$$S = V(0) + o_p(1)\ .$$

Now var $V(0) \rightarrow 0$, so by Chebyshev's inequality,

$$\begin{aligned} S &= EV(\mathbf{0}) + o_p(1) \\ &= E(\mathbf{Z}_i - \boldsymbol{\delta}(\alpha))^2 I(\xi_1 \le \mathbf{Z}_i \le \xi_2) + o_p(1)\ . \end{aligned}$$

Furthermore, for $j = 1, 2$,

$$c_j = \xi_j - \boldsymbol{\delta}(\alpha) + o_p(1)$$

by Corollary 2 and (5.2), and Theorem 5 follows.

Proof of Theorem 6: This follows in a straightforward manner from (5.2), Theorem 5, and Theorem 4.4 of Billingsley (1968).

[*Received July 1978. Revised May 1980.*]

REFERENCES

Andrews, D.F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523–531.

——— et al. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press.

Bahadur, R.R. (1966), "A Note on Quantiles in Large Samples," *Annals of Mathematical Statistics*, 37, 577–580.

Bickel, Peter J. (1973), "On Some Analogues to Linear Combinations of Order Statistics in the Linear Model," *Annals of Statistics*, 1, 597–616.

——— (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 70, 428–433.

Billingsley, Patrick (1968), *Convergence of Probability Measures*, New York: John Wiley & Sons.

Brownlee, K.A. (1965), *Statistical Theory and Methodology in Science and Engineering* (2nd ed.), New York: John Wiley & Sons.

Carroll, Raymond J. (1979), "On Estimating Variances of Robust Estimators When the Errors Are Asymmetric," *Journal of the American Statistical Association*, 74, 674–679.

Carroll, Raymond J., and Ruppert, David (1979), "Almost Sure Properties of Robust Regression Estimates," *Institute of Statistics Mimeo Series*, No. 1240, University of North Carolina at Chapel Hill, Dept. of Statistics.

deWet, T., and Venter, J.H. (1974), "An Asymptotic Representation of Trimmed Means With Applications," *South African Statistics Journal*, 8, 127–134.

Ghosh, J.K. (1971), "A New Proof of the Bahadur Representation of Quantiles and an Application," *Annals of Mathematical Statistics*, 42, 1957–1961.

Gross, Alan M. (1973), "A Monte-Carlo Swindle for Estimators of Location," *Applied Statistics*, 22, 347–356.

——— (1976), "Confidence Interval Robustness With Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association*, 71, 409–416.

Hogg, Robert V. (1974), "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Application and Theory," *Journal of the American Statistical Association*, 69, 909–927.

Huber, Peter J. (1970), "Studentizing Robust Estimates," in *Nonparametric Techniques in Statistical Inference*, ed. M.L. Puri, Cambridge, England: Cambridge University Press, 453–463.

——— (1977), *Robust Statistical Procedures*, Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Jaeckel, Louis A. (1971), "Robust Estimates of Location: Symmetry and Asymmetric Contamination," *Annals of Mathematical Statistics*, 42, 1020–1034.

Jurecková, Jana (1977), "Asymptotic Relations of *M*-Estimates and *R*-Estimates in Linear Regression Model," *Annals of Statistics*, 5, 464–472.

Koenker, Roger, and Bassett, Gilbert, Jr. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

McKeown, P.G., and Rubin, D.S. (1977), "A Student Oriented Preprocessor for MPS/360," *Computers and Operations Research*, 4, 227–229.

Ruppert, David, and Carroll, Raymond J. (1978), "Robust Regression by Trimmed Least Squares Estimation, *Institute of Statistics Mimeo Series*, No. 1186, University of North Carolina at Chapel Hill, Dept. of Statistics.

Stigler, Stephen M. (1973), "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885–1920," *Journal of the American Statistical Association*, 68, 872–879.

——— (1977), "Do Robust Estimators Work With Real Data?" *Annals of Statistics*, 5, 1055–1098.

The Annals of Statistics
1982, Vol. 10, No. 2, 429–441

ROBUST ESTIMATION IN HETEROSCEDASTIC LINEAR MODELS

BY RAYMOND J. CARROLL[1] AND DAVID RUPPERT[2]

We consider a heteroscedastic linear model in which the variances are given by a parametric function of the mean responses and a parameter θ. We propose robust estimates for the regression parameter β and show that, as long as a reasonable starting estimate of θ is available, our estimates of β are asymptotically equivalent to the natural estimate obtained with known variances. A particular method for estimating θ is proposed and shown by Monte-Carlo to work quite well, especially in power and exponential models for the variances. We also briefly discuss a "feedback" estimate of β.

1. Introduction. We consider the heteroscedastic linear model

$$Y_i = \tau_i + \sigma_i \varepsilon_i, \qquad \tau_i = x_i \beta, \qquad i = 1, \cdots, N, \tag{1.1}$$

where $\{x_i\}$ and $1 \times p$ design constants, β is a $p \times 1$ regression parameter, $\{\varepsilon_i\}$ are independent and identically distributed with mean zero and unknown symmetric distribution function F, and $\{\sigma_i\}$ are scaling constants which express the possible heteroscedasticity. Our primary interest is in inference about the unknown regression parameter β.

Of course, one could ignore the $\{\sigma_i\}$ and use classical methods such as least squares or M-estimation (Huber, 1981), but such estimates are not efficient. In order to make more efficient inference about β, it is necessary to get information about the $\{\sigma_i\}$. In one approach to the problem, the $\{\sigma_i\}$ are assumed completely unknown, but replication is assumed feasible so that the $\{Y_i\}$ occur in groups of equal variance. Recent results in this direction are due to Fuller and Rao (1978). Their results are complicated, and the delicate calculations involved seem to depend very heavily on an assumption of Gaussian errors, which is undesirable from the viewpoint of efficiency robustness; see Huber (1981) for details and further references.

The second approach to the estimation problem for (1.1) avoids the replication assumption by positing a known form for the error variance, i.e.,

$$\sigma_i = H(x_i, \beta, \theta), \tag{1.2}$$

where θ is an $r \times 1$ vector of unknown coefficients and H is smooth and known. A model such as (1.2) is behind the tests for homoscedasticity developed by Anscombe (1961), Bickel (1978), and Carroll and Ruppert (1981). Of course, in many real problems we suspect a heteroscedastic model because the dispersion of the residuals increases with the magnitude of the fitted values. Thus, it has become quite common to simplify (1.2) by assuming σ_i is a function of τ_i or $|\tau_i|$, e.g.,

$$\begin{aligned} &\sigma_i = \sigma(1 + |\tau_i|)^\lambda; \qquad \sigma_i = \sigma|\tau_i|^\lambda \quad \text{(Box and Hill, 1974);} \\ &\sigma_i = \sigma \exp(\lambda \tau_i) \quad \text{(Bickel, 1978);} \\ &\sigma_i = \sigma(1 + \lambda \tau_i^2)^{1/2} \quad \text{(Jobson and Fuller, 1980).} \end{aligned} \tag{1.3}$$

(See also Dent and Hildreth, 1977.) Following these examples, we will thus assume that for

Received August 1979; revised October 1981.

[1] University of North Carolina and the National Heart, Lung and Blood Institute. Supported by the Air Force Office of Scientific Research under grant AFOSR-80-0080.

[2] University of North Carolina. Supported by National Science Foundation grant MCS78-01240.

AMS 1970 *subject classifications.* Primary 62J05, secondary 62G35.

Key words and phrases. Feedback, M-estimates, non-constant variances, random coefficient models, weighted least squares.

some known H,

$$\sigma_i = \sigma H_*(\tau_i, \lambda) = H(\tau_i, \theta) \quad \text{with} \quad \theta = (\sigma, \lambda). \tag{1.4}$$

Our results can be generalized to the model (1.2), but the statements of results and assumptions then become extremely complicated.

Box and Hill (1974) and Jobson and Fuller (1980) both suggest a form of generalized weighted least squares. One obtains estimates of (θ, β), constructs estimated weights $\hat{\sigma}_i$, and then performs ordinary weighted least squares. Their methods are constructed from a normal error assumption, and their efficiency depends on this assumption. The maximum likelihood estimates for θ under the normality assumption have a quadratic influence curve and may be particularly non-robust. As argued above, the recent literature demonstrates some acceptance to the notion that estimators should be robust against departures from normality. One purpose of this article is to provide a set of such robust estimates.

Implicit in the work of Box and Hill (1974) and Jobson and Fuller (1980) is the notion that this problem is adaptable, i.e., the generalized weighted least squares methods are asymptotically equivalent to the "optimal" weighted least squares estimate for the true $\{\sigma_i\}$. Our second major aim is to show that there is a wide class of robust estimates of β which are adaptable for many distribution functions F and models (1.4).

2. A class of weighted robust estimates. Suppose we have estimates of (θ, β) which are $N^{1/2}$-consistent, i.e.,

$$N^{1/2}(\hat{\theta} - \theta) = O_p(1), \qquad N^{1/2}(\hat{\beta}_0 - \beta) = O_p(1). \tag{2.1}$$

The existence of such estimates is discussed in the next section. We then form the estimated σ_i as follows,

$$\hat{\sigma}_i = H(t_i, \hat{\theta}), \qquad t_i = x_i\hat{\beta}_0. \tag{2.2}$$

If the $\{\sigma_i\}$ were known, robustness considerations discussed by Huber (1973, 1981) suggest a general class of weighted M-estimates formed by solving the minimization problem in β;

$$\Sigma\rho\left\{\frac{Y_i - x_i\beta}{\sigma_i}\right\} = \text{minimum}. \tag{2.3}$$

Here ρ is taken to be a convex function. If, for example, $\rho(x) = x^2/2$, we get the "optimal" weighted least squares estimate with known weights. In general, the unknown solution to (2.3) is denoted $\hat{\beta}_{\text{opt}}$.

The class of estimates we suggest are very simply generated by substituting $\{\hat{\sigma}_i\}$ into (2.3). Taking derivatives, we suggest solving the equation

$$\textstyle\sum_{i=1}^N \left(\frac{x_i'}{\hat{\sigma}_i}\right)\psi\left\{\frac{Y_i - x_i\beta}{\hat{\sigma}_i}\right\} = 0, \tag{2.4}$$

with solution denoted by $\hat{\beta}$. Throughout we take ψ to be an odd, continuous function. The non-robust generalized weighted least squares estimates suggested by Box and Hill (1974) and Jobson and Fuller (1980) fall under the special case of (2.4) when $\psi(x) = x$; both propose possibilities for $\hat{\beta}_0$ and $\hat{\theta}$ of (2.1). As suggested by the literature, choosing a bounded ψ can result in reasonably efficient and robust estimates of β.

Define $d_i = x_i/\sigma_i$ and assume that for a positive definite matrix S,

$$S_N = N^{-1}\textstyle\sum_{i=1}^N d_i' d_i \to S. \tag{2.5}$$

Then by formal Taylor series arguments, the optimal robust weighted estimate $\hat{\beta}_{\text{opt}}$, which solves (2.3), satisfies

$$N^{1/2}(\hat{\beta}_{\text{opt}} - \beta) = N^{-1/2}\textstyle\sum_{i=1}^N S^{-1} d_i' \frac{\psi(\varepsilon_i)}{E\psi'(\varepsilon_1)} + o_p(1) \to_{\mathscr{L}} N(0, E\psi^2 S^{-1}(E\psi')^{-2}). \tag{2.6}$$

Our main result concerning adaptation is that when (2.1) holds, and hence we have a reasonable estimate of θ, then our estimate $\hat{\beta}$ is asymptotically equivalent to $\hat{\beta}_{\text{opt}}$. In stating assumptions and proofs, we simplify (1.4) to

$$\sigma_i = \exp\{h(\tau_i)\theta\}, \tag{2.7}$$

where h is a function from R to R^r. The model (2.7) includes the first three models in (1.3), but it is not strictly necessary for the validity of our results. Our reason for considering only (2.7) in the formal aspects of this section is to avoid making already cumbersome notation needlessly complicated. Generalizations to the model (1.4) required that $H(\cdot, \cdot)$ be smooth. Formally, we have the following.

THEOREM 1. *Assume* (2.1), (2.5), (2.7), *the smoothness conditions* B6 *through* B8 *listed in Section* 7, *and*

B1. *ψ monotone and odd, F symmetric,* $0 < E\psi^2(\varepsilon_1) < \infty$, $E\psi' > 0$.

B2. $\lim_{N\to\infty}\sup_{i\le N}(\| x_i \| + \| h(\tau_i)\|)N^{-1/2} = 0$.

B3. $\sup_N\{N^{-1}\sum_{i=1}^N (\| x_i \|^2 + \| h(\tau_i)\|^2)\} < \infty$.

B4. *The σ_i are bounded away from zero.*

B5. *On an open interval I (possibly infinite) containing all the $\{\tau_i\}$, h is Lipschitz continuous.*

Then

$$N^{1/2}(\hat{\beta} - \hat{\beta}_{\text{opt}}) \to_p 0. \tag{2.8}$$

That $\hat{\beta}$ is robust against outliers in the errors when ψ is bounded can be seen by combining (2.6) and (2.8). The resulting influence curve is strikingly similar to the unweighted case in homoscedastic models.

The proof is given in Section 7. Conditions B1 through B3 and B6 through B8 are similar to those used by Bickel (1975) in his study of one-step M-estimates in the *homoscedastic* model. Condition B4 ensures that we do not have infinite weights, and condition B5 assures us that when $\sigma_i = H(\tau_i, \theta) = \exp\{h(\tau_i)\theta\}$, the function H is sufficiently smooth.

3. Estimation of θ.

In the previous section we have shown that, except for certain technical conditions, one can construct robust weighted estimates of β as long as one has available estimates of θ and β which satisfy (2.1). Preliminary estimates $\hat{\beta}_0$ satisfying (2.1) are readily available and include (under reasonable assumptions) ordinary least squares estimates and ordinary M-estimates; details of sufficient conditions for this are available from the authors. Bounded influence regression estimates could also be used; see, e.g., Krasker and Welsch (1981). In this section, we propose a class of estimates of θ which are robust and satisfy (2.1). There are, of course, many possible ways to construct such estimates, but our method has the necessary theoretical properties as well as encouraging small sample properties; see the next section for details.

To motivate our estimates, suppose that the $\{\tau_i\}$ were known, that the $\{\sigma_i\}$ satisfy (1.4), and that the density f is proportional to $\exp\{-\rho(x)\}$, where ρ and $\rho' = \psi$ are as in the previous section. This device is common in robustness studies; see Huber (1981), Bickel and Doksum (1981), and Carroll (1980) for examples. In this instance, the log-likelihood for θ is, up to a constant,

$$\ell(\theta) = \sum_{i=1}^N \log H(\tau_i, \theta) - \sum_{i=1}^N \rho\left\{\frac{Y_i - \tau_i}{H(\tau_i, \theta)}\right\}. \tag{3.1}$$

Taking derivatives in θ suggests that we solve

$$0 = \ell'(\theta) = \sum_{i=1}^N [z_i(\theta)\psi\{z_i(\theta)\} - 1]\frac{\partial}{\partial\theta}H(\tau_i, \theta)/H(\tau_i, \theta), \tag{3.2}$$

where $z_i(\theta) = \dfrac{(Y_i - \tau_i)}{H(\tau_i, \theta)}$. Because the term in square brackets in (3.2) is not bounded and hence would, in general, lead to an unbounded influence function for the estimated θ and an overall lack of robustness in our estimation procedure, we follow the common device used in the homoscedastic case by Huber (1981) and Bickel and Doksum (1981) of replacing $x\psi(x) - 1$ by a function $\chi(\cdot)$, as well as replacing τ_i by $t_i = x_i\hat{\beta}_0$, thus leading to estimates obtained by solving

$$0 = G_N(\theta) = \textstyle\sum_{i=1}^N \chi\left\{\dfrac{Y_i - t_i}{H(t_i, \theta)}\right\} \dfrac{\partial}{\partial\theta} H(t_i, \theta)/H(t_i, \theta). \tag{3.3}$$

Probably the most common choice of $\chi(\cdot)$ in the homoscedastic case is

$$\chi(y) = \chi^2(y) - \int \psi^2(x)\phi(x)\, dx. \tag{3.4}$$

This choice of $\chi(\cdot)$ gives bounded influence to our estimates of θ, and thus might reasonably be preferred in our problem to $y\psi(y) - 1$, just as it is in the homoscedastic case; see Huber (1981, Section 11.1) for certain optimality properties of this choice. In the case of the special model (2.7), we have

$$G_N(\theta) = \textstyle\sum_{i=1}^N \chi\{(Y_i - t_i)e^{-h(t_i)\theta}\}h(t_i). \tag{3.5}$$

We make the assumptions that $\chi(\cdot)$ is an even function with $\chi(0) < 0$, $\chi(\infty) > 0$. In the model (1.4), σ is a free parameter defined so that

$$E\chi\left(\frac{Y_1 - \tau_1}{\sigma_1}\right) = 0. \tag{3.6}$$

In the first model of (1.3), we have

$$\theta = (\log \sigma, \lambda)^T, \qquad h(\tau) = \log(1 + |\tau|).$$

In many models (such as the first three models in (1.3), the third with $\tau_i > 0$), one can show that solutions to the equation $G_N(\theta) = G_N(\sigma, \lambda) = 0$ exist. We have been unable to show that the solutions are unique, although in all of our examples, unique solutions have been obtained. More importantly, one may not wish to consider all possible values of θ, e.g., in the first three models of (1.3), one may reasonably wish to restrict $|\theta| \le 1.5$ if one assumes that the variances will be no larger than the cubes of the means. For these reasons, we suggest the following procedure:

(3.7) Minimize $\| G_N(\theta)\| = \| G_N(\sigma, \lambda)\|$ on the interval $\lambda \in J$.
If the solution is not unique, choose the one with smallest $\| \lambda \|$.

The solution to (3.7) is thus well-defined. In all of our examples when θ is unrestricted, the solutions to (3.3) and (3.7) have coincided. In the examples in which we have restricted θ, (3.7) has always had a unique solution even when (3.3) has not had a solution in the restricted space.

An appealing feature of these estimates is that they are natural generalizations of the classical Huber Proposal 2 for the homoscedastic case.

THEOREM 2. *Assume* (2.5), (2.7), (3.6), *and* B2 *through* B5. *Further assume that* $N^{1/2}(\hat{\beta}_0 - \beta) = O_p(1)$. *Finally, make the assumptions*

C1. $0 < E\chi^2(\varepsilon_1) < \infty$, and χ is non-decreasing on $[0, \infty)$.
C2. *As* $r, s \to 0$, *for* $A(\chi) > 0$, $E\chi\{(\varepsilon_1 + r)(1 + s)\} = A(\chi)s + o(|r| + |s|)$.
C3. *Condition* B7 *holds for* χ.
C4. *Condition* B8 *holds for* χ.
C5. *If* λ_N *is the minimal eigenvalue of* $H_N = N^{-1} \sum_{i=1}^N h(\tau_i)^T h(\tau_i)$, *then* $\liminf \lambda_N = \lambda_\infty > 0$.

Then if $\hat{\theta}$ *solves* (3.7), *we have*

$$\hat{\theta} - \theta = O_p(N^{-1/2}). \tag{3.8}$$

The proof is given in Section 7. The conditions are similar to those of Bickel (1975), with only C5 affected by hetreroscedasticity. Further details of implementation are discussed in the next section.

One can also introduce redescending M-estimates by using ψ redescending to zero. Estimates for θ and β can be obtained by doing one or two steps of Newton-Raphson for (2.4) and (3.3) from any estimate satisfying (2.1). Proofs are similar to those given in the appendices.

4. A Monte-Carlo study. Because Theorem 1 is an asymptotic result, we performed a small Monte-Carlo study to assess the small sample properties of $\hat{\beta}$. The model was simple linear regression, given by

$$Y_i = \beta_0 + \beta_1 c_i + \sigma_i \epsilon_i = \tau_i + \sigma_i \epsilon_i, \quad i = 1, \cdots, N. \tag{4.1}$$

In the study, the $\{c_i\}$ were equally spaced between -2 and $+2$, and we chose to study the model

$$\sigma_i = \sigma(1 + |\tau_i|)^{\lambda}.$$

The experiments were each repeated two hundred times under the following circumstances:

(a) $N = 21$, $\{\epsilon_i\}$ are $N(0, 1)$, $\sigma = .25$, $\beta_0 = 2$, $\beta_1 = 1$.

(b) $N = 41$, $\{\epsilon_i\}$ are $N(0, 1)$ with probability $p = .90$ and $N(0, 9)$ with $p = .10$, $\sigma = .25$, $\beta_0 = 4$, $\beta_1 = 2$.

We made two choices for ψ. First was $\psi(x) = x$, which yields the usual weighted least squares estimate $\hat{\beta}_L$, and the second was Huber's $\psi(x) = \max\{-2.0, \min(x, 2.0)\}$. This gives a version $\hat{\beta}_R$ of our robust weighted estimates. In constructing $\hat{\sigma}_i$, we defined χ as in equation (3.4).

Both $\hat{\beta}_L$ and $\hat{\beta}_R$ were constructed as follows:

Step (i). Let β_* be the unweighted Huber Proposal 2 estimate ($\lambda = 0$) with χ given by (3.4) and $\psi(x) = \max\{-2.0, \min(x, 2.0)\}$.

Step (ii). Solve (3.7) for (σ_*, λ_*) and form inverse "weights"

$$w_i^2 = (1 + |t_i|)^{2\lambda}, \quad t_i = x_i\beta_{*}.$$

Step (iii). Solve a weighted Huber Proposal 2 by simultaneously solving (2.4) with the desired function ψ and the part of (3.6) given by

$$\textstyle\sum_{i=1}^{N} \chi\left(\frac{Y_i - x_i\beta}{\sigma w_i}\right) = 0. \tag{4.2}$$

The result is $\hat{\beta}_0$.

Step (iv). Repeat steps (ii) and (iii) to obtain$t_i = x_i\hat{\beta}_0$, $\hat{\lambda}$, $\hat{\sigma}$, $\hat{\beta}$.

The algorithm given here was chosen so as to reproduce Huber's Proposal 2 in the homoscedastic case $\hat{\lambda} = 0$. Direct application of the results of Section 2 involves only solving (2.4) in Step (iii) and gave results essentially indistinguishable from those reported here. In solving for $(\hat{\lambda}, \hat{\sigma})$, we used the subroutine ZXGSN of the IMSL library.

In Table 1, we list part of the results of the study. The values listed are ratios of mean square errors for estimating β_1 in model (4.1), the ratio being with respect to the "optimal" robust method one would use if $w_i^* = (1 + |\tau_1|)^{\lambda}$ were known, i.e., solve (2.4) and (4.2) simultaneously with the known weights. The study is fairly small, but it does seem to indicate that our robust weighted estimate will work in situations in which heteroscedasticity is suspected.

TABLE 1
Monte-Carlo MSE *ratio for simple linear regression under* (6.1).

	Sample Size N = 21 $\beta_0 = 2.0, \beta_1 = 1.0$ Normal Errors			Sample Size N = 41 $\beta_0 = 4.0, \beta_1 = 2.0$ Contaminated Errors		
Estimator	**$\lambda = 0.0$**	**$\lambda = .5$**	**$\lambda = 1.0$**	**$\lambda = 0.0$**	**$\lambda = .5$**	**$\lambda = 1.0$**
Unweighted LSE	.98	1.18	1.67	1.24	1.51	2.31
"Optimal" WLSE, known weights	.98	.98	.98	1.24	1.19	1.18
Our WLSE, estimated weights	1.14	1.13	1.11	1.29	1.25	1.26
Unweighted robust estimate	1.00	1.18	1.66	1.00	1.21	1.79
Our weighted robust estimate, estimated weights	1.14	1.13	1.10	1.03	1.04	1.07

It is important to note that our estimate has MSE never more than 15% larger than the unknown estimate formed with the correct weights and seems to do better than unweighted estimates when $\lambda \neq 0$. Note also the robustness feature; the efficiency of the weighted least squares estimates (even the "optimal" one) depends heavily on the normality assumption and is not very high in the contaminated case. All of the results tend to support the applicability of Theorem 1.

We repeated the experiment, but with the model

$$\sigma_i = \sigma \exp(\lambda |\tau_i|),$$

and obtained similar results, which seem to indicate that our theory is applicable for a variety of models for the $\{\sigma_i\}$.

For testing and interval estimation, we use the following generalization of methods suggested by Huber (1973) for the homoscedastic case. Using (2.6) and Theorem 1, we estimate the covariance of$N^{1/2}(\hat\beta - \beta)$ by

$$K(\widehat{E\psi^2})\widehat{S^{-1}}(\widehat{E\psi'})^{-2}, \tag{4.2}$$

where

$$\widehat{E\psi'} = N^{-1}\Sigma\psi'\left(\frac{Y_i - x_i\hat\beta}{\hat\sigma_i}\right), \quad K = 1 + (p+2)\frac{1-\lambda}{N\lambda}, \quad \hat S = N^{-1}\,\Sigma x_i' x_i \hat\sigma_i^{-2}$$

and $\widehat{E\psi^2}$ is defined similarly to $\widehat{E\psi'}$. In our Monte-Carlo experiment, we constructed confidence intervals for the slope parameter β_1 in (4.1), using (4.3) and t-percentage points with $N - p - r = N - 4$ degrees of freedom. The intended coverage probability was 95%. In none of these cases did the achieved coverage probability fall below 92%, and in the majority of the cases it was at least 94%.

We also attempted to solve equations (2.4) and (3.5) simultaneously using the IMSL routine ZSYSTM. Our experience was much like that of Froehlich (1973) in that the algorithm converged most of the time but not always. Dent and Hildreth (1977) were able to show that the difficulties experienced by Froehlich could be overcome by sophisticated optimization techniques. We suspect that the same holds for our problem.

The particular method for estimating $\theta = (\sigma, \lambda)$ outlined in Section 3 and explored in this section is recommended for models such as the first three in (1.3), which satisfy (2.7). In the fourth model of (1.3), an alternative procedure is preferable because we can exploit the relationship

$$\sigma_i^2 = \alpha_1 + \alpha_2 \tau_i^2.$$

Here one would obtain initial estimates of (α_1, α_2) by robust regression techniques, as long as the lines of Jobson and Fuller (1980), working with the squares of the residuals from a preliminary fit. One would then do one-step of a Newton-Raphson towards solving versions of (3.3) which are obtained by working with (α_1, α_2) and following the line of reasoning in (3.1) through (3.3). Monte-Carlo work, which will be reported elsewhere, indicates that this technique can be quite successful.

5. Feedback. In the case of normal errors, Jobson and Fuller have suggested using the information about β in the terms $\sigma_1 = H(\tau_1, \theta)$. This essentially reduces to maximizing (3.1) jointly in (θ, β). In a very nice result, they show that if the error distribution is exactly normal and if (1.2) is exactly correct, then improvement over the weighted least squares estimate can be achieved. It is clear that such feedback procedures will be adversely affected by outliers or non-normal error distributions, and it is not clear how to robustly modify them.

In cases where using feedback is contemplated, a second form of robustness must also be considered, i.e., robustness against misspecification of the functions H in (3.1). Carroll and Ruppert (1981, unpublished) have shown that as long as H is correctly specified to order $O(N^{-1/2})$, the asymptotic properties of the weighted estimates ((2.4), (3.5)) are the same as if H were correctly specified; in this sense, our weighted estimates are robust against small errors in specifying H. They also show that such robustness is not the case for feedback estimates. In fact, any gain from feedback can be more than offset by slight errors in specifying H. Since our primary interest is in β, and $\sigma_i = H(\tau_i, \theta)$ is at best an approximation, we suggest that feedback should not be automatically preferred in practical use.

6. An example. In Figure 1, we plot the outcomes of 113 observations of Total Esterase $\{C_i\}$ and Radioimmunoassay - RIA $\{Y_i\}$, made available to us by Drs. D. Horowitz and D. Proud of the National Heart, Lung and Blood Institute. The data are clearly heteroscedastic, so we fit the model (4.1) with variance model

$$\sigma_i = \sigma(1 + |\tau_i|)^\lambda \tag{6.1}$$

and estimation done as in the previous section. The results are summarized in Table 2. Since λ appears to be fairly large, the results of the Monte-Carlo indicate that weighting should be of real benefit. The confidence limits on $\hat{\lambda}$ were obtained by bootstrapping (using 60 simulations). In the weighted cases, the standard errors for β_0 and β_1 were obtained from (4.3); similar standard errors not reported here were found by bootstrapping. The weighted results are fairly close together. While our purpose in presenting the numbers is merely illustrative, we note that the values of λ suggest that a logarithmic or square root transformation might stabilize the variances (Box and Hill, 1974). A random coefficient model might also be contemplated (Dent and Hildreth, 1977). We fit a quadratic model to the data with little change.

A program has been written by Neal Thomas to solve equations (2.4) and (3.5) simultaneously when the second model in (1.3) is used. Since the program utilizes the IMSL package's Levenberg-Marquardt algorithm, it can be used on non-linear regression models. The program is now being tested on simulated data and has been used in a study of migration patterns of the Atlantic menhaden, where it was tried on a data set exhibiting heteroscedasticity and numerous outliers. There it produced estimates which, from a biological viewpoint, seemed more credible than estimates from three other procedures: least squares, least squares after a log transformation applied simultaneously to both the dependent variable and the regression function, and Huber's M-estimator with the MAD estimate of scale (Deriso, Reish, Ruppert, and Carroll, manuscript in preparation). Since menhaden are relatively rare in the northern part of their range (New England), catch data from that region exhibit small values but also low variability. Apparently, a weighted estimator is needed in order to obtain reasonable estimates of migration rates to and from northern waters.

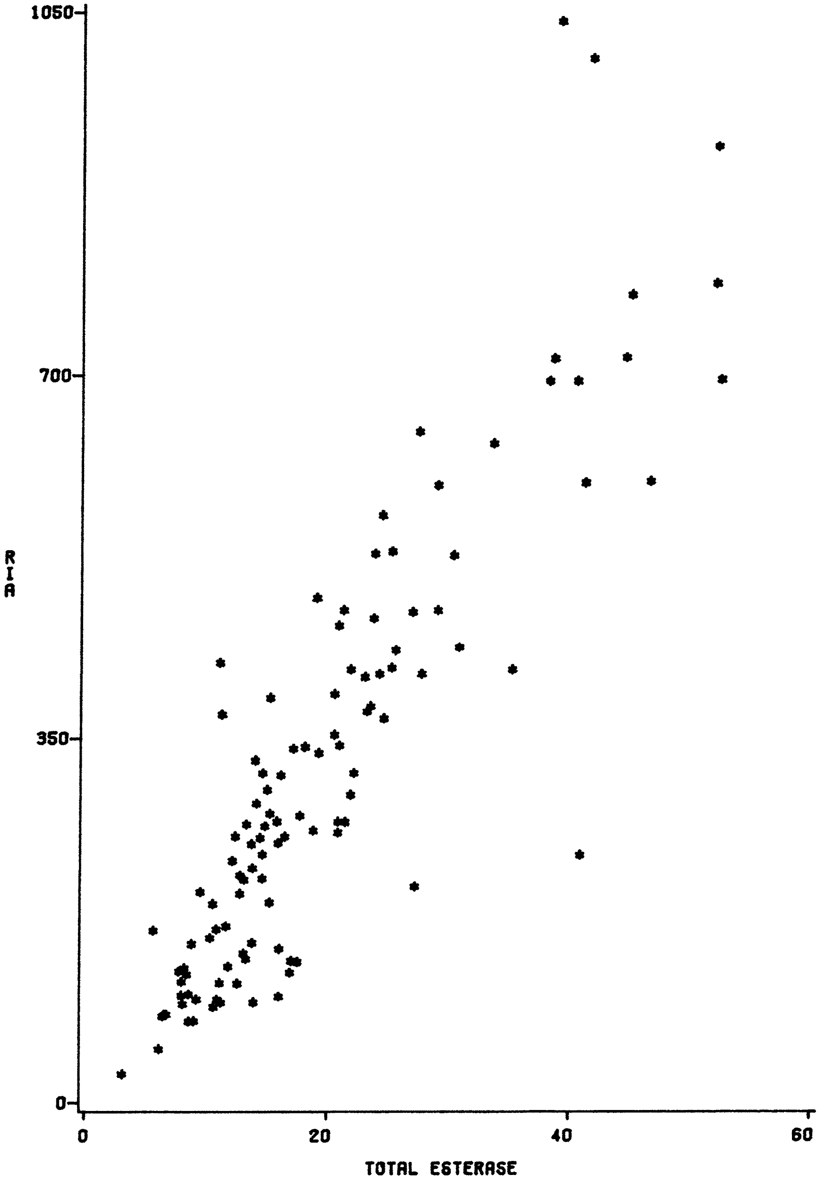


FIG. 1. *Scatterplot of* 113 *observations on* x = *total esterase and* y = *radioimmunoassay RIA.*

TABLE 2
Results of the analysis on the data for Figure 1 *assuming* (6.1).

Method	$\hat{\beta}_0$	Standard Error	$\hat{\beta}$	Standard Error	$\hat{\lambda}$	90% Confidence Limits for λ
Unweighted least squares	−6.30	20.0	16.73	.89	—	—
Our weighted least squares	−19.22	14.1	17.42	.94	.68	(0.4,0.9)
Unweighted robust	−6.54	17.4	16.67	.77	—	—
Our weighted robust	−26.99	11.8	17.73	.88	.85	(0.7,1.1)

Proofs of theorems. The smoothness conditions mentioned in Sections 2 and 3 are as follows:

B6. As $r \to 0$ and $s \to 0, E\psi\{(\varepsilon_1 + r)(1 + s)\} = rE\psi'(\varepsilon_1) + o(|r| + |s|)$.

B7. There exist $K > 0$ and $C_0 > 0$ such that when $0 < \delta < 1$, $|r| \leq K$, and $|s| \leq K$,

$$\text{E}\sup[|\psi\{(\varepsilon_1 + r)(1 + s)\} - \psi\{(\varepsilon_1 + r')(1 + s')\}: |r - r'| \leq \delta \text{ and } |s - s'| \leq \delta] \leq C_0\delta.$$

B8. $\lim_{\delta\to 0} E \sup([\psi\{(\varepsilon_1 + r)(1 + s)\} - \psi(\varepsilon_1)]^2: |r|, |s| \leq \delta) = 0.$

The following general theorem will be used when studying $\hat{\beta}_0$, $\hat{\theta}$, and $\hat{\beta}$.

THEOREM 7.1. *Let g_i, k_i, and $A(\phi, i)$ standing for g_{iN}, k_{iN}, and $A(\phi, i, N)$, be sequences of positive constants such that*

(7.1) $$\lim_{N\to\infty}\sup_{i\leq N}(k_i + k_i g_i) = 0, \quad \sup_N\sup_{i\leq N}A(\phi, i) < \infty,$$

and

(7.2) $$\sup_N \textstyle\sum_{i=1}^N (k_i^2 + k_i^2 g_i^2 + N^{-1/2} g_i k_i) = C_1 < \infty.$$

Let ϕ_{i} *be a function from* R^3 *to* R^1 *satisfying*

(7.3) $$E\phi_i(\varepsilon_1, 0, 0) = 0 \quad \textit{for all} \quad i.$$

Suppose that there exists $K > 0$ and $C_0 > 0$ such that for all i,

(7.4) $$E \sup\{|\phi_i(\varepsilon_1, r, s) - \phi_i(\varepsilon_1, r', s')|: |r - r'|, |s - s'| \leq \delta\} \leq C_0 g_i\delta$$

whenever $0 < \delta < 1$, $|r| \leq K$, *and* $|s| \leq K$,

(7.5) $$\sup_N\sup_{i\leq N} g_i^{-1} E\{\phi_i(\varepsilon_1, r, s) - \phi_i(\varepsilon_1, 0, 0) - A(\phi, i)r\} = o(|r| + |s|) \textit{ as } r, s, \to 0,$$

(7.6) $$\lim_{\delta\to 0} \sup_N \sup_{i\leq N} E[\sup_{|r|\leq\delta, |s|\leq\delta} g_i^{-2}\{\phi_i(\varepsilon_1, r, s) - \phi_i(\varepsilon_1, 0, 0)\}^2] = 0,$$

and $\sup_N \sup_{i\leq N} g_i^{-2} E\phi_i^2(\varepsilon_i, 0, 0) < \infty$. *Let $a_i^{(1)}$, $\alpha_i^{(2)}$, and $\boldsymbol{\alpha}_i^{(3)}$ be functions from R^m to R^1, R^1, and R^n respectively, and let $\mathbf{z}_i (=\mathbf{z}_{iN})$ be elements of R^n satisfying*

(7.7) $$\alpha_i^{(\ell)}(0) = 0, \quad \ell = 1, 2, 3,$$

and for each compact set S there exists K such that

(7.8) $$|\alpha_i^{(\ell)}(\mathbf{x}) - \alpha_i^{(\ell)}(\mathbf{y})| \leq k_i \|\mathbf{x} - \mathbf{y}\| K, \ell = 1, 2,$$

and $\|\boldsymbol{\alpha}_i^{(3)}(\mathbf{x}) - \boldsymbol{\alpha}_i^{(3)}(\mathbf{y})\| \leq k_i \|\mathbf{z}_i\| \|\mathbf{x} - \mathbf{y}\| K$ *for all x and y in S*, $i = 1, \cdots, N$, *and*

(7.9) $$N^{-1/2}\|\mathbf{z}_i\| \leq k_i.$$

For $\boldsymbol{\Delta} \in R^m$, *define the process*

$$U_N(\boldsymbol{\Delta}) = N^{-1/2} \textstyle\sum_{i=1}^N \phi_i\{\varepsilon_i, \alpha_i^{(1)}(\boldsymbol{\Delta}), \alpha_i^{(2)}(\boldsymbol{\Delta})\}\{\mathbf{z}_i + \boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta})\}.$$

Then, for all $M > 0$,

$$\sup_{\|\boldsymbol{\Delta}\| \le M} \| U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0}) - N^{-1/2} \textstyle\sum_{i=1}^N A(\phi, i)\alpha_i^{(1)}(\boldsymbol{\Delta})\mathbf{z}_i \| = o_p(1). \tag{7.10}$$

PROOF OF THEOREM 7.1. For convenience, take $M = 1$. For $0 < \delta < 1$, define

$$S_N(\boldsymbol{\Delta}, \delta) = \sup\{\| U_N(\boldsymbol{\Delta}') - U_N(\boldsymbol{\Delta})\| : \|\boldsymbol{\Delta}' - \boldsymbol{\Delta}\| \le \delta\}.$$

We will show that

$$E\{U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0})\} = N^{-1/2} \textstyle\sum_{i=1}^N A(\phi, i)\alpha_i^{(1)}(\boldsymbol{\Delta})\mathbf{z}_i + o(1), \tag{7.11}$$

$$U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0}) - E\{U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0})\} = o_p(1) \tag{7.12}$$

for each fixed $\boldsymbol{\Delta}$, and that there exists K depending upon M but not δ such that for all $0 < \delta < 1$, all N, and all $\|\boldsymbol{\Delta}\| \le 1$,

$$S_N(\boldsymbol{\Delta}, \delta) - ES_N(\Delta, \delta) = o_p(1) \text{ and } ES_N(\Delta, \delta) \le K\delta, \tag{7.13}$$

where K does not depend upon δ. Since for any δ, we can cover the ball of radius 1 in R^m with a finite number of balls of radius δ, (7.11), (7.12) and (7.13) prove the theorem.

To prove (7.11), note that by (7.3),

$$E(U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0})) = N^{-1/2} \textstyle\sum_{i=1}^N E[\phi_i\{\varepsilon_i, \alpha_i^{(1)}(\boldsymbol{\Delta}), \alpha_i^{(2)}(\boldsymbol{\Delta})\} - \phi_i(\varepsilon_i, 0, 0)]\{\mathbf{z}_i + \boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta})\}.$$

We next have by (7.1), (7.7) and (7.8) that, for all large N,

$$\|\mathbf{z}_i + \boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta})\| \le 2\|\mathbf{z}_i\| \tag{7.14}$$

(for simplicity, take $K = 1$ in (7.8)), and also by (7.5), (7.7) and (7.8),

$$E[\phi_i\{\varepsilon_i, \alpha_i^{(1)}(\boldsymbol{\Delta}), \alpha_i^{(2)}(\boldsymbol{\Delta})\} - \phi_i(\varepsilon_i, 0, 0)] = A(\phi, i)\alpha_i^{(1)}(\boldsymbol{\Delta}) + o(g_i k_i)$$

uniformly in i. Therefore,

$$E\{U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0})\} = N^{-1/2} \textstyle\sum_{i=1}^N A(\phi, i)\alpha_i^{(1)}(\boldsymbol{\Delta})\mathbf{z}_i + o\{N^{-1/2} \textstyle\sum_{i=1}^N g_i k_i \|\mathbf{z}_i\|$$
$$+ N^{-1/2} \textstyle\sum_{i=1}^N A(\phi, i)\alpha_i^{(1)}(\boldsymbol{\Delta})\boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta})\}.$$

By (7.1), (7.7), (7.8), and (7.9), the last term on the RHS is $o(1)$. By (7.2), (7.9) and the Cauchy-Schwarz inequality, the second term is bounded by

$$o(\textstyle\sum_{i=1}^N g_i k_i^2) = o(1),$$

so that (7.11) holds. Then, using (7.14), we have that for N large,

$$\tfrac{1}{2} \operatorname{Var}\{U_N(\boldsymbol{\Delta}) - U_N(\mathbf{0})\} \le (2N^{-1} \textstyle\sum_{i=1}^N g_i^2 \|\mathbf{z}_i\|^2) \sup_{i \le N} g_i^{-2} E[\phi_i\{\varepsilon_i, \alpha_i^{(1)}(\boldsymbol{\Delta}), \alpha_i^{(2)}(\boldsymbol{\Delta})\}$$
$$- \phi_i(\varepsilon_1, 0, 0)]^2 + E\| N^{-1/2} \textstyle\sum_{i=1}^N \phi_i(\varepsilon_i, 0, 0)\boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta})\|^2.$$

The second term on the RHS is $o(1)$ by (7.7) and (7.8). It also follows from (7.6) through (7.8) that

$$\sup_{i \le N} g_i^{-2} E[\phi_i\{\varepsilon_i, \alpha_i^{(1)}(\boldsymbol{\Delta}), \alpha_i^{(2)}(\boldsymbol{\Delta})\} - \phi_i(\varepsilon_i, 0, 0)]^2 = o(1).$$

Therefore, (7.12) is proved by applying (7.2) and (7.9). Finally, by (7.14) $ES_N(\boldsymbol{\Delta}, \delta)$ is less than or equal to

$$2N^{1/2} \textstyle\sum_{i=1}^N E \sup_{\|\boldsymbol{\Delta} - \boldsymbol{\Delta}'\| \le \delta} |\phi_i\{\varepsilon_1, \alpha_i^{(1)}(\boldsymbol{\Delta}), \alpha_i^{(2)}(\boldsymbol{\Delta})\} - \phi_i\{\varepsilon_1, \alpha_i^{(1)}(\boldsymbol{\Delta}'), \alpha_i^{(2)}(\boldsymbol{\Delta}')\}| \, \|\mathbf{z}_i\|$$
$$+ N^{-1/2} \textstyle\sum_{i=1}^N \sup_{\|\boldsymbol{\Delta} - \boldsymbol{\Delta}'\| \le \delta} \|\boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta}) - \boldsymbol{\alpha}_i^{(3)}(\boldsymbol{\Delta}')\| E\,|\phi_i\{\varepsilon_1, \alpha_i^{(1)}(\boldsymbol{\Delta}'), \alpha_i^{(2)}(\boldsymbol{\Delta}')\}|.$$

Thus by (7.2), (7.4), (7.6), and (7.9),

$$ES_N(\mathbf{\Delta}, \delta) \le K\delta$$

for some K which is independent of δ. By (7.1), (7.2), (7.6), (7.8), and (7.9),

$$\operatorname{Var} S_N(\mathbf{\Delta}, \delta) \to 0.$$

Therefore, (7.13) is verified.

PROOF OF THEOREM 1. For $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_3$ in R^p, $\mathbf{\Delta}_2$ in R^r, and $\mathbf{\Delta} = (\mathbf{\Delta}_1, \mathbf{\Delta}_2, \mathbf{\Delta}_3)$, define

$$\begin{aligned} \alpha_i^{(1)}(\mathbf{\Delta}) &= N^{-1/2} d_i \mathbf{\Delta}_1 \\ h_i(\mathbf{\Delta}) &= h(\tau_i + x_i \mathbf{\Delta}_3 N^{-1/2}) \\ \alpha_i^{(2)}(\mathbf{\Delta}) &= \exp[-h_i(\mathbf{\Delta})\mathbf{\Delta}_2 N^{-1/2} + \{h_i(\mathbf{0}) - h_i(\mathbf{\Delta})\}\theta] - 1 \end{aligned}$$

and

$$\alpha_i^{(3)}(\mathbf{\Delta}) = d_i \alpha_i^{(2)}(\mathbf{\Delta}).$$

Define the process

$$U_N(\mathbf{\Delta}) = N^{-1/2} \textstyle\sum_{i=1}^N \psi[\{\varepsilon_i - \alpha_i^{(1)}(\mathbf{\Delta})\}\{1 + \alpha_i^{(2)}(\mathbf{\Delta})\}]\{d_i + \alpha_i^{(3)}(\mathbf{\Delta})\}.$$

Note that (2.4) can be rewritten as

$$U_N(N^{1/2}(\hat\beta - \beta), N^{1/2}(\hat\theta - \theta), N^{1/2}(\hat\beta_0 - \beta)) = 0. \tag{7.15}$$

Letting $g_i \equiv 1$, $k_i = N^{-1/2}\{1 + \|d_i\| + \|h(\tau_i)\|\}$, $\phi_i(\varepsilon_i, r, s) = \psi\{(\varepsilon_i - r)(1 + s)\}$, $d_i = z_i$, and $A(\phi, i) = A(\psi) = E\psi'$, the conditions of Theorem 7.1 are implied by (2.5) and B1 through B8, so for all $M > 0$,

$$\sup_{\|\mathbf{\Delta}\| \le M} \| U_N(\mathbf{\Delta}) - U_N(\mathbf{0}) + A(\psi) S \mathbf{\Delta}_1 \| = o_p(1). \tag{7.16}$$

Now by Chebyshev's theorem, B1 and B2,

$$U_N(\mathbf{0}) = O_p(1).$$

In proving the theorem, we will not assume that $\hat\beta$ actually solves (2.4), but rather that the l.h.s. of (2.4) evaluated at $\hat\beta$ is less than twice its infimum over all β. However, as noted by Huber (1981, page 165), (2.4) will have a unique solution if ψ is strictly monotone. From the last equation, we have that if

$$\mathbf{\Delta}_1^* = -\{A(\psi)S\}^{-1} U_N(0) = O_p(1),$$

then by (7.16), $U(\mathbf{\Delta}^*) = o_p(1)$. Consequently, by the equivalence of (2.4) and (7.15),

$$\| U_N(N^{1/2}(\hat\beta - \beta), N^{1/2}(\hat\theta - \theta), N^{1/2}(\hat\beta_0 - \beta)) \| \le 2 \| U_N(\mathbf{\Delta}^*) \| = o_p(1). \tag{7.17}$$

By (2.1), we need only establish that

$$\hat\beta - \beta = O_p(N^{-1/2}) \tag{7.18}$$

to conclude from (7.15) and (7.16) that (2.8) holds. But by (7.17), (7.18) holds if for each $\eta > 0$, $\varepsilon > 0$ and M_1, there exists M_2 satisfying

$$P\{\inf_{\|\Delta_1\| \ge M_2} \inf_{\|\Delta_2\| \le M_1} \inf_{\|\Delta_3\| \le M_1} \| U_N(\mathbf{\Delta}) \| > \eta\} > 1 - \varepsilon. \tag{7.19}$$

Now (7.19) follows from (7.16) in a manner quite similar to Jurečková's (1977) proof of her Lemma 5.2. □

PROOF OF THEOREM 2. For $\mathbf{\Delta}_1$ in R^p, $\mathbf{\Delta}_2$ in R^q, and $\mathbf{\Delta}' = (\mathbf{\Delta}_1', \mathbf{\Delta}_2')$, define

$$h_i(\mathbf{\Delta}) = h(\tau_i + x_i \mathbf{\Delta}_1 N^{-1/2}),$$

$$\alpha_i^{(1)}(\mathbf{\Delta}) = \exp[-h_i(\mathbf{\Delta})\mathbf{\Delta}_2 N^{-1/2} + \{h_i(\mathbf{0}) - h_i(\mathbf{\Delta})\}\theta] - 1, \tag{7.20}$$

$$\alpha_i^{(2)}(\mathbf{\Delta}) = N^{-1/2} d_i \mathbf{\Delta}_1, \tag{7.21}$$

and

$$\alpha_i^{(3)}(\mathbf{\Delta}) = h_i(\mathbf{0}) - h_i(\mathbf{\Delta}). \tag{7.22}$$

Then let $\phi(x, y, z) = \chi\{(x - z)(1 + y)\}$ and define the process

$$W_N(\mathbf{\Delta}) = -N^{-1/2} \textstyle\sum_{i=1}^N \phi\{\varepsilon_i, \alpha_i^{(1)}(\mathbf{\Delta}), \alpha_i^{(2)}(\mathbf{\Delta})\}\{h(\tau_i) - \alpha_i^{(3)}(\mathbf{\Delta})\}.$$

Note that (3.7) can be written as

$$\| W_N\{N^{1/2}(\hat{\beta}_0 - \beta), N^{1/2}(\hat{\theta} - \theta)\}\| = \text{minimum}.$$

However, by (3.6), C1 and Chebyshev's inequality,

$$W_N(\mathbf{0}) = O_p(1) \tag{7.23}$$

so that

$$W_N\{N^{1/2}(\hat{\beta}_0 - \beta), N^{1/2}(\hat{\theta} - \theta)\} = O_p(1).$$

We can therefore prove (3.8) by showing that for each $M_1 > 0$, $\varepsilon > 0$ and $Q > 0$, there exists $M_2 > 0$ such that

$$P[\inf\{\| W_N(\mathbf{\Delta})\| : \|\mathbf{\Delta}_1\| \le M_1, \|\mathbf{\Delta}_2\| \ge M_2\} > Q] \ge 1 - \varepsilon. \tag{7.24}$$

We will prove (7.24) by modifying the proof of Jurečková's (1977) Lemma 5.2. We first apply Theorem 7.1 with $z_i = h_i(\mathbf{0})$, $g_i \equiv 1$, $A(\phi, i) = A(\chi)$, and $k_i = N^{-1/2}\{\| h(\tau_i)\| + \| x_i\| + \| d_i\|\}$. Then

$$\sup_{\|\mathbf{\Delta}\| \le M} \| W_N(\mathbf{\Delta}) - W_N(\mathbf{0}) + A(\chi) N^{-1/2} \textstyle\sum_{i=1}^N h(\tau_i)\alpha_i^{(1)}(\mathbf{\Delta})\| = o_p(1).$$

By a Taylor series expansion,

$$\alpha_i^{(1)}(\mathbf{\Delta}) = -N^{-1/2} h(\tau_i)\mathbf{\Delta}_2 + \{h_i(\mathbf{0}) - h_i(\mathbf{\Delta})\}\theta + o(N^{-1/2}).$$

Thus, by C5 setting

$$G_N(\mathbf{\Delta}) = N^{-1/2} \textstyle\sum_{i=1}^N \{h_i(\mathbf{0}) - h_i(\mathbf{\Delta})\}\theta h(\tau_i),$$

we obtain

$$\sup_{\|\mathbf{\Delta}\| \le M} \| W_N(\mathbf{\Delta}) - W_N(\mathbf{0}) - A(\chi)\mathbf{\Delta}_2^T H_N + G_N(\mathbf{\Delta})\| = o_p(1). \tag{7.25}$$

Now fix $\varepsilon > 0$, $M_1 > 0$, $Q > 0$. Use C1 to choose γ such that

$$P\{\| W_N(\mathbf{0})\| \ge \gamma/2\} < \varepsilon/2.$$

Define

$$D = \sup_N \sup_{\|\mathbf{\Delta}_1\| \le M_1} \| G_N(\mathbf{\Delta})\|.$$

Then $D < \infty$ (G_N depends only on $\mathbf{\Delta}_1$). Define M_2 by $\{A(\chi)\lambda_\infty M_2/2 - \gamma - D\} = Q$. Using C5 and (7.25), find N_0 such that $\lambda_N \ge \lambda_\infty/2$ and

$$P\{\sup_{\|\mathbf{\Delta}_2\| = M, \|\mathbf{\Delta}_1\| \le M_1} \| W_N(\mathbf{\Delta}) - W_N(\mathbf{0}) - A(\chi)\mathbf{\Delta}_2^T H_N - G_N(\mathbf{\Delta})\| < \gamma/2\} \ge 1 - \varepsilon/2 \ (N \ge N_0).$$

If $\|\mathbf{\Delta}_2\| = M_2$, $\|\mathbf{\Delta}_1\| \le M_1$, and $N \ge N_0$, then with probability at least $1 - \varepsilon$,

$$\begin{aligned} W_N(\mathbf{\Delta})\mathbf{\Delta}_2 &\ge -M_2\| W_N(\mathbf{0})\| + \mathbf{\Delta}_2^T H_N \mathbf{\Delta}_2 A(\chi) - M_2 D - M_2\gamma/2 \\ &\ge \{A(\chi)\lambda_\infty M_2/2 - \gamma - D\} M_2 = Q M_2. \end{aligned}$$

Since χ is nondecreasing on $[0, \infty)$ by C1, $W_N(\mathbf{\Delta}_1, \mathbf{\Delta}_2 s)\mathbf{\Delta}_2$ is a nondecreasing function of s. Thus, $\|\mathbf{\Delta}_2\| \geq M_2$ implies

$$W_N(\mathbf{\Delta})\mathbf{\Delta}_2 \geq (\|\mathbf{\Delta}_2\|/M_2)\{M_2\|\mathbf{\Delta}_2\|^{-1}W_N(\mathbf{\Delta}_1, M_2\mathbf{\Delta}_2\|\mathbf{\Delta}_2\|^{-1})\mathbf{\Delta}_2\} \geq \|\mathbf{\Delta}_2\|Q.$$

Thus,

$$P\left\{\inf_{\|\Delta_1\| \leq M_1, \|\Delta_2\| \geq M_2} \frac{W_N(\mathbf{\Delta})\mathbf{\Delta}_2}{\|\mathbf{\Delta}_2\|} \geq Q\right\} \geq 1 - \varepsilon,$$

which with the Cauchy-Schwarz inequality proves (3.8). □

REFERENCES

Anscombe, F. J. (1961). Examination of residuals. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* (J. Neyman, editor) 1-36. University of California Press, Berkeley.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428-434.

Bickel, P. J. (1978). Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *Ann. Statist.* **6** 266-291.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296-311.

Box, G. E. P. and Hill, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics* **16** 385-389.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *J. Royal Statist. Soc. Ser. B* 71-78.

Carroll, R. J. and Ruppert, D. (1981). On robust tests for heteroscedasticity. *Ann. Statist.* **9** 206-210.

Dent, W. T. and Hildreth, C. (1977). Maximum likelihood estimation in random coefficient models. *J. Amer. Statist. Assoc.* **72** 69-72.

Froehlich, B. R. (1973). Some estimators for a random coefficient regression model. *J. Amer. Statist. Assoc.* **68** 329-335.

Fuller, W. A. and Rao, J. N. K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Ann. Statist.* **6** 1149-1158.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte-Carlo. *Ann. Statist.* **5** 799-821.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Jobson, J. D. and Fuller, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *J. Amer. Statist. Assoc.* **75** 176-181.

Jurečková, J. (1977). Asymptotic relations of M-estimates and R-estimates in linear regression model. *Ann. Statist.* **5** 464-472.

Krasker, W. S. and Welsch, R. E. (1981). Efficient bounded-influence regression estimation using alternative definitions of sensitivity. To appear in *J. Amer. Statist. Assoc.*

9810 Parkwood Drive
Bethesda, Maryland 20814

Dept. of Statistics
University of North Carolina
at Chapel Hill
321 Philipps Hall 039A
Chapel Hill, North Carolina 27514

Biometrika (1986), **73**, 2, *pp.* 413–24
Printed in Great Britain

Optimally bounded score functions for generalized linear models with applications to logistic regression

BY LEONARD A. STEFANSKI

Department of Economic and Social Statistics, Cornell University, Ithaca, New York 14853, U.S.A.

RAYMOND J. CARROLL AND DAVID RUPPERT

Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27514, U.S.A.

SUMMARY

We study optimally bounded score functions for estimating regression parameters in a generalized linear model. Our work extends results obtained by Krasker & Welsch (1982) for the linear model and provides a simple proof of Krasker & Welsch's first-order condition for strong optimality. The application of these results to logistic regression is studied in some detail with an example given comparing the bounded-influence estimator with maximum likelihood.

Some key words: Bounded influence; Generalized linear model; Influential point; Logistic regression; Outlier; Robustness.

1. INTRODUCTION

In this paper we study robust estimation of θ in generalized linear models (McCullagh & Nelder, 1983, Ch. 2) when the conditional density of $Y|X$ has the form

$$f(y|x) = \exp[\{y - h(x^{\mathrm{T}}\theta)\}q(x^{\mathrm{T}}\theta) + c(y)],$$

where $h(\,.\,)$, $q(\,.\,)$ and $c(\,.\,)$ are known functions and θ is a vector of regression parameters. Models of this type include logistic and probit regression, Poisson regression, linear regression with known variance, and certain models for lifetime data.

Our motivation for seeking robust estimators is the same as that in the linear model; maximum likelihood estimation is sometimes sensitive to outlying data. For logistic regression, Pregibon (1981, 1982) has documented the nonrobustness of the maximum likelihood estimator and expounded the benefits of diagnostics as well as robust or resistant fitting procedures; see also Johnson (1985).

Much of the work on robust estimation concerns finding estimators which sacrifice little efficiency at the assumed model while providing protection against outliers and model violations. We follow this course finding bounded-influence estimators minimizing certain functionals of the asymptotic covariance matrix. Related work includes that of Hampel (1978), Krasker (1980) and Krasker & Welsch (1982).

When fitting models to data, two important issues are identification of outliers and influential cases and accommodation of these observations. Frequently when influential cases are present, the fitted model is not representative of the bulk of the data. To rectify this, one can simply delete influential cases and refit via standard methods, but this approach lacks a theory for inference and testing; the effects of case deletion upon the distributions of estimators is not well understood, even asymptotically.

The robust techniques studied here provide a method of accommodating anomalous data. They allow continuous downweighting of influential cases and are amenable to asymptotic inference. Also, together with more direct diagnostics, residuals and weights from a bounded-influence fit can be used to detect exceptional observations.

In § 2 we present some general theory; this is specialized to the case of logistic regression in § 3; proofs of theorems are given in an Appendix.

2. THE GENERAL THEORY

2·1. *The regression model*

We study regression models in which the dependent variable Y and explanatory p-vector X have a density of the form

$$g(y, x;\, \theta_0) = f(y;\, x^{\mathrm{T}}\theta_0)u(x). \tag{2·1}$$

The conditional density of Y given $X = x$ is $f(y;\, x^{\mathrm{T}}\theta_0)$ and depends on the unknown parameter θ_0 only through $x^{\mathrm{T}}\theta_0$; $u(x)$ is the marginal density of X. Expectation with respect to $g(y, x;\, \theta)$ is denoted by E_θ while $E_{\theta,x}$ indicates conditional expectation corresponding to $f(y;\, x^{\mathrm{T}}\theta)$. Model (2·1) includes many generalized linear models (McCullagh & Nelder, 1983, Ch. 2).

Suppose (Y_i, X_i) $(i = 1, \ldots, n)$ are independent copies of (Y, X). Under regularity conditions the maximum likelihood estimator of θ_0 satisfies

$$\sum_{i=1}^{n} l(Y_i, X_i, \hat{\theta}_{\mathrm{ML}}) = 0, \tag{2·2}$$

where $l(y, x, \theta) = (\partial/\partial\theta)[\log\{f(y;\, x^{\mathrm{T}}\theta)\}]$ and $n^{\frac{1}{2}}(\hat{\theta}_{\mathrm{ML}} - \theta_0)$ converges in distribution to a p-dimensional normal random variate with mean zero and covariance matrix $V(\theta_0) = [E_{\theta_0}\{l(Y, X, \theta_0)l^{\mathrm{T}}(Y, X, \theta_0)\}]^{-1}$.

2·2. *M-estimators and their influence curves*

We generalize (2·2) by considering estimators $\hat{\theta}_\psi$ satisfying

$$\sum_{i=1}^{n} \psi(Y_i, X_i, \hat{\theta}_\psi) = 0,$$

for suitably chosen functions ψ from $R \times R^p \times R^p$ to R^p. We require that ψ be unbiased, i.e.

$$E_\theta\{\psi(Y, X, \theta)\} = 0. \tag{2·3}$$

Under regularity conditions (Huber, 1967), $\hat{\theta}_\psi$ is consistent and asymptotically normal with influence curve

$$\mathrm{IC}_\psi(y, x, \theta) = D_\psi^{-1}(\theta)\psi(y, x, \theta), \tag{2·4}$$

where

$$D_\psi(\theta) = -(\partial/\partial\beta)[E_\theta\{\psi(Y, X, \beta)\}]_{\beta=\theta}. \tag{2·5}$$

Write $\psi(\theta)$ and $l(\theta)$ for $\psi(Y, X, \theta)$ and $l(Y, X, \theta)$ respectively. Assuming that integration and differentation can be interchanged in (2·3) and (2·5) it is easy to show that

$$D_\psi(\theta) = E_\theta\{\psi(\theta) l^{\mathrm{T}}(\theta)\}. \tag{2·6}$$

Now let

$$W_\psi(\theta) = E_\theta\{\psi(\theta)\psi^{\mathrm{T}}(\theta)\}. \tag{2·7}$$

It then follows (Huber, 1967) that the asymptotic variance of $n^{\frac{1}{2}}(\hat{\theta}_\psi - \theta_0)$ is

$$V_\psi(\theta_0) = D_\psi^{-1}(\theta_0)\, W_\psi(\theta_0)\{D_\psi^{-1}(\theta_0)\}^{\mathrm{T}}.$$

For robustness we want IC_ψ to be bounded; for efficiency V_ψ should be small. In § 2·3 we define a norm for IC_ψ and outline a theory which suggests efficient bounded score functions.

2·3. *A scalar measure of influence and an optimal score function*

As a scalar measure of maximum influence we employ a definition of sensitivity introduced by W. A. Stahel in his Swiss Federal Institute of Technology Ph.D. thesis, and by Krasker & Welsch (1982). The self-standardized sensitivity of the estimator $\hat{\theta}_\psi$ is defined as

$$s(\psi) = \sup_{(y,x)} \sup_{\lambda \neq 0} \frac{|\lambda^{\mathrm{T}} \mathrm{IC}_\psi|}{(\lambda^{\mathrm{T}} V_\psi \lambda)^{\frac{1}{2}}} = \sup_{(y,x)} (\mathrm{IC}_\psi^{\mathrm{T}} V_\psi^{-1} \mathrm{IC}_\psi)^{\frac{1}{2}} = \sup_{(y,x)} (\psi^{\mathrm{T}} W_\psi^{-1} \psi)^{\frac{1}{2}}. \tag{2·8}$$

For a generalized linear model $s(\psi)$ has a natural interpretation in terms of the link function, e.g. in logistic regression $s(\psi)$ measures the maximum normalized influence of (y, x) on an estimated logit in that $\lambda^{\mathrm{T}}\mathrm{IC}_\psi$ is the influence curve for $\lambda^{\mathrm{T}}\hat{\theta}_\psi$ and $\lambda^{\mathrm{T}} V_\psi \lambda$ is the asymptotic variance of $\lambda^{\mathrm{T}}\hat{\theta}_\psi$. Although this paper studies only the self-standardized sensitivity we believe that useful estimators can also be obtained by bounding other measures of influence, such as fitted values; see Johnson (1985) for measures of influence relative to the determination of fitted values in logistic regression.

For maximum likelihood $\psi = l$ and, in general, $s(l) = +\infty$. To obtain robustness we limit attention to only those estimators $\hat{\theta}_\psi$ for which $s(\psi) \leq b < \infty$. Such an estimator is said to have bounded influence with bound b.

Consider the score function

$$\psi_{\mathrm{BI}}(y, x, \theta) = (l - C) \min{}^{\frac{1}{2}}[1, b^2/\{(l-C)^{\mathrm{T}} B^{-1}(l-C)\}], \tag{2·9}$$

where $l = l(y, x, \theta)$ and the $p \times 1$ vector $C = C(\theta)$ and the $p \times p$ matrix $B = B(\theta)$ are functions of θ defined implicitly by the equations

$$E_\theta\{\psi_{\mathrm{BI}}(Y, X, \theta)\} = 0, \quad B(\theta) = E_\theta\{\psi_{\mathrm{BI}}\psi_{\mathrm{BI}}^{\mathrm{T}}\}. \tag{2·10}$$

With $C(\theta)$ and $B(\theta)$ so defined, ψ_{BI} is unbiased and $W_{\psi_{\mathrm{BI}}}(\theta) = B(\theta)$, so that, by (2·8), ψ_{BI} has bounded sensitivity.

The vector $C(\theta)$ and matrix $B(\theta)$ are analogous to robust multivariate location and scatter functionals for $l(Y, X, \theta)$ (Maronna, 1976). For sufficiently large b solutions $C(\theta)$ and $B(\theta)$ satisfying (2·10) exist, and as b tends to infinity these tend to zero and $E(ll^{\mathrm{T}})$ respectively. Equation (2·9) shows that ψ_{BI} is similar to a weighted maximum likelihood score with weights depending on the distance $(l-C)^{\mathrm{T}}B^{-1}(l-C)$; as b tends to infinity the weighting factor tends to one and ψ_{BI} to l.

For the normal theory linear model ψ_{BI} is the score function found by Krasker & Welsch (1982), who show that if there exists a score ψ_{opt} satisfying (2·3) and $s(\psi) \leq b < \infty$

which minimizes V_ψ in the strong sense of positive-definiteness, that is $V_\psi - V_{\psi_{\text{opt}}} \geq 0$ for all ψ, then it must be of the form (2·9). That ψ_{BI} possesses similar optimality properties is seen in Corollary 1·1 below.

THEOREM 1. *If for a given choice of* $b>0$ *equations* (2·10) *possess the solution* $\{C(\theta), B(\theta)\}$, *then* ψ_{BI} *minimizes* $\operatorname{tr}(V_\psi V_{\text{BI}}^{-1})$ *among all* ψ *satisfying* (2·3) *and*

$$\sup_{(y,x)} (\mathrm{IC}_\psi^{\mathrm{T}} V_{\text{BI}}^{-1} \mathrm{IC}_\psi) \leq b^2. \tag{2·11}$$

With the exception of multiplication by a constant matrix, ψ_{BI} *is unique almost surely.*

Any score function ψ_{opt} for which $V_\psi - V_{\psi_{\text{opt}}} \geq 0$ for all ψ will be called strongly efficient; we now state the following corollary.

COROLLARY 1·1. *If there exists an unbiased, strongly efficient score* ψ_{opt} *satisfying* $s(\psi) \leq b < \infty$, *then* ψ_{opt} *is equivalent to* ψ_{BI} *whenever the latter is defined.*

In Theorem 1 the conditions for optimality of ψ_{BI} depend on ψ_{BI} itself through V_{BI}^{-1}. This is somewhat disconcerting. Nevertheless ψ_{BI} does satisfy an optimality property and this result allows us to prove Corollary 1·1.

Working within the class of score functions of the form $l(y, x, \theta)\omega(y, x, \theta)$ where ω is a scalar weight function, Krasker & Welsch (1982) find the optimal form of ω. Theorem 1 and its corollary show that ψ_{BI} is optimal over a much larger class of functions and hence yield a technically stronger result than Krasker & Welsch's. Also our proof is somewhat simpler than Krasker & Welsch's.

Ruppert (1985) has shown that a strongly efficient score need not exist, in which case Corollary 1·1 is vacuous. In fact, we know of no case with $p \geq 2$ where a strongly efficient score has been shown to exist. However, the result given in Corollary 1·1 is still of interest; Ruppert (1985) uses it in his counter-example.

The proofs of Theorem 1 and its corollary are presented in the Appendix.

2·4. *A one-step estimator*

Write $\psi_{\text{BI}} = \psi_{\text{BI}}(y, x, \theta, B, C)$ to indicate dependence on B and C. Theorem 1 suggests the estimator $\hat{\theta}_{\text{BI}}$ obtained by solving

$$\sum_{i=1}^{n} \hat{\psi}_i(\hat{\theta}_{\text{BI}}) = 0, \tag{2·12}$$

where $\hat{\psi}_i(\theta) = \psi_{\text{BI}}\{Y_i, X_i, \theta, \hat{B}(\theta), \hat{C}(\theta)\}$ and $\hat{C}(\theta)$ and $\hat{B}(\theta)$ are defined implicitly by the equations

$$\sum_{i=1}^{n} E_{\theta, X_i}\{\hat{\psi}_i(\theta)\} = 0, \quad \hat{B}(\theta) = n^{-1} \sum_{i=1}^{n} E_{\theta, X_i}\{\hat{\psi}_i(\theta)\hat{\psi}_i^{\mathrm{T}}(\theta)\}. \tag{2·13}$$

In general the task of simultaneously solving (2·12) and (2·13) is formidable. For the linear model with known variance, $B(\theta)$ does not depend on θ and symmetry implies $C(\theta) \equiv 0$, thus solving for $\hat{\theta}_{\text{BI}}$ is greatly simplified. The case with variance unknown is only slightly more difficult (Krasker & Welsch, 1982). Our attempts at solving (2·12) and (2·13) for logistic regression have not yet been sufficiently successful to recommend a general computational scheme. When an M-estimator is difficult to compute, it is a standard practice to substitute a one-step approximation. Under mild regularity conditions, using a one-step approximation is asymptotically equivalent to iterating until convergence. Thus, the following one-step approximation to $\hat{\theta}_{\text{BI}}$ will have the same asymptotic optimality properties as $\hat{\theta}_{\text{BI}}$.

Let $\tilde{\theta}$ be an initial root-n consistent estimator of θ_0 and suppose $\hat{B}(\tilde{\theta})$ and $\hat{C}(\tilde{\theta})$ satisfy (2·13). Define

$$\hat{\theta}_{\mathrm{BI}}^{(1)} = \tilde{\theta} + n^{-1} \sum_{i=1}^{n} \hat{D}^{-1}(\tilde{\theta}) \hat{\psi}_i(\tilde{\theta}),$$

where

$$\hat{D}(\theta) = n^{-1} \sum_{i=1}^{n} E_{\theta, X_i}\{\hat{\psi}_i(\theta) l^{\mathrm{T}}(Y_i, X_i, \theta)\}.$$

This construction is similar to Bickel's (1975) type II one-step procedure. Under regularity conditions $\hat{\theta}_{\mathrm{BI}}^{(1)}$ is consistent and asymptotically normal with covariance matrix $V_{\mathrm{BI}}(\theta_0) = D_{\mathrm{BI}}^{-1}(\theta_0) B(\theta_0) \{D_{\mathrm{BI}}^{-1}(\theta_0)\}^{\mathrm{T}}$, which is consistently estimated by $\hat{V} = \hat{D}^{-1}(\tilde{\theta}) \hat{B}(\tilde{\theta}) \{\hat{D}^{-1}(\tilde{\theta})\}^{\mathrm{T}}$.

To implement the one-step procedure a method of solving (2·13) for $C(\theta)$ and $B(\theta)$ when θ remains fixed is required. Define

$$J_1(C, B) = \frac{\sum E_{\theta, X_i}(l_{i,\theta} \min^{\frac{1}{2}}[1, b^2/\{(l_{i,\theta} - C)^{\mathrm{T}} B^{-1} (l_{i,\theta} - C)\}])}{\sum E_{\theta, X_i}(\min^{\frac{1}{2}}[1, b^2/\{(l_{i,\theta} - C)^{\mathrm{T}} B^{-1} (l_{i,\theta} - C)\}])},$$

$$J_2(C, B) = n^{-1} \sum E_{\theta, X_i}\{\psi_{\mathrm{BI}}(Y_i, X_i, \theta, B, C) \psi_{\mathrm{BI}}^{\mathrm{T}}(Y_i, X_i, \theta, B, C)\},$$

where the sums are over $i = 1, \ldots, n$, and where $l_{i,\theta} = l(Y_i, X_i, \theta)$. Then (2·13) are, equivalently, $C = J_1(C, B)$ and $B = J_2(C, B)$. Let (C_0, B_0) be an initial guess and define recursively $C_{k+1} = J_1(C_k, B_k)$ and $B_{k+1} = J_2(C_k, B_k)$. If the sequence (C_k, B_k) converges the limiting value is $\{C(\theta), B(\theta)\}$ satisfying (2·13). Alternatively $\{C(\theta), B(\theta)\}$ can be computed using the iteration employed by Maronna (1976). Neither iteration is guaranteed to converge. Using the former method to compute one-step estimates in logistic regression the algorithm's success depended on the magnitude of the bound b; for larger choices of b no problems were encountered. However, computational difficulties did arise for small values of b. Particular choices for b in logistic regression are discussed in the next section. For the initial guess (C_0, B_0) we took $C_0 = 0$ and $B_0 = n^{-1} \sum l_{i,\theta} l_{i,\theta}^{\mathrm{T}}$.

Since robustness is our primary concern the initial estimator $\tilde{\theta}$ employed in the one-step approximation to $\hat{\theta}_{\mathrm{BI}}$ should also be resistant to outliers. In the next section an appropriate initial estimator for logistic regression is presented which is computationally simpler and has some interesting optimality properties of its own.

3. Application to logistic regression

3·1. *The logistic model*

Logistic regression is a special case of (2·1) in which Y is an indicator variable such that

$$\mathrm{pr}(Y = 1 \mid X = x) = F(x^{\mathrm{T}} \theta_0), \quad F(t) = (1 + e^{-t})^{-1}.$$

The general applicability of this form of binary regression is discussed by Berkson (1951), Cox (1970) and Efron (1975). The likelihood score is $l(y, x, \theta) = \{y - F(x^{\mathrm{T}} \theta)\} x$ and the maximum likelihood estimator is consistent and asymptotically normal with covariance matrix $V(\theta_0) = [E_{\theta_0}\{F^{(1)}(X^{\mathrm{T}} \theta_0) X X^{\mathrm{T}}\}]^{-1}$, where $F^{(1)}(t) = (d/dt) F(t)$.

3·2. *A bounded-leverage estimator for the logistic model*

The one-step approximation to $\hat{\theta}_{\mathrm{BI}}$ requires a more easily computed, robust, root-n consistent estimator to initiate the one-step procedure. Although many such estimators are possible, with some being computationally simpler than others, we solve this problem by finding an efficient estimator from among a class of estimators chosen to facilitate computation. As a result we obtain an estimator which is not only appropriate for use in the one-step procedure but which is optimal within its class, thus making it a reasonable competititor to $\hat{\theta}_{\mathrm{BI}}$. More specifically we find an optimal score function from among the class,

$$\mathcal{M} = [\psi\colon \psi(y, x, \theta) = \{y - F(x^{\mathrm{T}}\theta)\}\omega(x, \theta)],$$

where $\omega(x, \theta)$ is a $p \times 1$ vector-valued function of x and θ but not of y. The advantage, in terms of computational simplicity, of restricting attention to score functions in $\mathcal{M}$ is that condition (2·3) is automatically satisfied and it is not necessary to estimate a robust location functional.

The estimator we propose, and call a bounded-leverage estimator, corresponds to the score

$$\psi_{\mathrm{BL}} = \{y - F(x^{\mathrm{T}}\theta)\}x \min^{\frac{1}{2}}[1, b^2/\{m^2(x^{\mathrm{T}}\theta)x^{\mathrm{T}}Q^{-1}(\theta)x\}], \qquad (3{\cdot}1)$$

where $Q = Q(\theta)$ is an implicitly-defined $p \times p$ matrix function of θ satisfying

$$Q(\theta) = E_\theta(F^{(1)}(X^{\mathrm{T}}\theta)XX^{\mathrm{T}} \min[1, b^2/\{m^2(X^{\mathrm{T}}\theta)X^{\mathrm{T}}Q^{-1}X\}]), \qquad (3{\cdot}2)$$

and $m(t) = \max\{F(t), 1 - F(t)\}$. In the first author's University of North Carolina Ph.D. thesis it is shown that in order for (3·2) to possess a solution $Q > 0$, it is necessary that

$$b^2 > p/E_\theta\{F^{(1)}(X^{\mathrm{T}}\theta)/m^2(X^{\mathrm{T}}\theta)\}. \qquad (3{\cdot}3)$$

Condition (3·3) is generally not sufficient however. Note that with Q satisfying (3·2), $W_{\psi_{\mathrm{BL}}} = Q$ and, by (2·8), $\hat{\theta}_{\mathrm{BL}}$ has bounded influence.

We are able to restrict attention to scores in $\mathcal{M}$ and still obtain bounded influence simply because the absolute residual $|y - F(x^{\mathrm{T}}\theta)|$ is bounded. However, ψ_{BL} takes a pessimistic view in downweighting observations in accordance with their maximum potential influence determined by their position in the design space and by θ. The term leverage is often used to denote potential influence (Cook & Weisberg, 1983) and hence the name bounded leverage. Potential influence is often far greater than the actual influence when the observation is well fit by the model. Although downweighting such points results in a loss of efficiency for $\hat{\theta}_{\mathrm{BL}}$ this will not affect the efficiency of our one-step estimator. Also, as the following results show, ψ_{BL} is the most efficient score in $\mathcal{M}$.

THEOREM 2. *If for a given choice of $b > 0$ equation* (3·2) *possesses the solution $Q > 0$, then ψ_{BL} minimizes* $\operatorname{tr}(V_\psi V_{\mathrm{BL}}^{-1})$ *among all ψ in $\mathcal{M}$ satisfying*

$$\sup_{(y,x)} (\mathrm{IC}_\psi^{\mathrm{T}} V_{\mathrm{BL}}^{-1} \mathrm{IC}_\psi) \leq b^2.$$

With the exception of multiplication by a constant matrix, ψ_{BL} is unique almost surely.

COROLLARY 2·1. *If there exists a strongly efficient score ψ_{opt} in $\mathcal{M}$, then ψ_{opt} is equivalent to ψ_{BL} whenever the latter is defined.*

Proofs are similar to those of Theorem 1 and its corollary and will not be given. The extent to which Theorem 2 generalizes to other regression models is limited, since it requires that $l(y, x, \theta)$ be a bounded function of y.

The bounded-leverage estimator is obtained by solving $\Sigma\, \psi_{\mathrm{BL}}\{Y_i, X_i, \theta, \tilde{Q}(\theta)\} = 0$, where the sum is over $i = 1, \ldots, n$, and where ψ_{BL} is given in (3·1) and $\tilde{Q}(\theta)$ satisfies

$$\tilde{Q}(\theta) = n^{-1} \sum_{i=1}^{n} F^{(1)}(X_i^{\mathrm{T}}\theta)X_iX_i^{\mathrm{T}} \min[1, b^2/\{m^2(X_i^{\mathrm{T}}\theta)X_i^{\mathrm{T}}\tilde{Q}^{-1}(\theta)X_i\}]. \qquad (3{\cdot}4)$$

The algorithm used to compute $\tilde{\theta}_{\mathrm{BL}}$ for the example in § 3·3 is now described. Let θ_0 be an initial guess at $\tilde{\theta}_{\mathrm{BL}}$ and define recursively

$$\theta_{k+1}=\theta_k+\tilde{D}^{-1}(\theta_k)n^{-1}\sum_{i=1}^{n}\psi_{\mathrm{BL}}\{Y_i, X_i, \theta_k, \tilde{Q}(\theta_k)\}.$$

Since $\tilde{D}(\theta)$ is an approximation to $-(\partial/\partial\theta)\sum\psi_{\mathrm{BL}}\{Y_i, X_i, \theta, \tilde{Q}(\theta)\}$ this is a quasi-Newton-Raphson iteration. At each step of the iteration it is necessary to compute the matrix $\tilde{Q}(\theta_k)$ satisfying (3·4). For a fixed θ the matrix $\tilde{Q}(\theta)$ can be found as follows. Let

$$Q_0=n^{-1}\sum F^{(1)}(X_i^{\mathrm{T}}\theta)X_iX_i^{\mathrm{T}},$$

$$J(Q)=n^{-1}\sum F^{(1)}(X_i^{\mathrm{T}}\theta)X_iX_i^{\mathrm{T}}\min[1, b^2/\{m^2(X_i^{\mathrm{T}}\theta)X_i^{\mathrm{T}}Q^{-1}X_i\}],$$

where the sums are over $i=1,\ldots,n$, and define recursively $Q_{k+1}=J(Q_k)$. For fixed θ, $J(Q)$ is an increasing function of Q in the sense of positive-definiteness, i.e. for positive-definite matrices $A_1\leqslant A_2$ we have $J(A_1)\leqslant J(A_2)$. Since $Q_1=J(Q_0)\leqslant Q_0$ an inductive argument shows that the sequence (Q_k) is decreasing in the sense of positive-definiteness. As it is bounded below, the sequence necessarily converges. The limiting value is $\tilde{Q}(\theta)$ provided it is positive-definite. To specify fully the algorithm one must determine the bound b. For $\tilde{\theta}_{\mathrm{BL}}$ this was chosen as a constant multiple of $b(\tilde{\theta}_{\mathrm{BL}})$, where

$$b^2(\theta)=p\bigg/\left[n^{-1}\sum_{i=1}^{n}\{F^{(1)}(X_i^{\mathrm{T}}\theta)/m^2(X_i^{\mathrm{T}}\theta)\}\right];$$

see (3·3). For the example in § 3·3 we took the bound to be $\frac{3}{2}b(\tilde{\theta}_{\mathrm{BL}})$; this same bound was then used for the one-step estimator $\hat{\theta}_{\mathrm{BI}}^{(1)}$. The choice $\frac{3}{2}b(\theta)$ was suggested by experience; it is sufficiently small to provide protection from extreme observations yet large enough to avoid computational problems.

For the one-step construction in § 2·4 to work it is necessary that $\tilde{\theta}_{\mathrm{BL}}$ be root-n consistent. In the first author's Ph.D. thesis it is shown that $n^{\frac{1}{2}}(\tilde{\theta}_{\mathrm{BL}}-\theta_0)$ is asymptotically normal with covariance matrix $V_{\mathrm{BL}}(\theta_0)=D_{\mathrm{BL}}^{-1}(\theta_0)Q(\theta_0)\{D_{\mathrm{BL}}^{-1}(\theta_0)\}^{\mathrm{T}}$ provided:

(i) b is sufficiently large;

(ii) $E(\|X\|^2)<\infty$;

(iii) $E\{F^{(1)}(X^{\mathrm{T}}\theta)XX^{\mathrm{T}}\|\mathrm{X}\|^{-1}\}$ is positive-definite;

(iv) $(\partial/\partial Q)E\{J(X, \theta, Q)\}$ is nonsingular where

$$J(X, \theta, Q)=Q-F^{(1)}(X^{\mathrm{T}}\theta)XX^{\mathrm{T}}\min[1, b^2/\{m^2(X^{\mathrm{T}}\theta)X^{\mathrm{T}}Q^{-1}X\}].$$

The key assumptions are (iii) and (iv) which are similar to Assumption 7 of Krasker & Welsch (1982).

As an estimate of V_{BL} we use $\tilde{V}=\tilde{D}^{-1}(\tilde{\theta}_{\mathrm{BL}})\tilde{Q}(\tilde{\theta}_{\mathrm{BL}})\{\tilde{D}^{-1}(\tilde{\theta}_{\mathrm{BL}})\}^{\mathrm{T}}$, where

$$\tilde{D}(\theta)=n^{-1}\sum_{i=1}^{n}E_{\theta,X_i}[\psi_{\mathrm{BL}}\{Y_i, X_i, \theta, \tilde{Q}(\theta)\}l^{\mathrm{T}}(Y_i, X_i, \theta)].$$

3·3. *Example*

We now apply our results to fit a model relating participation in the U.S. Food Stamp Program to various socioeconomic indicators. The 150 observations used in our analyses were randomly selected from a data set containing information on over 2000 elderly citizens. The larger data set is part of the 1977–78 Nationwide Food Consumption Survey; see Rizek (1978) for a discussion of the data collection procedure. The covariates we selected for study are: tenancy, indicating home ownership; supplemental income, indicating whether some form of supplemental security income is received; and monthly income. In our sample of 150 there were 24 cases of participation.

The researcher who provided these data to us had been using probit regression with monthly income entering linearly in the model. We first fit the logistic model with covariates tenancy, supplemental income and (monthly income)/10. The maximum likelihood, $\hat{\theta}_{\mathrm{ML}}$, bounded-leverage, $\tilde{\theta}_{\mathrm{BL}}$, and one-step bounded-influence, $\hat{\theta}^{(1)}_{\mathrm{BI}}$, estimates for this model appear in Table 1(a). The one-step estimate was obtained using $\tilde{\theta}_{\mathrm{BL}}$ to initiate the one-step procedure; see § 2·4.

Both $\tilde{\theta}_{\mathrm{BL}}$ and $\hat{\theta}_{\mathrm{BI}}$ are similar to weighted maximum likelihood estimators with data-dependent weights; see (3·1) and (2·9). Empirical weights obtained in the course of computing $\tilde{\theta}_{\mathrm{BL}}$ and $\hat{\theta}^{(1)}_{\mathrm{BI}}$ can be used as diagnostics to identify extreme or ill-fitting observations. For $\hat{\theta}^{(1)}_{\mathrm{BI}}$ empirical weights less than one indicate influential observations while for $\tilde{\theta}_{\mathrm{BL}}$ the corresponding weights indicate potentially influential points. Since potentially influential points need not be influential, and this is particularly true in logistic regression due to the discrete response variable, the bounded-influence weights are generally more informative. For the analysis in Table 1(a) the only bounded-influence weights less than one were $\omega_{40}=0{\cdot}69$, $\omega_{66}=0{\cdot}40$, $\omega_{95}=0{\cdot}98$ and $\omega_{109}=0{\cdot}62$. The bounded-leverage weights for these same four observations were $\omega_{40}=0{\cdot}68$, $\omega_{66}=0{\cdot}39$, $\omega_{95}=0{\cdot}96$ and $\omega_{109}=0{\cdot}61$, which explains the similarity of $\hat{\theta}^{(1)}_{\mathrm{BI}}$ and $\tilde{\theta}_{\mathrm{BL}}$. There were many more bounded-leverage weights less than one, although again, these corresponded to observations which were well fitted by the model; thus downweighting these points had very little effect on the fit.

Table 1 (a). *Estimates for the logistic regression model with covariates tenancy, supplemental income, and (monthly income)/10; p-values in parentheses.*

	Intercept	Tenancy	Supplemental income	(Monthly income)/10
$\hat{\theta}_{\mathrm{ML}}$	−0·34 (0·5287)	−1·76 (0·0009)	0·78 (0·1259)	−0·01 (0·1122)
$\tilde{\theta}_{\mathrm{BL}}$	−0·16 (0·7872)	−1·75 (0·0014)	0·77 (0·1360)	−0·02 (0·0826)
$\hat{\theta}^{(1)}_{\mathrm{BI}}$	−0·20 (0·6006)	−1·76 (0·0012)	0·78 (0·1300)	−0·02 (0·0922)

Table 1 (b). *Estimates for the logistic regression model with covariates tenancy, supplemental income, and log (monthly income + 1); p-values in parentheses.*

	Intercept	Tenancy	Supplemental income	log (monthly income + 1)
$\hat{\theta}_{\mathrm{ML}}$	0·93 (0·5681)	−1·85 (0·0005)	0·90 (0.0737)	−0·33 (0·2228)
$\tilde{\theta}_{\mathrm{BL}}$	4·14 (0·1030)	−1·81 (0·0007)	0·75 (0·1444)	−0·86 (0·0430)
$\hat{\theta}^{(1)}_{\mathrm{BI}}$	4·02 (0·1100)	−1·81 (0·0006)	0·76 (0·1416)	−0·84 (0·0465)
$\hat{\theta}^{*}_{\mathrm{ML}}$	6·88 (0·0160)	−2·02 (0·0004)	0·76 (0·1586)	−1·33 (0·0062)
$\hat{\theta}^{**}_{\mathrm{ML}}$	6·29 (0·0374)	−1·96 (0·0007)	0·75 (0·1612)	−1·23 (0·0169)

* With cases 5 and 66 removed.

** With cases 5, 66 and thirty additional points removed.

Since all four observations with bounded-influence weights less than one correspond to the four largest incomes among those receiving food stamps a transformation of income is indicated.

Table 1(b) gives the analysis with log (monthly income + 1) replacing (monthly income)/10. This transformation substantially reduces the leverage of large income values but increases the leverage of small income values. For this model the bounded-influence estimator downweighted only two observations with $\omega_5 = 0{\cdot}19$ and $\omega_{66} = 0{\cdot}70$ and the bounded-leverage weights for the same two points were $\omega_5 = 0{\cdot}19$ and $\omega_{66} = 0{\cdot}69$ explaining the similarity of the estimates. As above, there were many more bounded-leverage weights less than one, although again, these corresponded to observations which were well fitted by the model and thus downweighting them had little effect on the fit. Case 66 has the largest income among those participating while case 5 has the smallest income among those not participating. Apparently cases 5 and 66 are influencing the maximum likelihood fit; this is indicated to a great extent by the bounded influence analysis and even more so by the maximum likelihood fit with the two outlying cases removed.

The estimates and p-values for the coefficients of tenancy and supplemental income are not altered much by the removal of cases 5 and 66 although the opposite is true of the estimate and p-value for the coefficient of log (monthly income + 1). Together cases 5 and 66 work to mask the significance of income as a predictor of participation.

An advantage of robust methods over maximum likelihood is that residual plots are more reliable for uncovering outliers. This is illustrated in Fig. 1. Standardized residuals (Cox, 1970, p. 96; Pregibon, 1981) are plotted for both the maximum likelihood and bounded-influence fits; residuals from the bounded-leverage fit are similar to those from the bounded-influence fit and are not plotted.

In this example the bounded-influence and bounded-leverage estimates are similar and our experience suggests that this is not uncommon at least for logistic regression. The similarity appears to arise because both estimators downweight influential points more

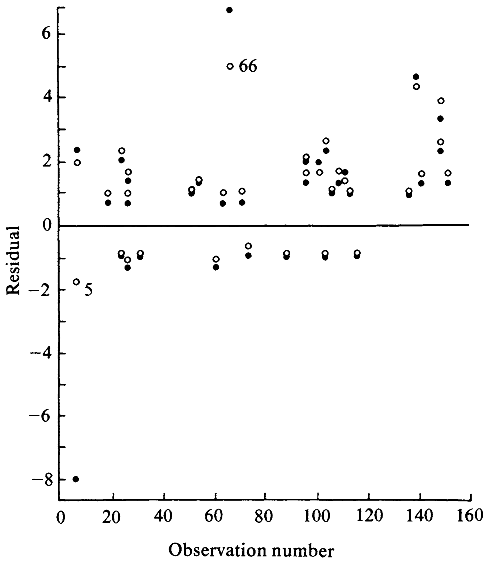


Fig. 1. Residual plots for food stamp data. Maximum likelihood residuals indicated by open circles; residuals from bounded-influence fit by darkened circles. Both residuals defined as by Cox (1970, p. 96). Negligible residuals omitted for clarity.

or less equally and, although the bounded-leverage estimator downweights many more points, these are typically ones which are very well fitted by the model and thus contribute little to its determination, i.e. they are noninfluential. As evidence for this claim we computed the maximum likelihood estimate for the model with log (monthly income +1) after removing from the data observations 5, 66 and thirty additional points corresponding to those observations with the thirty smallest absolute residuals, $|y - F(x^{\mathrm{T}}\hat{\theta})|$, from the model fitted by maximum likelihood with only cases 5 and 66 removed. The estimated coefficients appear in the bottom row of Table 1(b). The similarity of $\hat{\theta}^{*}_{\mathrm{ML}}$ and $\hat{\theta}^{**}_{\mathrm{ML}}$ is remarkable considering that the latter estimate is based on over one-fifth fewer observations. Apparently the thirty observations removed in the latter fit are noninfluential. These points correspond roughly to the majority of those downweighted by the bounded-leverage estimate.

4. CONCLUSIONS

Our bounded-influence and bounded-leverage procedures provide methods of fitting meaningful models in the presence of anomalous data. The bounded-leverage estimator was originally intended only as a starting value for the more efficient bounded-influence estimator. Our experience with data suggests that the two estimators may be rather similar in practice for logistic regression. However, until more experience is accumulated, it is premature to recommend exclusive use of the bounded-leverage estimator.

Robust procedures also supply useful diagnostic tools for model building. Variable selection, as well as estimation, can be influenced by anomalous data; Pregibon (1982) cites such an example. Often robust methods suggest variables appropriate for modelling the bulk of the data which would otherwise go undetected in a standard maximum likelihood analysis. Conversely, with nonresistant fitting, a variable might be used in the model simply to accommodate a single outlier. In addition to variable selection, the weights and residuals from a robust fit provide useful supplements to more direct diagnostics. For example, with the food stamp data, an analyst, seeing the impact of case five, might question the validity of that observation or the appropriateness of the model over the full range of incomes.

ACKNOWLEDGEMENTS

The research of the first two authors was supported by the Air Force Office of Scientific Research while that of the third author was supported by the National Science Foundation. We also acknowledge several helpful comments from the referees.

APPENDIX

Proofs of Theorem 1 *and Corollary* 1

Theorem 1 is a generalization of Appendix A of Hampel (1978) and the proof given here uses techniques given by Krasker (1980). To prove Theorem 1, let ψ be any competitor to ψ_{BI}. Without loss of generality assume that $\psi = \mathrm{IC}_{\psi}$, that is that ψ is in canonical form in the sense of Hampel (1974). This is equivalent to assuming

$$E_{\theta}\{\psi(Y, X, \theta) l^{\mathrm{T}}(Y, X, \theta)\} = I, \qquad (\mathrm{A}{\cdot}1)$$

and implies $V_{\psi}(\theta) = E_{\theta}\{\psi(Y, X, \theta)\psi^{\mathrm{T}}(Y, X, \theta)\}$. Now write l for $l(Y, X, \theta)$ and ψ for $\psi(Y, X, \theta)$.

If ψ satisfies (A·1) and (2·3) then

$$E_\theta[\{D_{\mathrm{BI}}^{-1}(l-C)-\psi\}\{D_{\mathrm{BI}}^{-1}(l-C)-\psi\}^{\mathrm{T}}]=D_{\mathrm{BI}}^{-1}E_\theta\{(l-C)(l-C)^{\mathrm{T}}\}(D_{\mathrm{BI}}^{-1})^{\mathrm{T}}-D_{\mathrm{BI}}^{-1}-(D_{\mathrm{BI}}^{-1})^{\mathrm{T}}+V_\psi(\theta).$$

Therefore $\operatorname{tr}(V_\psi V_{\mathrm{BI}}^{-1})$ is, neglecting an additive constant independent of ψ, proportional to

$$E_\theta[\{D_{\mathrm{BI}}^{-1}(l-C)-\psi\}^{\mathrm{T}}V_{\mathrm{BI}}^{-1}\{D_{\mathrm{BI}}^{-1}(l-C)-\psi\}]. \qquad (\mathrm{A}{\cdot}2)$$

Define $\phi = V_{\mathrm{BI}}^{-\frac{1}{2}}\psi$; in terms of ϕ, expression (A·2) becomes

$$E_\theta\{\|\phi-V_{\mathrm{BI}}^{-\frac{1}{2}}D_{\mathrm{BI}}^{-1}(l-C)\|^2\}. \qquad (\mathrm{A}{\cdot}3)$$

Note that $\|\phi\|^2=\psi^{\mathrm{T}}V_{\mathrm{BI}}^{-1}\psi$ and thus, subject to (2·11), equation (A·3) is minimized, as a function of ϕ, by

$$\phi=V_{\mathrm{BI}}^{-\frac{1}{2}}D_{\mathrm{BI}}^{-1}(l-C)\min\nolimits^{\frac{1}{2}}[1, b^2/\{(l-C)^{\mathrm{T}}D_{\mathrm{BI}}^{-1}V_{\mathrm{BI}}^{-1}(D_{\mathrm{BI}}^{-1})^{\mathrm{T}}(l-C)\}]. \qquad (\mathrm{A}{\cdot}4)$$

Condition (A·1) ensures that ϕ is unique almost surely. Equations (2·4), (2·6), (2·7) and (2·10) imply $D_{\mathrm{BI}}^{-1}V_{\mathrm{BI}}^{-1}(D_{\mathrm{BI}}^{-1})^{\mathrm{T}}=B^{-1}$; thus in terms of ψ, (A·4) becomes $\psi=D_{\mathrm{BI}}^{-1}\psi_{\mathrm{BI}}$. □

To prove Corollary 1·1, again assume that all scores are in canonical form and satisfy (2·3). Define

$$\mathscr{S}=\{\psi\colon \sup_{(y,x)}\psi^{\mathrm{T}}V_\psi^{-1}\psi\leq b^2\},\quad \mathscr{S}_{\mathrm{BI}}=\{\psi\colon \sup_{(y,x)}\psi^{\mathrm{T}}V_{\mathrm{BI}}^{-1}\psi\leq b^2\}.$$

We must show that if there exists ψ_{opt} in $\mathscr{S}$ such that $V_{\psi_{\mathrm{opt}}}\leq V_\psi$ for all ψ in $\mathscr{S}$, then ψ_{opt} is equivalent to $D_{\mathrm{BI}}^{-1}\psi_{\mathrm{BI}}$. Clearly $D_{\mathrm{BI}}^{-1}\psi_{\mathrm{BI}}$ is in $\mathscr{S}$; thus by assumption $V_{\psi_{\mathrm{opt}}}\leq V_{\mathrm{BI}}$. From this it follows that

$$\psi_{\mathrm{opt}}^{\mathrm{T}}V_{\mathrm{BI}}^{-1}\psi_{\mathrm{opt}}\leq\psi_{\mathrm{opt}}^{\mathrm{T}}V_{\psi_{\mathrm{opt}}}^{-1}\psi_{\mathrm{opt}}\leq b^2,$$

and hence ψ_{opt} is in $\mathscr{S}_{\mathrm{BI}}$. Let $I=\mathscr{S}\cap\mathscr{S}_{\mathrm{BI}}$. The set I is nonempty; it contains $D_{\mathrm{BI}}^{-1}\psi_{\mathrm{BI}}$ and ψ_{opt}. For any ψ in I we know $V_{\psi_{\mathrm{opt}}}\leq V_\psi$ and hence

$$\operatorname{tr}(V_{\psi_{\mathrm{opt}}}V_{\mathrm{BI}}^{-1})\leq\operatorname{tr}(V_\psi V_{\mathrm{BI}}^{-1})$$

for all ψ in I. But Theorem 1 proves that $D_{\mathrm{BI}}^{-1}\psi_{\mathrm{BI}}$, when defined, is the almost everwhere unique minimizer of $\operatorname{tr}(V_\psi V_{\mathrm{BI}}^{-1})$ among all ψ in I. The equivalence of ψ_{opt} and ψ_{BI} follows. □

References

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics* **7**, 327-39.

Bickel, P. A. (1975). One-step Huber estimates in the linear model. *J. Am. Statist. Assoc.* **70**, 428-34.

Cook, D. R. & Weisberg, S. (1983). Comment on paper by P. J. Huber "Minimax aspects of bounded-influence regression". *J. Am. Statist. Assoc.* **78**, 74-5.

Cox, D. R. (1970). *Analysis of Binary Data.* London: Methuen.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Statist. Assoc.* **70**, 892-8.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69**, 383-94.

Hampel, F. R. (1978). Optimally bounding the Gross-Error-Sensitivity and the influence of position in factor space. In *Proc. Statist. Comp. Sect., Am. Statist. Assoc.*, pp. 59-64.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proc. 5th Berkeley Symp.* **1**, 221-33.

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika* **72**, 59-66.

Krasker, W. S. (1980). Estimation in linear regression models with disparate data points. *Econometrica* **48**, 1333-46.

Krasker, W. S. & Welsch, R. E. (1982). Efficient bounded influence regression estimation using alternative definitions of sensitivity. *J. Am. Statist. Assoc.* **77**, 595-605.

Maronna, R. A. (1976). Robust *M*-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51-67.

McCullagh, P. & Nelder, J. A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705-24.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38**, 485-98.

RIZEK, R. L. (1978). The 1977-78 Nationwide Food Consumption Survey. *Family Econ. Rev.* Fall, 3-7.
RUPPERT, D. (1985). On the bounded-influence regression estimator of Krasker and Welsch. *J. Am. Statist. Assoc.* **80**, 205-8.

[*Received July* 1984. *Revised October* 1985]

Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, With Applications to Generalized Linear Models

HANS R. KÜNSCH, LEONARD A. STEFANSKI, and RAYMOND J. CARROLL*

In this article robust estimation in generalized linear models for the dependence of a response y on an explanatory variable x is studied. A subclass of the class of M estimators is defined by imposing the restriction that the score function must be conditionally unbiased, given x. Within this class of conditionally Fisher-consistent estimators, optimal bounded-influence estimators of regression parameters are identified, and their asymptotic properties are studied. The estimators studied in this article and the efficient bounded-influence estimators studied by Stefanski, Carroll, and Ruppert (1986) depend on an auxiliary centering constant and nuisance matrix. The centering constant can be given explicitly for the conditionally Fisher-consistent estimators, and thus they are easier to compute than the estimators studied by Stefanski et al. (1986). In addition, estimation of the nuisance matrix has no effect on the asymptotic distribution of the conditionally Fisher-consistent estimators; the same is not true of the estimators studied by Stefanski et al. (1986). Logistic regression is studied in detail. The nature of influential observations in logistic regression is discussed, and two data sets are used to illustrate the methods proposed.

KEY WORDS: Asymptotic bias; Bounded influence; Breakdown point; Generalized linear models; Linear models; Linear regression; Logistic regression; Robust regression.

1. INTRODUCTION

The basic generalized regression model states that, given the values of a p-dimensional explanatory variable x, the response y has a distribution function $P_\theta(y \mid x)$. We are interested in estimating the parameter θ from N independent observations (y_i, x_i). In such a general model, M estimators are defined implicitly by an equation of the form

$$\sum_{i=1}^{N} \psi(y_i, x_i, \hat{\theta}_N) = 0. \tag{1.1}$$

In Equation (1.1), θ and ψ have the same dimension p. Of course, maximum likelihood estimators are (nonrobust) M estimators. Assume that x is also a random variable with distribution function F. For $\hat{\theta}_N$ to be consistent, standard theory requires that the estimating equation (1.1) be unbiased; that is,

$$E_\theta(\psi(y, x, \theta)) = \int\int \psi(y, x, \theta)P_\theta(dy \mid x)F(dx) = 0 \qquad \text{for all } \theta. \quad (1.2)$$

Requiring (1.2) is the same as saying that ψ is Fisher-consistent. Certainly, Fisher consistency is a minimal requirement, but in linear and generalized linear regression it is too weak and even unpalatable because it involves the distribution of the predictors x. In a regression context, the x_i may not be random variables. Furthermore, when the x_i are random it is customary to condition on the observed values of x. We say that an M estimator is *conditionally Fisher-consistent* if it satisfies

$$E_\theta(\psi(y, x, \theta) \mid x) = \int\int \psi(y, x, \theta)P_\theta(dy \mid x) = 0 \qquad \text{for all } \theta \text{ and } x. \quad (1.3)$$

In linear and generalized linear regression, maximum likelihood estimators are conditionally Fisher-consistent whenever the distribution of x does not depend on θ.

Conditional Fisher consistency is an appealing concept because it does not depend on the x's being random, and even if they are, it does not involve the distribution of the x's. This does not mean that the properties of conditionally Fisher-consistent M estimators are independent of the design, of course; simply remember the formula for the covariance of least squares estimates. Similarly, the influence function and the sensitivity defined in the following depend on the design.

In the linear model with symmetric errors, essentially all M estimators in the literature (including least squares) satisfy (1.3). Nevertheless, it can be shown that there are some bounded-influence M estimators that are not conditionally Fisher-consistent when the errors are asymmetric. This is particularly true of the class popularly known as Schweppe-type estimators. The Mallows-type estimates, including ordinary M estimators, are Fisher-consistent. For the definition of these two types, see Krasker and Welsch (1982) or Hampel, Ronchetti, Rousseeuw, and Stahel (1986, pp. 315–316).

In generalized linear models, it is also possible to define Schweppe- and Mallows-type estimators [see Stefanski, Carroll, and Ruppert (1986) for the former and Pregibon (1981) for the latter (in logistic regression)]. The optimal robust estimators of Stefanski et al. (1986) are not con-

* Hans R. Künsch is Associate Professor, Seminar für Statistik, ETH-Zentrum, CH-8092, Zürich, Switzerland. Leonard A. Stefanski is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Raymond J. Carroll is Professor and Head, Department of Statistics, Texas A&M University, College Station, TX 77843. The work of Stefanski was supported by the National Science Foundation. The work of Carroll was supported by the Air Force Office of Scientific Research. The authors thank two referees for their helpful and thought-provoking comments, which improved the clarity of presentation.

Journal of the American Statistical Association
June 1989, Vol. 84, No. 406, Theory and Methods

ditionally Fisher-consistent, although they do satisfy (1.2). These estimators are difficult to compute, and even when computable their asymptotic distributions are difficult to understand because of a dependence on an auxiliary nuisance matrix B. In Sections 2 and 3, we study optimal robust conditionally Fisher-consistent estimates for generalized linear models. There is a practical payoff to restriction on this narrower class. In contrast to the estimates of Stefanski et al. (1986), our estimates are relatively easy to compute, and the asymptotic distribution theory is straightforward. In Section 5, we apply our methods to two data sets involving logistic regression.

We now review some general results and definitions from robust statistics (see Hampel et al. 1986). The influence function of an M estimator is

$$IC_\psi(y, x, \theta) = D(\psi, \theta)^{-1}\psi(y, x, \theta), \qquad (1.4)$$

where

$$D(\psi, \theta) = -\frac{\partial}{\partial\beta}\int\int \psi(y, x, \beta)P_\theta(dy \mid x)F(dx)\,|_{\beta=\theta}. \qquad (1.5)$$

The influence function measures the effect of an infinitesimal contamination at (y, x), standardized by the mass of the contamination. It thus gives an approximation to the effect of the inclusion or deletion of a single observation. Moreover, it gives the asymptotic covariance matrix. Under regularity conditions, $N^{1/2}(\hat{\theta}_N - \theta)$ is asymptotically normally distributed, with mean 0 and covariance matrix $V(\psi, \theta)$ [also written $V(\psi)$]:

$$\begin{aligned} V(\psi, \theta) &= E_\theta[IC_\psi(y, x, \theta)IC_\psi(y, x, \theta)^T] \\ &= D(\psi, \theta)^{-1}W(\psi, \theta)D(\psi, \theta)^{-T}, \qquad (1.6) \end{aligned}$$

where

$$W(\psi, \theta) = E_\theta[\psi(y, x, \theta)\psi(y, x, \theta)^T]. \qquad (1.7)$$

The main idea of bounded-influence estimation is to impose a bound on the influence function (1.4), and then find an estimator that has small variance subject to a chosen bound. The usual operational difficulty is that the influence function is a vector, and we need to reduce this to a scalar measure. This scalar is called the *sensitivity* of the influence function. This problem is not too different from what happens with regression diagnostics. For example, consider case-deletion diagnostics, where we try to understand the effect of deleting an observation on the parameter estimates. Changes in parameter estimates are necessarily p-dimensional. The beauty of a diagnostic such as Cook's distance (Cook and Weisberg 1982) is that the p-dimensional change in the parameters is summarized by a scalar. The same issue arises in bounded-influence regression, and as in deletion diagnostics we have to decide on a summary measure. Just as we can define diagnostics such as Cook's distance or DFFITS, we can define different methods of measuring the sensitivity of the influence function. The most common method, used with success by Krasker and Welsch (1982), is the *self-standardized sensitivity,* defined as

$$\begin{aligned} s(\psi)^2 &= \sup_{y,x}\sup_{\lambda\neq 0}\frac{(\lambda^T IC_\psi)^2}{\lambda^T V(\psi)\lambda} \\ &= \sup_{y,x}\psi(y, x, \theta)^T W(\psi, \theta)^{-1}\psi(y, x, \theta). \qquad (1.8) \end{aligned}$$

This definition of sensitivity measures the maximum influence an observation can have on a linear combination of parameters, with a standardization by the asymptotic standard deviation of this linear combination. Integrating (1.8) and taking the trace shows that $s(\psi)^2 \geq p$. Other measures of sensitivity were considered by Hampel et al. (1986, p. 317) and Giltinan, Carroll, and Ruppert (1986). For example, the latter authors considered bounding the influence on predicted values rather than the parameter estimates. The estimators resulting from the definition of sensitivity used by Giltinan et al. (1986) tend to downweight suspect observations much more severely than those considered here.

A referee asked about the meaning of the parameter θ if deviations from the model are considered. For instance, in logistic regression one cannot have errors with fatter tails, but this is not the only deviation robustness protects for. Whenever the true distribution is a *small* deviation from the parametric model with $\theta = \theta_0$, then the robust estimator is asymptotically close to that value θ_0. We may consider a gross error model with the amount of contamination depending on the regressor x: The conditional distribution of y (given x) is $P_\theta(dy \mid x)$ with probability $1 - \varepsilon(x)$ and arbitrary with probability $\varepsilon(x)$. This is a small deviation if the total proportion of contamination $E[\varepsilon(x)]$ is small. It also makes sense for logistic regression, and the meaning of the parameter is clear.

A somewhat different concept is the requirement that the estimators should not change much if a few observations are included or deleted. This is clearly desirable for any type of data analysis. Because of the interpretation of the influence function given previously, our robust estimators should be less affected by the inclusion or deletion of a few observations than the classical ones. A more rigorous investigation of this point is given in Section 4.

2. GENERALIZED LINEAR MODELS

We consider a generalized linear model with canonical link function

$$P_\theta(dy \mid x) = \exp\{yx^T\theta - G(x^T\theta) - S(y)\}\mu(dy) \qquad (2.1)$$

(see McCullagh and Nelder 1983). If g is the derivative of G, the likelihood score function is

$$l(y, x, \theta) = \{y - g(x^T\theta)\}x. \qquad (2.2)$$

Note that (2.2) satisfies (1.3), so the score is conditionally unbiased. Because l is proportional to x, the influence is unbounded; that is, $s(l) = \infty$.

We are looking for M estimators satisfying (1.3), and $s(\psi) \leq b$ that minimize $V(\psi)$ in some sense. Motivated by a general principle for constructing optimal robust estimators satisfying (1.2) (Hampel et al. 1986, sec. 4.3a), we consider the following score function:

$$\psi_{\text{cond}}(y, x, \theta, B) = d(y, x, \theta, B) w_b(|d(y, x, \theta, B)|(x^T B^{-1} x)^{1/2}) x, \quad (2.3)$$

where $d(y, x, \theta, B) = y - g(x^T\theta) - c(x^T\theta, b/(x^T B^{-1} x)^{1/2})$, and $w_b(a) = H_b(a)/a$, where H_b is the Huber function $H_b(a) = \max(-b, \min(a, b))$.

We work within the context of the Schweppe-type, although related results are obtainable for the Mallows-type (as in Stefanski et al. 1986). The major difference is that w_b in (2.3) factors into two parts. The first depends only on x and is of the form $w_1((x^T B^{-1} x)^{1/2})$. The other depends solely on

$$d(y, x, \theta) = y - g(x^T\theta) - c(x^T\theta, b/(x^T B^{-1} x)^{1/2})$$

and has the form $w_2(|d(y, x, \theta)|)$.

The scalar function c and the matrix B in (2.3) are chosen so that the side conditions (1.3) and $s(\psi_{\text{cond}}) = b$ are satisfied. By the definition of ψ_{cond}, (1.3) holds iff for all β and $a > 0$,

$$\int (y - g(\beta) - c(\beta, a)) w_a(|y - g(\beta) - c(\beta, a)|) \times \exp(y\beta - G(\beta) - S(y))\mu(dy) = 0. \quad (2.4)$$

First, we discuss the existence of a solution to (2.4).

Lemma 2.1. For any $a > 0$ and β, there is a solution $c = c(\beta, a)$ to (2.4).

Proof. For fixed y, β, and a, the function $c \to (y - g(\beta) - c) w_a(|y - g(\beta) - c|)$ is continuous, bounded, and monotone-nonincreasing, with limits $\pm a$. Hence the existence follows from dominated convergence and the intermediate value theorem.

A practical advantage here is that often the function c can be calculated in closed form. This is particularly important compared to the optimal ψ satisfying (1.2), where c is a vector (depending only on θ) whose computation is quite difficult (see Stefanski et al. 1986, sec. 2.4). In the following examples $c(\beta, a)$ can be calculated explicitly.

Example 2.1: Logistic Regression. In this case μ puts equal mass at 0 and 1, $S(y) = 0$, and $G(\beta) = \log\{1 + \exp(\beta)\}$. Write $F(\beta) = \exp(\beta)/(1 + \exp(\beta))$ and $\bar{F}(\beta) = 1 - F(\beta)$. It is easily checked that

$$\begin{aligned} c(\beta, a) &= aF(\beta)/\bar{F}(\beta) - F(\beta) && \text{if } \beta < 0, a < \bar{F}(\beta) \\ &= \bar{F}(\beta) - a\bar{F}(\beta)/F(\beta) && \text{if } \beta > 0, a < F(\beta) \\ &= 0 && \text{otherwise} \end{aligned}$$

satisfies (2.4).

Example 2.2: Negative Exponential Regression. In this case μ is Lebesgue measure on $[0, \infty)$, $G(\beta) = -\log(-\beta)$, $S(y) = 0$, and $\beta < 0$. Two cases occur: If the bound is large, the Huberization in ψ_{cond} is one-sided (for large y's only), whereas for small a's both large and small y's are Huberized. It can be checked by straightforward calculations that the cut point between the two cases is given by the equation $e^{2\beta a} = 1 + \beta a$, so $\beta a \approx -.797$. In the former case $c(\beta, a) = -\beta^{-1}$ times the smaller solution of $\exp(x + \beta a - 1) = x$, and in the latter case $c(\beta, a) = -\beta^{-1}(1 + \log(\beta a/(\exp(\beta a) - \exp(-\beta a))))$.

Turn now to the matrix B. First, note that the estimator is conditionally Fisher-consistent and has bounded influence for any choice of B. If we want $s(\psi_{\text{cond}}) = b$, however, B depends on both the design and θ. In linear regression B depends on the design, but not on θ. It follows from the definition of ψ_{cond} that $s(\psi_{\text{cond}}) = b$, provided

$$E_\theta[\psi_{\text{cond}}(y, x, \theta, B)\psi_{\text{cond}}(y, x, \theta, B)^T] = B. \quad (2.5)$$

Equation (2.5) is used to define $B = B(\theta, F)$. Because $s(\psi)^2 \geq p$, a necessary condition for (2.5) to have a solution is $b^2 \geq p$, but we do not know if it is also sufficient.

The estimators we have defined are intuitively appealing because they downweight observations according to their leverage and "outlyingness." It is reasonable to ask if they satisfy any optimality criterion. The discussion of optimality within a bounded-influence class started with Krasker and Welsch (1982), but the results of Ruppert (1985) suggest that there is no estimator that has uniformly smallest covariance subject to a bound on the influence. The best-known optimality result seems to be that of Stefanski et al. (1986). It is not completely satisfactory, because the criterion to be minimized depends on the solution. Nevertheless, it implies that no other estimator satisfying the same bound on $s(\psi)$ can have a uniformly smaller covariance. We can achieve the same optimality result within the class of conditionally Fisher-consistent estimates. We state this in the following theorem.

Theorem 2.1. Suppose that for a given b, (2.5) has solution $B(\theta)$. Then, ψ_{cond} minimizes $\text{tr}\{V(\psi)V(\psi_{\text{cond}})^{-1}\}$ among all ψ that satisfy both (1.3) and $\sup_{y,x} IC_\psi^T V(\psi_{\text{cond}})^{-1} IC_\psi \leq b^2$.

Theorem 2.1 is a corollary of the following analog to theorem 1 of Stefanski et al. (1986). Note that the following theorem applies to any kind of model with explanatory variables.

Theorem 2.2. Let $l(y, x, \theta)$ be the likelihood score function. Define the score function as

$$\psi_{\text{cond}}(y, x, \theta) = (l - c)\min(1, b/\{(l - c)^T B^{-1}(l - c)\}^{1/2}), \quad (2.6)$$

where $c = c(x, \theta)$ and $B = B(\theta)$ are assumed to exist and satisfy $E(\psi_{\text{cond}}(y, x, \theta) \mid x) = 0$ and $E\{\psi_{\text{cond}}(y, x, \theta)\psi_{\text{cond}}(y, x, \theta)^T\} = B$. Then, (3.6) minimizes $\text{tr}\{V(\psi) V(\psi_{\text{cond}})^{-1}\}$ among all ψ satisfying (1.3) and $\sup_{(y,x)} IC_\psi V(\psi_{\text{cond}})^{-1} IC_\psi \leq b^2$. With the exception of multiplication by a constant matrix, ψ_{cond} is unique almost surely.

Proof. The proof is almost identical to that of theorem 1 in Stefanski et al. (1986), once one notes that for any conditionally unbiased score function ψ, $E[c(x, \theta)\psi(y, x, \theta)] = E[c(x, \theta)E(\psi(y, x, \theta) \mid x)] = 0$.

The computational simplicity of the conditional Fisher-consistent estimator is not particular to the canonical model (2.1). For instance, consider a generalized linear model with arbitrary link function h; that is, $x^T\theta$ in (2.1) is replaced by $h(x^T\theta)$. Then, we have to replace $d(y, x, \theta, B)$ in (2.3) with

$$h'(x^T\theta)\{y - g(h(x^T\theta)) \\ - c(h(x^T\theta), b/((x^TB^{-1}x)^{1/2}|h'(x^T\theta)|))\},$$

where $c(\beta, a)$ is still defined by (2.4).

In applications, the distribution F of the $\{x_i\}$ is unknown. It is common to replace F by its empirical distribution. From (2.3) and (2.5), this means that we solve

$$\sum_{i=1}^{N} \psi_{\text{cond}}(y_i, x_i, \hat{\theta}_N, \hat{B}_N) = 0 \tag{2.7}$$

and

$$N^{-1} \sum_{i=1}^{N} x_i x_i^T v(x_i^T\hat{\theta}_N, b/(x_i^T\hat{B}_N^{-1}x_i)^{1/2}) = \hat{B}_N, \tag{2.8}$$

where

$$v(\beta, a) = \int (y - g(\beta) - c(\beta, a))^2 w^2(y, \beta, a) \\ \times \exp(y\beta - G(\beta) - S(y))\mu(dy) \tag{2.9}$$

and

$$w(y, \beta, a) = \min(1, a/|y - g(\beta) - c(\beta, a)|). \tag{2.10}$$

In many applications, one wants to reject extreme outliers completely. This can be done by replacing the Huber function H_b in (2.3), (2.4), and (2.10) with any of the redescenders, such as Hampel's three-part function or the Tukey biweight. The calculation of $c(\beta, a)$ is of the same complexity as it is with the Huber function.

3. THE EFFECT OF ESTIMATING THE MATRIX B

In Section 2 we derived the estimator defined by (2.7) and (2.8) as an approximation to the optimal estimator that uses $\psi_{\text{cond}}(y, x, \theta, B(\theta))$. We consider (2.7) and (2.8) an M estimator for both θ and a nuisance parameter B. The ψ function defining this M estimator is $(\psi_{\text{cond}}(y, x, \theta, B)^T, \chi(x, \theta, B)^T)^T$, where $\chi(x, \theta, B) = xx^Tv(x^T\theta, b/(x^TB^{-1}x)^{1/2}) - B$. The influence function of this estimator is [compare (1.4) and (1.5)]

$$IC_{\psi,\chi}(y, x, \theta, B) \\ = D_{\psi,\chi}(\theta)^{-1}(\psi_{\text{cond}}(y, x, \theta, B)^T, \chi(x, \theta, B)^T)^T, \tag{3.1}$$

where

$$D_{\psi,\chi} = \\ \begin{bmatrix} -\frac{\partial}{\partial\beta} E_\theta[\psi_{\text{cond}}(y, x, \beta, B)]\,|_{\beta=\theta} & -\frac{\partial}{\partial A} E_\theta[\psi_{\text{cond}}(y, x, \theta, A)]\,|_{A=B(\theta)} \\ -\frac{\partial}{\partial\beta} E_\theta[\chi(x, \beta, B)]\,|_{\beta=\theta} & -\frac{\partial}{\partial A} E_\theta[\chi(x, \theta, A)]\,|_{A=B(\theta)} \end{bmatrix}. \tag{3.2}$$

By the definition of ψ_{cond} and $c(\beta, a)$ in (2.3) and (2.4), $\psi_{\text{cond}}(y, x, \theta, A)$ satisfies (1.3) for arbitrary A. Hence $E_\theta[\psi_{\text{cond}}(y, x, \theta, A)] = 0$ for all A, and the upper-right block of $D_{\psi,\chi}$ is 0. This means that the θ part of the influence function for (2.7) and (2.8) is equal to

$$\left\{-\frac{\partial}{\partial\beta} E_\theta[\psi_{\text{cond}}(y, x, \beta, B)]\,|_{\beta=\theta}\right\}^{-1} \psi_{\text{cond}}(y, x, \theta, B). \tag{3.3}$$

On the other hand, the influence function for the optimal $\psi_0(y, x, \theta) = \psi_{\text{cond}}(y, x, \theta, B(\theta))$ is also equal to (3.3), because by the same argument

$$\begin{aligned} D_{\psi_0} &= -\frac{\partial}{\partial\beta} E_\theta[\psi_{\text{cond}}(y, x, \beta, B)]\,\bigg|_{\beta=\theta} \\ &\quad - \frac{\partial}{\partial A} E_\theta[\psi_{\text{cond}}(y, x, \beta, A)]\,\bigg|_{A=B} \frac{\partial}{\partial\theta} B(\theta) \\ &= -\frac{\partial}{\partial\beta} E_\theta[\psi_{\text{cond}}(y, x, \beta, B)]\,|_{\beta=\theta}. \end{aligned}$$

We have thus shown the following theorem.

Theorem 3.1. The θ part of the influence function when θ and B are simultaneously estimated by (2.7) and (2.8) is the same as the influence function when θ alone is estimated using the optimal $\psi_0(y, x, \theta) = \psi_{\text{cond}}(y, x, \theta, B(\theta))$. As a consequence, the asymptotic covariance matrix of $\hat{\theta}_N$ is the same in both cases.

Remark 1. $\hat{\theta}_N$ and $\hat{B}_N$ are not asymptotically independent: $E_\theta[\psi_{\text{cond}}\chi^T] = 0$ by (1.3), but $(\partial/\partial\beta)E_\theta[\chi(x, \beta, B)]\,|_{\beta=\theta} \neq 0$ in general.

Remark 2. Because in linear regression with symmetric errors χ does not depend on θ, an analog to Theorem 3.1 is obvious. In addition, estimation of the scale of the errors does not change the asymptotic covariance either, and $\hat{\theta}_N$ is asymptotically independent of all nuisance parameters.

Remark 3. From the finite-sample interpretation of the influence function, (3.3) means the following: To the first order of approximation the change in $\hat{\theta}_N$ caused by adding or deleting an observation at (x, y) is

$$\left(\sum_{i=1}^{N} \frac{\partial}{\partial\beta} E_{\hat{\theta}_N}[\psi_{\text{cond}}(y, x_i, \beta, \hat{B}_N) \mid x_i]\,\bigg|_{\beta=\hat{\theta}_N}\right)^{-1} \\ \times \psi_{\text{cond}}(y, x, \hat{\theta}_N, \hat{B}_N);$$

that is, the change in $\hat{B}_N$ has approximately no effect on the change in $\hat{\theta}_N$. In this sense the estimator (2.7)–(2.8) is reasonably stable.

Remark 4. For the Fisher-consistent estimator (2.12)–(2.13) of Stefanski et al. (1986), there is no analog of Theorem 3.1. The θ part of the influence function is generally a linear combination of ψ_{BI}, $E_\theta[\psi_{BI} \mid x]$, and $E_0[\psi_{BI}\psi_{BI}^T \mid x] - B$, because all blocks in D are generally different from 0.

Remark 5. As both referees have pointed out, estimation of B makes no difference to asymptotic arguments, but almost certainly will have some effects in small samples. The analog to ordinary M estimation in linear regression is the problem of simultaneous estimation of scale, say by mean absolute deviation or Huber's proposal 2 (see Hampel et al. 1986, p. 234), which does have some effect on small-sample properties. One way to investigate this difference, at least in principle, is through the use of second-order expansions. Such expansions are extremely tedious, even for trying to understand the effect of scale in linear regression, and they are likely to be prohibitively difficult and complex for understanding the effect of estimating B. We even doubt if formal second-order expansion gives a much better approximation to the small-sample effects. Small-sample asymptotics (see Hampel et al. 1986, sec. 8.5) look more promising, but are even harder. In any case, Theorem 3.1 suggests that the small-sample effect of estimating B is smaller for our estimator than for the one studied by Stefanski et al. (1986). In Example 5.2 (Sec. 5) we check how good the approximation described in Remark 3 is in practice.

4. INFLUENTIAL OBSERVATIONS IN LOGISTIC REGRESSION

We are interested in the effect of small changes in a sample on the maximum likelihood estimate $\hat{\theta}_N$ in a logistic regression model: $P_\theta[Y = 1 \mid x] = F(x^T\theta)$, where $F(\beta) = \exp(\beta)/(1 + \exp(\beta))$. If we delete a small group of observations from a sample, perfect or almost-perfect discrimination between the two responses may become possible. In other words, the model becomes indeterminate, or the remaining data are nonoverlapping (Santner and Duffy 1986). In such a situation the deleted observations are influential, but any estimation procedure has to use them. We cannot expect any estimator, robust or otherwise, to produce a completely satisfactory model, for none exists. The situation is similar to one that occurs in linear regression when the $X'X$ matrix approaches singularity upon the removal of one or a few observations (e.g., see Chatterjee and Hadi 1986, fig. 2; Draper and Smith 1981, p. 258). Although data analysts should be aware of these points, they cannot hope to obtain meaningful parameter estimates under these circumstances; this fact would be reflected by dramatic increases in standard errors when these points are downweighted or removed.

To avoid this problem in our discussion of influence, we investigate what happens when a few observations are added to a sample with sufficient overlap. The influence function of the maximum likelihood estimator is $D(\theta)^{-1}x(y - F(x^T\theta))$. This suggests that

$$\hat{\theta}_{N+k} - \hat{\theta}_N \approx N^{-1} \sum_{i=N+1}^{N+k} D(\hat{\theta}_N)^{-1}x_i(y_i - F(x_i^T\hat{\theta}_N)).$$

This approximation is not uniform in x. To investigate the effect of extreme leverage points, more refined methods have to be used. We have the following lower bound for $\|\hat{\theta}_{N+k} - \hat{\theta}_N\|$ if all additional observations are equal.

Theorem 4.1. If $p \geq 3$, then $\sup\{\|\hat{\theta}_{N+k} - \hat{\theta}_N\|^2; x_{N+1} = \cdots = x_{N+k}, x_{N+1}^T\hat{\theta}_N = \beta\} \geq h(\beta)4k/\sum_{i=1}^N \|x_i\|^2$, where $h(\beta) = \sup_\xi \xi F(|\beta| - \xi)$.

Proof. Because zeros and ones can be exchanged, we may assume $\beta \geq 0$. Let ξ_0 be such that $h(\beta) = \xi_0 F(\beta - \xi_0)$, and set $\delta^2 = h(\beta)4k/\sum \|x_i\|^2$. We chose $y_{N+i} = 0$ and $\|x_{N+i}\| = \xi_0/\delta$. Suppose that $\|\hat{\theta}_{N+k} - \hat{\theta}_N\| < \delta$. We show that this leads to a contradiction by splitting $0 = \sum_{i=1}^{N+k} x_i(y_i - F(x_i^T\hat{\theta}_{N+k}))$ into two parts and bounding the first one from above and the second one from below.

Because $F'(\beta) = F(\beta)(1 - F(\beta)) \leq \frac{1}{4}$, we obtain from the definition of $\hat{\theta}_N$ and the mean-value theorem the following:

$$\begin{aligned}
&\left\|\sum_{i=1}^{N} x_i(y_i - F(x_i^T\hat{\theta}_{N+k}))\right\| \\
&\quad = \left\|\sum_{i=1}^{N} x_i(F(x_i^T\hat{\theta}_N) - F(x_i^T\hat{\theta}_{N+k}))\right\| \\
&\quad \leq \sum_{i=1}^{N} \|x_i\| \frac{1}{4} |x_i^T(\hat{\theta}_N - \hat{\theta}_{N+k})| \\
&\quad \leq \frac{1}{4} \sum \|x_i\|^2\|\hat{\theta}_{N+k} - \hat{\theta}_N\| < \frac{1}{4} \sum \|x_i\|^2\delta.
\end{aligned}$$

On the other hand, by the monotonicity of F,

$$\begin{aligned}
&\left\|\sum_{i=N+1}^{N+k} x_i(y_i - F(x_i^T\hat{\theta}_{N+k}))\right\| \\
&\quad = k\xi_0\delta^{-1}F(\beta - x_{N+1}^T(\hat{\theta}_N - \hat{\theta}_{N+k})) \\
&\quad \geq k\xi_0\delta^{-1}F(\beta - \xi_0\delta^{-1}\|\hat{\theta}_N - \hat{\theta}_{N+k}\|) \\
&\quad > k\xi_0\delta^{-1}F(\beta - \xi_0) = kh(\beta)\delta^{-1}.
\end{aligned}$$

Hence $kh(\beta)\delta^{-1} < \frac{1}{4} \sum \|x_i\|^2\delta$, which contradicts the definition of δ.

The condition $p \geq 3$ ensures that expected response and leverage can be varied independently. Taking $\xi = \beta$ shows that $h(\beta) > \frac{1}{2}\beta$. Hence by the aforementioned theorem $\|\hat{\theta}_{N+1} - \hat{\theta}_N\|$ can be arbitrarily large if the additional observation has an unexpected response. This means that the finite-sample breakdown point in the sense of Donoho and Huber (1983) is $1/(N + 1)$, the lowest possible value. The breakdown properties of our robust estimators remain to be investigated.

There remains the question of what can happen if we keep the expected response of an additional observation fixed and vary its leverage. We have an example where the effect on the estimate itself is bounded, but the effect on the estimated standard error is unbounded. We are currently investigating whether this phenomenon occurs more generally. Details will be given elsewhere.

5. EXAMPLES

To illustrate our estimators we consider two examples of logistic regression that have appeared elsewhere in the context of robust regression (see Pregibon 1981, 1982; Stefanski et al. 1986). Both data sets are difficult because they contain outliers, and without these outliers they are rather close to indeterminacy. The first example is particularly extreme, whereas in the second it is still possible to fit a meaningful model.

Example 5.1: Skin Vaso-Constriction Data (Pregibon 1981, 1982). These data consist of 39 observations on three variables: the occurrence of vaso-constriction in the skin of the digits, and the rate and volume of air inspired. The model to be fit regresses the occurrence of vaso-constriction on the logarithms of the remaining two variables. We took rate 32 = .30 (see Pregibon 1981).

Pregibon (1981) established that observations 4 and 18 are enormously influential in determining the maximum likelihood fit. It is not as evident from his analysis that without these two observations the model is nearly indeterminate; that is, we are in the situation described at the beginning of Section 4. In such a case the main advantage of robust procedures lies in their diagnostic capability. The near-indeterminacy is reflected by dramatic increases in standard errors when the influential points are downweighted or removed.

Table 1 contains parameter estimates, estimated standard errors, and weights for three robust fits, two using the Huber weight function with choices of $bp^{-1/2}$ equal to 3.7 and 3.2, and a biased analysis, performed by using the Huber weight function with $bp^{-1/2} = 3.7$ and setting $c(\beta, \alpha) = 0$. Results from the maximum likelihood fit are also given. When we attempted a fit using the Hampel function, observations 4 and 18 were immediately assigned 0 weight, and computational difficulties arose as a consequence of the near-indeterminacy.

Three comments are worth making. First, as b decreases the weights assigned observations 4 and 18 decrease rapidly, clearly indicating their anomalous nature. This fact would also manifest itself in a simple residual plot (see Stefanski et al. 1986). All other observations received weight 1 in all cases. Second, the estimated standard errors increase significantly as observations 4 and 18 are downweighted. Although some loss of efficiency is to be expected with the use of robust methods, the sizeable increase in standard errors for these data reflects the problem with indeterminacy mentioned before. Finally, the choice of b is crucial, but this is not surprising in view of the particular nature of the data. The biased estimator with $c \equiv 0$ seems to be more robust than the conditional unbiased one with the same b. We have no explanation for this.

Example 5.2: Food-Stamp Data (Stefanski et al. 1986). For these data the response indicates participation in the federal food-stamp program, and the predictor variables employed include two dichotomous variables (tenancy and supplemental income) and a logarithmic transformation of monthly income [log(monthly income + 1)]. The data consist of observations on 150 persons, of whom 24 participated in the program.

Table 2 displays results from several robust fits, as well as maximum likelihood estimation. In computing the Hampel estimator a concession was made for computational convenience: Rather than solving (2.4) to define $c(\beta, \alpha)$, we chose to use the same formula as in Example 2.1. The conclusions drawn by Stefanski et al. (1986) apply equally well to the estimators here. Observations 5 and 66 are most influential for the maximum likelihood estimator. As b decreases, these observations are downweighted. This results in an increase in perceived significance of the monthly income, accompanied by a decrease of the importance of supplemental income. Besides these two outliers, there are seven other atypical observations that are downweighted, but to a smaller extent.

Unlike the previous example, the estimated standard errors remain relatively stable, suggesting a greater degree of overlap in the data. A closer look shows that there are only six persons with tenancy participating in the food-stamp program. Once these are eliminated the parameter for tenancy can no longer be estimated. Moreover, without tenancy as a predictor the data become completely separated after elimination of 17 observations. All nine downweighted observations belong to this group, so any reasonable estimator has to use these outliers to some extent.

In Table 3 we give the changes in the estimates due to deletion of observation 5 or 66 and compare it with the change predicted by the influence function. Although we used a robust estimator, the changes are rather big. This is due to the peculiarity of the data set, mentioned previously. Still, the changes are much smaller than for the maximum likelihood estimator. For instance, if observation 5 is deleted, the maximum likelihood estimator for the coefficient of log(monthly income + 1) changes by .73, compared to .26 with our estimator. For B kept constant, the predictions by the influence function are excel-

Table 1. Maximum Likelihood and Robust Estimators for the Skin Vaso-Constriction Data

	Maximum likelihood estimator, $b = \infty$	Huber $c(\beta, a) = 0$, $b = 3.7p^{1/2}$	Huber conditional unbiased, $b = 3.7p^{1/2}$	Huber conditional unbiased, $b = 3.2p^{1/2}$
Intercept	−2.92	−5.71	−2.98	−6.41
	(1.29)	(2.45)	(1.35)	(2.84)
log(volume)	5.22	9.13	5.27	9.98
	(1.93)	(3.73)	(1.93)	(4.38)
log(rate)	4.63	8.09	4.67	8.85
	(1.79)	(3.31)	(1.86)	(3.82)
Weights				
Observation 4		.38	>.80	.25
Observation 18		.44	>.80	.29

NOTE: For selected observations, the weights w_b in Equation (2.3) are given.

Table 2. Maximum Likelihood and Robust Estimators for the Food-Stamp Data

	Maximum likelihood estimator, $b = \infty$	*Huber $c(\beta, a) = 0$, $b = 3.5p^{1/2}$*	*Huber conditional unbiased, $b = 3.5p^{1/2}$*	*Huber conditional unbiased, $b = 2.75p^{1/2}$*	*Hampel conditional unbiased, bends at $(3, 7, 16)p^{1/2}$*
Intercept	.93 (1.62)	4.26 (2.55)	4.51 (2.54)	5.49 (2.66)	6.00 (2.76)
Tenancy	−1.85 (.53)	−1.85 (.54)	−1.78 (.54)	−1.76 (.51)	−1.80 (.54)
Supplemental income	.90 (.50)	.75 (.52)	.74 (.51)	.62 (.52)	.70 (.52)
Log(monthly income + 1)	−.33 (.27)	−.89 (.43)	−.93 (.43)	−1.10 (.45)	−1.18 (.47)
Weights					
Observation 5		.21	.16	.13	.0
Observation 66		.76	.60	.41	.54

NOTE: For selected observations, the weights w_b in Equation (2.3) are given.

lent. By the results of Section 3 the influence function predicts no effect of reestimating B. In this example this is not quite true, but as a first-order approximation it is acceptable, in particular because the effective sample size is less than 150.

Our estimators are suitable for inferences based on the majority of the data. Moreover, they can also be used as a diagnostic with the following strategy suggested by the examples. Choose a large b (say $b = 5p^{1/2}$) and decrease it (e.g., in steps by $.5p^{1/2}$). Looking at the weights allows one to identify outliers. At the same time, it should be checked how close the data are to indeterminacy. A possible indication is how fast the estimated standard errors change, but it would be interesting to have other criteria. In this way, one can either fit a meaningful model to the good observations or identify the data set as problematic.

6. CONCLUSIONS

Conditionally unbiased score functions are appealing because their definition does not depend on the distribution of the predictors. In the context of robustness, there is an optimality theory for this class analogous to that already developed for unconditionally unbiased score functions. The optimal estimator depends on the unknown distribution of the predictors, and thus one has to estimate a matrix B. Nevertheless, consistency holds for any B, and asymptotically the uncertainty about B does not matter. In addition, conditionally unbiased score functions are often far easier to define. Although ignoring the bias and setting $c \equiv 0$ did not matter much in the examples considered, one can construct situations where this bias is large. With our estimator, we avoid this problem with little additional complexity.

Table 3. Effect of Deletion of Selected Observations in the Food-Stamp Data

	B constant	*B reestimated*	*Approximation by the influence function*
Deletion of observation 5			
Intercept	.48	.61	.47
Tenancy	−.05	−.04	−.04
Supplemental income	−.08	−.06	−.08
log(monthly income + 1)	−.51	−.62	−.47
Deletion of observation 66			
Intercept	.37	.65	.32
Tenancy	.26	.29	.24
Supplemental income	.05	−.10	.05
log(monthly income + 1)	−.40	−.67	−.33

NOTE: Changes in the Huber conditional unbiased estimates ($b = 2.75p^{1/2}$) divided by the estimated standard deviations.

[*Received April 1987. Revised December 1988.*]

REFERENCES

Chatterjee, S., and Hadi, A. S. (1986), "Influential Observations, High Leverage Points, and Outliers in Linear Regression," *Statistical Science*, 1, 379–393.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman & Hall.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich L. Lehmann*, eds. P. J. Bickel, K. Doksum, and J. L. Hodges, Jr., Belmont, CA: Wadsworth, pp. 157–184.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986), "Some New Methods for Weighted Regression When There Are Possible Outliers," *Technometrics*, 28, 219–230.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

Krasker, W. S., and Welsch, R. E. (1982), "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595–604.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman & Hall.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705–724.

——— (1982), "Resistant Fits for Some Commonly Used Logistic Model With Applications," *Biometrics*, 38, 485–498.

Ruppert, D. (1985), "On the Bounded-Influence Regression Estimator of Krasker and Welsch," *Journal of the American Statistical Association*, 80, 205–208.

Santner, T. J., and Duffy, D. E. (1986), "A Note on A. Albert and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 73, 755–758.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986), "Optimally Bounded Score Functions for Generalized Linear Models With Applications to Logistic Regression," *Biometrika*, 73, 413–425.

On One-Step GM Estimates and Stability of Inferences in Linear Regression

D. G. SIMPSON, D. RUPPERT, and R. J. CARROLL*

The folklore on one-step estimation is that it inherits the breakdown point of the preliminary estimator and yet has the same large sample distribution as the fully iterated version as long as the preliminary estimate converges faster than $n^{-1/4}$, where n is the sample size. We investigate the extent to which this folklore is valid for one-step GM estimators and their associated standard errors in linear regression. We find that one-step GM estimates based on Newton–Raphson or Scoring inherit the breakdown point of high breakdown point initial estimates such as least median of squares provided the usual weights that limit the influence of extreme points in the design space are based on location and scatter estimates with high breakdown points. Moreover, these estimators have bounded influence functions, and their standard errors can have high breakdown points. The folklore concerning the large sample theory is correct assuming the regression errors are symmetrically distributed and homoscedastic. If the errors are asymmetric and homoscedastic, Scoring still provides root-n consistent estimates of the slope parameters, but Newton–Raphson fails to improve on the rate of convergence of the preliminary estimates. If the errors are symmetric and heteroscedastic, Newton–Raphson provides root-n consistent estimates, but Scoring fails to improve on the rate of convergence of the preliminary estimate. Our primary concern is with the stability of the inferences associated with the estimates, not merely with the point estimates themselves. To this end we define the notion of standard error breakdown, which occurs if the estimated standard deviations of the parameter estimates can be driven to zero or infinity, and study the large sample validity of the standard error estimates. A real data set from the literature illustrates the issues.

KEY WORDS: Asymmetry; Heteroscedasticity; Least median of squares; Minimum volume ellipsoid; Robust inference; Standard error breakdown.

Consider the linear model $y_i = z_i^t\beta + \varepsilon_i$, for $i = 1, \ldots, n$, where $z_i = (1x_i^t)^t$, x_i is a known $(p-1)$-dimensional vector of explanatory variables, and y_i is an observed response. Two standard assumptions are: (1) $\varepsilon_1, \ldots, \varepsilon_n$ are identically distributed according to some F, and (2) $F = N(0, \sigma^2)$ for some $\sigma^2 > 0$. The earlier robust regression estimators—for example, M estimators (Andrews 1974; Bickel 1975; Huber 1973), rank estimators (Hettmansperger and McKean 1977; Jaeckel 1972), and trimmed least squares (Ruppert and Carroll 1980)—were designed to maintain efficiency under violations of (2), especially when the error distribution is heavy-tailed. However, it is as important to protect against violations of (1), particularly at outlying x observations, where heteroscedasticity or nonlinearity is likely. The generalized M estimators (GM estimators), such as the proposals of Mallows (1975), Hampel (1978), Krasker (1980), and Krasker and Welsch (1982), and the weighted trimmed least squares estimators of De Jongh, DeWet, and Welsh (1988) were intended to produce stable results when there are possible response outliers at outlying values of x, as can occur when (1) fails. In particular, they have influence functions bounded in both x and y. Unfortunately these bounded-influence estimators have breakdown points of at most $1/(p+1)$, where p is the number of predictor variables (Maronna, Bustos, and Yohai 1979), suggesting that they can be overwhelmed by a cluster of outliers; see, for example, Rousseeuw (1984).

The low breakdown point of the GM estimators has been viewed as a serious deficiency, particularly for multidimensional problems and exploratory data analysis. Several high breakdown point (HBP) estimators have been proposed that achieve breakdown points near $\frac{1}{2}$ for each p, including the least median of squares estimator of Rousseeuw (1984), the S estimators of Rousseeuw and Yohai (1984), and the estimators of Yohai (1987) and Yohai and Zamar (1988), which combine good asymptotic efficiency under the normal linear model with HBP. These estimators do not have bounded influence functions.

The HBP property provides some confidence that one will not be completely fooled by a cluster of poorly fit data. In practice, however, one would like the inferences to be robust to outliers, leverage points, and so on. If a few points can change the estimate by many standard errors or change drastically the standard error, it is small consolation that the change in the estimate is bounded. Routine data are thought to contain 1%–10% gross errors (Hampel, Ronchetti, Rousseeuw, and Stahel 1986). Although this is below the breakdown point of HBP estimators currently available, such a fraction of anomalous data can have a substantial effect if the influence function is unbounded. See, for instance, table I of Yohai and Zamar (1988), in which the bias of the Krasker–Welsch bounded-influence estimator is considerably less than that of the HBP unbounded-influence estimators if the level of contamination is 5%. We therefore contend that the

* D. G. Simpson is Associate Professor, Department of Statistics and Institute for Environmental Studies, University of Illinois at Urbana-Champaign, Champaign, IL 61820. D. Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. R. J. Carroll is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Simpson's research was supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship and Air Force Office of Scientific Research Grant AFOSR-87-0041. Ruppert's research was supported by NSF Grant DMS-88-0029 and the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell. Carroll's research was supported by Air Force Office of Scientific Research Grant AFOSR-89-0240. The authors thank the participants of the IMA Conference on Robustness, Diagnostics, Computation, and Graphics for their useful comments; the IMA for providing travel funds and housing; and the referees, who suggested strengthening the technical conditions to substantially simplify the proofs.

Journal of the American Statistical Association
June 1992, Vol. 87, No. 418, Theory and Methods

local stability associated with the bounded-influence property is as important as the global stability suggested by a high breakdown point. Moreover, the stability of the standard errors themselves is important and worthy of investigation.

To construct regression estimators that have bounded influence functions and high breakdown points, we follow a strategy that exists in the folklore: Start with a high breakdown point estimator and perform one iteration of a Newton–Raphson-type algorithm towards solution of the GM estimating equations. Hampel et al. (1986, p. 330) mentioned the possibility of using a one-step GM estimator but gave no details. We find one detail to be crucial for a high breakdown point, namely, the x-dependent weights associated with the GM iteration need to be based on high breakdown point location and scatter estimates rather than on the customary multivariate M estimates. Section 1 provides the specific definitions of our one-step GM estimates. Section 2 provides the breakdown analysis. Clearly one can iterate a fixed finite number of times and retain the breakdown point of the one-step. As a rough measure of the stability of inferences based on the estimates, we consider breakdown of the standard errors as well as the parameter estimates. The influence functions are derived in Section 3.

The large sample theory of the one-step GM estimators requires some care, as one natural initial estimator (least median of squares) converges only like $n^{-1/3}$ rather than the $n^{-1/2}$ rate usually associated with parametric estimation (Davies 1990; Kim and Pollard 1990; Rousseeuw 1984). However, results presented in Section 4 establish that both Newton–Raphson and Scoring versions of the one-step GM estimators converge at the root-n rate provided that the preliminary estimate is better than fourth root-n consistent and that the regression errors are symmetric and homoscedastic. Using a different method of proof, Jureckova and Portnoy (1987) established this kind of result for certain one-step Huber estimators. We find that if the errors are asymmetric and homoscedastic, Scoring still provides root-n consistent estimates of the slope parameters, whereas Newton–Raphson fails to improve on the rate of convergence of the preliminary estimate. On the other hand, if the errors are heteroscedastic and symmetric then Newton–Raphson provides root-n consistent estimates, whereas Scoring fails to improve on the rate of convergence of the preliminary estimate. We study asymptotic validity of the standard errors as well.

A potential objection to bounded-influence estimators is their low efficiency in cases where most of the sample information about β is contained in a few high leverage points. However, Morganthaler (1988) and Stefanski (1991) have shown that no estimator with a breakdown point greater than $1/n$ can have high finite-sample efficiency in the presence of extreme leverage points. In such instances, which involve a kind of extrapolation, it requires considerable faith in the linear model to take seriously the efficiency under the model. Our principal motivation for requiring a bounded-influence function as well as a high breakdown point is stability of inference. Section 5 illustrates some of the issues with a particularly vexing data set.

1. ONE-STEP MALLOWS ESTIMATES

Define residuals, $r_i = y_i - z_i^t \hat{\beta}_0$, where $\hat{\beta}_0$ is a high breakdown preliminary estimate with breakdown value at least m/n. For instance, a modified least median of squares (LMS) estimate has $m = [(n-p)/2] + 1$ (Rousseeuw and Leroy 1987). Let $\hat{\sigma}_0 = \text{med}\{|r_i|\}/\kappa$, where κ is a standardizing constant, and let m_x and C_x be multivariate location and scatter for the $\{x_i\}$ with breakdown point at least m/n. A possible choice for (m_x, C_x), the minimum volume ellipsoid (MVE) estimator, is given by the center and covariance of the smallest ellipsoid containing at least $[(n+p+1)/2]$ points. It has $m = [(n-p+1)/2]$, the best possible for affine equivariant covariance estimators (Rousseeuw and van Zomeren 1990). Cook and Hawkins (1990) discussed certain difficult computational issues associated with MVE.

The estimators we use are one-step estimators taking the form

$$\hat{\beta} = \hat{\beta}_0 + H_0^{-1} g_0, \qquad g_0 = \hat{\sigma}_0 \sum_{i=1}^{n} \psi(r_i/\hat{\sigma}_0) w_i z_i,$$

where there are two viable choices for H_0:

$$\text{Newton–Raphson:} \quad H_0 = \sum_{i=1}^{n} w_i z_i z_i^t \psi^{(1)}(r_i/\hat{\sigma}_0);$$

$$\text{Scoring:} \quad H_0 = n^{-1} \sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}_0) \sum_{j=1}^{n} w_j z_j z_j^t.$$

In the regression we employ Mallows weights,

$$w_i = \min\left[1, \left\{\frac{b}{(x_i - m_x)^t C_x^{-1}(x_i - m_x)}\right\}^{\alpha/2}\right]. \qquad (1.1)$$

The case $\alpha = 0$ is the one-step Huber estimate discussed by Bickel (1975) and Jureckova and Portnoy (1987). Jureckova and Portnoy (1987) imposed a nonequivariant bound on the step size to get HBP when $\alpha = 0$. We show that if $\alpha \geq 1$, the Mallows weights automatically bound the step size. The case $\alpha = 1$ is usual for GM estimators, whereas $\alpha = 2$ was used by Giltinan, Carroll, and Ruppert (1986) to force a bounded change of variance function, indicating local stability of the asymptotic variance. Ronchetti and Rousseeuw (1985) gave the form of the change of variance function for GM estimators. An even more extreme case, $\alpha = \infty$, deletes any observation in which the robust Mahalanobis distance from m_x exceeds b. Rousseeuw and van Zomeren (1990) discussed this possibility. We set b equal to the $(1-\gamma)$ quantile of the chi-squared distribution on $p - 1$ degrees of freedom, where $\gamma = .1$ or $.05$.

Scoring and Newton–Raphson are asymptotically equivalent if the errors $\{\varepsilon_i\}$ are independent and identically and symmetrically distributed; see Section 4. Another common choice for H_0 is based on iterative weighted least squares, but the resulting one-step estimator has a different asymptotic distribution that depends on that of the initial estimate; we forego the details. For either Newton–Raphson or Scoring, the large sample theory estimate of the covariance matrix of $\hat{\beta}$ is $D = H_0^{-1} M_0 H_0^{-1}$, where M_0 has one of two forms:

Nonexchangeable: $M_0 = \hat{\sigma}_0^2 \sum_{i=1}^{n} w_i^2 z_i z_i^t \psi^2(r_i/\hat{\sigma}_0)$;

Exchangeable: $M_0 = n^{-1}\hat{\sigma}_0^2 \sum_{i=1}^{n} \psi^2(r_i/\hat{\sigma}_0) \sum_{j=1}^{n} w_j^2 z_j z_j^t$.

If the ε_i are heteroscedastic, then in general D is consistent only if H_0 is "Newton–Raphson" and M_0 is "Nonexchangeable."

2. BREAKDOWN ANALYSIS

The finite sample breakdown point was introduced by Donoho and Huber (1983). Let $X = \{(x_i, y_i): i = 1, \ldots, n\}$ and let T be an estimator of β. Then the breakdown point of T at X is given by

$$\mathrm{BP}(T, X) = \min\{m/n : \sup_{X^*} \|T(X) - T(X^*)\| = \infty\},$$

where the supremum is over all choices of X^* consisting of $(n - m)$ points from X and m arbitrary points. A HBP estimator like Rousseeuw's (1984) LMS estimator has BP $\approx \frac{1}{2}$ for any data set where the z_i's are in general position; that is, any p of them are linearly independent. For the scatter matrix, C_x, breakdown is defined as driving $\lambda_{\max}(C_x) + \{\lambda_{\min}(C_x)\}^{-1}$ to infinity, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of the matrix A. We obtain breakdown points for the one-step GM estimators defined in Section 1 and then consider breakdown of their covariance estimates.

2.1 Breakdown of Estimates

In what follows, we will assume that the first $n - m$ observations are the "good" ones and that the remaining m observations are free to roam. We assume that $n - m \geq n/2 + 1 \geq p$, and, without loss of generality, that the first p observations are such that $(z_1, \ldots, z_p)$ are linearly independent. As usual, $\psi(v)$ is odd and bounded. We make use of the following additional assumptions:

A. Assume that ψ is nondecreasing with the properties

$$\psi(v)/v \geq d_0 > 0 \qquad \text{if } 0 \leq |v| \leq a; \tag{2.1}$$

$$\psi^{(1)}(v) \geq d_1 > 0 \qquad \text{if } 0 \leq |v| \leq a; \tag{2.2}$$

and

$$a > \kappa. \tag{2.3}$$

B. If ψ is redescending, assume (2.1)–(2.3) as well as

$$\sup_{|v| \geq a} |\psi^{(1)}(v)| = d_2, \qquad \text{where } d_1 > d_2. \tag{2.4}$$

C. Assume that any set of $n - m - n/2$ "good" points has a linearly independent subset of size p.

Theorem 2.1. Either of Assumptions A or B suffice for the breakdown value of the one-step Mallows to be at least m/n under Scoring. For Newton–Raphson, the breakdown value is at least m/n under Assumptions A and C taken together.

Remark 2.1. If ψ is redescending, then $\psi^{(1)}(r_i/\hat{\sigma}_0)$ can go negative. We conjecture that in this case it is possible to manipulate p data points so that the Newton–Raphson version of H_0 equals 0.

2.2 Standard Error Breakdown

Let D be the covariance estimate of $\hat{\beta}$ given in Section 1. Standard-error breakdown occurs if either $\lambda_{\max}(D) \rightarrow \infty$ or $\lambda_{\min}(D) \rightarrow 0$. The former usually is the only concept considered, as in Hampel et al. (1986), but the latter is important as well. For instance, even if the estimate does not break down, the Wald-type tests for the parameters can break down if D breaks down to 0. He, Simpson, and Portnoy (1990) have discussed breakdown of tests in general. A simple analysis shows that $\lambda_{\max}(D) \leq \lambda_{\max}(M_0)/\lambda_{\min}^2(H_0)$, and we show in the Appendix that $\lambda_{\min}(H_0) > 0$. It is clear that, because $\alpha \geq 1$, $\lambda_{\max}(M_0)$ has a finite upper bound under any arrangement of the "bad" points, and hence the same holds for $\lambda_{\max}(D)$.

Unfortunately, breakdown to 0 may occur unless $\alpha \geq 2$. Because $\lambda_{\min}(D) \rightarrow 0$ if $\det(D) \rightarrow 0$, this breakdown occurs if either $\det(M_0) \rightarrow 0$ or $\det(H_0^{-1}) \rightarrow 0$, the latter occurring if $\lambda_{\max}(H_0) \rightarrow \infty$. A detailed analysis as presented in Section 2.1 shows that under any arrangement of the "bad" points, $\lambda_{\min}(M_0) > 0$. Thus $\lambda_{\min}(D) \rightarrow 0$ if we can show that $\lambda_{\max}(H_0) \rightarrow \infty$. This may happen if $\alpha < 2$.

Lemma 2.1. Define $d_i = z_i/\|z_i\|$ and let $\|z_j\| \rightarrow \infty$ for $j \geq n - m + 1$ in such a way that for a positive definite matrix S, $\sum_{i=n-m+1}^{n} d_i d_i^t \rightarrow S$. Then $\lambda_{\max}(H_0) \rightarrow \infty$ if $\alpha < 2$, whereas $\lambda_{\max}(H_0) = O(1)$ if $\alpha \geq 2$.

3. INFLUENCE ANALYSIS

Influence analysis is a method of studying the local stability of estimators in terms of the effect of point-mass perturbations of the data or the underlying distribution. Two approaches to influence analysis of linear regression are in common use: (1) treat $\{(x_i, y_i)\}$ as a random sample and define the influence function on the space of distributions for (x, y) (Hampel et al. 1986) and (2) define the influence function via asymptotic linearity of the estimator (Krasker and Welsch 1982). We show that in either case the influence function of the one-step Mallows estimator is bounded when evaluated at the model. Method (1) requires that the preliminary estimates have influence functions, but they need not be bounded. Method (2) requires only a rate of convergence. Method (2) is perhaps more appropriate for regression, because it yields an influence function even when an iid assumption on the x_i's is inappropriate.

First act as if $\{(x_i, y_i)\}$ is a random sample from a distribution F_0 and consider the effect of perturbation of F_0. We suppose that the preliminary estimates and the location and scatter functionals for x have influence functions, but the influence functions need not be bounded. For instance, the preliminary regression estimate might be a regression S estimate (Rousseeuw and Yohai 1984), and the location scatter estimate might be a multivariate S estimate (Davies 1987; Lopuhaä 1989). The alternative definition of the influence function via asymptotic linearity allows treatment of the minimum volume ellipsoid.

Consider a generic matrix-valued functional $T(F)$ defined on the space of distributions for (x, y). Let F_0 be a fixed distribution representing the target model and let F_λ be a point-mass contamination of F_0: $F_\lambda = (1 - \lambda)F_0 + \lambda\Delta_{x,y}$, for $0 \le \lambda \le 1$. Following Hampel et al. (1986), T has an influence function, which we shall denote by $IF(x, y; T)$, if it has a directional derivative at $\lambda = 0$:

$$IF(x, y; T) = \lim_{\lambda \downarrow 0} \{T(F_\lambda) - T(F_0)\}/\lambda.$$

The IF operation preserves matrix dimensions and satisfies the multiplication and chain rules of scalar differentiation.

The one-step Newton–Raphson estimators described in the preceding section correspond to the functional $\hat\beta(F) = \hat\beta_0(F) + \{H(F)\}^{-1}g(F)$, where

$$g(F) = \hat\sigma_0(F)E_F\left[\psi\left(\frac{Y - Z^t\hat\beta_0(F)}{\hat\sigma_0(F)}\right)w(X, m(F^x), C(F^x))Z\right]$$

and

$$H(F) = E_F\left[\psi^{(1)}\left(\frac{Y - Z^t\hat\beta_0(F)}{\hat\sigma_0(F)}\right)w(X, m(F^x), C(F^x))ZZ^t\right].$$

For Scoring, $H(F)$ instead takes the form

$$H(F) = E_F\left[\psi^{(1)}\left(\frac{Y - Z^t\hat\beta_0(F)}{\hat\sigma_0(F)}\right)\right] \times E_{F^x}[w(X, m(F^x), C(F^x))ZZ^t].$$

Here E_F denotes expectation with respect to F, $\hat\beta_0(F)$ and $\hat\sigma_0(F)$ are the functionals corresponding to the preliminary regression and scale estimates, F^x is the marginal distribution for x, $m(F^x)$ and $C(F^x)$ are the location and scatter functionals for x, and the weight function w is of the same form as in (1.1). Assuming F_0 is such that the conditional distribution of $(Y - Z^t\hat\beta_0(F_0))$ given $Z = z$ is independent of z, Newton–Raphson and Scoring reduce to the same functional at F_0. If F_n is the empirical distribution of $\{(x_i, y_i)\}$, then statistics and functionals are related as follows: $\hat\beta = \hat\beta(F_n)$, $\hat\beta_0 = \hat\beta(F_n)$, $g_0 = ng(F_n)$, and $H_0 = nH(F_n)$.

For both Newton–Raphson and Scoring the multiplication rule yields

$$IF(x, y; \hat\beta) = IF(x, y; \hat\beta_0) + \{H(F_0)\}^{-1}IF(x, y; g), \quad (3.1)$$

because $g(F_0) = 0$. In the following we suppress the dependence of IF on (x, y). Fisher consistency and symmetry of the residual distribution for F_0 yield

$$\begin{aligned}IF(g) = {} & \sigma\psi\left(\frac{y - z^t\beta}{\sigma}\right)w(x, m(F_0^x), C(F_0^x))z \\ & - E_{F_0}\left[\psi^{(1)}\left(\frac{Y - Z^t\beta}{\sigma}\right)w(X, m(F_0^x), C(F_0^x))ZZ^t\right] \\ & \times IF(\hat\beta_0) + \sigma^{-1}g(F_0)IF(\hat\sigma_0) \\ & + \sigma E_{F_0}\left[\psi\left(\frac{Y - Z^t\beta}{\sigma}\right)IF(w(X, m(\cdot), C(\cdot)))Z\right] \\ & + IF(\hat\sigma_0)E_{F_0}\left[\left(\frac{Y - Z^t\beta}{\sigma}\right)\psi^{(1)}\left(\frac{Y - Z^t\beta}{\sigma}\right)\right. \\ & \left. \times w(X, m(F_0^x), C(F_0^x))Z\right] \\ = {} & \sigma\psi\left(\frac{y - z^t\beta}{\sigma}\right)w(x, m(F_0^x), C(F_0^x))z \\ & - \mathrm{H}(F_0)IF(\hat\beta_0).\end{aligned}$$

Inserting the latter expression in (3.1) gives

$$IF(x, y; \hat\beta) = \{H(F_0)\}^{-1}\sigma\psi\left(\frac{y - z^t\beta}{\sigma}\right)w(x, m(F_0^x), C(F_0^x))z.$$

This expression agrees with the influence function of the fully iterated GM estimate with weight function w (Hampel et al. 1986).

Alternatively, observe that by Theorem 4.1 the one-step GM estimator has the following asymptotic representation:

$$\hat\beta = \beta + n^{-1}\sum_{i=1}^{n} Q^{-1}z_iw_i\sigma\psi\left(\frac{y_i - z_i'\beta}{\sigma}\right) + o_p(n^{-1/2}), \quad (3.2)$$

where Q is as in D1 of Section 4.3. The summand in (3.2) shows the contributions of the observations to the deviation of $\hat\beta$ from β. Following Krasker and Welsch (1982) we call the corresponding function, $Q^{-1}zw(z)\sigma\psi((y - z'\beta)/\sigma)$, the influence function.

Remark 3.1. If an estimator has an influence function, then general results of He and Simpson (in press) imply that the bounded-influence property is necessary rather than sufficient for local stability of the estimator. A stronger result would be to establish that the bias sensitivity is bounded. Martin, Yohai, and Zamar (1989) studied bias properties of certain S estimators and GM estimators.

The Huber estimates, which used bounded ψ but $w(\cdot) = 1$, bound the residuals but not the influence of the position in the design space. These estimators are susceptible to leverage points; that is, to outliers in the design space. On the other hand, if ψ and $\|z\|w(z)$ are both bounded, then the Mallows estimators bound the joint influence of the residuals and the position in the design space.

4. LARGE SAMPLE THEORY

To provide a rigorous large sample theory on which to base precision estimates and other inferences, we derive asymptotic representations for the one-step GM estimators. The preliminary estimates $(\hat{\beta}_0, \hat{\sigma}_0)$ need only be n^τ-consistent for some $\tau \in (\frac{1}{4}, \frac{1}{2}]$. For instance, $\hat{\beta}_0$ might be the LMS estimate, which converges at the rate $n^{-1/3}$ (Davies 1990; Kim and Pollard 1990; Rousseeuw 1984), or the least trimmed sum of squares estimate (Rousseeuw 1984), which converges at the rate $n^{-1/2}$.

The rate of convergence of the remainder in the asymptotic representation depends on the rate of convergence of the preliminary estimator. Although any rate better than $n^{-1/4}$ suffices for the one-step estimator to be root-n consistent and asymptotically normal, a better rate of convergence for the preliminary estimator implies a better rate of convergence for the remainder. In the following let

$$Q_n = \sum_{i=1}^{n} E[\psi^{(1)}(\varepsilon_i/\sigma)] w_i z_i z_i^t. \qquad (4.1)$$

Theorem 4.1. Assume conditions A1–D2 of Section 4.3. Suppose $\hat{\beta}_0 - \beta = O_p(n^{-\tau})$ and $\hat{\sigma}_0 - \sigma = O_p(n^{-\tau})$ for some $\tau \in (\frac{1}{4}, \frac{1}{2}]$. Then for Newton–Raphson, $n^{-1}(H_0 - Q_n) = O_p(n^{-\tau})$ and

$$n^{-1/2} H_0(\hat{\beta} - \beta)$$
$$= n^{-1/2}\sigma \sum_{i=1}^{n} \psi(\varepsilon_i/\sigma) w_i z_i + O_p(n^{1/2-2\tau}). \qquad (4.2)$$

The same is true of Scoring if $n^{-1} \sum_{i=1}^{n} \|z_i\| = O(1)$.

Theorem 4.1 implies that $H_0(\hat{\beta} - \beta)$ is asymptotically normal with mean 0 and covariance A_n, where

$$A_n = \sigma^2 \sum_{i=1}^{n} \operatorname{var}[\psi(\varepsilon_i/\sigma)] w_i^2 z_i z_i^t. \qquad (4.3)$$

In practice we estimate A_n by M_0. The following result shows that this works.

Theorem 4.2. Assume conditions A1–D2 and suppose $\hat{\beta}_0 - \beta = O_p(n^{-\tau})$ and $\hat{\sigma}_0 - \sigma = O_p(n^{-\tau})$. If $n^{-1} \sum_{i=1}^{n} w_i^4 \|z_i\|^4 = O(1)$, then nonexchangeable M_0 satisfies

$$n^{-1}(M_0 - A_n) = O_p(n^{-\tau}), \qquad (4.4)$$

and hence $M_0^{-1/2} H_0(\hat{\beta} - \beta) = Z_n + O_p(n^{1/2-2\tau})$, where Z_n has mean 0, covariance I, and is asymptotically normal. The same is true for exchangeable M_0 if instead $n^{-1} \sum_{i=1}^{n} \|z_i\| = O(1)$.

4.1 Effect of Asymmetry

Condition D2 of Theorem 4.1 is essentially symmetry of the error distribution. Carroll and Welsh (1988) and Welsh (1989) noted that the Huber and Mallows GM estimates of the slope are consistent even when the errors are asymmetric. This kind of result extends to the one-step versions as well. We show that if the errors are iid, then the asymptotic bias introduced by asymmetry is absorbed in the intercept, and we provide asymptotic expansions for the slope estimates. Asymmetry implies that the Scoring and Newton–Raphson estimators have different limiting behavior. In particular, the Scoring estimate of the slope vector is root-n consistent, whereas the Newton–Raphson estimate fails to improve on the rate of convergence of the preliminary estimate.

Partition $\beta^t = (\eta, \gamma^t)$ into intercept η and slope vector γ, and do the same for $\hat{\beta}_0$ and $\hat{\beta}$. Even if the error distribution is asymmetric, γ is identifiable as the value such that the distribution of $y_i - x_i^t\gamma$ is independent of x_i (Carroll and Welsh 1988). Hence it is reasonable to expect $n^\tau(\hat{\gamma}_0 - \gamma) = O_p(1)$ even in the asymmetric case, as long as the errors are homoscedastic. For a fully iterated GM estimate the intercept η may be defined by the condition

$$E\psi\left(\frac{\varepsilon_i}{\sigma}\right) = E\psi\left(\frac{y_i - \eta - x_i^t\gamma}{\sigma}\right) = 0. \qquad (4.5)$$

As different choices of ψ give different values of η in the asymmetric case, we can expect only that $n^\tau(\hat{\eta}_0 - \eta_0) = O_p(1)$ for some η_0 not necessarily the same as η.

Let $\beta_0 = (\eta_0, \gamma^t)^t$ be the limiting value of the preliminary estimator and define $u_i = y_i - z_i^t\beta_0 = \varepsilon_i + \eta - \eta_0$ for $i = 1, \ldots, n$. Replace ε_i by u_i in the definition of Q_n. In correspondence with the partition of β, we partition the Hessian matrix and Q_n:

$$H_0 = \begin{bmatrix} h_{11} & h_{(1)}^t \\ h_{(1)} & H_{22} \end{bmatrix}, \qquad Q_n = \begin{bmatrix} q_{11} & q_{(1)}^t \\ q_{(1)} & Q_{22} \end{bmatrix}.$$

Here h_{11} and q_{11} are scalars and H_{22} and Q_{22} are $(p - 1) \times (p - 1)$ symmetric matrices. Define $H_{22\cdot 1} = H_{22} - h_{(1)}h_{(1)}^t/h_{11}$ and similarly define $Q_{22\cdot 1}$. To simplify the analysis, we center the x's by their Mallows-weighted means so that

$$\sum_{i=1}^{n} x_i w_i = 0. \qquad (4.6)$$

This centering implies that $Q_{22\cdot 1} = Q_{22}$ and, for Scoring, $H_{22\cdot 1} = H_{22}$.

Lemma 4.1. Assume conditions A1–C2. Assume D1, replacing $\{\varepsilon_i\}$ by $\{u_i\}$. Suppose $\hat{\beta}_0 - \beta_0 = O_p(n^{-\tau})$ and $\hat{\sigma}_0 - \sigma = O_p(n^{-\tau})$. Then for Scoring, $n^{-1}(H_{22} - Q_{22}) = O_p(n^{-\tau})$ and

$$n^{-1/2} H_{22}(\hat{\gamma} - \gamma) = n^{-1/2}\sigma \sum_{i=1}^{n} x_i w_i \{\psi(u_i/\sigma)$$
$$- E[\psi(u_1/\sigma)]\} + O_p(n^{1/2-2\tau}).$$

Assume also that $\psi^{(2)}$ has derivative $\psi^{(3)}$ with $\|\psi^{(3)}\|_{\sup}$ and $\|(\cdot)^2\psi^{(3)}(\cdot)\|_{\sup}$ both finite. Then for Newton–Raphson, $n^{-1}(H_{22\cdot 1} - Q_{22}) = O_p(n^{-\tau})$ and

$$\hat{\gamma} = \gamma + Q_{22}^{-1}\sigma \sum_{i=1}^{n} x_i w_i \{\tilde{\psi}(u_i/\sigma) - E[\tilde{\psi}(u_1/\sigma)]\}$$
$$+ \frac{a_0 a_2}{a_1^2}(\hat{\gamma}_0 - \gamma) + O_p(n^{1/2-2\tau}),$$

where $\tilde{\psi}(t) = \psi(t) - a_0 a_1^{-1}\psi^{(1)}(t)$ and $a_k = E[\psi^{(k)}(u_1/\sigma)]$ for $k = 0, 1, 2$.

Remark 4.1. If the preliminary estimate converges more slowly than $n^{-1/2}$, then the expansion for Newton–Raphson implies $n^{\tau}(\hat{\gamma} - \gamma) = (a_0 a_2 / a_1^2) n^{\tau}(\hat{\gamma}_0 - \gamma) + o_p(1)$, and the asymptotic relative efficiency of Newton–Raphson versus $\hat{\gamma}_0$ is $a_1^4 a_0^{-2} a_2^{-2}$. This approaches infinity as the error distribution approaches symmetry.

Remark 4.2. Both the Scoring and Newton–Raphson versions of $\hat{\eta}$ converge in probability to $\eta_0 + \sigma a_0 / a_1$, which is one step of a Newton–Raphson algorithm for solving (4.5). Hence, iteration can drive a_0 to 0. In theory, iterating k_n times to achieve $a_0 = o(n^{\tau - 1/2})$ implies that the Newton–Raphson k_n step has the same asymptotic distribution as does the fully iterated version.

Remark 4.3. In the asymmetric case, asymptotically valid Wald-type inferences on the slope parameters may be obtained by the Scoring method coupled with the following modification of the exchangeable M_0:

$$M_{22} = n^{-1} \hat{\sigma}_0^2 \sum_{i=1}^{n} \{\psi(r_i / \hat{\sigma}_0) - \bar{\psi}\}^2 \sum_{j=1}^{n} w_i^2 x_i x_i^t,$$

where $\bar{\psi} = n^{-1} \sum_{i=1}^{n} \psi(r_i / \hat{\sigma}_0)$. In this case $M_{22}^{-1/2} H_{22}(\hat{\gamma} - \gamma) = Z_{n2} + O_p(n^{1/2 - 2\tau})$, where Z_{n2} has mean 0 and covariance I_{p-1} and is asymptotically normal.

4.2. Effect of Heteroscedasticity

We next consider the large sample behavior of one-step estimators when the errors are symmetrically distributed but heteroscedastic. We show that Newton–Raphson and the nonexchangeable version of M_0 provide valid large sample inferences, whereas Scoring fails to improve on the rate of convergence of the initial estimator.

Lemma 4.2. Suppose the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with $\varepsilon_i \sim F_i$. Assume A2–D1 of Section 4.3, and assume D2 holds for each F_i. Suppose $n^{\tau}(\hat{\beta}_0 - \beta) = O_p(1)$ and $n^{\tau}(\hat{\sigma}_0 - \sigma) = O_p(1)$. Then both Newton–Raphson and Scoring have expansions of the form

$$n^{-1} H_0(\hat{\beta} - \beta) = n^{-1} \sigma \sum_{i=1}^{n} \psi(\varepsilon_i / \sigma) w_i z_i + T_n + O_p(n^{-2\tau}).$$

For Newton–Raphson $T_n = 0$, whereas for Scoring T_n is asymptotically equivalent to $\Gamma_n(\hat{\beta}_0 - \beta)$ for a symmetric nonstochastic matrix Γ_n.

Because of the heteroscedasticity, the limiting value of $\hat{\sigma}_0$ depends on the estimator. Although σ has an effect on the efficiency of $\hat{\beta}$, the Newton–Raphson covariance estimate $H_0^{-1} M_0 H_0^{-1}$ is asymptotically correct.

Theorem 4.3. Assume the conditions of Lemma 4.2. For the Newton–Raphson version of H_0 and nonexchangeable M_0 we have $M_0^{-1/2} H_0(\hat{\beta} - \beta) = Z_n + O_p(n^{1/2 - 2\tau})$, where Z_n has mean 0 and covariance I and is asymptotically normal.

4.3 Technical Conditions and Remarks

A1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with distribution function F.

A2. The score function ψ is bounded and continuous.

B1. ψ has derivative $\psi^{(1)}$ such that (a) $\|\psi^{(1)}\|_{\sup} < \infty$ and (b) $\|(\cdot)\psi^{(1)}(\cdot)\|_{\sup} < \infty$, where $\|\cdot\|_{\sup}$ is the supremum norm.

B2. $\psi^{(1)}$ has derivative $\psi^{(2)}$ such that (a) $\|\psi^{(2)}\|_{\sup} < \infty$, (b) $\|(\cdot)\psi^{(2)}(\cdot)\|_{\sup} < \infty$, and (c) $\|(\cdot)^2\psi^{(2)}(\cdot)\|_{\sup} < \infty$.

C1. As $n \to \infty$ the design satisfies (a) $n^{-1} \sum_{i=1}^{n} \|z_i\|^4 \times w_i^2 = O(1)$ and (b) $n^{-1} \sum_{i=1}^{n} \|z_i\|^3 w_i = O(1)$.

C2. The design satisfies $\lim_{n \to \infty} \max_{1 \le i \le n} \|z_i\|^2 w_i^2 / \sum \|z_j\|^2 w_j^2 = 0$.

D1. $\lim_{n \to \infty} n^{-1} A_n = A$ and $\lim_{n \to \infty} n^{-1} Q_n = Q$ for some symmetric positive definite matrices A and Q.

D2. $E_F[\psi(\varepsilon v)] = 0$ and $E_F[\varepsilon v \psi^{(1)}(\varepsilon v)] = 0$ for any nonnegative scalar v. For example, ψ is odd and F has a density symmetric about 0.

Remark 4.4. We place heavy conditions on ψ but weak conditions on F. In the context of robust inference it seems appropriate to place conditions on ψ (which is under our control) rather than on F. The differentiability of $\psi^{(1)}$ given in B2 can be weakened by Lipschitz-type conditions, as indicated in Lemma A.1.

Remark 4.5. For appropriately chosen Mallows weights the present design conditions are weaker than the standard conditions for Huber regression. In particular, taking $\alpha = 2$ in (2.1) ensures that $\|z_i\|^2 w_i \le \lambda_{\max}(C_x)$, so it is sufficient that $\lambda_{\max}(C_x) = O(1)$, $n^{-1} \sum \|z_i\| = O(1)$, and $\sum_{i=1}^{n} w_i^2 \|z_i\|^2 \to \infty$. The asymptotics of the preliminary estimator may require additional conditions; for instance, the conditions given by Kim and Pollard (1990) or Davies (1990) for least median of squares.

Remark 4.6. The conditions on ψ exclude piecewise linear score functions such as Hampel's three-part redescender. Simpson, Ruppert, and Carroll (1989) gave an alternative proof for such estimators. Discontinuities in $\psi^{(1)}$ can lead to instability in the large sample variance if there is substantial discreteness in the data (Simpson, Carroll, and Ruppert 1987).

5. LAND USE/WATER QUALITY

Haith (1976) collected data relating land use to water quality. Each case was a river basin in New York State. Basins were selected by two criteria: independence (no basin in the sample being a tributary of another basin in the sample) and completeness of the data. All 20 basins satisfying these criteria were included in the sample. The data, which also were given in Allen and Cady (1982, table 2.1), include five variables, nitrogen concentration and four land use variables given as a percentage of total land usage: N = total nitrogen; AC = active agriculture; FR = forest, brushland, or plantation; RS = residential; and CI = commercial/industrial. Haith (1976) developed linear regression models relating N to subsets of the four other variables. Because the purpose of modeling was to attribute nonpoint source pollution to the various types of land use, the parameter estimates and their standard errors were of primary interest.

The covariates exhibit sizeable linear dependencies, and there are design outliers. *AC* and *FR* have a negative association, except for case #5 (the Hackensack River), which is an outlier in the design space with low *AC* and *FR* values and high *RS* and *CI* values. Much of the variation in *RS* and *CI* is due to five rivers, and the observed *RS* and *CI* values exhibit a strong positive association. Their sample correlation is .86; their sample correlation excluding the rivers with the two highest *RS* values is .93. With such a design it is difficult to disentangle the residential and commercial effects reliably. To alleviate the collinearity, we replace *RS* and *CI* by their sum, $UR := RS + CI =$ percent urban land usage. If the goal were to predict *N*, one might instead use stepwise regression to select a subset of the variables; this was Haith's strategy. However, simpler models do not attain a goal of relating all land uses to water quality. Aggregating *RS* and *CI* is a compromise made necessary by the design that still allows us to relate all land uses to pollution.

Case #5 (the Hackensack River) is such a severe design outlier that data analysts likely would set this point aside rather than including it in a linear least squares analysis. We shall present results both with and without case #5. Although inferences that rely heavily on this point are too unstable to be trusted, it would be of interest to determine whether the Hackensack River conforms roughly to the model suggested by the other rivers or whether it points to some alternative phenomenon in urban rivers. The Mallows weights that we use essentially delete case #5 in the fitting algorithm. Such downweighting of design outliers and response outliers is meant to limit their influence on the fitted model and associated inferences, but it also has the benefit of accentuating the inadequacy of the model for these points, possibly making it easier to discover alternative and more satisfactory models. Thus, although outliers may be downweighted or even deleted during the fitting of the model, this does not imply that they are "discarded" in the analysis of the data. They are in fact emphasized.

For the full data, ordinary least squares (OLS) regression of *N* on the land use variables yields (with standard errors in parentheses):

$$\hat{N} = 1.43(\pm 1.29) + .0085(\pm .016)AC - .0084(\pm .015)FR + .029(\pm .028)UR.$$

Omitting case #5 yields instead

$$\hat{N} = 1.70(\pm .76) + .0021(\pm .0094)AC - .014(\pm .0086)FR + .16(\pm .028)UR.$$

Table 1. Linear Model Parameter Estimates and Standard Errors for New York Rivers Data

	OLS	*LMS*	*M*	*GM*
AC	.0028 (.0043)	.0157	.0175 (.0021)	.0164 (.0030)
FR	.0058 (.0020)	.00019	.0022 (.00096)	.0026 (.0014)
UR	.0437 (.016)	.171	.179 (.0077)	.203 (.046)
OTHER	.0143 (.013)	.0364	.0251 (.0063)	.0239 (.0077)

Table 2. Linear Model Parameter Estimates and Standard Errors Excluding the Hackensack River

	OLS	*LMS*	*M*	*GM*
AC	.0191 (.0026)	.0175	.0177 (.0018)	.0162 (.0029)
FR	.00322 (.0013)	.00114	.0023 (.00089)	.0024 (.0013)
UR	.173 (.025)	.136	.156 (.018)	.195 (.052)
OTHER	.0170 (.0076)	.0335	.0276 (.0054)	.0263 (.0070)

It is clear that case #5 would have considerable effect on the OLS inferences about urban effects were it included.

The parameter estimates and standard errors for the covariates in the above model are somewhat difficult to interpret, because the parameters represent incremental effects over other land uses not measured. We therefore reparameterize the model by replacing the intercept with the constructed variable $OTHER := 100 - AC - FR - UR$. This reparameterization leaves the design space intact but provides directly interpretable parameters. For instance, the *AC* parameter is the nitrogen that can be attributed to each percentage of agricultural use.

Tables 1 and 2 give estimates and standard errors using several methods: OLS; LMS; a three-step Huber estimator (M)—that is, Mallows with $\alpha = 0$, starting from LMS; and a three-step Mallows estimator (GM) with $\alpha = 2$, starting from LMS. The three-step estimates used the scoring method, exchangeable standard errors, and a three-part redescending Hampel ψ function with tuning constants $(a, b, c) = (1.5, 3, 8)$. The normalizing constant in the scale estimate was set equal to $\kappa = .6745$, but standard errors were inflated by the factor $\{W/(W-p)\}^{1/2}$, where W is the number of observations with nonzero weight. The Mallows weights for GM used $b = \chi^2(.95; p-1)$. MVE estimates of location and scatter for the covariates were computed using a FORTRAN program supplied by B. van Zomeren. LMS was computed via the S-plus (Statistical Sciences, Inc.) function, LMSREG. S functions for the GM steps and diagnostics are available from the authors on request.

On deletion of case #5, the nitrogen concentration attributed to urban use by OLS quadruples and the standard errors become considerably smaller. It is clear that the nitrogen concentration for case #5 is much less than was predicted by linear extrapolation from the remaining data. The LMS and M parameter estimates are not affected drastically by the presence or absence of case #5; however, the M standard error for *UR* is more than doubled by the deletion. The GM parameter estimates and standard errors show little change on deletion of case #5. The M standard error for *UR* seems overly optimistic, even after deleting case #5, given the change in the estimates induced by the deletion and the differences among the estimates. The standard error associated with GM is perhaps more realistic.

Table 3 provides diagnostics for selected rivers based on the full data: diagonals of the OLS projection matrix (h_{ii}); OLS studentized residuals (t_i^{OLS}); standardized residuals for LMS (s_i^{LMS}), M (s_i^{M}), and GM (s_i^{GM}); and the Mallows weights (w_i). The standardized residuals s_i were scaled by $\text{median}\{|\text{residual}|\}/.6745$. McKean, Sheather, and Hett-

Table 3. Diagnostics for Selected Observations From the New York Rivers Data

i	h_{ii}	t_i^{OLS}	s_i^{LMS}	s_i^M	s_i^{GM}	w_i
3	.365	.726	0	.206	.302	.286
4	.170	.588	−.712	−1.16	−2.08	.0662
5	.957	−3.29	−44.6	−34.1	−41.2	.000585
6	.053	.839	−1.35	−1.10	−1.76	.0637
7	.063	2.89	0	−.041	−1.15	.0175
19	.315	−2.12	−5.38	−3.52	−3.62	1.00

mansperger (1990) developed a method of studentizing rather than standardizing robust residuals that likely will be helpful in studying outliers.

Case #5 is an OLS leverage point in the full data, and it exhibits a moderately large OLS studentized residual. Clearly this point will have a large effect on the OLS fit (Cook 1977). The OLS residuals not shown were all smaller than 1 in magnitude, perhaps a clue that case #5 has inflated the scale estimate. The extreme discordance of case #5 is obvious from the more robust standardized residuals, and the MVE-based Mallows weight also identifies it as extremely outlying in the design space. The Mallows weights not shown were all equal to 1. The corresponding MVE-based Mahalanobis distances (Rousseeuw and van Zomeren 1990) provide a clear identification of several urban rivers (cases #3–7). The robust residuals also point to case #19 (the Oswegatchie River) as a possible response outlier. It is suggestive that case #19 is the largest river basin and case #5 the smallest (Haith 1976, table 2).

Table 4 presents the same diagnostics after exclusion of case #5. Only case #19 remains as a response outlier. Case #7 emerges as a moderate OLS leverage point. The Mallows weights excluding case #5 are unchanged, because the resampling algorithm (Rousseeuw and van Zomeren 1990) selects the same subsample. Is there a pattern in the residuals? Figure 1 shows plots of residuals versus *UR* for OLS, LMS, M, and GM after excluding case #5. The plot for GM reveals a pattern of negative residuals for the more urban rivers. Coupled with the huge negative residual of the much more urban Hackensack River, there is evidence of nonlinearity for large values of *UR*. The pattern fails to emerge in the other plots, for which the estimators do not have the bounded-influence property. It is clear, however, that additional leverage points could influence the fit in the plots for OLS, LMS, and M.

The nonlinearity revealed by the GM plot suggests that an alternative mechanism might come into play in urban areas. Perhaps urban areas have more efficient waste treatment, which would mitigate the effects of urbanization on water quality. One could attempt to introduce nonlinearity into the model to account for such diminishing effects; however, because nearly all information about the nonlinearity is provided by the four most urban rivers, the effect will be difficult to model reliably.

Table 4. Diagnostics for Selected Observations Excluding the Hackensack River (#5)

i	h_{ii}	t_i^{OLS}	s_i^{LMS}	s_i^M	s_i^{GM}	w_i
3	.374	.577	.153	.251	.293	.286
4	.279	−1.09	0	−.952	−2.20	.0662
6	.178	−.650	−.783	−.976	−1.96	.0637
7	.640	.865	2.56	.812	−1.07	.0175
19	.323	−3.021	−7.92	−4.68	−4.51	1.00

The preceding analysis leads us to some tentative conclusions, with caveats about the hazards of interpreting observational data. The significant contribution of agricultural use to nitrogen content persists across estimators, so this appears to be a reliable attribution. Forestland also persists as a minor, marginally significant contributor. Urbanization of rural rivers is associated with relatively large increases in nitrogen content, but there is evidence that further urbanization of substantially urban rivers has less effect. Given the size of the data set and the collinearity, attribution of nitrogen to sources is very difficult; we would not be surprised if others discovered analyses that they prefer to ours.

6. CONCLUSIONS

We have examined the behavior of one-step Mallows type robust regression methods in the linear model using either Scoring or Newton–Raphson. Two major general conclusions have emerged:

1. Under reasonably general conditions, the regression parameter estimates inherit the breakdown properties of the preliminary estimates of the regression parameters and the multivariate location and scale estimates of the design *x*'s.

2. It makes little sense to confine attention to regression parameter estimation and to completely ignore the associated problem of inference. Even when regression parameter estimates have reasonable breakdown properties, their estimated standard errors may change radically with the deletion of a single observation.

We have shown how to construct Mallows regression parameter estimates with the same breakdown properties as their standard error estimates. The Mallows weights depend on a user-chosen parameter α in (1.1). When using a redescending ψ function, the Scoring method with $\alpha \geq 2$ is recommended for inference; $\alpha \geq 1$ suffices for point estimation.

In our analysis of the New York rivers data, we used LMS as the preliminary regression parameter estimate and the MVE scatter matrix estimate for the design. Both have high breakdown points, but they are extremely inefficient estimates and might have undesirable small sample performance; see, for example, Cook and Hawkins (1990). In our example this was not a problem. In other settings, however, one might be more successful in lowering the breakdown requirement from 50% to something less ambitious, such as 20%, to avoid the exact fit property (Rousseeuw and Yohai 1984). Moreover, although any rate of convergence better than $n^{-1/4}$ is sufficient for the one-step GM estimator to be root-*n* consistent and asymptotically normal, this approximation is more accurate if the preliminary estimator has a better rate of convergence. Hence improved performance

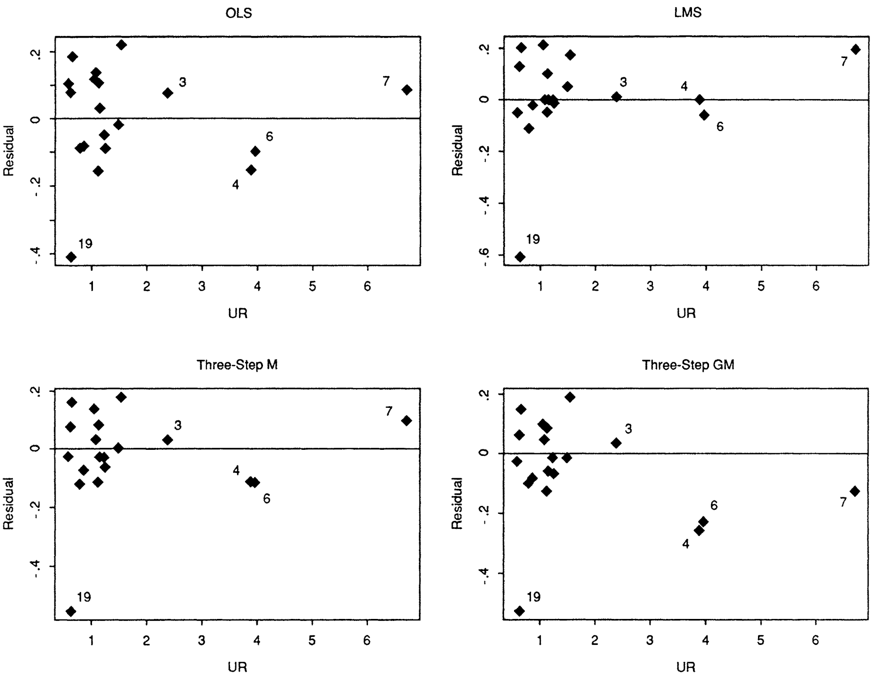


Figure 1. Residuals Versus Percent Urban Usage for Least Squares (OLS), Least Median Squares (LMS), a Three-Step M Estimator, and a Three-Step GM Estimator, Excluding the Hackensack River.

may occur using more efficient preliminary estimates such as S estimates (Rousseeuw and Yohai 1984; Davies 1987). Another way to improve the starting value is to iterate more than once, as in our analysis of the New York rivers data; we have observed empirically that three-step GM estimates starting from LMS or MVE are somewhat more stable than the one-step versions.

The behavior of one-step regression estimators with asymmetric and heteroscedastic errors deserves further study. If the regression errors are iid and symmetrically distributed, then both Scoring and Newton-Raphson have the standard large sample theory of fully iterated GM estimates. If the errors are asymmetric, however, then only Scoring improves on the rate of convergence of the preliminary estimate. On the other hand, if the errors are symmetric and heteroscedastic, then only Newton-Raphson shows this improvement. Fully iterated Mallows estimates work in either case, but they may give up the high breakdown point (Maronna et al. 1979).

The complexities encountered in the analysis of the land use data suggest that several important areas of research need further development, including stability of inference, robust model selection, and robust diagnostics.

APPENDIX: TECHNICAL PROOFS AND LEMMA

Proof of Theorem 2.1. First observe that $\|H_0^{-1}g_0\| \le \|g_0\| / |\lambda_{\min}(H_0)|$. Because $\alpha \ge 1$, we have

$$\begin{aligned}
&\hat{\sigma}^{-2}\|g_0\|^2 \\
&\quad \le \|\psi^2\|_{\text{sup}} \sum_{i=1}^{n} \|z_i\|^2 w_i^2 \le \|\psi^2\|_{\text{sup}} \\
&\qquad \times \sum_{i=1}^{n} \{1 + \|m_x\|^2 + \|x_i - m_x\|^2\} w_i^2 \\
&\quad \le \|\psi^2\|_{\text{sup}} \left\{ n(1 + \|m_x\|^2) + b \sum_{i=1}^{n} \frac{\|x_i - m_x\|^2}{(x_i - m_x)' C_x^{-1} (x_i - m_x)} \right\} \\
&\quad \le n\|\psi^2\|_{\text{sup}} \{1 + \|m_x\|^2 + b\lambda_{\max}(C_x)\}.
\end{aligned}$$

Because ψ is bounded and C_x has breakdown m/n, $\|g_0\|^2$ has breakdown at least m/n. We now must show that no matter what one does with the "bad" points, $\lambda_{\min}(H_0) > 0$.

Scoring. We have that

$$\lambda_{\min}\left(\sum_{i=1}^{n} w_i z_i z_i^t\right) \geq \lambda_{\min}\left(\sum_{i=1}^{p} w_i z_i z_i^t\right)$$

$$\geq (\inf_{1\leq j\leq p} w_j)\lambda_{\min}\left(\sum_{i=1}^{p} z_i z_i^t\right). \qquad (A.1)$$

Because by convention the first p of the $\{z_i\}$ are linearly independent, we need only consider the first factor on the right side in (A.1). This term is 0 only if $\sup_{1\leq j\leq p}\{(x_i - m_x)^t C_x^{-1}(x_i - m_x)\} = \infty$, which cannot happen because C_x has breakdown m/n and the first p observations are "good." It thus suffices to show that

$$\sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}) > 0. \qquad (A.2)$$

By (2.3), there are at least $n/2$ observations with $|r_i|/\hat{\sigma} \leq a$; so that if ψ is nondecreasing, application of (2.2) suffices to prove (A.2). Under Assumption B we find that the left side of (A.2) is at least $n/2(d_1 - d_2)$, and (A.2) then follows from (2.4).

Newton–Raphson. We must show that under arbitrary manipulation of the "bad" points

$$\lambda_{\min}\left\{\sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}) w_i z_i z_i^t\right\} > 0. \qquad (A.3)$$

When ψ is nondecreasing, $\psi^{(1)}(v) \geq 0$ and $\psi^{(1)}(r_i/\hat{\sigma}) \geq d_1 > 0$ for at least $n - m - n/2$ "good" points. Thus (A.3) follows from Assumption C.

Proof of Lemma 2.1. For the first part of the lemma, it suffices to replace H_0 by $\sum_1^n w_i z_i z_i^t$. As in (A.1), $\lambda_{\max}(\sum_{i=1}^n w_i z_i z_i^t) \geq \lambda_{\max}(\sum_{i=n-m+1}^n w_i z_i z_i^t)$. Now letting $\|z_j\| \rightarrow \infty$ for $j \geq n - m + 1$, we have $w_j \sim b^{\alpha/2}(x_j^t C_x^{-1} x_j)^{-\alpha/2} \sim b^{\alpha/2}\|x_j\|^{-\alpha}(g_j C_x^{-1} g_j)^{-\alpha/2} \sim b^{\alpha/2}\|z_j\|^{-\alpha}(g_j C_x^{-1} g_j)^{-\alpha/2}$, where $g_i = x_i/\|x_i\|$, because C_x and m_x have breakdown m/n. But because $g_j^t C_x^{-1} g_j \leq \lambda_{\max}(C_x^{-1}) = \{\lambda_{\min}(C_x)\}^{-1}$, it then follows that in the limit, as $\|z_j\| \rightarrow \infty$, $(\inf_{n-m+1\leq j\leq n} w_j) \geq \frac{1}{2}\{b\lambda_{\min}(C_x)\}^{\alpha/2}\|z_j\|^{-\alpha}$, and hence that $\lambda_{\max}(\sum_{i=1}^n w_i z_i z_i^t) \geq \frac{1}{2}\{b\lambda_{\min}(C_x)\}^{\alpha/2}\, \lambda_{\max}(\sum_{i=n-m+1}^n d_i d_i^t \|z_i\|^{2-\alpha})$. This can be made to diverge to ∞ if $\alpha < 2$. If $\alpha \geq 2$, then $\lambda_{\max}(\sum_{i=1}^n w_i z_i z_i^t) \leq \sum_{i=1}^n \|z_i\|^2 w_i \leq \sum_{i=1}^n (1 + \|m_x\|^2 + \|x_i - m_x\|^2) w_i \leq n\{1 + \|m_x\|^2 + b\lambda_{\max}(C_x)\}$, the last step following because $\alpha \geq 2$.

Proof of Theorem 4.1. We derive a more general result that holds even if the errors are asymmetric, as in Section 4.2. Let η_0 be the limiting value of the preliminary estimate of the intercept. Let $\beta_0 = (\eta_0, \gamma^t)^t$, $u_i = y_i - z_i^t\beta_0$, and $G(\beta, \sigma) = \sigma \sum_{i=1}^n \psi(u_i/\sigma) w_i z_i = \sigma \sum_{i=1}^n \psi((r_i + z_i'(\hat{\beta}_0 - \beta_0))/\sigma) w_i z_i$.

Newton–Raphson. Conditions B1 and B2 and the mean value theorem yield

$$G(\beta, \hat{\sigma}_0) = \hat{\sigma}_0 \sum_{i=1}^{n} \psi(r_i/\hat{\sigma}_0) w_i z_i + \sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}_0) w_i z_i z_i'(\hat{\beta}_0 - \beta_0)$$

$$+ \frac{1}{2}\hat{\sigma}_0^{-1} \sum_{i=1}^{n} \psi^{(2)}\left(\frac{r_i + z_i'(\tilde{\beta}_0 - \beta_0)}{\hat{\sigma}_0}\right) w_i z_i (z_i'(\hat{\beta}_0 - \beta_0))^2$$

$$= H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0) + O\left(\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2 \sum_{i=1}^{n} w_i \|z_i\|^3\right), \qquad (A.4)$$

where $\tilde{\beta}_0$ is on the line segment between β_0 and $\hat{\beta}_0$. On the other hand, applying the mean value theorem to $g(s) = G(\beta, s)$ yields, after some simplification,

$$G(\beta, \hat{\sigma}_0) = \hat{\sigma}_0 \sum_{i=1}^{n} \psi(u_i/\sigma) w_i z_i$$

$$- (\hat{\sigma}_0 - \sigma) \sum_{i=1}^{n} \psi^{(1)}(u_i/\sigma)(u_i/\sigma) w_i z_i$$

$$+ \frac{1}{2}(\hat{\sigma}_0 - \sigma)^2 \tilde{\sigma}^{-1} \sum_{i=1}^{n} \psi^{(2)}(u_i/\tilde{\sigma})(u_i/\tilde{\sigma})^2 w_i z_i,$$

where $\tilde{\sigma}$ is between $\hat{\sigma}_0$ and σ. Equating (A.4) and (A.5),

$$H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0) = \hat{\sigma}_0 \sum \{\psi(u_i/\sigma) - a_0\} w_i z_i$$

$$- (\hat{\sigma}_0 - \sigma) \sum \{\psi^{(1)}(u_i/\sigma)(u_i/\sigma) - b_1\} w_i z_i$$

$$+ \{a_0\hat{\sigma}_0 + b_1(\sigma - \hat{\sigma}_0)\} \sum w_i z_i$$

$$+ O(\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2 \sum w_i \|z_i\|^3)$$

$$+ O((\hat{\sigma}_0 - \sigma)^2 \tilde{\sigma}^{-1} \sum w_i \|z_i\|),$$

with $a_0 = E[\psi(u_1/\sigma)]$ and $b_1 = E[\psi^{(1)}(u_1/\sigma)(u_1/\sigma)]$.

The assumption on $\hat{\sigma}_0$, Conditions A1, A2, B1(b), C1, and Chebyshev's inequality imply $n^{-1/2}(\hat{\sigma}_0 - \sigma) \sum \{\psi(u_i/\sigma) - a_0\} w_i z_i = O_p(n^{-\tau})$ and $n^{-1/2}(\hat{\sigma}_0 - \sigma) \sum \{\psi^{(1)}(u_i/\sigma)(u_i/\sigma) - b_1\} w_i z_i = O_p(n^{-\tau})$. Moreover, by C1, $n^{-1/2}\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2 \sum w_i \|z_i\|^3 = O_p(n^{1/2-2\tau})$ and $n^{-1/2}(\hat{\sigma}_0 - \sigma)^2 \tilde{\sigma}^{-1} \sum w_i \|z_i\| = O_p(n^{1/2-2\tau})$. Observing also that $\tau \leq \frac{1}{2}$ implies $2\tau - \frac{1}{2} \leq \tau$, we have

$$n^{-1/2} H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0)$$

$$= n^{-1/2}\sigma \sum_{i=1}^{n} \{\psi(u_i/\sigma) - a_0\} w_i z_i + B_n + O_p(n^{1/2-2\tau}), \qquad (A.6)$$

where the bias term is $B_n = n^{-1/2}\{a_0\hat{\sigma}_0 + b_1(\sigma - \hat{\sigma}_0)\} \sum_{i=1}^n w_i z_i$. Condition D2 implies $B_n = 0$, which establishes (4.2) for Newton–Raphson.

Scoring. Observe that

$$H_{\mathrm{S}}(\hat{\beta}_{\mathrm{S}} - \beta_0) = H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0) + (H_{\mathrm{S}} - H_{\mathrm{NR}})(\hat{\beta}_0 - \beta_0). \qquad (A.7)$$

Hence if we show that the components of $(H_{\mathrm{S}} - H_{\mathrm{NR}})$ are of order $O_p(n^{1-\tau})$, it will then follow that expansion (A.6) holds with $\hat{\beta}_{\mathrm{NR}}$ and H_{NR} replaced by $\hat{\beta}_{\mathrm{S}}$ and H_{S}. Setting $g(t) = \psi^{(1)}(t)$ and $c_i = n^{-1} w_i z_{ij} z_{ik}$ in Lemma A.1 shows that

$$n^{-1}H_{\mathrm{NR}} = n^{-1}Q_n + O_p\left(\frac{1}{n^{1+\tau}} \sum_{i=1}^{n} w_i \|z_i\|^2 (1 + \|z_i\|)\right.$$

$$\left. + \left\{\frac{1}{n^2} \sum_{i=1}^{n} w_i^2 \|z_i\|^4\right\}^{1/2}\right), \qquad (A.8)$$

replacing ε_i by u_i in the definition of Q_n. On the other hand, setting $c_i = n^{-1}$ shows that $n^{-1} \sum \{\psi^{(1)}(r_i/\hat{\sigma}_0) - E[\psi^{(1)}(u_i/\sigma)]\} = O_p(n^{-\tau}(1 + n^{-1} \sum \|z_i\|) + n^{-1/2})$, from which it follows that

$$n^{-1}H_{\mathrm{S}} = n^{-1}Q_n$$

$$+ \mathrm{O}_p\left(n^{-\tau}\left(1 + n^{-1} \sum_{i=1}^{n} \|z_i\|\right) n^{-1} \sum_{i=1}^{n} w_i \|z_i\|^2\right). \qquad (A.9)$$

Comparing (A.8) and (A.9) shows that $n^{-1}(H_{\mathrm{S}} - H_{\mathrm{NR}}) = O_p(n^{-\tau})$, whence

$$n^{-1/2} H_{\mathrm{S}}(\hat{\beta}_{\mathrm{S}} - \beta_0)$$

$$= n^{-1/2}\sigma \sum_{i=1}^{n} \{\psi(u_i/\sigma) - a_0\} w_i z_i + B_n + O_p(n^{1/2-2\tau}). \qquad (A.10)$$

Lemma A.1. Suppose $\{u_i\}$ are independent, $n^{\tau}(\hat{\beta}_0 - \beta_0) = O_p(1)$, and $n^{\tau}(\hat{\sigma}_0 - \sigma) = O_p(1)$ with $\sigma > 0$. Let $\{c_i\}$ be a sequence of finite constants. If a measurable function g satisfies the Lipschitz condition

$$|g(s)-g(t)| \leq L|s-t|/(1+|t|), \quad (\text{all } s, t) \qquad (\text{A.11})$$

for a finite constant L, then

$$\sum_{i=1}^{n} c_i\left\{g\left(\frac{u_i + z_i^t(\beta_0 - \hat{\beta}_0)}{\hat{\sigma}_0}\right) - E[g(u_i/\sigma)]\right\}$$
$$= O_p\left(n^{-\tau}\sum_{i=1}^{n}|c_i|(1+\|z_i\|) + \left\{\sum_{i=1}^{n} c_i^2\right\}^{1/2}\right).$$

Proof. Condition (A.11) implies

$$\left|g\left(\frac{u_i + z_i^t(\beta_0 - \hat{\beta}_0)}{\hat{\sigma}_0}\right) - g(u_i/\hat{\sigma}_0)\right| \leq L\frac{|z_i^t(\beta_0 - \hat{\beta}_0)|}{\hat{\sigma}_0}$$
$$\leq L\frac{\|z_i\|\|\beta_0 - \hat{\beta}_0\|}{\hat{\sigma}_0}$$

and $|g(u_i/\hat{\sigma}_0) - g(u_i/\sigma)| \leq L|u_i\|\hat{\sigma}_0^{-1} - \sigma^{-1}|/(1+|u_i|\sigma^{-1}) \leq L\hat{\sigma}_0^{-1}|\sigma - \hat{\sigma}_0|$. Hence

$$\sum_{i=1}^{n}|c_i|\left|g\left(\frac{u_i + z_i^t(\beta_0 - \hat{\beta}_0)}{\hat{\sigma}_0}\right) - g(u_i/\sigma)\right|$$
$$\leq L\left(\frac{\|\beta_0 - \hat{\beta}_0\| + |\sigma - \hat{\sigma}_0|}{\hat{\sigma}_0}\right)\sum_{i=1}^{n}|c_i|(1+\|z_i\|).$$

Condition (A.11) also implies that g is continuous and bounded between $g(0) \pm L$. Hence the sum $\Delta_n = \sum_{i=1}^n c_i\{g(u_i/\sigma) - E[g(u_i/\sigma)]\}$ has mean 0 and variance bounded by $\{|g(0)| + L\}^2 \sum c_i^2$. Chebyshev's inequality implies $\Delta_n = O_p(\{\sum c_i^2\}^{1/2})$.

Proof of Theorem 4.2. To prove the result for nonexchangeable M_0, use Lemma A.1 with $g = \psi^2$ and $c_i = n^{-1}w_i^2 z_{ik}z_{il}(k, l \in \{1, \ldots, n\})$. For exchangeable M_0 set $c_i = n^{-1}$ to show $n^{-1}\sum \psi^2(r_i/\hat{\sigma}_0) = E\psi^2(\varepsilon_1/\sigma) + O_p(n^{-\tau}(1 + n^{-1}\sum\|z_i\|))$.

Proof of Lemma 4.1. The expansion for Scoring follows from (A.10), because $\sum_{i=1}^n w_i x_i = 0$ and H_S is block diagonal.

For Newton–Raphson, first recall that $n^{-1}(H_{\text{NR}} - Q_n) = O_p(n^{-\tau})$ by (A.8), and $q_{(1)} = 0$ due to the centering in (4.6). It follows that $h_{(1)}h_{11}^{-1} = O_p(n^{-\tau})$ and

$$n^{-1}H_{22\cdot 1} = n^{-1}H_{22} + O_p(n^{-2\tau}) = n^{-1}Q_{22} + O_p(n^{-\tau}). \qquad (\text{A.12})$$

Next rearrange (A.6) to obtain

$$H_{22\cdot 1}(\hat{\gamma} - \gamma) = h_{(1)}h_{11}^{-1}(b_n + S_1) + S_2 + O_p(n^{1-2\tau}), \qquad (\text{A.13})$$

where $b_n = \{a_0\hat{\sigma}_0 + b_1(\sigma - \hat{\sigma}_0)\}\sum w_i = \{a_0\hat{\sigma}_0 + O_p(n^{-\tau})\}\sum w_i$, $S_1 = \sigma\sum_{i=1}^n\{\psi(u_i/\sigma) - a_0\}w_i = O_p(\{\sum w_i^2\}^{1/2})$, and $S_2 = \sigma\sum_{i=1}^n\{\psi(u_i/\sigma) - a_0\}w_i x_i$. In (A.13) the term $h_{(1)}h_{11}^{-1}S_1 = O_p(n^{1/2-\tau}) = o_p(n^{1-2\tau})$, which can be absorbed into the remainder. Further we have $h_{11}^{-1}b_n = a_1^{-1}a_0\hat{\sigma}_0 + O_p(n^{-\tau})$, so it remains to detail the large sample behavior of $\hat{\sigma}_0 h_{(1)}$. An application of the mean value theorem yields

$$\hat{\sigma}_0 h_{(1)} = \hat{\sigma}_0\sum\psi^{(1)}(r_i/\hat{\sigma}_0)w_i x_i$$
$$= \hat{\sigma}_0\sum\psi^{(1)}(u_i/\hat{\sigma}_0)w_i x_i + \sum\psi^{(2)}(u_i/\hat{\sigma}_0)w_i x_i z_i^t(\hat{\beta}_0 - \beta_0)$$
$$+ O(\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2\|\psi^{(3)}\|_{\text{sup}}\sum w_i\|z_i\|^3). \qquad (\text{A.14})$$

Further expansion of the first term in (A.14) yields

$$\hat{\sigma}_0\sum\psi^{(1)}(u_i/\hat{\sigma}_0)w_i x_i$$
$$= \hat{\sigma}_0\sum\psi^{(1)}(u_i/\sigma)w_i x_i - (\hat{\sigma}_0 - \sigma)\sum\psi^{(2)}(u_i/\sigma)(u_i/\sigma)w_i x_i$$
$$+ \mathrm{O}\left(\frac{(\hat{\sigma}_0 - \sigma)^2}{\tilde{\sigma}}\|(\cdot)^2\psi^{(3)}(\cdot)\|_{\text{sup}}\sum w_i\|x_i\|\right),$$

where $\tilde{\sigma}$ is between σ and $\hat{\sigma}_0$. Because of the centering, Chebyshev's inequality and the conditions on ψ and x imply that $\sum\psi^{(1)}(u_i/\hat{\sigma}_0))w_i x_i = O_p(n^{1/2})$ and $\sum\psi^{(2)}(u_i/\sigma)(u_i/\sigma)w_i x_i = O_p(n^{1/2})$. Hence $\hat{\sigma}_0\sum\psi^{(1)}(u_i/\hat{\sigma}_0)w_i x_i = \sigma\sum\{\psi^{(1)}(u_i/\sigma) - a_1\}w_i x_i + O_p(n^{1/2-\tau}) + O_p(n^{1-2\tau})$. To handle the second term in (A.14), note that $\sum\psi^{(2)}(u_i/\hat{\sigma}_0)w_i x_i z_i^t = a_2\sum w_i x_i z_i^t + O_p(n^{-\tau}\sum w_i) + O_p(\{\sum w_i^2\}^{1/2}) = a_1^{-1}a_2[q_{(1)}; Q_{22}] + O_p(n^{1-\tau})$. Thus we have

$$\hat{\sigma}_0 h_{(1)} = \sigma\sum\{\psi^{(1)}(u_i/\sigma) - a_1\}w_i x_i$$
$$+ a_1^{-1}a_2 Q_{22}(\hat{\gamma}_0 - \gamma) + O_p(n^{1-2\tau}). \qquad (\text{A.15})$$

Combining (A.12), (A.13), and (A.15) completes the proof.

Proof of Lemma 4.2. The proof of Theorem 4.1 for Newton–Raphson extends immediately to the present case. To handle Scoring, use (A.7) and observe that, by Lemma A.1,

$$n^{-1}(H_{\text{NR}} - H_S)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{E\left[\psi^{(1)}\left(\frac{\varepsilon_i}{\sigma}\right)\right] - \frac{1}{n}\sum_{j=1}^{n}E\left[\psi^{(1)}\left(\frac{\varepsilon_j}{\sigma}\right)\right]\right\}w_i z_i z_i^t + O_p(n^{-\tau}).$$

As an example where this matrix fails to vanish asymptotically, let the empirical covariance between $E[\psi^{(1)}(\varepsilon_i/\sigma)]$ and $w_i z_{ip}^2$ converge to unity as $n \to \infty$.

Proof of Theorem 4.3. This follows from Lemma 4.2 and an application of Lemma A.1 to M_0.

[*Received January 1990. Revised February 1991.*]

REFERENCES

Allen, D. M., and Cady, F. B. (1982), *Analyzing Experimental Data by Regression,* Belmont, CA: Lifetime Learning Publications.

Andrews, D. F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics,* 16, 523-531.

Bickel, P. J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association,* 70, 428-434.

Carroll, R. J., and Welsh, A. H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician,* 42, 285-287.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R. D., and Hawkins, D. M. (1990), on "Unmasking Multivariate Outliers and Leverage Points" by P. J. Rousseeuw and B. C. van Zomeren, *Journal of the American Statistical Association,* 85, 640-644.

Davies, P. L. (1987), "Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics,* 15, 1269-1244.

Davies, L. (1990), "The Asymptotics of S-Estimators in the Linear Regression Model," *The Annals of Statistics,* 18, 1651-1675.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," In *A Festchrift for Erich L. Lehmann,* eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges Jr., Belmont, CA: Wadsworth, pp. 157-184.

De Jongh, P. J., De Wet, T., and Welsh, A. H. (1987), "Mallows-Type Bounded-Influence Trimmed Means," *Journal of the American Statistical Association,* 84, 805-810.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986), "Some New Estimation Methods for Weighted Regression When There Are Possible Outliers," *Technometrics,* 28, 219-230.

Haith, D. A. (1976), "Land Use and Water Quality in New York Rivers," *Journal of the Environmental Engineering Division, Proceedings of the American Society of Civil Engineers,* 102, 1-15.

Hampel, F. R. (1978), "Optimally Bounding the Gross-Error-Sensitivity and the Influence of Position in Factor Space," in *Proceedings of the Statistical Computing Section, American Statistical Association,* pp. 59-64.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions,* New York: John Wiley.

He, X., Simpson, D. G., and Portnoy, S. (1990), "Breakdown Robustness of Tests," *Journal of the American Statistical Association,* 85, 446-452.

He, X., and Simpson, D. G. (in press), "Robust Direction Estimation," *The Annals of Statistics.*

Hettmansperger, T. P., and McKean, J. P. (1977), "A Robust Alternative Based on Ranks to Least Squares in Analyzing Linear Models," *Technometrics,* 19, 275-284.

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics,* 1, 799-821.

Jaeckel, L. A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of Residuals," *The Annals of Mathematical Statistics,* 43, 1449–1458.

Jureckova, J., and Portnoy, S. (1987), "Asymptotics for One-Step M-Estimators in Regression With Application to Combining Efficiency and High Breakdown Point," *Communications in Statistics, Theory, and Methods,* 16(8), 2187–2199.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191–219.

Krasker, W. S. (1980), "Estimation in Linear Regression Models With Disparate Data Points," *Econometrica,* 48, 1333–1346.

Krasker, W. S., and Welsch, R. E. (1982), "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association,* 77, 595–604.

Lopuhaä, H. P. (1989), "On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance," *The Annals of Statistics,* 17, 1662–1683.

Mallows, C. L. (1975), "On Some Topics in Robustness," technical memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R. A., Bustos, O. H., and Yohai, V. J. (1979), "Bias- and Efficiency-Robustness of General M-Estimators for Regression With Random Carriers," in *Smoothing Techniques for Curve Estimation,* eds. T. Gasser and M. Rosenblatt, New York: Springer-Verlag, pp. 91–116.

Martin, R. D., Yohai, V. J., and Zamar, R. H. (1989), "Min-Max Bias Robust Regression," *The Annals of Statistics,* 17, 1608–1630.

McKean, J. W., Sheather, S. J., and Hettmansperger, T. P. (1990), "On the Use of Standardized Residuals From a High Breakdown GM-Fit of a Linear Model," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association,* pp. 242–247.

Morgenthaler, S. (1989), "Comment on Yohai and Zamar," *Journal of the American Statistical Association,* 84, 636.

Ronchetti, E., and Rousseeuw, P. J. (1985), "Change-of-Variance Sensitivities in Regression Analysis," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 68, 503–519.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871–880.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection,* New York: John Wiley.

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633–639.

Rousseeuw, P. J., and Yohai, V. (1984), "Robust Regression by Means of S-Estimators," in *Robust and Nonlinear Time Series Analysis,* eds. J. Franke, W. Hardle, and R. D. Martin, New York: Springer-Verlag, pp. 256–272.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828–838.

Simpson, D. G., Carroll, R. J., and Ruppert, D. (1987), "M-Estimation for Discrete Data: Asymptotic Distribution Theory and Implications," *The Annals of Statistics,* 15, 657–669.

Simpson, D. G., Ruppert, D., and Carroll, R. J. (1989), "One-Step GM-Estimates for Regression With Bounded Influence and High Breakdown Point," Technical Report No. 859, Cornell University, School of Operations Research and Industrial Engineering.

Stefanski, L. A. (1991), "A Note on High-Breakdown Estimators," *Statistics and Probability Letters,* 11, 353–358.

Welsh, A. H. (1989), "On M-Processes and M-Estimation," *The Annals of Statistics,* 17, 337–361.

Yohai, V. J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics,* 15, 642–656.

Yohai, V. J., and Zamar, R. H. (1988), "High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association,* 83, 406–413.

Chapter 7
Other Work

By George Casella

About the Author. George Casella was a long-time friend of Ray's and leader in the field of Statistics who passed away in 2012. At the time of his death, George was Distinguished Professor in the Department of Statistics at the University of Florida. He was active in many aspects of statistics, having contributed to theoretical statistics in the areas of decision theory and statistical confidence, to environmental statistics, and has more recently to statistical genomics and political science methodology. He also had strong research interests in the theory and application of Monte Carlo and other computationally intensive methods. He was elected a Foreign Member of the Spanish Royal Academy of Sciences (2010) and a Fellow of the American Association for the Advancement of Science (2012). In other capacities, he served as Theory and Methods Editor of the *Journal of the American Statistical Association*, 1996–1999, Executive Editor of *Statistical Science*, 2002–2004, and Joint Editor the *Journal of the Royal Statistical Society, Series B*, 2009–2012. He authored seven textbooks, including *Statistical Inference, Second Edition*, 2001, with Roger Berger, and *Monte Carlo Statistical Methods, Second Edition*, 2004, with Christian Robert. His friendship with Ray Carroll went back to their graduate studies at Purdue in the seventies. Although their statistical expertise tended to be in different areas, they kept familiar with each other's work and saw each other regularly, both professionally and on the golf course.

This commentary was written in the year before George's passing.

Selected Papers on Other Work

[OW-1]-[52] Carroll, R. J. and Lombard, F. (1985). A note on N-estimators for the binomial distribution. *Journal of the American Statistical Association*, 80, 423–426.

[OW-2]-[339] Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175–188.

[OW-3]-[117] Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96, 1387–1396.

[OW-4]-[115] Molenberghs, G., Thijs, H., Kenward, M. G., Carroll, R. J., Mallinckrodt, C., Jansen, I., and Beunckens, C. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5, 445–464.

Imagine your first day of graduate school, and one of the first people you meet is Ray Carroll. It was Ray's second year (and almost final year, as he only took two-and-a-half years to get a PhD), and right then you know that you have no chance. How on earth can someone compete in the Carroll arena? The secret? There is no competition. As the rest of this introduction shows, Ray Carroll is not only one of the best statisticians on the planet, he is one of the sweetest people you will ever meet. Much of his career has been spent mentoring young statisticians, and showing them how to solve real problems in meaningful ways.

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_7,

Ray and I have been friends for over 30 years. Although we never lived or worked in the same place, we are always in touch. Meetings and university visits have helped, and throughout we have always found time to play a round of golf. There are many stories, but the only one I will share took place in North Carolina in the 1980s, when I was visiting Ray. He made reservations at Pinehurst (one of the most famous, and expensive, golf courses in the world). Reservations were almost impossible to get if you were not a member, but Ray convinced them that a very famous professor was visiting North Carolina, and they should accommodate him (me—ha!) for the benefit of the university. They not only let us on the course, but also since Ray was not a member they had no way of billing us, so we played for free!

This section chapter is, in many ways, quintessential Ray. Although the other chapters show his deep contributions in many areas, this chapter highlights his personality and approach to statistics, and how his influence on researchers, both young and old, transcends the field. Ray is always positive, upbeat, and dedicated to good science and solving good problems. The overall theme of these works is that Ray talks with someone—often a young researcher—understands the problem, suggests a solution, and produces a substantial contribution. Consider the range of these four papers, from the applied treatment of wildlife data and informative censoring, to the efficiency of sandwich estimators and a deep look at handling missing data in longitudinal studies.

Binomial N

Carroll and Lombard (1985, [OW-1]) is my personal favorite, and I still recall first reading it and thinking—what a cool solution! It spurred me to also write a paper on n estimation (nowhere near as influential as Ray's). This problem actually has a long history, going back to a disagreement between Fisher and Haldane (references in the paper). Ray's coauthor Fred Lombard recalls, "Going back 30 years is a bit of a stretch, however, I can recall the following details: On a visit to the Kruger Park with Ray in 1982 I mentioned to him that I had been involved with the Park authorities in trying to extract useful information about animal numbers from counts they had done. The data were of the binomial(n,p) type—p unknown and n to be estimated. Maximum likelihood estimation didn't yield much because the likelihood functions were 'flat as pancakes' in a large neighborhood of the maximum. He suggested that the nuisance parameter p be integrated out—an idea which led eventually to the paper."

This problem, and the type of solution provided (marginal likelihood), has actually become more important. Ray notes, "I have recently been using the thinking from that problem to analyze the number of significant SNPs in genome wide association studies, together with Nilanjan Chatterjee." I can also attest that these unknown n models have become very useful in modeling *RNA-seq data*.

Informative Censoring

Wu and Carroll (1988, [OW-2]) is another paper motivated by a real problem, that of assessing the effectiveness of certain therapies for treating lung disease, and the possibility of there being informative censoring. Ray and Margaret Wu look at the effect of informative censoring on the usual estimates, in particular bias and loss of efficiency, and propose a solution based on a probit model for the informative censoring. The impact of this paper is clear, it has been cited over 350 times.

Ray notes, "This paper was, I think, the first paper in biostatistics to actually think about informative censoring,[1] which is now also called Missing Not at Random or MNAR, and it had a major impact at the time. Margaret and I had talked about the problem when I was at NHLBI in 1980–1832, and in 1986 she called me and suggested we finish the initial work, which I had completely forgotten but for which she had kept notes. We did, and submitted it to Biometrics, where it was quickly accepted. Neither of us thought this was a big deal paper, and both of us were obviously wrong."

Margaret Wu has retired from NHLBI, and attempts to contact her for comments have failed.

Sandwich

Of Kauermann and Carroll (2001, [OW-3]), Ray notes "I like to think of this as a little gem of a paper," and I concur. Every once in a while we see a short paper in a top journal that has an important message. This is one, where Ray and Göran Kauermann show that the sandwich variance estimator suffers from increased variability and that the resulting confidence intervals do not attain their nominal level. They propose an adjustment that mediates this problem.

Göran Kauermann notes "I was at a pretty early stage of my career and I remember that Ray was giving me a manuscript he had just finished with coauthors where he showed in simulations and on theoretical grounds that the use of sandwich variance estimation leads to undercoverage of confidence intervals. I found the result interesting and started working on a correction for the undercoverage." This joint work lead to the sandwich paper, which has had a major impact.

Missing Data

Ray's approach to science is totally professional, with the goal of doing and promoting good science. Molenberghs et al. (2004, [OW-4]) grew out of a talk on Ray's website entitled, "Last Observation Carried Forward Is A Stupid Method For Handling Missing Data in Longitudinal Studies." Understand that there is no malice or insult here; it is a statement that this is not good science, and must be fixed. This paper is a very careful look at missing data methods in longitudinal studies, characterizing their properties and recommending good procedures. Although the authors do not actually say that Last Observation Carried Forward is stupid, they show that it is biased, can distort the variance and covariance structure of a model, and can almost certainly lead to incorrect inferences.

Ray notes that this paper "... is one of a series inspired by Craig Mallinckrodt of Eli Lilly, who made the professionally bold decision to push for modern mixed

[1] No, but real close. A Web of Science search on "nonignorable nonresponse" or "informative missingness" or "informative censoring" finds 520 articles, with only three older than Wu and Carroll (but not by much). Moreover, of the 520 articles, Wu and Carroll is number 4 in citations.

model statistical methods for handling missing data." Geert Molenberghs, speaking for the "team," notes "Early in the first decade of this century, Ray came up to talk to me during the Joint Statistical Meetings. He had been working with Eli Lilly and Company, and wanted to make sure that they would be using proper, modern, adequate missing data methodology and had been referring to our book (Verbeke and Molenberghs, 2000), in which we spent one paragraph on the infamous Last Observation Carried Forward method. Ray forged a satisfying, socially pleasant, and lasting relationship and, over time, it started to involve many other people in our department. The 2004 Biostatistics paper is a direct testimony to this collaboration and a first milestone of it. Nothing of this would have happened without Ray."

Endnote

These papers solve a variety of real problems and provide sophisticated solutions that are not only applicable but that also move the theory forward. To this I say, "Bravo, Ray." Perhaps this is the most important lesson we can learn from Ray—let the real problems guide you, and bring the theory to bear to produce a useable solution.

References

Publications by other authors cited in this chapter.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

A Note on *N* Estimators for the Binomial Distribution

RAYMOND J. CARROLL and F. LOMBARD*

Consider k success counts from a binomial distribution with unknown N and success probability p. We examine the problem of estimating N. By integrating the likelihood for N and p over a beta density for p, we obtain the beta-binomial distribution resulting in stable and reasonably efficient estimators of N, which compare favorably with and are often better than the estimates introduced by Olkin et al. (1981).

KEY WORDS: Analysis of count data; Maximum likelihood; Method of moments; Unstable estimators.

1. INTRODUCTION

Olkin, Petkau, and Zidek (OPZ; 1981) considered the problem of estimating the parameter N based on independent success counts $s_1, \ldots, s_k$ from a binomial distribution with unknown parameters N and p. They showed that the method of moments estimator (MME; see Haldane 1942) and the maximum likelihood estimator (MLE; see Fisher 1942) of N can be extremely unstable in the sense that changing an observed success count s to $s + 1$ can result in a massive change in the estimate of N. The difficulty arises when the method of moments estimates of mean and variance, $\hat{\mu}$ and $\hat{\sigma}^2$, are nearly equal, so the success probability p is apparently small.

To overcome the instability of the MME and MLE of N, OPZ introduced two estimators that they showed to be stable. The first is MLE:S, which is either the ordinary MLE or a jackknifed version of the maximum success count, depending on whether $\hat{\mu}/\hat{\sigma}^2 \geq 1 + 1/\sqrt{2}$. The stabilized MME:S also varies: if $\hat{\mu}/\hat{\sigma}^2 \geq 1 + 1/\sqrt{2}$, the usual MME is used; otherwise a ridge tracing method is employed (Hoerl and Kennard 1970).

Both MME:S and MLE:S are reasonably stable, and OPZ demonstrated in a convincing Monte Carlo Study that these estimators dominate the ordinary MME and MLE. They also showed that the ridge-stabilized MME:S is generally a better estimator of N than is the jackknife-stabilized MLE:S, except in unstable cases in which p is large. The purpose of this article is to describe a simple, stable estimator that is closely related to the MLE and seems to be competitive with and often superior to both MLE:S and MME:S in terms of mean squared error.

2. A NEW CLASS OF ESTIMATORS

The instability of the MME and MLE arises when p is apparently near zero. OPZ cited a case in which $N = 75$, $p = .32$, and the success counts are 16, 18, 22, 25, and 27. Even though p is not small, the natural estimate of it from the observed counts is $1 - \hat{\sigma}^2/\hat{\mu} = .21$. This is an example of an unstable case, since $\hat{\mu}/\hat{\sigma}^2 = 1.27$ even though $E\hat{\mu}/E\hat{\sigma}^2 = 1.84$. The extreme instability of the MLE and MME of N in this case was noted by OPZ. In the examples that motivated this research, namely counting the number of impala herds and individual waterbuck in the Kruger National Park, South Africa, it is fairly certain that p is much different from zero (see Sec. 4 for details). We reasoned from these examples that a stable procedure ought to be obtained if one smoothly builds in automatic discounting of data for which p is apparently near zero. In particular, it seemed that fairly stable procedures with good frequentist properties could be obtained by pretending that p had a beta distribution with parameters (a, b) and then looking at the likelihood obtained after integrating out p. Specifically, for $0 < p < 1$ and $N \geq s_{\max} = \max(s_1, \ldots, s_k)$, write the likelihood of the data as

$$L(N, p) = \left\{ \prod_{i=1}^{k} \binom{N}{s_i} \right\} p^{\Sigma_1^k s_i} (1 - p)^{kN - \Sigma_1^k s_i}. \tag{1}$$

Suppose for the moment that the density of p is proportional to

$$p^a (1 - p)^b, \tag{2}$$

where a and b are integers. To eliminate the nuisance parameter, multiply (1) and (2) and integrate over p to obtain an integrated likelihood for N:

$$\mathcal{L}(N) = \left\{ \prod_{i=1}^{k} \binom{N}{s_i} \right\} \times \left[(kN + a + b + 1) \binom{kN + a + b}{a + \sum_1^k s_i} \right]^{-1} \quad \text{for} \quad N \geq s_{\max}. \tag{3}$$

The estimate Mbeta (a, b) of N is obtained by maximizing (3) as a function of $N \geq \max(s_1, \ldots, s_k)$. Of course, in the standard terminology, (3) is the beta-binomial likelihood.

The idea of maximizing (3) as a function of N was justified in a Bayesian context by Draper and Guttman [1971; our Eq. (3) is equivalent to their (2.8)]. A non-Bayesian justification for eliminating nuisance parameters by integrating them out is given by Barnard et al. (1962; see pp. 348–350 in particular).

For every $a \geq 0$ and $b \geq 0$, the integrated likelihood (3) is maximized at some finite N. This follows because $\mathcal{L}(N) \to 0$ as $N \to \infty$, using Stirling's formula. We do not know if (3) always has a unique maximum when considered as a continuous function of N. In our calculations, however, we always found that (3) was either decreasing or first increasing and then decreasing in N, suggesting that (3) does have a unique maximum. DeRiggi (1983) showed that the likelihood function (1) evaluated at $p = \Sigma s_i/kN$ is unimodal; we have been unable to prove a similar result for (3).

* Raymond J. Carroll is Professor of Statistics, University of North Carolina, 315 Phillips Hall 039 A, Chapel Hill, NC 27514; and F. Lombard is Professor of Statistics, University at South Africa, P.O. Box 392, Pretoria, South Africa. Carroll's work was supported by Air Force Office of Scientific Research Contract F 49620 82 C 0009. Lombard's work was supported by research grants from the University of South Africa and the Council for Scientific and Industrial Research. The authors thank the associate editor for detailed and helpful comments and the referees for their suggestions.

© 1985 American Statistical Association
Journal of the American Statistical Association
June 1985, Vol. 80, No. 390, Theory and Methods

Of course, one need not be restricted to having the distribution of p given N be beta (a, b). Indeed, different but fairly unnatural choices of distributions for p given N lead to some familiar, unstable estimators. For example, the MLE is formed by supposing that given N, p has a point mass distribution at $\hat{p}(N) = \Sigma_1^k s_i/N$. Following the prescription that led from (1) to (3) gives

$$\mathcal{L}(N : \text{MLE}) = \left\{\prod_{i=1}^{k}\binom{N}{s_i}\right\}\hat{p}(N)^{\Sigma_1^k s_i}(1 - \hat{p}(N))^{kN-\Sigma_1^k s_i}$$

Part of the instability of the MLE may be due to the fact that as $N \to \infty$, it does not follow that $\mathcal{L}(N : \text{MLE}) \to 0$.

A second, rather strange choice is to pretend that the density of p is proportional to $1/p$ $(0 < p < 1)$. This gives extreme weight to small values of p and, as might be expected, leads to a very unstable estimator. It turns out that this unstable estimator is equivalent to the one obtained by maximizing the conditional likelihood for N given $\Sigma_1^k s_i$; the conditional likelihood for N differs from (3) here by a multiplier not dependent on N.

In contrast to the unrestricted and conditional MLE, our method uses a proper and natural distribution for p. We have chosen to fix the choice of (a, b) in line with our experience, but one could reasonably attempt to use the data to estimate (a, b). A Bayesian might also wish to construct a proper prior distribution for the parameter (N, p), the result of which might be an estimator with good frequentist properties.

Our method differs from that of Blumenthal and Dahiya (1981), who multiplied (1) and (2) together and then maximized this product jointly in N and p. They did not give any guidelines on how to choose (a, b) or on the stability of the result. From our limited calculations, it is clear that their choice of (a, b) must operate entirely differently from ours. In fact, our integration over p seems to induce stability for much smaller values of (a, b) than is the case with the Blumenthal and Dahiya method.

3. NUMERICAL WORK

The estimate of N obtained from maximizing (3) is reasonably stable. In Table 1 we analyze the examples listed in table 2 of OPZ, who computed MME, MME:S, MLE, and MLE:S for some particularly difficult cases; the MLE differs slightly from that of OPZ in cases 2 and 6 because of the extreme flatness of the likelihood in these cases. In addition, we provide the estimator Mbeta (0, 0) obtained from (3) with $a = b = 0$ (the uniform distribution) and the estimator Mbeta (1, 1) with $a = b = 1$. It is clear from these examples that MME and MLE are highly unstable. In addition, MME:S, MLE:S, Mbeta (0, 0), and Mbeta (1, 1) are clearly stable, with MME:S, Mbeta (0, 0), and Mbeta (1, 1) giving rather similar results. Cases 6 and 8 are particularly illustrative. Case 6 is an unstable case with large p, and here MLE:S dominates MME:S, with our Mbeta (0, 0) and Mbeta (1, 1) falling somewhere in between. Case 8 is unstable with small p, and MLE:S is now much worse than MME:S; again our estimators fall between the two, although they are more efficient in this case. The different behavior in unstable cases is reflected in the Monte Carlo study we now describe.

Table 1. N Estimates for Selected and Perturbed Samples

	Parameters			Estimators					
Sample	N	p	K	MME	MME:S	MLE	MLE:S	Mbeta (0, 0)	Mbeta (1, 1)
1	75	.32	5	102 195	70 80	19 190	29 30	51 57	49 52
2	34	.57	4	507 <0	77 91	514 ∞	31 32	52 59	47 52
3	37	.17	20	65 154	25 27	66 159	11 13	26 29	23 25
4	48	.06	15	18 135	10 12	15 125	7 9	9 12	8 10
5	40	.17	12	32 61	26 32	40 79	21 22	27 33	25 29
6	74	.68	12	210 259	153 162	213 266	67 69	135 144	125 131
7	55	.48	20	71 79	69 74	71 81	43 45	64 70	63 67
8	60	.24	15	67 88	49 53	67 90	24 26	45 49	41 45

NOTE: The exact samples are given in table 2 of OPZ. For each sample number, the first entries are the N estimates for the original sample, and the second entries are the N estimates for perturbed samples obtained by adding one to the largest success count.

In Table 2 we expand the Monte Carlo study of OPZ, comparing MME:S and MLE:S with Mbeta (0, 0) and Mbeta (1, 1). All random numbers were generated by using the IMSL generators GGBTR and GGBN. The basic study was as in OPZ, so k was randomly chosen such that $3 \le k \le 22$, p was uniformly chosen such that $0 < p < 1$, and $1 \le N \le 100$ was uniformly chosen. There were 2,000 randomly generated cases. A case was called stable if $\hat{\mu} \ge (1 + 1/\sqrt{2})\hat{\sigma}^2$ and unstable otherwise. We also considered subcases in which $.2 \le p \le .8$, $0 < p \sqrt{2} - 1$ and $\sqrt{2} - 1 \le p < 1$.

Readers of an earlier version of this article pointed out that our study seemed biased in favor of our estimators, since p was uniformly distributed on $0 < p < 1$. To avoid this criticism we redid the study completely, generating beta (A, B) random variables [see (2)] with the asymmetric choices $(A, B) = (0, 1)$, $(1, 2)$, $(1, 3)$, $(1, 0)$, $(2, 1)$, and $(3, 1)$. Since in each of these studies, our estimators performed as well as or better than

Table 2. Relative Mean Square Error Efficiencies of the N Estimates Relative to MME:S

	Stable Cases					Unstable Cases				
Range	No.	MME:S	MLE:S	Mbeta (0, 0)	Mbeta (1, 1)	No.	MME:S	MLE:S	Mbeta (0, 0)	Mbeta (1, 1)
$0 < p < 1$	1,367	1.00	.99	.99	1.03	633	1.00	.86	1.16	1.18
$.2 < p < .8$	863	1.00	.98	.99	1.08	336	1.00	1.18	1.96	2.79
$p < \sqrt{2} - 1.0$	281	1.00	.99	.96	.99	519	1.00	.66	.96	.92
$p > \sqrt{2} - 1.0$	1,086	1.00	.99	1.08	1.20	114	1.00	5.70	2.90	5.16

in the case $(A, B) = (0, 0)$ reported in Table 2, we do not report them.

For the unrestricted case in which $0 < p < 1$, the actual relative mean squared error efficiencies and the percentages of stable cases were similar to the results of OPZ. For the stable cases, the four estimators performed equally well. For the unstable cases, MLE:S was the clear loser, with the other three estimators being similar in performance.

When we consider the special cases $.2 < p < .8$, interesting results emerge. For the unstable cases, MME:S still beats MLE:S, but our estimators Mbeta (0, 0) and Mbeta (1, 1) are vastly superior to the other two. An intuitive reason for this may be that our estimators downweight the possibility that p is near zero.

Following OPZ, we also consider the cases of "small" p ($0 < p < \sqrt{2} - 1$) and "large" p ($\sqrt{2} - 1 \leq p < 1$). The former case is, as expected, least favorable to our estimators, which discount the possibility that p is small. However, our estimators still perform well; for example, Mbeta (0, 0) is only 4% less efficient in terms of relative mean squared error than MME:S, and Mbeta (1, 1) loses only 1% efficiency.

More striking results emerge when p is large ($\sqrt{2} - 1 \leq p < 1$). For the stable cases (90%) in this subset, our estimators have a definite advantage over MME:S and MLE:S, especially Mbeta (1, 1). For the few unstable cases, MLE:S is much better than MME:S (a fact noted by OPZ); even in these cases, our estimators perform competitively, and overall Mbeta (1, 1) emerges as the clear winner.

We found that the stabilized MLE:S was much more negatively biased than the three other stabilized estimates. Though all were negatively biased in general, the stabilized MLE:S had a bias in the unstable cases of almost 60% of the true value of N, versus 20% for the moments estimate MME:S and 30%–35% for our suggestions Mbeta (0, 0) and Mbeta (1, 1). Interestingly, for the 82% of unstable cases with true probability less than $\sqrt{2} - 1$, our suggestions were negatively biased, and for the other 18% of unstable cases, the bias was positive.

4. EXAMPLES

This research was motivated by the following two examples, the second of which is especially difficult. The counts of impala herds and individual waterbucks were obtained on five successive cloudless days in a small area of the Kruger Park. Counting was done from a light aircraft by five highly trained and experienced wildlife officials. The assumption of independent binomial counts with approximately equal success probabilities seems reasonable in this example, but the assumption is of course not absolutely indisputable.

Example 1. The observed number of herds of at least 25 impala were given as 15, 20, 21, 23, and 26. This is an unstable case, since $\hat{\mu}/\hat{\sigma}^2 = 1.59$. The various estimators are MME = 57, MLE:S = 28, MLE = 53, Mbeta (0, 0) = 42, MME:S = 54, and Mbeta (1, 1) = 42. When we changed the largest count from 26 to 27, the estimators MME:S, MLE:S, Mbeta (0, 0), and Mbeta (1, 1) exhibited little change. The moments estimator MME, however, changed from 57 to 77 and the MLE changed from 53 to 74. Note how the stabilized MLE:S is the smallest here, which is in line with the extreme negative-bias results found in the Monte Carlo study in the previous section. The conditional maximum likelihood estimator Mbeta (−1, 0) was fairly unstable here—95 for the original data but 215 for the perturbed data.

Table 3. N Estimates in the Waterbuck Data

a	Mbeta (a, 0) *Original Data*	 *Perturbed Data*
0	146	155
−.25	159	168
−.50	179	193
−.75	225	251
−.90	311	367
−1.00	1,545	>4,000

Example 2. The observed number of waterbucks was 53, 57, 66, 67, and 72. Since $\hat{\mu}/\hat{\sigma}^2 = 1.32$, this is a highly unstable case. For the observed data, the estimates of N are MME = 272, MLE:S = 72, MLE = 265, Mbeta (0, 0) = 146, MME:S = 199, and Mbeta (1, 1) = 140. When we changed the largest count from 72 to 73, the estimates became MME = 362, MLE:S = 78, MLE = 355, Mbeta (0, 0) = 155, MME:S = 215, and Mbeta (1, 1) = 146. Note again the apparent extreme bias of the stabilized MLE:S.

The conditional MLE was again very unstable here—1,545 for the original data and >4,000 for the perturbed data. Because of the bias observed in the Monte Carlo study, we did some experimentation with the estimator Mbeta (a, 0), with $a \leq 0$. The results are displayed in Table 3. Whether a reasonable, perhaps data-based choice of the value of a in Mbeta (a, 0) would improve on the estimators we have studied remains to be seen. As noted in the previous section, it is in the highly unstable cases such as these waterbuck data for which negative bias is of most concern. Casella (1984) discusses an interesting graphical device for assessing the degree of instability of a given set of data. It seems that he implicitly suggests a data-dependent choice of (a, b), something along the lines of using Mbeta (0, 0) for stable cases but smoothly adjusting to Mbeta (a, 0) as the instability increases; the waterbuck data suggest that we must stay strictly away from the conditional maximum likelihood estimate Mbeta (−1, 0).

5. ASYMPTOTIC THEORY

An illuminating general asymptotic theory for this problem awaits development. The stabilized method of moments and maximum likelihood estimators of OPZ have not been fully studied. Of course, as $k \to \infty$ for fixed N, all estimators discussed in this article are consistent. We have also considered an asymptotic theory for the estimators Mbeta (a, b) in the case of fixed (a, b), $N \to \infty$, $k \to \infty$, and $\sqrt{k}/N \to 0$. The results of this asymptotic theory are not too interesting because we find that regardless of the choice of (a, b),

$$k^{1/2}(\text{Mbeta }(a, b) - N)/N \xRightarrow{\mathcal{L}}$$

$$\text{normal}\left(\text{mean} = 0, \text{ variance} = \frac{2(1-p)^2}{p^2}\right). \quad (4)$$

Note that (4) suggests a large effect in variance for smaller values of p.

We have obtained two theoretical results that shed light on the behavior of our procedures. Both results involve fairly intricate calculations, which will not be presented here.

The first theoretical result illustrating the role of (a, b) in (3) and in the estimators Mbeta (a, b) occurs when we fix k—the number of observers—and let $N \xrightarrow{p} \infty$. In this case, none of the estimators will be consistent in the sense that $\hat{N}/N \xrightarrow{p} 1$. If we fix $k = 2$ and let the chi-squared limit distribution of $\{2p(1 - p)\}^{-1}(s_1 - s_2)^2/N$ be denoted by χ_1^2, then $\hat{\mu}/N \rightarrow p$ and $\{4p(1 - p)\}^{-1}\hat{\sigma}^2/N \overset{\mathcal{L}}{\Rightarrow} \chi_1^2$. We can thus expect interesting results here because there will still be a positive probability of an unstable case. In fact, we obtain the following:

Lemma 1. In (3) take $a = b$ and $k = 2$. Let the chi-squared limit distribution of $\{2p(1 - p)\}^{-1}(s_1 - s_2)^2/N$ be denoted by χ_1^2. Then as $N \rightarrow \infty$,

$$\frac{\text{Mbeta}(b, b)}{N} \overset{\mathcal{L}}{\Rightarrow} \frac{p}{(4b + 4)}\{6b + 3 + (1 - p)\,\chi_1^2 + \sqrt{(6b + 3 + (1 - p)\chi_1^2)^2 - 4(b + 1)(8b + 2)}\}. \quad (5)$$

The lemma illustrates one interesting facet of our estimators Mbeta (b, b) obtained from maximizing (3). The unstable cases are those in which $(s_1 - s_2)^2$ is large relative to $(s_1 + s_2)$. This corresponds to the situations in (5) in which χ_1^2 is large. Taking the limit of the right side of (5) as $\chi_1^2 \rightarrow \infty$, we obtain the proportionality

$$(5) \propto \chi_1^2 p(1 - p)/(2b + 2). \quad (6)$$

Equation (6) shows that the effect of increasing the smoothing parameter $a = b$ in (3) is a type of shrinkage. This agrees with the intuitive notion that the effect of larger (a, b) is to discount the possibility that p is small. This simple asymptotic theory helps explain why in Table 2 the most severe unstable cases are better handled by $(a = 1, b = 1)$ than by Mbeta $(a = 0, b = 0)$.

Our second useful asymptotic theoretical result illustrating the role of (a, b) in our estimator Mbeta (a, b) occurs under the following specifications:

$$N \rightarrow \infty, \qquad k \rightarrow \infty, \qquad k^{1/2}/N \rightarrow 0$$
$$a = \alpha k^{1/2}, \qquad b = \beta k^{1/2}. \quad (7)$$

Lemma 2. Consider the assumptions (7) with $\alpha > 0$ and $\beta > 0$. Then

$$k^{1/2}(\text{Mbeta}\ (a, b) - N)/N \overset{\mathcal{L}}{\Rightarrow} \text{normal}\left(\frac{2(1 - p)^2\,\{(\alpha + \beta)p - \alpha\}}{p^2}, \frac{2(1 - p)^2}{p^2}\right). \quad (8)$$

Consider what Lemma 2 says qualitatively for the case $a = b$ and $\alpha = \beta$, which was examined in our Monte Carlo study. Equation (8) suggests that in our Monte Carlo we should have found more negative bias for the case in which the true success probability $p < \sqrt{2} - 1$ than for the case $p \geq \sqrt{2} - 1$. This we found to be true for stable cases taken together as well as unstable cases taken together.

6. DISCUSSION

By considering beta-binomial distributions, we have obtained stable estimates that are at least competitive with, and in some instances superior to, the stabilized MME and MLE introduced by OPZ. OPZ were primarily interested in easily computed stable estimators with good efficiency properties, and thus it is natural that they did not consider refinements of their methods. In particular, they noted that perhaps their definition of unstable also ought to depend on k. We think their work is an excellent step toward better understanding of this difficult problem. Our estimators are differently motivated than theirs, and we hope that they will provide some additional insight. The advantages of our method include the flexibility of choosing a and b and the modification of the likelihood by smooth handling of the nuisance parameter p. We believe that further progress is inevitable and that even better estimates can be found. For example, one might suppose a natural joint distribution for $(\hat{N}, p)$ that downweights small p and large N; an estimator with good frequentist properties might emerge from such as Bayesian analysis.

Finally, little is known about the shape of $\mathcal{L}(N)$ in (3), so the question of finding a confidence interval for N remains to be addressed.

[*Received May 1983. Revised October 1984.*]

REFERENCES

Barnard, G. A., Jenkins, G. M., and Winsten, C. B. (1962), "Likelihood Inference and Time Series" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 125, 321–372.

Blumenthal, S., and Dahiya, R. C. (1981), "Estimating the Binomial Parameter *N*," *Journal of the American Statistical Association*, 76, 903–909.

Casella, G. (1984), "Stabilizing Binomial *N* Estimators," manuscript submitted for publication.

DeRiggi, D. F. (1983), "Unimodality of Likelihood Functions for the Binomial Distribution," *Journal of the American Statistical Association*, 78, 181–183.

Draper, N., and Guttman, I. (1971), "Bayesian Estimation of the Binomial Parameter," *Technometrics*, 13, 667–673.

Fisher, R. A. (1942), "The Negative Binomial Distribution," *Annals of Eugenics*, 11, 181–187.

Haldane, J. B. S. (1942), "The Fitting of Binomial Distributions," *Annals of Eugenics*, 11, 179–181.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Olkin, I., Petkau, A. J., and Zidek, J. V. (1981), "A Comparison of *N* Estimators for the Binomial Distribution," *Journal of the American Statistical Association*, 76, 637–642.

BIOMETRICS 44, 175–188
March 1988

Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process

Margaret C. Wu

Biometrics Research Branch, National Heart, Lung, and Blood Institute,
Bethesda, Maryland 20892, U.S.A.

and

Raymond J. Carroll*

Department of Statistics, Texas A&M University,
College Station, Texas 77843, U.S.A.

SUMMARY

In the estimation and comparison of the rates of change of a continuous variable between two groups, the unweighted averages of individual simple least squares estimates from each group are often used. Under a linear random effects model, when all individuals have complete observations at identical time points, these statistics are maximum likelihood estimates for the expected rates of change. However, with censored or missing data, these estimates are no longer efficient when compared to generalized least squares estimates. When, in addition, the right-censoring process is dependent on the individual rates of change (i.e., informative right censoring), the generalized least squares estimates will be biased. Likelihood-ratio tests for informativeness of the censoring process and maximum likelihood estimates for the expected rates of change and the parameters of the right-censoring process are developed under a linear random effects model with a probit model for the right-censoring process. In realistic situations, we illustrate that the bias in estimating group rate of change and the reduction of power in comparing group differences could be substantial when strong dependency of the right-censoring process on individual rates of change is ignored.

1. Introduction

In clinical trials and longitudinal studies it is often of interest to estimate and compare the rates of change of one or more variables between groups, in, e.g., lung function or tumor growth. Furthermore, comparing the rates of change of a continuous response variable between two treatment groups is often the primary objective. Death or withdrawal may cause some observations of the primary variable to be right-censored.

Growth curve methods for comparing rates of change have been studied extensively (see Rao, 1965; Fearn, 1975; and Schlesselman, 1973). Most of these analyses assume that there are no right-censored or missing observations. Maximum likelihood and generalized weighted least squares provide alternative approaches to simple least squares for the analysis of a series of measurements when some observations are right-censored or missing. Koziol et al. (1981) proposed a distribution-free test for the comparison of growth curves with incomplete data. In order to be valid, these procedures require that the probabilities of

* Research supported by Air Force Office of Scientific Research Contract AFOSR-F-49620-85-C-0144.

Key words: Informative right censoring; Linear random effect; Probit right censoring; Rate of change.

right censoring or missing do not depend on the parameter values of the response under investigation, i.e., they are noninformative with respect to the response parameters.

In this paper we are primarily interested in right censoring caused by the participant's death or withdrawal, to be referred to as the primary right-censoring process. The primary right-censoring process could be informative with respect to the response parameters. In our development, staggered entry and other missing-value processes, if incorporated, are assumed to be noninformative and independent of the primary right-censoring process.

Under a linear random effects model, we propose a model that can depend both on the individual's initial value and slope. A likelihood-ratio test for informativeness and maximum likelihood estimates for the response parameters and the primary right-censoring process coefficients are derived under a probit model for the probability of primary right censoring.

The right censoring is considered to be noninformative with respect to the response parameters if the likelihood function can be factored into two independent parts, one corresponding to the response parameter and the other corresponding to censoring parameters.

We show that when the primary right censoring is noninformative, the maximum likelihood estimates for the average linear regression coefficients of the response are weighted linear combinations of the simple least squares estimates. In the case of complete observations at identical time points among all individuals, these estimates are just the unweighted averages of the individual simple least squares estimates.

The proposed method is applied to data on patients with PiZ phenotype, gathered by the Workshop on Natural History of PiZ Emphysema (1983). To illustrate the effect of informative right censoring, maximum likelihood and the weighted and unweighted least squares procedures are applied to a set of simulated clinical trials with primary right censoring generated from a noninformative probit process and then to another set of simulated trials with primary right censoring generated from an informative probit process. Mean squared error and power comparisons are made among the different statistical procedures and between these two sets.

2. Linear Random Effects and Informative Right Censoring

We assume that the participants of a longitudinal study are divided into two treatment groups of sample sizes n_k, for $k = 1, 2$. The combined sample size is $n = n_1 + n_2$. Let there be J identical mortality (and withdrawal status) follow-up time points, t_j, with $t_1 = 0$ and $t_J =$ the length of the study. Each participant can have at most R measurements of the response during the study. The measurement time need not be identical among individuals. Let $\nu_i =$ total number of measurements made for the ith individual. Let $Y_{i\nu}$ and $t_{i\nu}$ be the νth response and the corresponding measurement time for the ith participant in the combined sample for $\nu = 1, \ldots, \nu_i$ and $i = 1, \ldots, n$. With $t_{i1} = 0$, let $t_{i\nu_i} \leq t_j$ if death, withdrawal, or right censoring due to staggered entry occurred for the ith participant between time t_j and t_{j+1} (the jth interval); otherwise $t_{i\nu_i} = t_J$ if the ith participant was not right censored and $t_{i\nu_i} = t_{iR} = t_J$ if the ith participant had complete observations.

It is assumed that the serial measurements of the primary variable follow a linear function of time. Let $\boldsymbol{\beta}_i^{\mathrm{t}} = (\beta_{i1}, \beta_{i2})$ be the unobservable vector representing the true initial value and slope of the primary variable for the ith individual in the combined sample. For $i \in k$ and $k = 1, 2$;

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \varepsilon_i, \quad \text{where} \quad \mathbf{Y}_i^{\mathrm{t}} = (Y_{i1}, \ldots, Y_{i\nu_i}), \tag{2.1}$$

$$\boldsymbol{\beta}_i \sim \mathrm{N}(\mathbf{B}_k, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \quad \text{and} \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2\mathbf{I});$$

$$\mathbf{X}_i = \begin{bmatrix} 1, \ldots, 1 \\ t_{i1}, \ldots, t_{i\nu_i} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{bmatrix} \sigma_{\beta_1}^2 & \sigma_{\beta_1\beta_2} \\ \sigma_{\beta_1\beta_2} & \sigma_{\beta_2}^2 \end{bmatrix}, \quad \mathbf{B}_k^{\mathrm{t}} = (B_{k1}, B_{k2}).$$

The notation $i \in k$ is used to denote that the ith participant in the combined sample belonged to the kth treatment group.

We further suppose that the probability of being primarily right-censored due to death or withdrawal during a specified time interval $(0, t_j)$, given $\boldsymbol{\beta}_i$, is $\mathrm{M}(\boldsymbol{\alpha}^t\boldsymbol{\beta}_i, t_j)$. Here $\boldsymbol{\alpha}^t = (\alpha_1, \alpha_2)$ is the vector of "regression parameters" relating this probability to the primary variables $\boldsymbol{\beta}_i$. Examples of the logical choices for M are proportional hazards regression (Cox, 1972), logistic regression (Walker and Duncan, 1967), and probit regression (Halperin, Wu, and Gordon, 1979). For instance, under probit regression $\mathrm{M}(\boldsymbol{\alpha}^t\boldsymbol{\beta}, t_j) = \Phi(\boldsymbol{\alpha}^t\boldsymbol{\beta} + \alpha_{0j})$, where Φ is the cumulative probability of a standard normal variate, and α_{0j} for $j = 2, \ldots, J$ are censoring time parameters. The argument of the probit function is thus a linear combination of the regression parameters and a discrete time component. Note that we allow the α_0's to be arbitrary parameters for the different time intervals. A more restrictive time-dependent probit function can be obtained by restricting the α_0's to follow a linear or higher-order polynomial function of time.

Since for each $\boldsymbol{\beta}_i$, (i) the simple least squares estimates $\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^t\mathbf{X}_i)^{-1}(\mathbf{X}_i^t\mathbf{Y}_i)$, (ii) censoring time, and (iii) survival time are sufficient statistics for $\boldsymbol{\beta}_i$, it suffices to consider the joint distribution of $\hat{\boldsymbol{\beta}}_i$, $\boldsymbol{\beta}_i$ and the primary right-censoring process. The marginal likelihood for $\mathbf{B}_k$ and $\boldsymbol{\alpha}$ for the ith individual can be expressed as

$$L_i = D \int \phi_2(\hat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_i, \mathbf{C}_{1i})\phi_2(\boldsymbol{\beta}_i, \mathbf{B}_K, \boldsymbol{\Sigma}_\beta) \cdot \prod_{j=2}^{J} [(1 - M_{j-1})^{C(i,j-1)}(M_j - M_{j-1})^{Z(i,j-1)}](1 - M_J)^{(1-m(i))}\, d\boldsymbol{\beta}_i, \tag{2.2}$$

where $M_j = \mathrm{M}(\boldsymbol{\alpha}^t\boldsymbol{\beta}_i, t_j)$ for $i \in k$ ($k = 1, 2; j = 1, \ldots, J$); $M_1 = 0$. The dependency of M_j on i is suppressed here for notational simplicity. Here $C(i, j)$ is the indicator function that the ith individual was censored in the jth interval because of staggered entry, $Z(i, j)$ is the indicator function that death or withdrawal occurred in the jth interval for the ith individual, D is constant with respect to $\boldsymbol{\beta}_i$, $\boldsymbol{\alpha}$, and $\mathbf{B}_k$, and

$$\mathbf{C}_{1i} = \sigma_\varepsilon^2(\mathbf{X}_i^t\mathbf{X}_i)^{-1}, \quad m(i) = \sum_j [C(i, j) + Z(i, j)].$$

The notation $\phi_2(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ represents the bivariate normal density with mean vector $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$. On the right-hand side of equation (2.2), under the integration sign, the first factor represents the conditional probability distribution of $\hat{\boldsymbol{\beta}}_i$ given $\boldsymbol{\beta}_i$, $C(i, j)$, and $Z(i, j)$ for $j = 1, \ldots, J - 1$. The second factor is the probability distribution of $\boldsymbol{\beta}_i$. The third factor of products corresponds to the conditional probabilities that the ith participant survived the $(j - 1)$th time point and then was censored by staggered entry or death (or withdrawal) between the $(j - 1)$th and the jth time points, respectively, for $j = 2, \ldots, J$ given $\boldsymbol{\beta}_i$. The last factor represents the conditional probability that the ith participant survived the entire study, given $\boldsymbol{\beta}_i$. Therefore, the product of these four factors is proportional to the joint distribution of $\hat{\boldsymbol{\beta}}_i$, $\boldsymbol{\beta}_i$, $Z(i, j)$, and $C(i, j)$, because the staggered entry process and the missing-value process are assumed to be noninformative and independent of the primary right-censoring process. Hence, integration with respect to the vector $\boldsymbol{\beta}_i$ provides the marginal likelihood of $\hat{\boldsymbol{\beta}}_i$, $Z(i, j)$, and $C(i, j)$ with respect to $\mathbf{B}_k$ and $\boldsymbol{\alpha}$.

The marginal likelihood for those measured only at baseline is obtained from equation (2.2) by equating all elements except the (1, 1)th of $\mathbf{C}_{1i}$ and $\boldsymbol{\Sigma}_\beta$ to zero and letting the (1, 1)th element of $\mathbf{C}_{1i}$ equal σ_ε^2 and $\hat{\beta}_{i2} = \beta_{i2} = 0$. The marginal likelihood for all n individuals is the product of the individual likelihoods.

Joint estimation of the parameters depends on the ability to evaluate (2.2) and its derivatives. For this section we assume that $\boldsymbol{\Sigma}_\beta$ and σ_ε^2 are known. The more realistic case will be discussed in the next two sections. In principle, (2.2) can be evaluated by numerical integration. When the primary right-censoring process is a probit model, (2.2) can be evaluated explicitly: For $i \in k$ and $k = 1, 2$,

$$\ln(L_i) = \ln(D) + \ln(A_i) - .5(\hat{\boldsymbol{\beta}}_i - \mathbf{B}_k)^{\mathrm{t}}\mathbf{C}_{2i}^{-1}(\hat{\boldsymbol{\beta}}_i - \mathbf{B}_k) + T_i, \tag{2.3}$$

where

$$A_i = (2\pi\,|\,\mathbf{C}_{2i}\,|^{1/2})^{-1}, \quad \mathbf{C}_{2i} = \mathbf{C}_{1i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}},$$

$$T_i = \sum_{j=2}^{J} \{C(i, j-1)\ln[1 - \Phi(U_{ij-1})] + Z(i, j-1)\ln[\Phi(U_{ij}) - \Phi(U_{ij-1})]\}$$

$$+ \left[1 - \sum_j Z(i, j) - \sum_j C(i, j)\right]\ln[1 - \Phi(U_{iJ})],$$

$$U_{ij} = (\alpha_{0j} + \mathbf{d}_{ik}^{\mathrm{t}}\mathbf{C}_{3i}\boldsymbol{\alpha})(1 + \boldsymbol{\alpha}^{\mathrm{t}}\mathbf{C}_{3i}\boldsymbol{\alpha})^{-1/2},$$

$$\mathbf{d}_{ik} = \mathbf{C}_{1i}^{-1}\hat{\boldsymbol{\beta}}_i + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\mathbf{B}_k, \quad \mathbf{C}_{3i} = (\mathbf{C}_{1i}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}.$$

When there are right-censored or missing observations, $\mathbf{C}_{3i}$ will differ among individuals. Hence, the primary right-censoring contribution to the likelihood, T_i, is in the form of a nonlinear probit model. Maximum likelihood estimation of the parameters can be made in principle provided that the number of time intervals is small. Otherwise, some constraints could be imposed on the α_0's to reduce the number of parameters.

Likelihood-ratio tests for the hypothesis (H_0: $\alpha_1 = \alpha_2 = 0$) versus (H_1: $\alpha_2 = 0$ and $\alpha_1 \neq 0$) and the hypothesis H_1 versus (H_2: $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$) can be conducted. When H_0 is true, the primary right censoring will be noninformative with respect to $\mathbf{B}_k$ for $k = 1, 2$. However, when H_1 is true, it can be shown that the coefficient of B_{k2} in U_{ij} of (2.3) is nonzero even when $\sigma_{\beta_1\beta_2} = 0$. Hence, the primary right censoring will be informative with respect to B_{k2} for $k = 1, 2$. This is true because when the measurement error variance σ_ϵ^2 is nonzero, the true initial value of the response variable is not observable and the estimated initial value contains information about the expected slope. When H_1 is true and $\sigma_{\beta_2}^2 = \sigma_{\beta_1\beta_2} = 0$, or when H_1 is true and $\sigma_\epsilon = 0$, the primary right censoring is noninformative with respect to B_{k2}.

3. Estimation and Testing for Noninformative Censoring

When H_0 is true, the maximum likelihood estimate of $\mathbf{B}_k$ is

$$\hat{\mathbf{B}}_{\mathrm{GL},k} = \left[\sum_{i\in k} \mathbf{C}_{2i}^{-1}\right]^{-1} \sum_{i\in k} (\mathbf{C}_{2i}^{-1}\hat{\boldsymbol{\beta}}_i), \tag{3.1}$$

the weighted or generalized least squares estimate (GLSE). When all individuals have complete observations measured at identical time points, $\mathbf{C}_{2i}$ will be the same among individuals, in which case (3.1) reduces to

$$\hat{\mathbf{B}}_{\mathrm{UW},k} = \sum_{i\in k} \hat{\boldsymbol{\beta}}_i/n_k,$$

the unweighted least squares estimate (UWLE). In what follows, the unweighted least squares estimate is computed only on those individuals with at least two measurements. The covariance matrices are

$$\mathbf{C}_{\mathrm{GL},k} = \left[\sum_{i\in k} \mathbf{C}_{2i}^{-1}\right]^{-1} \quad \text{and} \quad \mathbf{C}_{\mathrm{UW},k} = \left[\sum_{i\in k} \mathbf{C}_{2i}\right] n_k^{-2}. \tag{3.2}$$

When $\boldsymbol{\Sigma}_\beta$ and σ_ε^2 are unknown, the following unbiased estimators can be substituted:

$$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 \Big/ \left[\sum_{i=1}^{n} (\nu_i - 2) \right], \quad \hat{\boldsymbol{\Sigma}}_\beta = \mathbf{s}_\beta/(n-1) - \sum_{i=1}^{n} \mathbf{C}_{1i}/n, \tag{3.3}$$

$$\mathbf{s}_\beta = \sum_{k=1}^{2} \sum_{i \in k} (\hat{\boldsymbol{\beta}}_i - \hat{\mathbf{B}}_{\mathrm{UW},k})(\hat{\boldsymbol{\beta}}_i - \hat{\mathbf{B}}_{\mathrm{UW},k})^{\mathrm{t}}, \quad \text{and} \quad s_\varepsilon^2 = \sum_{i=1}^{n} (\mathbf{Y}_i^{\mathrm{t}}\mathbf{Y}_i - \mathbf{Y}_i^{\mathrm{t}}\mathbf{X}_i\hat{\boldsymbol{\beta}}_i).$$

However, $\hat{\boldsymbol{\Sigma}}_\beta$ has the disadvantage that it is not necessarily positive definite. The procedure given by Bock and Petersen (1975) for constructing an estimate that is at least semidefinite will be used.

When the goal of a study is to compare differences in rate of change between two groups, we wish to test the null hypothesis, H_{N}: $B_{12} = B_{22}$ against the alternative hypothesis, H_{A}: $B_{12} < B_{22}$. The test statistic is of the form

$$(\hat{B}_{12} - \hat{B}_{22})/[(\sigma_{\hat{B}_{12}})^2 + (\sigma_{\hat{B}_{22}})^2]^{1/2},$$

with $\hat{B}_{k2} = \hat{B}_{\mathrm{GL},k2}$ or $\hat{B}_{\mathrm{UW},k2}$. For shifted alternatives, sample size, power, and significance level of the test can be related according to the approximate formula,

$$[(\sigma_{\hat{B}_{12}})^2 + (\sigma_{\hat{B}_{22}})^2](Z_\alpha + Z_\beta)^2 = \Delta^2, \tag{3.4}$$

where Δ is the difference in expected rates of change we wish to detect, α and β are the Type I and Type II error probabilities of the test, and Z_α and Z_β are the unit normal deviates corresponding to α and β.

Remarks: We have by assumption that an individual's coefficient estimate is unbiased, i.e.,

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_i \mid \boldsymbol{\beta}_i, \text{ censoring, death, and withdrawal pattern}] = \boldsymbol{\beta}_i.$$

Thus, the unweighted least squares estimate is unbiased for $\mathbf{B}_k$. There are two cases. When the primary right censoring is noninformative, the distribution of $\mathbf{C}_{2i}$ in (3.1) does not depend on $\boldsymbol{\beta}_i$, so that the GLSE and UWLE are both consistent and unbiased estimators of $\mathbf{B}_k$, although of course the UWLE is less efficient. Furthermore, the relative differences between the variances and hence the required sample sizes of the UWLE and the GLSE for the slope or initial value are a function of $[\sigma_\varepsilon^2/(\sigma_{\beta_2})^2]$ or $[\sigma_\varepsilon^2/(\sigma_{\beta_1})^2]$, respectively. When the primary right-censoring process is informative, the unweighted least squares estimate is still unbiased, although the GLSE is not because $\mathbf{C}_{2i}$ and $\boldsymbol{\beta}_i$ are dependent.

4. Examples and Simulations

This paper was motivated by design and analysis problems encountered in many clinical trials concerning lung diseases, e.g., the Intermittent Positive Pressure Breathing Trial (IPPB, 1983) for chronic pulmonary diseases. One specific example was the feasibility study of an antiproteolytic replacement therapy trial among individuals with PiZ phenotype, conducted by the Workshop on the Natural History of PiZ emphysema. The association between severe alpha$_1$-antitrypsin deficiency and lung diseases, particularly pulmonary emphysema, has been observed since the early 1960s (Laurell and Eriksson, 1963). Individuals with PiZ phenotype tend to develop severe alpha$_1$-antitrypsin deficiency and hence pulmonary emphysema and more rapid decline in lung function. The planned trial was designed to detect differences in rates of decline of a 1-second forced expiratory volume (FEV$_1$) between a control and a therapeutic group. Retrospective data on PiZ individuals were gathered from the ten participating institutions (see Workshop on Natural History of PiZ Emphysema, 1983) to provide crude estimates of parameter values required for sample size calculations.

4.1 *Estimation and Testing*

A FORTRAN program was developed for estimation when there is no staggered entry. The method of pseudo-maximum-likelihood estimation (PMLE; see Gong and Samaniego, 1981) was used. Estimates of σ_ϵ^2 and $\mathbf{\Sigma}_\beta$ were made according to (3.3) and the Bock and Petersen (1975) procedure and substituted into (2.3), which was then maximized by the Newton–Raphson method. The algorithm first calculates the simple least squares intercept and slope for each individual and estimates σ_ϵ^2 and $\mathbf{\Sigma}_\beta$. The UWLE of $\mathbf{B}_k$ is used as initial value for $\mathbf{B}_k$ in calculating $\mathbf{d}_{ik}$ and $\mathbf{C}_{3i}$ for each individual according to (2.3). Partial derivatives of the log-likelihood for the Newton–Raphson iterative procedure are then calculated using initial values for $\alpha_1, \alpha_2, \alpha_{02}, \ldots, \alpha_{0J}$. Formulae for these partial derivatives are presented in the Appendix. Note that the initial values for the α's can be chosen arbitrarily with the constraint $\alpha_{02} < \alpha_{03} < \cdots < \alpha_{0J}$.

This algorithm was applied to the PiZ emphysema data. Among the data gathered for 294 PiZ individuals by the ten U.S. institutions, initial and follow-up FEV_1 values (with the initial and the last measurements at least 6 months apart) were available on 117 individuals. The number of FEV_1 measurements ranged from 2 to 12 (mean number of measurements = 3.8, median = 3.0). The duration between the initial and the last measurements ranged from 6 to 227 months (mean duration = 52 months, median = 40 months). Since the proposed trial duration was between 3 and 6 years, an analysis corresponding to a 3-year follow-up study was first made using the initial and all follow-up FEV_1 measurements made within 3 years of the initial measurement. Since many did not have reported follow-up FEV_1s within 3 years of the initial measurement, only 81 individuals with 8 deaths were included in this analysis. A second analysis, corresponding to a 6-year follow-up study, was also made among those with a minimum follow-up of 6 years or a reported death within the first 6 years. Follow-up FEV_1s within 6 years of the initial measurement were used. This analysis included 65 individuals with 19 deaths. Because of the small number of deaths, mortality follow-ups were grouped into two equal-length intervals for both analyses. The average numbers of FEV_1 measurements were 2.9 and 3.6 (median = 3 for both) and the average (median) durations between the initial and the last FEV_1s were 28 (33) and 48 (55) months for the 3- and 6-year follow-ups, respectively. Those individuals with only one FEV_1 measurement were not included in these analyses. This has the effect of causing a slight bias in the unweighted least squares analysis and a slight loss of efficiency in the informative censoring analysis.

The purpose of these analyses was to test for informativeness of the right censoring caused by a participant's death with respect to FEV_1 initial value and slope, obtained from 3- and 6-year follow-ups, respectively, and to derive crude estimates of the primary right-censoring coefficients. The initial values used for the iterative procedure were $\alpha_{02} = -1.35$, $\alpha_{03} = -.90$, and $\alpha_1 = \alpha_2 = 0$. The algorithm converged after 12 and 10 iterations for the first and second analyses, respectively. The results are presented in Table 1. The estimated probit right-censoring coefficients for FEV_1 initial value and slope (α_1 and α_2) were -3.8, -11.3 and -4.6, -13.8 for the two analyses, respectively. The negative values of the estimates for both parameters indicate that low initial FEV_1 and rapid decline in FEV_1 lead to greater risk of death. Likelihood-ratio tests indicated that the coefficients for FEV_1 initial value (α_1) were statistically significantly different from zero in both analyses. Although the chi-squared statistic (with 1 degree of freedom) of 2.8 for the slope coefficient (α_2) of the first analysis was not statistically significant at a 5% level, the chi-squared statistic of 7.1 for the slope coefficient of the second analysis was statistically significant. The significance of the initial value coefficients, as well as the large negative slope coefficients obtained from both analyses and the significance of the slope coefficient from the second analysis, indicate that the right censoring by participants' deaths could be informative with respect to both FEV_1 initial value and slope.

Survival probability distributions estimated by the product limit method (Kaplan and Meier, 1958) for the entire data set of 294 individuals, for those individuals included in the first and second analyses of Table 1, respectively, and for the 117 individuals with two or more FEV$_1$ measurements, are displayed in Figure 1. Since these data were collected retrospectively, mortality follow-ups were not as complete and rigorous as one would like them to be for the proposed prospective study. Hence, survival probabilities in Figure 1 could be optimistic.

Table 1
Estimation for the expected FEV$_1$ *slope, the missing-value coefficients, and likelihood-ratio test statistics for the coefficients*

	Three-year mortality	Six-year mortality
Estimates and test statistics	Three-year FEV$_1$ follow-up	Six-year FEV$_1$ follow-up
Estimated FEV$_1$ change/year		
Unweighted	−.093 (.0164)[a]	−.078 (.0138)[a]
Weighted	−.090 (.0151)	−.076 (.0136)
Probit informative missing	−.095 (.0152)	−.085 (.0133)
Estimated missing-value coefficients		
FEV$_1$ Initial value	−3.80 (2.01)	−4.61 (1.70)
FEV$_1$ Slope	−11.30 (7.46)	−13.80 (6.76)
α_{02}	−.53 (1.10)	1.25 (.93)
α_{03}	.42 (1.10)	2.42 (1.05)
Likelihood-ratio test statistics		
Initial value (H_1 vs H_0)[b]: χ_1^2	11.02	26.97
Slope (H_2 vs H_1)[b]: χ_1^2	2.81	7.13
No. at risk at baseline	81	65
No. of deaths	8	19

[a] Numbers in parentheses are estimated standard errors.
[b] H_0: $\alpha_1 = \alpha_2 = 0$; H_1: $\alpha_1 \neq 0, \alpha_2 = 0$; H_2: $\alpha_1 \neq 0, \alpha_2 \neq 0$.

The estimates we have proposed are of course sensitive to model misspecification. When using the estimation and test procedures derived under the probit model, goodness of fit to the data should be checked. One approach is to note that the estimated probability for the ith individual being primarily right-censored in the jth time interval, for given $\hat{\boldsymbol{\beta}}_i$, is $\hat{P}_{ij} = \Phi(\hat{U}_{ij+1}) - \Phi(\hat{U}_{ij})$, for $j = 1, \ldots, J - 1$, where $\hat{U}_{ij}$ is U_{ij} of (2.3) with $\boldsymbol{\alpha}$, α_{0j}, and the expected intercept and slope for the primary variable being replaced by their maximum likelihood estimates. Therefore, group the $\Phi(\hat{U}_{iJ})$ into groups and compute for each group, $E_j = \Sigma_i \hat{P}_{ij}$. Then compare E_j with the observed deaths and dropouts between the jth and $(j + 1)$th time points.

For the PiZ 6-year follow-up data of Table 1, the observed number of deaths among those whose estimated probabilities of death in 6 years were above the 85th percentile, between the 70th and 85th percentiles, and below the 70th percentile (for the entire 65 individuals) were 4, 2, 2 and 3, 4, 4 for the first and second 3-year intervals, respectively. The corresponding expected numbers of death were 4.47, 1.82, .85 and 2.98, 3.70, 3.88 for the two time intervals, respectively. Hence, the probit model seems to fit the data reasonably well.

Graphical comparison of the actual versus expected (under the probit censoring model) cumulative numbers of death by the estimated probability of death in 6 years for the 6-year follow-up data is displayed in Figure 2. Comparisons of the actual versus expected (under the probit censoring model) cumulative numbers of deaths in the first and second

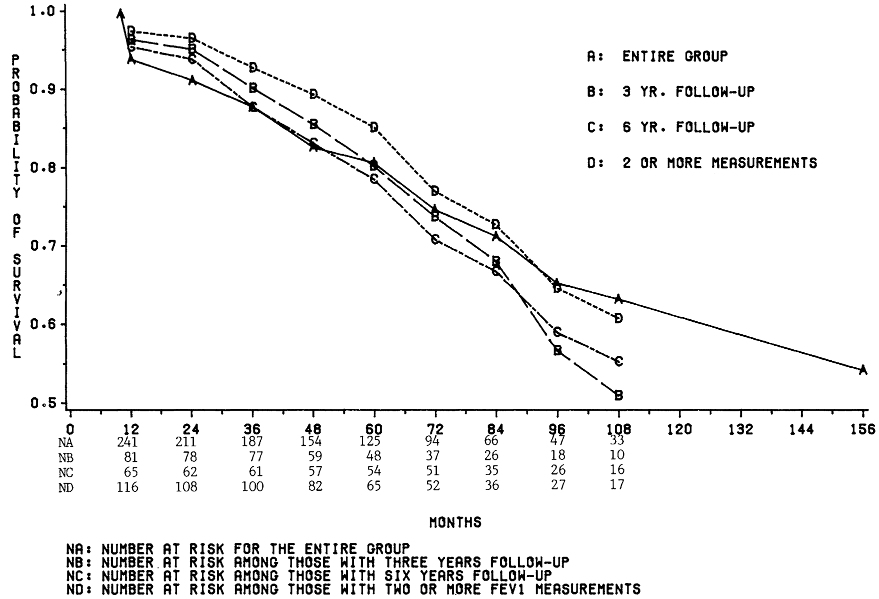


Figure 1. Survival curves: PiZ emphysema data.

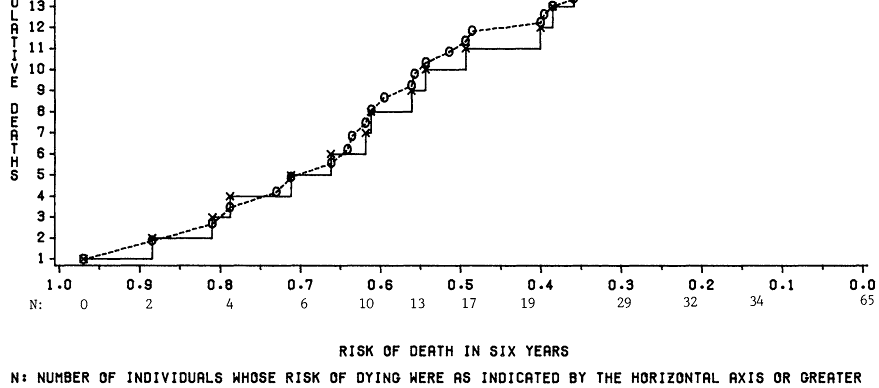


Figure 2. Actual vs expected deaths by expected risk of death in 6 years.

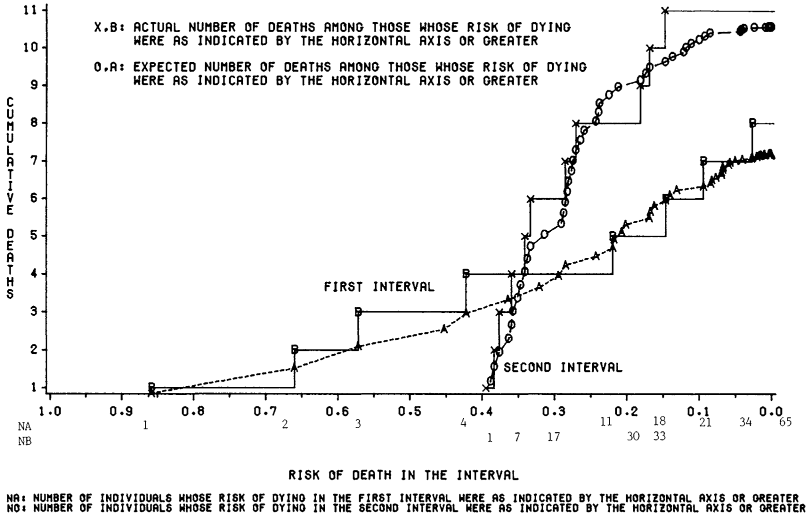


Figure 3. Actual vs expected deaths by expected risk of death in each of the two 3-year intervals.

3-year intervals by the estimated probability of death in the corresponding time intervals, for the same 6-year follow-up data, are shown in Figure 3. The overall fits of the data to the probit censoring model were again reasonably good in both figures.

4.2 *The Effect of Informative Censoring*

The UWLE, GLSE, and the PMLE were compared in simulated experiments based on the model (2.2) with the following primary right-censoring processes: (1) probit noninformative censoring with $\alpha_1 = \alpha_2 = 0$; (2) probit informative censoring with coefficients $\alpha_1 = -3.8$, $\alpha_2 = -11.3$; (3) probit informative censoring with coefficients $\alpha_1 = -4.6$, $\alpha_2 = -13.8$, corresponding to the two analyses of Section 4.1; and (4) probit informative censoring with $\alpha_1 = -3.8$ and $\alpha_2 = 0$. Similar to the IPPB trial (1983), the study duration was assumed to be 3 years with four FEV_1 measurements per year. The expected FEV_1 slope and initial value in the control group, and the within- and between-individual variances used were all estimated from the PiZ data. A 50% reduction in FEV_1 rate of decline was assumed in the treatment group. Equal sample sizes of 100 each were generated for the two groups. In the IPPB trial, similar to the proposed trial, patients were required to have their FEV_1 values less than 65% predicted [by age, sex, and height of the participants using regression coefficients given by Morris, Koski, and Johnson (1971)] at entry and the comparison of FEV_1 annual rates of decline between two randomized treatment groups was the primary objective of the trial. The primary right-censoring rate for the IPPB trial was more than 12% per year. For these illustrations the probability of primary right censoring was assumed to be 16% each year for all individuals under the noninformative right-censoring process. When the informative probit model was used, this probability was assumed to be 16% for an individual whose initial value and slope were equal to the expected values for the control group. It was further assumed that there is no correlation between the slope and the initial value ($\sigma_{\beta_1\beta_2} = 0$). The decision value used for rejecting the null hypothesis of no difference was $(\hat{B}_{12} - B_{22})/[(\hat{\sigma}_{B_{12}})^2 + (\hat{\sigma}_{B_{22}})^2]^{1/2} < -1.645$. Normal random numbers were generated

Table 2

Comparison of simulated results of different procedures under a linear random effects model with noninformative versus probit informative censoring with parameter values estimated from PiZ *emphysema data*[a]

Statistical procedures and treatment groups	Noninformative censoring $\alpha_1 = \alpha_2 = 0$		Informative censoring					
			$\alpha_1 = -3.8$ $\alpha_2 = -11.3$		$\alpha_1 = -4.6$ $\alpha_2 = -13.8$		$\alpha_1 = -3.8$ $\alpha_2 = 0.0$	
	FEV_1		FEV_1		FEV_1		FEV_1	
	Slope[b]	MSE	Slope[b]	MSE	Slope[b]	MSE	Slope[b]	MSE
Control group								
UWLE	−89.3	465.4	−88.9	645.7	−88.3	719.4	−89.2	570.4
GLE	−90.3	156.3	−68.3	634.9	−63.7	883.3	−90.7	164.6
PMLE	−89.4	200.6	−83.6	455.0	−81.2	597.8	−90.2	210.6
Treatment group								
UWLE	−46.5	442.3	−45.9	444.9	−46.2	489.8	−46.9	444.7
GLE	−45.9	148.2	−28.2	423.1	−24.4	576.0	−46.3	150.7
PMLE	−44.6	194.4	−39.7	288.4	−38.8	340.6	−44.3	162.9
Between-group differences								
UWLE	−42.8	935.8	−43.0	1,059.3	−42.11	1,210.9	−42.3	1,052.0
GLE	−44.5	361.8	−40.1	339.0	−39.3	377.8	−44.4	313.0
PMLE	−44.8	267.4	−43.9	297.1	−42.4	309.0	−44.1	297.3
Simulated power and significance level								
	Power	Signif	Power	Signif	Power	Signif	Power	Signif
UWLE	.45	.05	.39	.05	.36	.05	.40	.05
GLE (.85)[c]	.81	.06	.72	.07	.67	.07	.77	.06
PMLE	.80	.05	.80	.06	.79	.06	.78	.07

[a] The parameter values used were: Measurement error std. dev. σ_ϵ = .155L; FEV_1 initial value std. dev. σ_{β_1} = .39L; FEV_1 slope std. dev. σ_{β_2} = .091L/yr; $\sigma_{\beta_1,\beta_2} = 0$; expected FEV_1 initial value $B_{11} = B_{21}$ = .96L; control and treatment group expected FEV_1 slopes B_{12} = −.09L/yr and B_{22} = −.045/yr. For significance level, $B_{21} = B_{22}$ = −.09L/yr. The probability of missing = 16%/yr and $n_1 = n_2 = 100$.

[b] The FEV_1 slopes are presented in units of ML/yr.

[c] Expected power under noninformative missing process, calculated according to (3.4), using the actual parameter values.

by the IMSL routine GNPM. The experiments were repeated 600 times. In the simulation, every individual turned out to have at least two measurements.

The results in Table 2 indicate that the UWLE procedure remained relatively unbiased in estimating the mean FEV_1 slope for each group and the between-group difference in slopes. However, the PMLE clearly had much smaller mean squared errors in estimating the individual group mean slopes and the between-group differences, and much higher statistical power in detecting the between-group differences in all four censoring processes considered. The GLSE, although most efficient under noninformative censoring, resulted in large underestimations of individual group mean FEV_1 rates of decline (24–46%), under the two probit informative censoring processes with nonzero coefficients for FEV_1 slope. The underestimations for the between-group differences were much smaller (11–13%), because under the shifted alternative of equation (3.4), the biases in the two group estimates tend to cancel each other. The GLSE had smaller mean squared errors in estimating the between-group differences and higher statistical power to detect these differences than the UWLE in all four censoring processes considered. Compared to the PMLE, under the two probit informative censoring processes with nonzero slope coefficients, the GLSE had much larger mean squared errors in estimating the individual group mean slopes (39–69%) and in estimating the between-group differences (14–22%); and lower statistical power (10–15%) in detecting the between-group differences. The expected power for the proposed study, calculated according to (3.4) using the assumed parameter values, was .85 for the GLSE under the noninformative censoring process. The simulated power for the GLSE under noninformative censoring, using the estimated within- and between-individual variances according to (3.3) and the Bock and Petersen (1975) procedure for constructing covariance matrices that are at least semidefinite, was .81, and not very different from the expected power. Using the PMLE when the censoring process was noninformative or when the probit censoring slope coefficient was zero could result in larger mean squared errors than the GLSE, in estimating the group slopes. The simulated significance levels were not much different from the expected 5% level for all procedures in Table 2.

5. Discussion

The probit model used in Sections 2 and 4 is not necessarily meant to be biologically valid for describing the underlying right-censoring process. Indeed, the choice of the probit was made primarily on computational grounds, and because logistic and probit regressions give similar estimates of event probabilities (Halperin et al., 1979).

When the estimation and test procedures derived under the probit model are used, goodness of fit to the data should be checked as suggested in Section 4.1. However, the distribution of the chi-squared goodness-of-fit test statistic for this situation cannot be obtained from a straightforward application of the usual theory because (i) parameter estimates are determined using likelihood functions for ungrouped data; and (ii) cell boundaries are random. Moore and Spruill (1975) derived the large-sample distribution of the usual chi-squared goodness-of-fit statistics under these two problems. Their basic result is that under appropriate regularity conditions the large-sample distribution of the goodness-of-fit statistic is that of a central chi-squared with the usual reduction in degrees of freedom due to estimated parameters plus a weighted sum of independent chi-squared random variables each with 1 degree of freedom. Application of their result to this problem is under investigation. In work as yet unpublished, Wu and Bailey have developed estimation and test procedures to account for informative right censoring without modeling the censoring process. Their procedures are less dependent than ours on the underlying censoring model, but are also less efficient under the probit censoring model.

Although the estimation and test procedures of Sections 2 and 4 were developed for $k = 2$ groups, they could be extended easily to the case of $k > 2$ groups. To test for equality or linear trend among the expected slopes of the $k > 2$ groups, the likelihood-ratio chi-squared statistic could be used.

The standard errors provided in Table 1 for estimates based on the probit right-censoring model and those used in computing the test statistics for the PMLE in Table 2 were estimated from the sample information matrix based on the pseudo–maximum-likelihood (rather than the maximum likelihood), by assuming that the estimated between- and within-individual error variances were the true values. The bootstrap (Efron, 1979) could be used to improve these estimates. Alternatively, the maximum likelihood procedure, treating σ_ε^2, $\sigma_{\beta_1}^2$, and $\sigma_{\beta_2}^2$ as additional parameters, could also be used.

Acknowledgement

The authors wish to thank the referees for many helpful suggestions.

Résumé

On utilise souvent les moyennes non pondérées des moindres carrés ordinaires de chaque groupe quand on estime et compare les taux de changement d'une variable continue entre deux groupes. Sous un modèle linéaire à effets aléatoires, quand tous les individus ont été observés complètement aux mêmes instants, ces statistiques sont les estimations du maximum de vraisemblance des taux moyens de changement. Cependant, avec des données censurées ou manquantes, ces estimations ne sont plus efficaces, comparées aux estimations des moindres carrés généralisés. Quand, de plus, le processus de censure à droite dépend des taux individuels de changement (c.a.d. censure à droite informative), les estimations des moindres carrés sont biaisées. On développe des tests du rapport de vraisemblance sur la réalité du processus de censure, et l'estimation du maximum de vraisemblance des taux moyens de changements et les paramètres du processus de censure à droite, sous un modèle linéaire à effets aléatoires, avec un modèle probit pour le processus de censure à droite. Dans des estimations réelles, on montre que, d'une part, le biais obtenu en estimant un taux de changement par groupe et, d'autre part, la réduction de puissance dans la comparaison entre groupes peuvent être considérables quand on ne tient pas compte de la forte dépendance entre le processus de censure à droite et les taux de changement individuels.

References

Bock, R. D. and Petersen, A. C. (1975). A multivariate correction for attenuation. *Biometrika* **62,** 673–678.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Efron, B. (1979). Bootstrap methods. Another look at the jackknife. *Annals of Statistics* **7,** 1–26.

Fearn, T. (1975). A Bayesian approach to growth curves. *Biometrika* **62,** 89–100.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and application. *Annals of Statistics* **9,** 861–869.

Halperin, M., Wu, M., and Gordon, T. (1979). Genesis and interpretation of differences in distributions of baseline characteristics between cases and non-cases in cohort studies. *Journal of Chronic Diseases* **32,** 483–491.

Intermittent Positive Pressure Breathing Trial Group (1983). Intermittent positive pressure breathing therapy of chronic obstructive pulmonary disease. *Annals of Internal Medicine* **99,** 615–620.

Kaplan, L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53,** 457–481.

Koziol, J. A., Maxwell, D. A., Matsuro Fukushima, M. E. M., and Yosef, H. P. (1981). A distribution-free test for tumor-growth curve analyses with application to an animal immunotherapy experiment. *Biometrics* **37,** 383–390.

Laurell, B. and Eriksson, S. (1963). The electrophoretic alpha$_1$-3 lobulin pattern of serum in alpha$_1$-trypsin deficiency. *Scandinavian Journal of Clinical and Laboratory Investigation* **5,** 132.

Moore, D. S. and Spruill, M. C. (1975). Unified large-sample theory of general chi-squared statistics for test of fit. *Annals of Statistics* **3,** 599–616.

Morris, J. F., Koski, A., and Johnson, A. C. (1971). Spirometric standards for healthy nonsmoking adults. *American Review of Respiratory Disease* **103,** 57–68.

Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* **52,** 447–458.

Schlesselman, J. J. (1973). Planning a longitudinal study, II. Frequency of measurements and study duration. *Journal of Chronic Diseases* **26,** 561–570.

Walker, R. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54,** 167–179.

Workshop on Natural History of PiZ Emphysema (1983). The natural history of PiZ emphysema. Proteases and Inhibitors in the Lungs. *American Review of Respiratory Diseases* **127** (Supplement), 543–545.

Received December 1985; *revised April and September* 1987.

Appendix

Let

$$\mathbf{C}_{2i}=\begin{bmatrix}\sigma^2_{2i1} & \sigma_{2i12}\\ \sigma_{2i12} & \sigma^2_{2i2}\end{bmatrix},\quad \mathbf{C}_{3i}=\begin{bmatrix}\sigma^2_{3i1} & \sigma_{3i12}\\ \sigma_{3i12} & \sigma^2_{3i2}\end{bmatrix},$$

$\mathbf{d}^{\mathrm{t}}_{ik}=(d_{ik1},d_{ik2})$, $\rho_i=\sigma_{2i12}/(\sigma_{2i1}\sigma_{2i2})$, $\phi_{ij}=\phi(U_{ij})$, $\Phi_{ij}=\Phi(U_{ij})$, for $j=2,\ldots,J$ and $\phi_{i1}=\Phi_{i1}=0$. From (2.3), $U_{ij}=(\alpha_{0j}+\alpha_1 d_{ik1}+\alpha_2 d_{ik2})/D^{1/2}$, for $j=2,\ldots,J$, where $D=(1+\sigma^2_{3i1}\alpha^2_1+\sigma_{3i12}\alpha_1\alpha_2+\sigma^2_{3i2}\alpha^2_2)$. The parameter to be estimated is $\boldsymbol{\theta}^{\mathrm{t}}=(\theta_1,\ldots,\theta_{J+5})=(B_{11},B_{12},B_{21},B_{22},\alpha_1,\alpha_2,\alpha_{02},\ldots,\alpha_{0J})$. The partial derivatives of the log-likelihood (2.3) with respect to these parameters are as follows:

$$\frac{\partial\ \ln(L_i)}{\partial B_{kl}}=\begin{cases}\dfrac{[(\hat{\beta}_{il}-B_{kl})/\sigma^2_{2i2}-\rho_i(\hat{\beta}_{im}-B_{km})/(\sigma_{2i1}\sigma_{2i2})]}{(1-\rho^2_i)+[\partial T_i/\partial\beta_{kl}],} & \\ \qquad\qquad\qquad\text{for}\quad i\in k;\ k=1,2;\ l=1,2;\ m=3-l; & \\ 0\quad\text{otherwise} & \end{cases}$$

$$\frac{\partial^2\ln(L_i)}{\partial B_{k_1l}\partial B_{k_2m}}=\begin{cases}(-1)^{(l-m)}\rho^{|l-m|}/[\sigma_{2il}\sigma_{2im}(1-\rho^2_i)]+\partial T_i/\partial B_{k_1l}\partial B_{k_2m}, & \\ \qquad\qquad\text{for}\quad i\in k;\ k_1=k_2=k;\ l=1,2;\ \text{and}\ m=1,2; & \\ 0\quad\text{otherwise;} & \end{cases}$$

$$\partial\ \ln(L_i)/\partial\theta_l=\partial T_i/\partial\theta_l,\quad\text{for}\quad l=5,\ldots,J+5;$$

$$\frac{\partial^2\ln(L_i)}{\partial\theta_l\partial\theta_m}=\frac{\partial T_i}{\partial\theta_l\partial\theta_m},\quad\text{for}\quad l=5,\ldots,J+5;\quad m=1,\ldots,J+5.$$

When there is no staggered entry we have,

$$\begin{aligned}\frac{\partial T_i}{\partial\theta_l}&=\sum_{j=2}^{J}\left\{\frac{Z(i,j-1)[\phi_{ij}(\partial U_{ij}/\partial\theta_l)-\phi_{ij-1}(\partial U_{ij-1}/\partial\theta_l)]}{(\Phi_{ij}-\Phi_{ij-1})}\right\}\\ &\quad-\left[1-\sum_{j=2}^{J}Z(i,j-1)\right]\phi_{iJ}(\partial U_{iJ}/\partial\theta_l)/(1-\Phi_{iJ});\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 T_i}{\partial\theta_l\partial\theta_m}=\sum_{j=2}^{J}\Bigg\{&Z(i,j-1)\Bigg\{(\Phi_{ij}-\Phi_{ij-1})\\ &\times\left[-U_{ij}\phi_{ij}\left(\frac{\partial U_{ij}}{\partial\theta_l}\right)\left(\frac{\partial U_{ij}}{\partial\theta_m}\right)+U_{ij-1}\phi_{ij-1}\left(\frac{\partial U_{ij-1}}{\partial\theta_l}\right)\left(\frac{\partial U_{ij-1}}{\partial\theta_m}\right)+\phi_{ij}\left(\frac{\partial U_{ij}}{\partial\theta_l\partial\theta_m}\right)-\phi_{ij-1}\left(\frac{\partial U_{ij-1}}{\partial\theta_l\partial\theta_m}\right)\right]\\ &\qquad-\left[\phi_{ij}\left(\frac{\partial U_{ij}}{\partial\theta_l}\right)-\phi_{ij-1}\left(\frac{\partial U_{ij-1}}{\partial\theta_l}\right)\right]\left[\phi_{ij}\left(\frac{\partial U_{ij}}{\partial\theta_m}\right)-\phi_{ij-1}\left(\frac{\partial U_{ij-1}}{\partial\theta_m}\right)\right]\Bigg\}\Bigg/(\Phi_{ij}-\Phi_{ij-1})^2\Bigg\}\\ &-\left[1-\sum_{j=2}^{J}Z(i,j-1)\right]\\ &\times\left[(1-\phi_{iJ})U_{ij}\phi_{iJ}\left(\frac{\partial U_{iJ}}{\partial\theta_l}\right)\left(\frac{\partial U_{iJ}}{\partial\theta_m}\right)+\phi_{iJ}\left(\frac{\partial U_{iJ}}{\partial\theta_l\partial\theta_m}\right)+\phi^2_{iJ}\left(\frac{\partial U_{iJ}}{\partial\theta_l}\right)\left(\frac{\partial U_{iJ}}{\partial\theta_m}\right)\right]\Bigg/(1-\Phi_{iJ})^2,\\ &\qquad\qquad\text{for}\ l=1,\ldots,J+5\quad\text{and}\quad m=1,\ldots,J+5;\end{aligned}$$

with

$$\frac{\partial U_{ij}}{\partial B_{kl}} = \begin{cases} \alpha_1(\partial d_{ik1}/\partial B_{kl}) + \alpha_2(\partial d_{ik2}/\partial B_{kl})/D^{1/2}, \\ \qquad\qquad \text{for} \quad i \in k, \quad l = 1, 2; \\ 0 \quad \text{otherwise;} \end{cases}$$

$$\partial U_{ij}/\partial \alpha_l = [d_{ikl}D^{1/2} - U_{ij}C_l]/D, \quad \text{for} \quad i \in k; \quad l = 1, 2;$$

$$\partial U_{ij}/\partial \alpha_{0l} = \begin{cases} D^{-1/2} & \text{for} \quad l = j; \\ 0 & \text{otherwise;} \end{cases}$$

$$\frac{\partial^2 U_{ij}}{\partial B_{kl}\partial \alpha_m} = \left[D^{1/2}\left(\frac{\partial d_{ikl}}{\partial B_{kl}}\right) - C_m\left(\frac{\partial U_{ij}}{\partial B_{kl}}\right)\right] \bigg/ D, \qquad \text{for} \quad i \in k; \quad l = 1, 2; \quad \text{and} \quad m = 1, 2;$$

$$\frac{\partial^2 U_{ij}}{\partial \alpha_l \partial \alpha_m} = \left\{D\left[d_{ikl}C_m D^{-1/2} - a_{lm}U_{ij} - C_l\left(\frac{\partial U_{ij}}{\partial \alpha_m}\right)\right] - 2[D^{-1/2}d_{ikl} - C_l U_{ij}]C_m\right\} \bigg/ D^2;$$

$$\frac{\partial^2 U_{ij}}{\partial \alpha_l \partial \alpha_{0j}} = -\frac{C_l}{D^{3/2}} \quad \text{for} \quad l = 1, 2;$$

$$\frac{\partial^2 U_{ij}}{\partial B_{kl}\partial B_{km}} = \frac{\partial^2 U_{ij}}{\partial B_{kl}\partial \alpha_{0j}} = \frac{\partial^2 U_{ij}}{\partial \alpha_{0j_1}\partial \alpha_{0j_2}} = 0,$$

for $l = 1, 2;$ $m = 1, 2;$ $i \in k;$ $j_1 = 2, \ldots, J;$ $j_2 = 2, \ldots, J;$ and $j = 2, \ldots, J;$

where $C_\gamma = \sigma^2_{3i\gamma}\alpha_\gamma + \sigma_{3i12}\alpha_{3-\gamma}$ for $\gamma = 1, 2;$

$$a_{lm} = \begin{cases} \sigma^2_{3il} & \text{for} \quad l = m \\ \sigma_{3i12} & \text{for} \quad l \neq m. \end{cases}$$

A Note on the Efficiency of Sandwich Covariance Matrix Estimation

Göran KAUERMANN and Raymond J. CARROLL

The sandwich estimator, also known as robust covariance matrix estimator, heteroscedasticity-consistent covariance matrix estimate, or empirical covariance matrix estimator, has achieved increasing use in the econometric literature as well as with the growing popularity of generalized estimating equations. Its virtue is that it provides consistent estimates of the covariance matrix for parameter estimates even when the fitted parametric model fails to hold or is not even specified. Surprisingly though, there has been little discussion of properties of the sandwich method other than consistency. We investigate the sandwich estimator in quasi-likelihood models asymptotically, and in the linear case analytically. We show that under certain circumstances when the quasi-likelihood model is correct, the sandwich estimate is often far more variable than the usual parametric variance estimate. The increased variance is a fixed feature of the method and the price that one pays to obtain consistency even when the parametric model fails or when there is heteroscedasticity. We show that the additional variability directly affects the coverage probability of confidence intervals constructed from sandwich variance estimates. In fact, the use of sandwich variance estimates combined with t-distribution quantiles gives confidence intervals with coverage probability falling below the nominal value. We propose an adjustment to compensate for this fact.

KEY WORDS: Coverage probability; Generalized estimating equation; Generalized linear model; Heteroscedasticity; Linear regression; Marginal model; Quasi-likelihood; Robust covariance estimator; Sandwich estimator.

1. INTRODUCTION

The *heteroscedasticity-consistent covariance matrix estimator* is a common tool used for variance estimation of parameter estimates. Originally introduced by Huber (1967), Eicker (1967), and White (1980), the estimate has become popular in the econometric literature. In the last decade, the method has also been widely used in the context of generalized estimating equations (see, e.g., Diggle, Liang, and Zeger 1994; Liang and Zeger 1986; Liang, Zeger and Qaqish 1992, where it was introduced as the *sandwich variance estimator*. Whereas in econometric models the estimate is used to cope for heteroscedastic errors, in generalized estimating equations its objective is consistent variance estimation for dependent data. In the latter setting, efficient estimation of parameters requires specification of the correlation structure among the observations—which, however, typically is unknown. Therefore, a so-called working covariance matrix is used in the estimation step, which for variance estimation is combined with its corresponding empirical version in a sandwich form. This approach yields consistent estimates of the covariance matrix under misspecified working covariances as well as under heteroscedastic errors. Because of this desirable model-robustness property, the sandwich estimator is also sometimes called the *robust covariance matrix estimator* or the *empirical covariance matrix estimator*. We use the term *sandwich variance estimator* throughout the article.

The argument in favor of the sandwich estimate is that asymptotic normality and asymptotic coverage of confidence intervals require only a consistent variance estimate, so there is no direct need to construct a highly accurate covariance matrix estimate. But the consistency of the sandwich variance estimate has its price in increased variability; that is, sandwich variance estimators generally have a larger variance than model-based classical variance estimates. In his discussion of the article by Wu (1986), Efron (1986) gave simulation evidence of this phenomenon. Breslow (1990) demonstrated this in a simulation study of overdispersed Poisson regression. Firth (1992) and McCullagh (1992) both raised concerns that the sandwich estimator may be particularly inefficient. Diggle et al. (1994, p. 77) suggest that it is best used when the data come from "many experimental units." We clarify and refine these statements. An earlier discussion about small sample improvements for the sandwich estimate in the econometric literature was given by MacKinnon and White (1985), who proposed jackknife sandwich estimates. The performance of this estimate compared to other approaches was recently investigated by means of simulations by Long and Ervin (2000).

The objectives of this article are twofold. First we investigate the sandwich estimate in terms of efficiency; and second, we analyze the effect of the increased variability of the sandwich estimate on the coverage probability of confidence intervals. For efficiency, we derive asymptotic and fairly precise small-sample properties, neither of which appear to have been quantified before. For example, the sandwich method in simple linear regression when estimating the slope has an asymptotic efficiency equal to the inverse of the sample kurtosis of the design values. This inefficiency also holds for quasi-likelihood estimation and in generalized linear models. For example, in simple linear logistic regression, at the null value where there is no effect from the predictor, the sandwich method's asymptotic relative efficiency is again the inverse of the kurtosis of the predictors. In Poisson regression, the sandwich method has even less efficiency. The problem of undercoverage of confidence intervals was shown through simulation studies by Wu (1986) and by Breslow (1990), who reported somewhat

Göran Kauermann is Lecturer, Department of Statistics & Robertson Centre, University of Glasgow, University Gardens, Glasgow G12 8QW, Scotland (E-mail: *goeran@stats.gla.ac.uk*). Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *carroll@stat.tamu.edu*). Carroll's research was supported by National Cancer Institute Grant CA-57030, and by the Texas A&M Center for Environmental and Rural Health National Institute of Environmental Health Sciences Grant P30-ESO9106. The authors also acknowledge the support received from the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 386 at the Ludwig-Maximilans-Universität München). Parts of this article contain material in an unpublished technical report written with Douglas Simpson, Arny Stromberg, Suojin Wang, and David Ruppert, whose help is gratefully acknowledged.

Journal of the American Statistical Association
December 2001, Vol. 96, No. 456, Theory and Methods

elevated levels of Wald-type tests based on the sandwich estimator. Rothenberg (1988) derived an adjusted distribution function for the t-statistic calculated from sandwich variance estimates. We give a different theoretical justification for the empirical fact that confidence intervals calculated from sandwich variance estimates and t-distribution quantiles are generally too small; that is, the coverage probability falls below the nominal value. We show that undercoverage is determined mainly by the variance of the variance estimate. To correct this deficit, we present an adjustment that depends on normal distribution quantiles and the variance of the sandwich variance estimate.

The article is organized as follows. In Section 2 we compare the sandwich estimator with the usual parametric regression estimator in the linear regression model. In Section 3 we discuss the sandwich estimate for quasi-likelihood and generalized estimating equations (GEEs). We provide proofs and other general statements in the Appendix.

2. LINEAR REGRESSION

2.1 Properties of the Sandwich Estimator

First, consider the simple homoscedastic linear regression model

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \quad \text{with } \epsilon_i \sim N(0, \sigma^2), \tag{1}$$

where $\mathbf{x}_i^{\mathrm{T}}$ are $1 \times p$-dimensional vectors of covariates and $i = 1, \dots, n$. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ be the ordinary least squares estimator of $\boldsymbol{\beta}$, where $\mathbf{Y}^{\mathrm{T}} = (Y_1, \dots Y_n)$ and $\mathbf{X}^{\mathrm{T}} = (\mathbf{x}_1, \dots \mathbf{x}_n)$. Assume now that we are interested in inference about the linear combination $\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, where $\mathbf{z}^{\mathrm{T}}$ is a $1 \times p$-dimensional contrast vector of unit length, that is, $\mathbf{z}^{\mathrm{T}}\mathbf{z} = 1$. The variance of $\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is given by $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}$, which can be estimated by the classical model-based variance estimator $V_{\mathrm{model}} = \hat{\sigma}^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}$, where $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2/(n-p)$ with $\hat{\epsilon}_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ as fitted residuals. A major assumption used implicitly in the calculation of V_{model} is that the errors ϵ_i are homoscedastic. This assumption is often not very plausible, particularly in econometric models, where one is faced with heteroscedasticity, so that the model

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \quad \text{with } \epsilon_i \sim N(0, \sigma_i^2) \tag{2}$$

holds. In this case V_{model} does not provide a consistent variance estimate. In contrast, the sandwich variance estimate,

$$V_{\mathrm{sand}} = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\Big(\sum_i \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\hat{\epsilon}_i^2\Big)(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} = \sum_{i=1}^n a_i^2\hat{\epsilon}_i^2, \tag{3}$$

consistently estimates $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, where $a_i = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_i$. Estimate (3) is called the *sandwich variance estimator* because of its sandwich structure, even though the terms *robust variance estimator, heteroscedasticity-consistent covariance estimator*, and *empirical covariance estimator* are more common in the econometric literature.

In linear regression, (3) is often multiplied by $n/(n-p)$ (Hinkley 1977) to reduce the bias. Let h_{ii} be the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} = (h_{ij})$. Under homoscedasticity, one finds $E(\hat{\epsilon}_i^2) = \sigma^2(1-h_{ii})$, so that

$$E(V_{\mathrm{sand}}) = \sigma^2\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}(1-b_n), \tag{4}$$

where $b_n = \sum_{i=1}^n h_{ii}a_i^2 / \sum_{i=1}^n a_i^2 \le \max_{1\le i\le n} h_{ii}$. Because $b_n \ge 0$, one obtains that in general the sandwich estimator is biased *downward*, as was shown by Chesher and Jewitt (1987), (see also MacKinnon and White 1985). The bias therefore depends on the design of $\mathbf{x}_i$ and can be substantial when there are leverage points. Bias problems can be avoided by replacing $\hat{\epsilon}_i$ in (3) by $\tilde{\epsilon}_i = \hat{\epsilon}_i/(1-h_{ii})^{1/2}$. The resulting estimator is referred to as the *unbiased* sandwich variance estimator and is denoted by $V_{\mathrm{sand},\,u}$ (Wu 1986, eq. 2.6). It is easily seen that $E(V_{\mathrm{sand},\,u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, whereas under heteroscedasticity of model (2), the estimate is still consistent but with an asymptotic of order $O(n^{-1})$. Because $\mathrm{var}(\tilde{\epsilon}_i^2) = 2\sigma^4$ and $\mathrm{cov}(\tilde{\epsilon}_i^2, \tilde{\epsilon}_j^2) = 2\widetilde{h}_{ij}^2\sigma^4$ for $i \ne j$, where $\widetilde{h}_{ij} = h_{ij}/\{(1-h_{ii})(1-h_{jj})\}^{1/2}$, it follows that

$$\begin{aligned}\mathrm{var}(V_{\mathrm{sand},\,u}) &= \sum_{i=1}^n a_i^4\mathrm{var}(\tilde{\epsilon}^2) + \sum_{i\ne j} a_i^2a_j^2\mathrm{cov}(\tilde{\epsilon}_i^2, \tilde{\epsilon}_j^2)\\ &= 2\sigma^4\left(\sum_{i=1}^n a_i^4 + \sum_{i\ne j} a_i^2a_j^2\widetilde{h}_{ij}^2\right).\end{aligned} \tag{5}$$

We now compare the variance (5) to the variance of the model-based variance estimator $\mathbf{V}_{\mathrm{model}}$, which equals $\mathrm{var}(V_{\mathrm{model}}) \approx 2\sigma^4\{\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z}\}^2/n = 2\sigma^4(\sum a_i^2)^2/n$.

Theorem 1. Under the homoscedastic linear model (1), the efficiency of the unbiased sandwich estimate $\mathbf{V}_{\mathrm{sand}.u}$ compared to the classical variance estimate $\mathbf{V}_{\mathrm{model}}$ for $\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ satisfies

$$\frac{\mathrm{var}(\mathbf{V}_{\mathrm{sand},\,u})}{\mathrm{var}(\mathbf{V}_{\mathrm{model}})} \ge \left\{n^{-1}\sum_{i=1}^n a_i^4\right\}\left\{n^{-1}\sum_{i=1}^n a_i^2\right\}^{-2} \ge 1. \tag{6}$$

The proof follows directly from the Cauchy–Schwarz inequality. Theorem 1 states that the sandwich estimate is less efficient when the model is correct, that is, when the errors are homoscedastic. Because of the vector $\mathbf{z}$, the loss of efficiency basically depends on the design of the covariates, as the following example shows.

Example 1 (The Intercept and the Slope in Simple Linear Regression). Assume that $\mathbf{x}_i^{\mathrm{T}} = (1, u_i)$, where $\sum u_i = 0$. Suppose that we are interested in the intercept, that is, $\mathbf{z}^{\mathrm{T}} = (1, 0)$. We then have $a_i = n^{-1}$, and the asymptotic relative efficiency in (6) is 1. Suppose now that $\mathbf{z} = (0, 1)$, so that $\hat{\beta}_1 = \mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is the slope estimate. Assuming $\max(|u_i|) = o(n^{1/4})$ for technical purposes, the asymptotic relative efficiency is κ_n^{-1}, where $\kappa_n = n^{-1}\sum u_i^4/(n^{-1}\sum_{i=1}^n u_i^2)^2 \ge 1$. Note that κ_n is the sample kurtosis of the design points u_i. For instance, if the design points $(u_1, \dots, u_n)$ are realizations of a normal distribution, then $\kappa_n \to 3$, and hence the sandwich estimator $V_{\mathrm{sand},\,u}$ has three times the variability of the usual model-based estimator V_{model}. If the design points are generated from a Laplace distribution, then the usual sandwich estimator is six times more variable.

The foregoing example shows that using sandwich variance estimates in linear models can lead to a substantial loss of efficiency. A similar phenomena occurs for quasi-likelihood estimation, as discussed in the next section. Note that Theorem 1 is formulated under the assumption of homoscedasticity. Even though it might be interesting to weaken this assumption and analyze the efficiency of $V_{\text{sand}, u}$ under heteroscedasticity theoretically, one should keep in mind that V_{model} is not a consistent estimate under heteroscedasticity, so its bias also must be taken into account. Instead, we investigate the behavior of the estimate under heteroscedastic errors empirically in the simulations studies of the following subsection.

2.2 Coverage Probability of Confidence Intervals

In this section we investigate how the additional variability affects the coverage probability of confidence intervals obtained from sandwich variance estimates. As one would expect, the excess variability of the sandwich estimate is directly reflected in undercoverage of confidence intervals. Let $\theta = \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$ be the unknown parameter of interest with $\hat{\theta} = \mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \sim N(\theta, \sigma^2/n)$ an unbiased estimate of θ based on a random sample of size n. The symmetric $1-\alpha$ confidence interval is given by $CI(\sigma^2, \alpha) := [\hat{\theta} \pm z_p \sigma/\sqrt{n}]$, where z_p is the $p = 1-\alpha/2$ quantile of the standard normal distribution. If σ^2 is estimated by an unbiased variance estimate $\hat{\sigma}^2$, it is well known that the confidence interval $CI(\hat{\sigma}^2, \alpha)$ shows undercoverage, and typically t-distribution quantiles are used instead of normal quantiles. The following theorem shows explicitly how the variance of $\hat{\sigma}^2$ affects the undercoverage.

Theorem 2. Let $\hat{\theta} \sim N(\theta, \sigma^2/n)$ and let $\hat{\sigma}^2$ be an unbiased estimate of σ^2 independent of $\hat{\theta}$. The coverage probability of the $1-\alpha$ confidence interval $CI(\hat{\sigma}^2, \alpha)$ equals

$$\Pr\{\theta \in CI(\hat{\sigma}^2, \alpha)\} = 1 - \alpha - c_p \frac{\operatorname{var}(\hat{\sigma}^2)}{\sigma^4} + O(n^{-2}), \tag{7}$$

where $c_p = \phi(z_p)(z_p^3 + z_p)/8$, with $\phi(\cdot)$ the standard normal distribution density.

The proof of Theorem 2 is given in the Appendix. Note that the assumption of independence of $\hat{\sigma}^2$ and $\hat{\theta} - \theta$ holds in a normal homoscedastic regression model if $\hat{\sigma}^2$ is calculated from fitted residuals; that is, it holds for sandwich variance estimates. Because $c_p > 0$ (for $p > 1/2$), undercoverage becomes obvious. In particular, the undercoverage increases linearly with the variance of the variance estimate $\hat{\sigma}^2$. Using the results of Theorem 1, we therefore conclude that confidence intervals based on sandwich variance estimators have lower coverage probability than confidence intervals based on model-based variance estimates. This also implies that t-distribution quantiles do not correct the undercoverage. The result stated in Theorem 2 resembles that given by Rothenberg (1988, p. 1005). He derived an adjustment for the distribution function of the t-statistic based on sandwich variance estimates. In contrast to Rothenberg, however, Theorem 2 points out the distinct role of the variance of $\hat{\sigma}^2$.

Coverage Adjustment. In normal linear regression models, formula (7) can be used directly to construct a coverage correction for confidence intervals. Instead of using quantile z_p, we suggest choosing $\tilde{p} > p$ and make use of the quantile $z_{\tilde{p}}$. The increased $\tilde{p}$ is then selected such that $\Pr(\theta \in [\hat{\theta} \pm z_{\tilde{p}}\hat{\sigma}/\sqrt{n}]) = p$ holds; that is, with (7), $\tilde{p}$ is defined as the numerical solution to

$$p = \tilde{p} - \phi(z_{\tilde{p}})\operatorname{var}(\hat{\sigma}^2)\frac{z_{\tilde{p}}^3 + z_{\tilde{p}}}{8\sigma^4}. \tag{8}$$

Example 2 (t-Distribution Quantiles). Before applying the correction to the sandwich variance estimate, we demonstrate the use of (8) in a setting where an exact solution is available. Let the random sample $Y_i \sim N(\mu, \sigma^2)$ be drawn from an univariate normal distribution. The centered mean estimate $n^{1/2}(\hat{\mu} - \mu)$ is distributed as a normal $(0, \sigma^2)$, with $\hat{\mu} = \sum_i^n Y_i/n$ and variance σ^2 estimated by $\hat{\sigma}^2 = \sum_i^n (Y_i - \hat{\mu})^2/(n-1)$. Exact quantiles for confidence intervals based on the estimates $\hat{\mu}$ and $\hat{\sigma}^2$ are available from t-distribution quantiles with $n-1$ degrees of freedom. Approximative quantiles $z_{\tilde{p}}$ follow from solving (8) using $\operatorname{var}(\hat{\sigma}^2) = 2\sigma^4/(n-1)$. It is a special feature of the normal distribution that the unknown variance component σ^4 in (8) cancels out and is not required for the calculation of $z_{\tilde{p}}$. In Table 1 we compare the exact quantiles based on a t-distribution with the corrected versions based on (8). Even for small sample sizes, the corrected quantiles $z_{\tilde{p}}$ are distinctly close to the exact t-distribution quantiles. This is also seen in the true one-sided coverage probability $\Pr(\hat{\theta} \leq \theta + z_{\tilde{p}}\hat{\sigma}/\sqrt{n})$ of the confidence intervals, and demonstrates that the adjustment applied in a standard setting behaves quite well.

Example 3 (Sandwich Variance Estimate). We now apply the corrected quantile $z_{\tilde{p}}$ to confidence intervals based on sandwich variance estimates. Inserting (5) in (8) shows again that the variance component σ^4 cancels out, so that the correction depends exclusively on the design of the covariates. We ran a small simulation study to demonstrate the behavior of the correction. Let $Y_i = \beta_0 + x_i\beta_x + \varepsilon_i$ with $\beta_0 = 0$, $\beta_x = 1$, and $\varepsilon_i \sim N(0, \sigma_i^2)$. The errors are drawn from the homoscedastic model (1), that is, σ_i constant with value .2 (model 1), as well as from the heteroscedastic model (2) with $\sigma_i = .2 + \exp(x_i/2)/2$ (model 2) and $\sigma_i = \sqrt{(.1 + x_i^2)}$ (model 3). The covariates x_i are chosen to be (a) uniformly, (b)

Table 1. Comparison of Coverage Probability Based on $z_{\tilde{p}}$ and t-Distribution Quantiles $t_{p,n-1}$ for $n-1$ Degrees of Freedom

p	$t_{p,n-1}$	$z_{\tilde{p}}$	$P(\hat{\theta} \leq \theta + z_{\tilde{p}}\hat{\sigma}/\sqrt{n})$
$n = 5$			
.90	1.533	1.551	.902
.95	2.132	2.095	.948
.975	2.776	2.543	.968
$n = 15$			
.90	1.345	1.346	.900
.95	1.761	1.761	.950
.975	2.145	2.137	.975

Table 2. Corrected Quantiles $z_{\tilde{p}}$ for Different Designs

	$p = .90$		$p = .95$	
Design	$t_{p,n-2}$	$z_{\tilde{p}}$	$t_{p,n-2}$	$z_{\tilde{p}}$
		$n = 20$		
(a)		1.81		2.21
(b)	1.73	1.86	2.10	2.27
(c)		1.94		2.36
		$n = 40$		
(a)		1.72		2.07
(b)	1.68	1.75	2.02	2.12
(c)		1.80		2.19

normally, and (c) Laplace distributed. The corrected quantiles, $z_{\tilde{p}}$, are listed in Table 2. The results shown in Figure 1 give the empirical coverage probability based on 2000 replicates with $n = 20$ (upper row) and $n = 40$ (bottom row) observations. For comparison, we also show the coverage probability for confidence intervals calculated from $\mathbf{V}_{\text{sand},u}$ and t-distribution quantiles. Moreover, we calculate confidence intervals based on the jackknife estimate as suggested by MacKinnon and White (1985, form. 13). This has the form

$$\mathbf{V}_{\text{jack}} = \frac{n-1}{n}\mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\Big(\sum_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i\hat{\epsilon}_i^{*2}\Big) \times (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{z} - \frac{n-1}{n^2}\hat{\gamma}^{\mathrm{T}}\hat{\gamma}, \qquad (9)$$

where $\hat{\gamma} = \mathbf{z}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\epsilon}^*$ and $\hat{\epsilon}_i^* = \hat{\epsilon}_i(1-h_{ii})$.

It appears that uncorrected intervals clearly suffer from undercoverage. This is corrected to a large extent by both the corrected quantiles and the jackknife estimate. In general, however, the corrected quantiles behave slightly better in all three models, for heteroscedastic as well as homoscedastic errors.

The foregoing examples show the benefits of correction (8). For practical purposes, it might be cumbersome to solve (8) explicitly, however. Instead, an approximate solution of (8) based on the relative efficiency $\text{var}(\mathbf{V}_{\text{sand},u})/\text{var}(\mathbf{V}_{\text{model}})$ given in (6) in Theorem 1 can be used. As shown in the Appendix, one easily gets

$$z_{\tilde{p}} = t_{p,n-p} + d_{p,n-p}\left(\frac{\text{var}(\mathbf{V}_{\text{sand},u})}{\text{var}(\mathbf{V}_{\text{model}})} - 1\right) + O(n^{-2}), \qquad (10)$$

where t_p is the t-distribution quantile with $n-p$ degrees of freedom and $d_{p,n-p} = \text{var}(\mathbf{V}_{\text{model}})(z_p^3 + z_p)/(8\sigma^4)$. As before, the variance term σ^4 cancels out when $\text{var}(\mathbf{V}_{\text{model}})$ is inserted. Formula (10) shows that the corrected quantiles depend linearly on the relative efficiency. The slope parameter is thereby decreasing with increasing sample size. Relation (10) is visualized in Figure 2, where we plot $z_{\tilde{p}}$ against the relative efficiency $\text{var}(\mathbf{V}_{\text{sand},u})/\text{var}(\mathbf{V}_{\text{model}})$ for $p = .95$ and $p = .975$. The linear shape is obvious, and it appears that the correction is substantial if the relative efficiency is large and/or the sample size is small. Figure 2 and (10) can also be used to provide confidence intervals with appropriate coverage probability by calculating the relative efficiency.

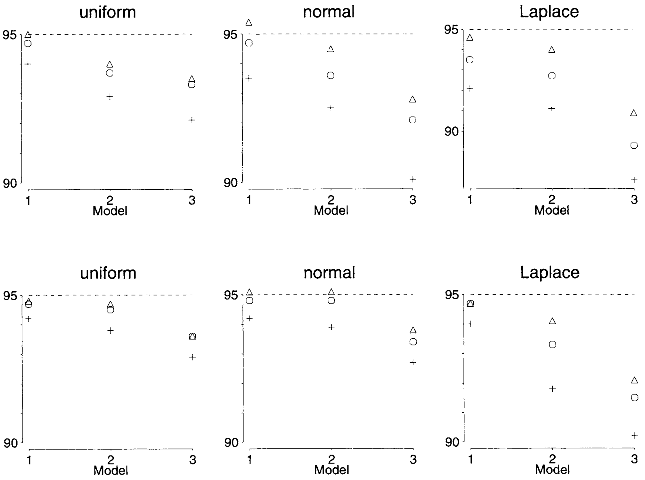


Figure 1. Coverage Probability of Confidence Intervals Based on $\mathbf{V}_{\text{sand},u}$ With Corrected Quantiles $z_{\tilde{p}}$ (△) as Well as With t-Distribution Quantiles $t_{p,n-1}$ (+) and Based on the Jackknife Estimate $\mathbf{V}_{\text{jack}}$(○). The upper row is for $n = 20$; the bottom row, for $n = 40$.

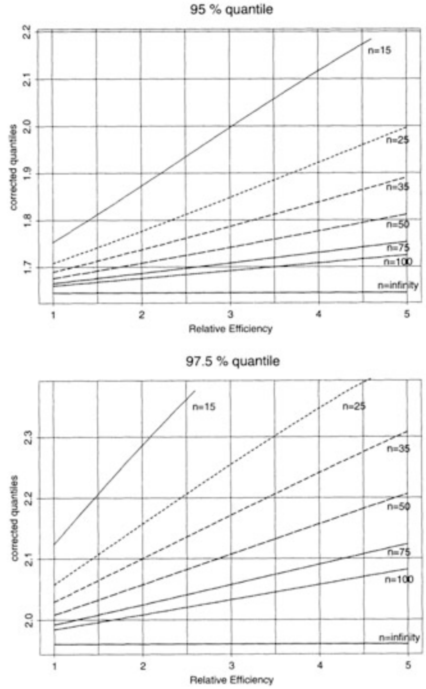


Figure 2. Corrected Quantiles $z_{\hat{p}}$ in Dependence of Sample Size n and Relative Efficiency $\mathrm{var}(\mathbf{V}_{\mathrm{sand},u})/\mathrm{var}(\mathbf{V}_{\mathrm{model}})$.

3. QUASI-LIKELIHOOD AND GENERALIZED ESTIMATING EQUATIONS

3.1 Properties of the Sandwich Estimate

In this section we consider the sandwich variance estimate for quasi-likelihood estimates from GEEs. Let $Y_i = (Y_{i1}, \ldots, Y_{im})^{\mathrm{T}}$, be a random vector taken at the ith unit for $i = 1, \ldots, n$ and $m \geq 1$. For $m > 1$, the components of Y_i are allowed to be correlated while observations taken at two different units are independent. Although in principle the number of observations per unit may vary from unit to unit, for ease of notation we take m as constant here. The case $m = 1$ is of course a special case in the formulas. The mean of Y_i given the $m \times p$-dimensional design matrix $\mathbf{X}_i^{\mathrm{T}}$ is given by the generalized linear model $E(Y_i|\mathbf{X}_i) = h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, where $h(\cdot)$ is an invertible m-dimensional link function. Efficient estimation of $\boldsymbol{\beta}$ requires knowledge of the covariance matrix of Y_i. This is typically unknown, and thus one specifies $\sigma^2 V(\mu_i) =: \sigma^2 \mathbf{V}_i$ as the so-called working covariance matrix, where $\mu_i = h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$, $V(\cdot)$ is a specified covariance variance function, and σ^2 is a dispersion scalar that is either unknown (e.g., for normal response) or a known constant, (e.g., $\sigma^2 \equiv 1$ for Poisson data). Models of this type are also called marginal models (see Diggle et al. 1994 and references therein). If Y_i is a scalar, (i.e., if $m = 1$), models of this type are better known as quasi-likelihood models (Wedderburn 1974) or generalized linear models (McCullagh and Nelder 1989). The parameter $\boldsymbol{\beta}$ can be estimated using the GEE (e.g., Gourieroux, Monfort, and Trognon 1984; Liang and Zeger 1986)

$$0 = \sum_i \frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(Y_i - \mu_i). \tag{11}$$

In the previous section, we were able to perform exact calculations. In quasi-likelihood models, such exact calculations are not feasible, and asymptotics are required. We do not write down formal regularity conditions, but essentially what is necessary is that sufficient moments of the components of $\mathbf{X}$ and Y exist. We also require sufficient smoothness of $h(\cdot)$. Under such conditions, a Taylor expansion of (11) about the true parameter $\boldsymbol{\beta}$ provides the first-order approximation

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\Omega}^{-1} \sum_i \frac{\partial \mu_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(Y_i - \mu_i) + O_p(n^{-1}), \tag{12}$$

where $\boldsymbol{\Omega} = \sum_i \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta}) \mathbf{V}_i^{-1} \partial \mu_i/(\partial \boldsymbol{\beta})$. Assume that we are interested in inference about $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}$. If $\mathbf{V}_i$ is correctly specified, that is, if $\sigma^2 \mathbf{V}_i = \mathrm{var}(Y_i|\mathbf{X}_i)$, then one gets $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{z}\sigma^2$ to a first-order approximation. Hence we can estimate $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ by $\mathbf{V}_{\mathrm{model}} := \hat{\sigma}^2 \mathbf{z}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{-1}\mathbf{z}$, where $\widehat{\boldsymbol{\Omega}}$ is a simple plug-in estimate of $\boldsymbol{\Omega}$ and $\hat{\sigma}^2$ is an estimate of the dispersion parameter if this is unknown. But in practice the covariance matrix may not be known—that is, $\mathbf{V}_i$ in (11) can be misspecified—which means that $\sigma^2 \mathbf{V}_i \neq \mathrm{var}(Y_i|\mathbf{X}_i)$ holds. In this case the variance $\mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$ can be estimated consistently by the sandwich formula

$$\mathbf{V}_{\mathrm{sand}} = \mathbf{z}^{\mathrm{T}} \widehat{\boldsymbol{\Omega}}^{-1} \left(\sum_i \frac{\partial \hat{\mu}_i^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \widehat{\mathbf{V}}_i^{-1} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^{\mathrm{T}} \widehat{\mathbf{V}}_i^{-1} \frac{\partial \hat{\mu}_i}{\partial \boldsymbol{\beta}} \right) \widehat{\boldsymbol{\Omega}}^{-1}\mathbf{z}, \tag{13}$$

where $\hat{\boldsymbol{\epsilon}}_i = Y_i - \hat{\mu}_i = Y_i - h(\mathbf{X}_i\hat{\boldsymbol{\beta}})$ are the fitted residuals and the hat notation refers to simple plug-in estimates. The fitted residuals can be expanded as $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{\epsilon}_i - \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\{1 + O_p(n^{-1/2})\}$, and assuming for the moment that $\mathbf{V}_i$ correctly specifies the covariance, that is, $E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}) = \sigma^2\mathbf{V}_i$, one finds via (12) that $E(\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^{\mathrm{T}}) = \sigma^2\mathbf{V}_i - \sigma^2 \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\Omega}^{-1} \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})\{1 + O(n^{-1})\}$. Because $\partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\Omega}^{-1} \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta})$ is positive definite, the sandwich estimate $\mathbf{V}_{\mathrm{sand}}$ appears to be biased downward with order $O(n^{-1})$, and, as in the previous section, the bias can be corrected. Thus let $\tilde{\boldsymbol{\epsilon}}_i = (\mathbf{I} - \mathbf{H}_{ii})^{-\frac{1}{2}}\hat{\boldsymbol{\epsilon}}_i$ define the leverage-adjusted residuals with $\mathbf{I}$ as identity matrix and $\mathbf{H}_{ii} = \partial \mu_i/(\partial \boldsymbol{\beta}^{\mathrm{T}}) \boldsymbol{\Omega}^{-1} \partial \mu_i^{\mathrm{T}}/(\partial \boldsymbol{\beta}) \mathbf{V}_i^{-1}$. Replacing $\hat{\boldsymbol{\epsilon}}$ in (13) by $\tilde{\boldsymbol{\epsilon}}$ gives the bias-reduced sandwich estimate $\mathbf{V}_{\mathrm{sand},u}$ that satisfies $E(\mathbf{V}_{\mathrm{sand},u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\{1 + O(n^{-2})\}$, assuming that the variance is correctly specified. If in contrast the variance is not correctly specified, that is, if $\mathbf{V}_i\sigma^2 \neq E(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{T})$ holds, then the first-order asymptotic bias remains, so that $E(\mathbf{V}_{\mathrm{sand},u}) = \mathrm{var}(\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\{1 + O(n^{-1})\}$. This means that the first-order bias reduction holds only if the variance is known.

In practice, however, it seems to be a plausible strategy to work with $\mathbf{V}_{\text{sand},u}$ instead of $\mathbf{V}_{\text{sand}}$, even if $\mathbf{V}_i$ is a working covariance and the true variance structure is unknown.

3.2 Examples

Theorem A.1 in Appendix Section A.2 extends Theorem 1 to the quasi-likelihood setting. The formulation and presentation of the result is cumbersome, however, and thus is deferred to the Appendix. The major reason for the additional complications is that in quasi-likelihood equations and GEEs, variance estimates have two different sources of stochastic variation. The first source is estimation of the dispersion parameter σ^2, if this is unknown; the second is the use of plug-in estimates, which are used if the variance function $V(\mu)$ depends on the mean. We demonstrate the loss of efficiency from the second source with Poisson and binomial data where the dispersion parameter is known. The variability of the model-based variance estimate occurs here solely from plug-in estimation.

Example 4 (Poisson Log-Linear Regression). We consider the univariate model $E(Y_i|\mathbf{x}) = \exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ where $\mathbf{x}_i = (1, u_i)$ with u_i scalar, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$, and Y_i Poisson distributed. The slope β_1 is the parameter of interest, and we investigate the null case $\boldsymbol{\beta} = (1, 0)^{\mathrm{T}}$. Then, as seen in the Appendix, if u has a symmetric distribution, then in limit as $n \to \infty$, $\text{var}(\mathbf{V}_{\text{sand}})/\text{var}(\mathbf{V}_{\text{model}}) = \kappa_n\{1 + 2\exp(\beta_0)\}$, where $\kappa_n = n^{-1}\sum_i u_i^4/(n^{-1}\sum_i u_i^2)^2$ is the sample kurtosis as in Example 3. The additional variability in the Poisson case is somewhat surprising—namely, that as the background event rate $\exp(\beta_0)$ increases, at the null case the sandwich estimator has efficiency decreasing to 0.

Example 5 (Logistic Regression). Let Y_i be binary with $E(Y_i|\mathbf{x}_i) = \text{logit}^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ with $\mathbf{x}_i$ as described before. Again, the slope β_1 is the parameter of interest. We vary β_1 while choosing β_0 so that marginally $E(y|\mathbf{x}) = .10$. With $\beta_1 = 0$, .5, 1.0, and 1.5, the asymptotic relative efficiency $\text{var}(\mathbf{V}_{\text{sand}})/\text{var}(\mathbf{V}_{\text{model}})$ varies for u_i standard normally distributed as 3.00, 2.59, 1.92, and 1.62. When u_i comes from a Laplace distribution (with unit variance), the corresponding efficiencies are 6.00, 4.36, 3.31, and 2.57. Note that in both cases, at the null case $\beta_1 = 0$ the efficiency of the sandwich estimator is exactly the same as the linear regression problem. This is no numerical fluke, and in fact can be shown to hold generally when u has a symmetric distribution.

The previous two examples show that the loss of efficiency of the sandwich variance estimate in nonnormal models differs from and can be worse than that occurring in normal models.

3.3 Coverage Probability

Undercoverage as pointed out in (7) of Theorem 2 extends asymptotically to quasi-likelihood or generalized linear models. For multivariate normal response models with correctly specified covariance matrices $\mathbf{V}_i$, $i = 1, \ldots, n$, Theorem 2 still holds exactly because variance estimates are independent of parameter estimates. But even if covariance matrices are misspecified, correction (8) derived from Theorem 2 can provide improved coverage, as demonstrated in our simulations. In the general case, however, exact calculation of the coverage probability and of the variance of the sandwich estimate are cumbersome, as seen from Examples 4 and 5. Because we concentrate on symmetric confidence intervals, which themselves are based on asymptotic normality arguments, it seems plausible to neglect the effect of plug-in estimates in the following. We show in simulations that for normal response and nonnormal response models, the coverage adjustment has a positive effect by compensating for undercoverage. To apply correction (8), we have to calculate the variance of the sandwich estimate. This can be done efficiently using matrix algebra.

Calculation of $\text{var}(\mathbf{V}_{\text{sand},u})$. We rewrite (13) in matrix form. Let $\mathbf{Y}$ denote the $(mn) \times 1$-dimensional vector $(Y_1^{\mathrm{T}}, \ldots, Y_n^{\mathrm{T}})^{\mathrm{T}}$ and set $\boldsymbol{\mu} = (\mu_1^{\mathrm{T}}, \ldots, \mu_n^{\mathrm{T}})^{\mathrm{T}}$. The residual vector is defined by $\boldsymbol{\epsilon} = \mathbf{Y} - \boldsymbol{\mu}$. Let $\mathbf{P}$ denote the projection-type matrix $\mathbf{P} = (\mathbf{I} - \mathbf{H})$, where $\mathbf{I}$ is the $(nm) \times (nm)$ identity matrix and $\mathbf{H}$ is the hat-type matrix

$$\mathbf{H} = \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\boldsymbol{\Omega}^{-1}\frac{\partial\boldsymbol{\mu}^{\mathrm{T}}}{\partial\boldsymbol{\beta}}\text{diag}_m(\mathbf{V}_i^{-1}),$$

with $\text{diag}_m(\mathbf{V}_i^{-1})$ denoting the block diagonal matrix with $\mathbf{V}_i^{-1}$ on its diagonal, $i = 1, \ldots n$. Note that for $m \equiv 1$, other versions of the hat matrix have been suggested (see Cook and Weisberg 1982, pp. 191–192, for logistic regression or Carroll and Ruppert 1988, p. 74, for other models). Let $\mathbf{W}$ be the block diagonal matrix $\mathbf{W} = \text{diag}_m(\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i)$ with $\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i$ on the block diagonals, where $\mathbf{a}_i = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\frac{\partial\mu_i^{\mathrm{T}}}{\partial\boldsymbol{\beta}}V_i^{-1}$. With $\mathbf{D} = \text{diag}_m(\mathbf{I} - \mathbf{H}_{ii})^{-1/2}$, we get the leverage-adjusted fitted residuals $\tilde{\boldsymbol{\epsilon}} = \mathbf{D}\{\mathbf{Y} - \hat{\boldsymbol{\mu}}\} = \mathbf{DP}(\mathbf{Y} - \boldsymbol{\mu})\{1 + O_p(n^{-1/2})\}$. As before, we use the hat notation to denote plug-in estimates. This allows us to write

$$\begin{aligned}\mathbf{V}_{\text{sand},u} &= \tilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\widehat{\mathbf{W}}\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{P}\widehat{\mathbf{D}}\widehat{\mathbf{W}}\widehat{\mathbf{D}}\mathbf{P})\boldsymbol{\epsilon} \\ &= \sigma^2\dot{\boldsymbol{\epsilon}}^{\mathrm{T}}\widehat{\mathbf{M}}\dot{\boldsymbol{\epsilon}}\{1 + O(n^{-1})\}, \qquad (14)\end{aligned}$$

where $\mathbf{M} = \text{diag}_m(\mathbf{V}_i^{1/2})\mathbf{P}\mathbf{D}\mathbf{W}\mathbf{D}\mathbf{P}\,\text{diag}_m(\mathbf{V}_i^{1/2})$ and $\dot{\boldsymbol{\epsilon}}^{\mathrm{T}} = (\dot{\boldsymbol{\epsilon}}_1^{\mathrm{T}}, \ldots, \dot{\boldsymbol{\epsilon}}_n^{\mathrm{T}})$ independent, homoscedastic residuals defined by $\dot{\boldsymbol{\epsilon}}_i = \mathbf{V}_i^{-1/2}\boldsymbol{\epsilon}_i$, where we assume again that $\sigma^2\mathbf{V}_i$ correctly specifies the variance of Y_i. The quadratic form now easily allows calculation of the variance of the sandwich variance. Let m_{kl} denote the k, lth element of $\mathbf{M}$ and let $\dot{\epsilon}_k$ be the elements of $\dot{\boldsymbol{\epsilon}}$, where $k, l = 1, 2, \ldots mn$. Neglecting the effect of plug-in estimates, we find

$$\text{var}(\mathbf{V}_{\text{sand},u}) = 2\sigma^4\text{trace}(\mathbf{MM}) + \sigma^4\sum_k\{E(\dot{\epsilon}_k^4) - 3\}m_{kk}^2. \qquad (15)$$

If the $(\dot{\epsilon}_k)$ are standard normal, then (15) simplifies to $\text{var}(\mathbf{V}_{\text{sand},u}) = 2\sigma^4\text{tr}(\mathbf{MM})$. The variance of the sandwich variance estimate again depends distinctly on the design of the covariates because of $\partial\mu_i^{\mathrm{T}}/\partial\boldsymbol{\beta} = \mathbf{X}_i\partial h(\eta)/\partial\eta$ with $\eta = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$.

The foregoing calculation of the variance depends on the covariance structure $\mathbf{V}_i$ used for fitting. In the calculations we implicitly assumed that $\mathbf{V}_i$ was specified correctly. Even though this appears to be a conceptional restriction, we demonstrate in simulations that correction (8) actually is rather robust against misspecified covariances. This means that even if $\mathbf{V}_i$ is misspecified, the corrected profiles $z_{\tilde{p}}$ show a positive effect.

Example 6 (Multivariate Normal Response). Let $Y_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I})$ with $\mathbf{X}_i = (\mathbf{1}_m, \mathbf{U}_i)$, where $\mathbf{1}_m$ is the $m \times 1$-dimensional unit vector and $\mathbf{U}_i$ is an $m \times 1$ covariate vector. We set $\boldsymbol{\beta} = (.5, .5)^{\mathrm{T}}$ and consider $\beta_1 = (0, 1)\boldsymbol{\beta}$ to be the parameter of interest. We simulate from the following designs for the covariates: Let $\mathbf{U}_i = \mathbf{1}_m u_i$ with scalar $u_i \in \Re$ chosen (a) uniformly, (b) normally, or (c) from a Laplace distribution. Inserting $\mathrm{var}(\mathbf{V}_{\mathrm{sand},u}) = 2\sigma^4\mathrm{tr}(\mathbf{MM})$ in (8) shows that σ^4 cancels out as before, so that the correction depends only on the design and the working covariance $\mathbf{V}_i$. We assume working independence (i.e., $\mathbf{V}_i = \mathbf{I}$) and simulate Y_i from three different settings: with correctly specified working covariance matrix, that is, $\mathrm{var}(Y_i) = \sigma^2\mathbf{I}$ (model 1); with misspecified working covariances, that is, $\mathrm{var}(Y_i) = \sigma^2(3/4\mathbf{I} + 1/4\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}})$ (model 2), and with autocorrelated errors $\mathrm{var}(Y_i)_{rs} = \sigma^2\rho^{|r-s|}$ with $\rho = .5$ (model 3). The corrected quantiles are listed in Table 3. Figure 3 shows simulated coverage probabilities for 2000 simulations for the $p = .9$ confidence interval. For comparison, we again report the coverage probabilities for t-distribution quantiles with $n - 2$ degrees of freedom and for the multivariate jackknife estimate. The proposed adjustment shows satisfactory behavior for all three designs. The misspecification of the covariance has only a small effect on the coverage probability, so the adjustment appears to work for misspecified models as well. In contrast, both $t_{p,n-2}$ distribution quantiles and jackknife estimates show undercoverage, although the jackknife approach behaves more accurately.

For nonnormal data, $\mathrm{var}(\mathbf{V}_{\mathrm{sand},u})$ depends not only on the design and the working covariance, but also on the unknown parameter $\boldsymbol{\beta}$. This implies that in practice matrix $\mathbf{M}$ must to be estimated by plug-in estimates. Moreover, the latter term in (15) does not vanish, and the kurtosis must be estimated. Even though at first glance this appears cumbersome, estimation is usually not too complicated when assuming an underlying probability model. We demonstrate this using a binomial model.

Table 3. Corrected Quantiles $z_{\tilde{p}}$ for Different Designs

	$p = .90$		$p = .95$	
Design	$t_{p,n-2}$	$z_{\tilde{p}}$	$t_{p,n-2}$	$z_{\tilde{p}}$
	$n = 10$ ($m = 4$)		$n = 20$ ($m = 4$)	
(a)		2.03		1.81
(b)	1.86	2.10	1.71	1.86
(c)		2.18		1.94

Example 7 (Logistic Regression). We simulate (independent) binomial data with predictor $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (0, .5)^T$ (model 1) and $\boldsymbol{\beta} = (1, 1)^{\mathrm{T}}$ (model 2). The covariates $\mathbf{X}_i$ are distributed as in Example 6, and we are interested in the slope parameter β_1. For comparison, we again compare our proposed correction with the jackknife estimate, which in this case is a weighted and multivariate version of (9). The results are given in Table 4. The general positive appearance of the corrected quantiles carries over to binomial data, even if the distribution is rather skew, as in model 2. A similar behavior was also observed for simulations with Poisson data, not reported here.

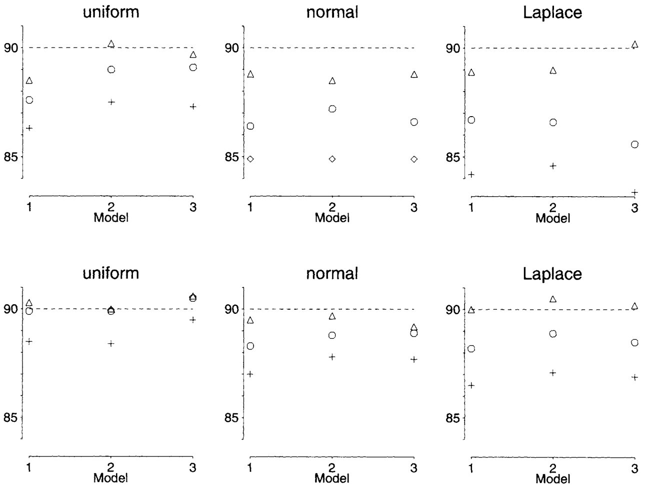


Figure 3. Coverage Probability of Confidence Intervals Based on $\mathbf{V}_{\mathrm{sand},u}$ With Corrected Quantiles $z_{\tilde{p}}$ (△) as Well as With t-Distribution Quantiles $t_{p,n-1}P(+)$ and Based on the Jackknife Estimate $\mathbf{V}_{\mathrm{jack}}$ (○). The upper Row is for $n = 10$, $m = 4$; the bottom row is for $n = 20$, $m = 4$.

Table 4. Coverage Probability of Confidence Based on $\mathbf{V}_{\text{sand},u}$ With $z_{\hat{p}}$ Calculated With True and Fitted Parameters and t-Distribution Quantiles $t_{p,n-1}$

			Coverage based on		
Design	$t_{p,n-2}$	$z_{\hat{p}}$	$\mathbf{V}_{\text{sand},u}$ $z_{\tilde{p}}$	$\mathbf{V}_{\text{sand},u}$ $t_{p,n-2}$	$\mathbf{V}_{\text{jack}}$ $t_{p,n-2}$
Logistic regression $n=30$ ($m=4$), $p=.9$					
(a)		1.74 (1.74)	89.9 (90.6)	87.3 (87.7)	89.8 (90.4)
(b)	1.70	1.77 (1.78)	89.5 (90.1)	85.3 (84.6)	88.5 (89.0)
(c)		1.82 (1.83)	91.1 (91.8)	85.6 (85.1)	89.6 (90.5)
Logistic regression $n=30$ ($m=4$), $p=.95$					
(a)		2.08 (2.11)	95.4 (95.5)	93.4 (92.0)	95.4 (95.3)
(b)	2.04	2.12 (2.16)	95.4 (95.6)	92.1 (91.1)	94.9 (95.1)
(c)		2.19 (2.22)	95.8 (96.0)	91.7 (89.6)	95.2 (94.7)

3.4 Balanced Design

Finally, we revisit the design issue. So far we have focused on undercoverage properties with sandwich estimates. This undercoverage basically occurs if the covariates differ between the units, as in the foregoing simulations. In contrast, as we shown later, if the covariate design is the same for all units, then undercoverage may not occur.

Example 8 (Balance Design). Consider again the multivariate normal model $Y_i \sim N(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, with $\mathbf{X}_i^{\mathrm{T}}$ as a $m \times p$ design matrix. We assume that the covariates are scaled and orthogonal such that $\boldsymbol{\Omega} = \sum_i \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = n\mathbf{I}$. This gives $\sum_i \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i = n$, and the variance is obtained from

$$\begin{aligned}\operatorname{var}\{\mathbf{V}_{\text{sand},u}\} &= 2\sigma^4\operatorname{tr}(\mathbf{MM}) = 2\operatorname{tr}(\mathbf{WW})\{1+O(n^{-1})\}\\ &= 2n^{-4}\sigma^4\sum_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i)^2\{1+O(n^{-1})\}\\ &\geq 2n^{-5}\sigma^4\Big(\sum_i \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i\Big)^2\{1+O(n^{-1})\}\\ &= 2n^{-3}\sigma^4\{1+O(n^{-1})\}.\end{aligned}$$

The lower bound is reached if the covariates are individually orthogonal or balanced in the sense $\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} = \mathbf{I}$ for all i. This is the case if, for instance, the individual design $\mathbf{X}_i$ does not differ among the individuals. A typical example is given by longitudinal data, where the covariates give the timepoint of measurement, that is, $\mathbf{X}_i = (\mathbf{1}, \mathbf{t})$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$ and $\mathbf{t} = (-T, -T+1, \ldots, T-1, T)^{\mathrm{T}}/(\sum_{t=-T}^{T} t^2)$ is a centered and standardized time vector. In this case one gets the lower bound $\operatorname{var}(\mathbf{V}_{\text{sand},u}) = 2\sigma^4/\{n^2(n-1)\}\{1+O(n^{-1})\}$, which equals the variance of the classical variance estimate discussed in Example 2. Hence one finds that in general $z_{\tilde{p}} \geq t_{p,n-1}$ holds asymptotically, where the lower bound is reached if the design is individually balanced. As a consequence undercoverage is not an issue in this case.

4. DISCUSSION

We have shown that sandwich variance estimates are typically less efficient than model-based variance estimates. The loss of efficiency depends mainly on the design; for standard cases, it is proportional to the inverse of the design kurtosis of the design points, and for nonnormal data, additional components beside the kurtosis influence the loss of efficiency. The variance of the sandwich variance estimate directly affects the coverage probability of confidence intervals. An adjustment has been suggested that depends on the design. The adjustment has shown promising behavior, although we expect it to be possible to break down the method.

Basically, the use of the sandwich variance estimate leads to undercoverage of confidence intervals if the covariates differ between the units. For individually balanced designs, as may occur in dynamic data, undercoverage does not occur. Therefore, we can refine the statement of Diggle et al. (1994, p. 77) that the sandwich variance estimate should be used with care if the data come from a small number of "experimental units" and the covariates differ between the units. In this case, the suggested corrected quantiles provide a small-sample adjustment for the confidence intervals.

APPENDIX: TECHNICAL DETAILS

A.1 Proof of Theorem 2

In general, the result can be proved by applying an Edgeworth series to $\hat{\theta} - \theta$ (see, e.g., Hall 1992, pp. 46–68). But we pursue a more direct proof here, which makes the result accessible for readers not too familiar with Edgeworth series.

Let $n^{1/2}(\hat{\theta} - \theta) \sim \text{normal}(0, \sigma^2)$ and $z_p = \Phi^{-1}(p)$, where $\Phi(\cdot)$ is the standard normal distribution function. We define $v_p = \sigma z_p$ and $\hat{v}_p = \hat{\sigma} z_p$ such that $F(v_p) = \Pr\{n^{1/2}(\hat{\theta} - \theta) \leq v_p\} = p$ with $F(v_p) = \Phi(z_p)$. The intention is to calculate $\Pr\{n^{1/2}(\hat{\theta} - \theta) \leq \hat{v}_p\}$. Let $H_{\hat{v}_p}(\cdot)$ denote the distribution function of $\hat{v}_p$, and take $\hat{\sigma}^2$ as the $\sqrt{n}$-consistent variance estimate independent of $\hat{\theta} - \theta$. This gives

$$\begin{aligned}\Pr\{(\hat{\theta} - \theta) \leq \hat{v}_p\} &= \int \Pr\{(\hat{\theta} - \theta) \leq (v|\hat{v}_p = v) dH_{\hat{v}_p}(v)\\ &= \int F(v) dH_{\hat{v}_p}(v) = E\{F(\hat{v}_p)\}.\end{aligned}$$

Hence we have to calculate the expectation of $F(\hat{v}_p)$ to obtain the coverage probability. Applying the delta method to the root function $g(v) = v^{1/2}$, we find that

$$\begin{aligned}\hat{\sigma} - \sigma &= g(\hat{\sigma}^2) - g(\sigma^2)\\ &= \frac{\hat{\sigma}^2 - \sigma^2}{2\sigma} - \frac{(\hat{\sigma}^2 - \sigma^2)^2}{8\sigma^3} + O_p(n^{-3/2}).\end{aligned}$$

This along with $\hat{v}_p = v_p + z_p(\hat{\sigma} - \sigma)$ implies that

$$\begin{aligned}F(\hat{v}_p) &= F\left\{v_p + z_p\frac{\hat{\sigma}^2 - \sigma^2}{2\sigma^2} - z_p\frac{(\hat{\sigma}^2 - \sigma^2)^2}{8\sigma^4}\right\} + O_p(n^{-3/2})\\ &= F(v_p) + F^{(1)}(v_p)\left\{z_p\frac{\hat{\sigma}^2 - \sigma^2}{2\sigma^2} - z_p\frac{(\hat{\sigma}^2 - \sigma^2)^2}{8\sigma^4}\right\}\\ &\quad + \frac{1}{2}F^{(2)}(v_p)\left\{z_p\frac{\hat{\sigma}^2 - \sigma^2}{2\sigma^2}\right\}^2 + O_p(n^{-3/2}).\end{aligned}$$

Because $F(v_p) = p$, this yields

$$\begin{aligned}E\{F(\hat{v}_p)\} &= p + \operatorname{var}(\hat{\sigma}^2)\left\{\frac{z_p^2F^{(2)}(z_p)}{8\sigma^4} - \frac{z_pF^{(1)}(z_p)}{8\sigma^4}\right\}\\ &\quad + O(n^{-3/2}).\end{aligned}$$

Inserting the derivatives for $F(v) = \Phi(v/\sigma)$ gives formula (7) in Theorem 2.

Note that $\tilde{p} - p = O(n^{-1})$, so that by simple Taylor expansion,

$$\begin{aligned} z_p &= \Phi^{-1}(p) = \Phi^{-1}(\tilde{p} + p - \tilde{p}) \\ &= z_{\tilde{p}} + (p - \tilde{p})/\phi(z_{\tilde{p}}) + O(n^{-2}), \end{aligned}$$

which provides $z_p - z_{\tilde{p}} = O(n^{-1})$. Let $\hat{\sigma}_t^2$ be the variance estimate such that $(\hat{\theta} - \theta)/\hat{\sigma}_t$ is t-distributed with the usual degrees of freedom. Denoting by t_p the t-distribution quantiles, calculations similar to the foregoing show that $t_p = z_p + (p - p_t)/\phi(t_p) + O(n^{-2})$, where p_t is defined through $z_{p_t} = t_p$. Using these equations and applying (8) provides

$$\begin{aligned} z_{\tilde{p}} &= t_p + \frac{z_{\tilde{p}}^3 + z_{\tilde{p}}}{8\sigma^4} \operatorname{var}(\hat{\sigma}^2) - \frac{t_p^3 + t_p}{8\sigma^4} \operatorname{var}(\hat{\sigma}_t^2) + O(n^{-2}) \\ &= t_p + \frac{z_p^3 - z_p}{8} \frac{\operatorname{var}(\hat{\sigma}_t^2)}{\sigma^4} \left(\frac{\operatorname{var}(\hat{\sigma}^2)}{\operatorname{var}(\hat{\sigma}_t^2)} - 1 \right) + O(n^{-2}), \end{aligned}$$

as claimed in Section 2.2.

A.2 Sandwich Estimates in Quasi-Likelihood and Generalized Estimating Equations

Here we derive the relative efficiency in quasi-likelihood models. For simplicity of notation, we consider univariate regression models of the form $E(Y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = h(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ with $\mathbf{x}_i^{\mathrm{T}}$ as a $1 \times p$ vector. The variance of Y_i is given by $\operatorname{var}(Y_i|\mathbf{x}_i) = \sigma^2 V\{\mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$, where $V(\cdot)$ is a known variance function. In some problems σ^2 is estimated, which we indicate by setting $\xi = 1$, whereas when σ^2 is known, we set $\xi = 0$. We denote the derivatives of functions by superscripts, for example, $\mu^{(l)}(\eta) = \partial^l \mu(\eta)/(\partial\eta)^l$. Let us assume that the variance is correctly specified, that is, $\operatorname{var}(Y_i|\mathbf{x}_i) = \sigma^2 V\{\mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}$, so that with expansion (12) we get $\operatorname{var}(n^{1/2}\hat{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\beta}) = \mathbf{V}_{\text{asymp}}\{1 + O(n^{-1})\}$, where $\mathbf{V}_{\text{asymp}} = \sigma^2 \mathbf{z}^{\mathrm{T}} \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$ and $\boldsymbol{\Omega}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} Q(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ with $Q(\eta) = \{\mu^{(1)}(\eta)\}^2/V(\eta)$. The model-based variance estimator for $n^{1/2}\hat{\mathbf{z}}^{\mathrm{T}}\boldsymbol{\beta}$ is $\mathbf{V}_{\text{model}} = \hat{\sigma}^2(\boldsymbol{\beta})\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\hat{\beta})\mathbf{z}$, where

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \xi n^{-1} \sum_{i=1}^n \{Y_i - \mu(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\}^2 / V(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + \sigma^2(1 - \xi).$$

Defining $\mathbf{B}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} M(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2$ and $M(\eta) = \{\mu^{(1)}(\eta)/V(\eta)\}^2$, the sandwich estimator for $n^{1/2}\mathbf{z}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is written as $V_{\text{sand}} = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}})\mathbf{B}_n(\hat{\boldsymbol{\beta}})\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}})\mathbf{z}$.

To derive the following theorem, we need some additional notation. Let $\mathbf{R}_n = \xi n^{-1}\sum_{i=1}^n g(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i^{\mathrm{T}}$, where $g(\eta) = (\partial/\partial\eta)\log\{V(\eta)\}$, $\epsilon_i = \{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}/V^{1/2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, $q_{in} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a_n = \mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\mathbf{C}_n = n^{-1}\sum_{i=1}^n q_{in}^2 Q^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i$,

$$\begin{aligned} \boldsymbol{\ell}_{in} &= \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{x}_i\mu^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/V^{1/2}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \\ v_i &= \{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2 M(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sigma^2 Q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \end{aligned}$$

and

$$\mathbf{K}_n = n^{-1}\sum_{i=1}^n q_{in}^2 V(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}_i.$$

In what follows, we treat $\mathbf{x}_i$ as a sample from a distribution. We assume that sufficient moments of the components of $\mathbf{x}$ and y exist, as does sufficient smoothness of $\mu(\cdot)$. Under the foregoing conditions, at least asymptotically there are no leverage points, so that the usual and unbiased sandwich estimators will have similar asymptotic behavior. We write $\overline{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = E\{\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}$, $q = \mathbf{x}^{\mathrm{T}}\overline{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $a = \mathbf{z}^{\mathrm{T}}\overline{\boldsymbol{\Omega}}^{-1}(\boldsymbol{\beta})\mathbf{z}$, $\overline{\mathbf{C}} = E\{q^2 Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}\}$, and so on—that is, the bar notation refers to asymptotic moments.

Theorem A.1. As $n \to \infty$, under the foregoing conditions we have

$$\begin{aligned} &n^{1/2}(\mathbf{V}_{\text{model}} - \mathbf{V}_{\text{asymp}}) \\ &\quad \Rightarrow \text{normal}\left[0, \Sigma_{\text{model}} := E\{a\xi(\epsilon^2 - \sigma^2) - \sigma^2(a\overline{\mathbf{R}} + \overline{\mathbf{C}})^{\mathrm{T}}\boldsymbol{\ell}\epsilon\}^2\right] \end{aligned}$$

and

$$\begin{aligned} &n^{1/2}(\mathbf{V}_{\text{sand}} - \mathbf{V}_{\text{asymp}}) \\ &\quad \Rightarrow \text{normal}\left[0, \Sigma_{\text{sand}} := E\{q^2 v + (\overline{\mathbf{K}} - 2\sigma^2\overline{\mathbf{C}})^{\mathrm{T}}\boldsymbol{\ell}\epsilon\}^2\right]. \end{aligned}$$

For the proof, recall that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx n^{-1/2}\sum_{i=1}^n \boldsymbol{\ell}_{in}\epsilon_i$, where $\approx$ means that the difference is of order $o_p(1)$. By a simple delta-method calculation we get $\xi n^{1/2}\{\widehat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\} \approx n^{-1/2}\sum_{i=1}^n \xi(\epsilon_i^2 - \sigma^2) - \sigma^2\mathbf{R}_n^{\mathrm{T}} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Thus

$$\begin{aligned} &n^{1/2}\{\mathbf{V}_{\text{model}} - \mathbf{V}_{\text{asymp}}\} \\ &\quad \approx \xi n^{1/2}\{\hat{\sigma}^2(\hat{\beta}) - \sigma^2\}a_n + n^{1/2}\sigma^2\mathbf{z}^{\mathrm{T}}\{\boldsymbol{\Omega}_n^{-1}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\}\mathbf{z} \\ &\quad \approx \xi n^{1/2}\{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\}a_n - \sigma^2 n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\{\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}}) \\ &\qquad - \boldsymbol{\Omega}_n(\boldsymbol{\beta})\}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z} \\ &\quad \approx \xi n^{1/2}\{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}) - \sigma^2\}a_n - \sigma^2\mathbf{C}_n^{\mathrm{T}} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \approx n^{-1/2}\sum_{i=1}^n \{a_n\xi(\epsilon_i^2 - \sigma^2) - \sigma^2(a_n\mathbf{R}_n + \mathbf{C}_n)^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i\}, \end{aligned}$$

which shows the first part of Theorem A.1.

We now turn to the sandwich estimator and note that $\mathbf{B}_n(\boldsymbol{\beta}) - \sigma^2\boldsymbol{\Omega}_n(\boldsymbol{\beta}) = O_p(n^{-1/2})$. Because of this, we have that

$$\begin{aligned} &n^{1/2}\{\mathbf{V}_{\text{sand}} - \mathbf{V}_{\text{asymp}}\} \\ &\quad \approx -2\sigma^2 n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\{\boldsymbol{\Omega}_n(\hat{\boldsymbol{\beta}}) - \boldsymbol{\Omega}_n(\boldsymbol{\beta})\}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z} \\ &\qquad + n^{1/2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\{\mathbf{B}_n(\hat{\boldsymbol{\beta}}) - \sigma^2\boldsymbol{\Omega}_n(\boldsymbol{\beta})\}\boldsymbol{\Omega}_n^{-1}(\boldsymbol{\beta})\mathbf{z} \\ &\quad \approx -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2[M(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \\ &\qquad \times \{Y_i - \mu(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})\}^2 - \sigma^2 Q(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})] \\ &\quad \approx -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2 v_i \\ &\qquad + n^{-1}\sum_{i=1}^n q_i^2 M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{X}_i\{Y_i - \mu(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}^2 n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \approx -2\sigma^2 n^{-1/2}\sum_{i=1}^n \mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + n^{-1/2}\sum_{i=1}^n q_{in}^2 v_i \\ &\qquad + n^{-1}\sum_{i=1}^n q_i^2 M^{(1)}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{X}_i V(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad \approx n^{-1/2}\sum_{i=1}^n (-2\sigma^2\mathbf{C}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i + q_i^2 v_i + \mathbf{K}_n^{\mathrm{T}}\boldsymbol{\ell}_{in}\epsilon_i), \end{aligned}$$

as claimed.

Theorem A.1 can now be used to prove the statements listed in Examples 4 and 5. For the logistic case, we have $V(\eta) = \mu^{(1)}(\eta) = Q(\eta) = \mu(\eta)\{1 - \mu(\eta)\}$, $\sigma^2 = 1, \xi = 0, \mathbf{R}_n = 0$, and $Q^{(1)}(\eta) = \mu^{(1)}(\eta)\{1 - 2\mu(\eta)\}$. All of the terms in Theorem A.1 can then be computed by numerical integration, which gives the numbers presented in Example 5.

For the Poisson case, it is easily verified that $\overline{\boldsymbol{\Omega}}(\boldsymbol{\beta}) = \exp(\beta_0)\mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Also, $q = U\exp(-\beta_0)$, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \beta_0$, $Q^{(1)}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) = \exp(\beta_0)$, $\overline{\mathbf{C}} = \exp(-\beta_0)(1, 0)^{\mathrm{T}}$,

$\boldsymbol{\ell} = \exp(-\beta_0/2)(1, U)^{\mathrm{T}}$, $\epsilon = \{Y - \exp(\beta_0)\}/\exp(\beta_0/2)$, and hence $\Sigma_{\mathrm{model}} = \exp(-3\beta_0)$.

Let $\theta = \exp(\beta_0)$. Then $E(Y^2) = \theta + \theta^2$, $E(Y^3) = \theta^3 + 3\theta^2 + \theta$, and $E(Y^4) = \theta^4 + 6\theta^3 + 7\theta^2 + \theta$. If we define $Z = Y - \theta$, then $E(Z) = 0$, $E(Z^2) = E(Z^3) = \theta$, and $E(Z^4) = 3\theta^2 + \theta$. Further, $M(\eta) = 1$, $M^{(1)}(\eta) = 0$, and $\overline{\mathbf{K}} = 0$. A detailed calculation then gives that $\Sigma_{\mathrm{sand}} = 2\kappa \exp(-2\beta_0) + \kappa \exp(-3\beta_0)$, which shows the relative efficiency given in Example 4.

[Received March 2000. Revised April 2001.]

REFERENCES

Breslow, N. (1990), "Test of Hypotheses in Overdispersion Regression and Other Quasi-Likelihood Models." *Journal of the American Statistical Association*, 85, 565–571.

Carroll, R. J., and Ruppert, D. (1998), *Transformation and Weighting in Regression*, London: Chapman and Hall.

Chesher, A., and Jewitt, I. (1987), "The Bias of a Heteroscedasticity Consistent Covariance Matrix Estimator," *Econometrica*, 55, 1217–1222.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Clarendon Press.

Efron, B. (1986), Discussion of "Jackknife, Bootstrap and Other Resampling Methods in Statistics," by C. F. J. Wu, *The Annals of Statistics*, 14, 1301–1304.

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimator for Families of Linear Regression," *Annals of Mathematical Statistics*, 34, 447–456.

Firth, D. (1992), Discussion of "Multivariate Regression Analysis for Categorical Data," by Liang, Zeger, and Qaqish, *Journal of the Royal Statistical Society*, Ser. B, 54, 24–26.

Gourieroux, C., Monfort, A., and Trognon, A. (1984), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 52, 701–720.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Hinkley, D. V. (1977), "Jackknifing in Unbalanced Situations," *Technometrics*, 19, 285–292.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, eds. L. M. LeCam and J. Neyman, Berkeley: University of California Press, pp. 221–233.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate Regression Analysis for Categorical Data," *Journal of the Royal Statistical Society*, Ser. B, 54, 3–40.

Long, J. S., and Ervin, L. H. (2000), "Using Heteroscedasticity-Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217–223.

MacKinnon, J. G., and White, H. (1985), "Some Heteroscedasticity-Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

McCullagh, P. (1987), *Tensor Methods in Statistics*, London: Chapman and Hall.

———. (1992). Discussion of "Multivariate Regression Analysis for Categorical Data," by Liang, Zeger, and Qaqish, *Journal of the Royal Statistical Society*, Ser. B, 54, 24–26.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall.

Rothenberg, T. J. (1988), "Approximative Power Functions for Some Robust Tests of Regression Coefficients," *Econometrica*, 56, 997–1019.

Wedderburn, R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss–Newton Method," *Biometrika*, 61, 439–447.

White, H. (1980), "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 48, 817–838.

Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Statistics," *The Annals of Staticstics*, 14, 1261–1350.

Biostatistics (2004), **5**, 3, *pp.* 445–464
doi: 10.1093/biostatistics/kxh001

Analyzing incomplete longitudinal clinical trial data

GEERT MOLENBERGHS*, HERBERT THIJS, IVY JANSEN, CAROLINE BEUNCKENS
Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium
geert.molenberghs@luc.ac.be

MICHAEL G. KENWARD
London School of Hygiene and Tropical Medicine, London, UK

CRAIG MALLINCKRODT
Eli Lilly and Company, Indianapolis, IN, USA

RAYMOND J. CARROLL
Department of Statistics, Texas A&M University, College Station, TX, USA

SUMMARY

Using standard missing data taxonomy, due to Rubin and co-workers, and simple algebraic derivations, it is argued that some simple but commonly used methods to handle incomplete longitudinal clinical trial data, such as complete case analyses and methods based on last observation carried forward, require restrictive assumptions and stand on a weaker theoretical foundation than likelihood-based methods developed under the missing at random (MAR) framework. Given the availability of flexible software for analyzing longitudinal sequences of unequal length, implementation of likelihood-based MAR analyses is not limited by computational considerations. While such analyses are valid under the comparatively weak assumption of MAR, the possibility of data missing not at random (MNAR) is difficult to rule out. It is argued, however, that MNAR analyses are, themselves, surrounded with problems and therefore, rather than ignoring MNAR analyses altogether or blindly shifting to them, their optimal place is within sensitivity analysis. The concepts developed here are illustrated using data from three clinical trials, where it is shown that the analysis method may have an impact on the conclusions of the study.

Keywords: Complete case analysis; Ignorability; Last observation carried forward; Missing at random; Missing completely at random; Missing not at random.

1. INTRODUCTION

In a longitudinal clinical trial, each unit is measured on several occasions. It is not unusual in practice for some sequences of measurements to terminate early for reasons outside the control of the investigator, and any unit so affected is called a dropout. It might therefore be necessary to accommodate dropout in the modeling process.

*To whom corespondence should be addressed.

Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended study design (Afifi and Elashoff, 1966; Hartley and Hocking, 1971). More recently, general algorithms such as expectation-maximization (EM) (Dempster *et al.*, 1977), and data imputation and augmentation procedures (Rubin, 1987), combined with powerful computing resources have largely solved the computational difficulties. There remains the difficult and important question of assessing the impact of missing data on subsequent statistical inference.

When referring to the missing-value, or non-response, process we will use terminology of Little and Rubin (1987, Chapter 6). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are *ignorable,* while a non-random process is non-ignorable.

Numerous missing data methods are formulated as selection models (Little and Rubin, 1987) as opposed to pattern-mixture modeling (PMM; Little, 1993, 1994). A selection model factors the joint distribution of the measurement and response mechanisms into the marginal measurement distribution and the response distribution, conditional on the measurements. This is intuitively appealing because the marginal measurement distribution would be of interest with complete data. Little and Rubin's taxonomy is most easily developed in the selection model setting. Parametrizing and making inference about treatment effects and their evolution over time is straightforward in the selection model context.

In many clinical trial settings, the standard methodology used to analyze incomplete longitudinal data is based on such methods as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of imputation. This is often done without questioning the possible influence of these assumptions on the final results, even though several authors have written about this topic. A relatively early account is given in Heyting *et al.* (1992). Mallinckrodt *et al.* (2003a,b) and Lavori *et al.* (1995) propose direct-likelihood and multiple-imputation methods, respectively, to deal with incomplete longitudinal data. Siddiqui and Ali (1998) compare direct-likelihood and LOCF methods.

As will be discussed in subsequent sections, it is unfortunate that such a strong emphasis is placed on methods like LOCF and CC in clinical trial settings, since they are based on strong and unrealistic assumptions. Even the strong MCAR assumption does not suffice to guarantee that an LOCF analysis is valid. In contrast, under the less restrictive assumption of MAR, valid inference can be obtained through a likelihood-based analysis without modeling the dropout process. One can then use linear or generalized linear mixed models (Verbeke and Molenberghs, 2000), without additional complication or effort. We will argue that such an analysis is more likely to be valid, and even easier to implement than LOCF and CC analyses.

Nevertheless, approaches based on MNAR need to be considered. In practical settings, the reasons for dropout are varied and it may therefore be difficult to justify the assumption of MAR. For example, in 11 clinical trials of similar design, considered by Mallinckrodt *et al.* (2003b), with the same drug and involving patients with the same disease state, the rate of and the reasons for dropout varied considerably. In one study, completion rates were 80% for drug and placebo. In another study, two-thirds of the patients on drug completed all visits, while only one-third did so on placebo. In yet another study, 70% finished on placebo but only 60% on drug. Reasons for dropout also varied, even within the drug arm. For example, at low doses more patients on drug dropped out due to lack of efficacy whereas at higher doses dropout due to adverse events was more common. At first sight, this calls for a further shift towards MNAR models. However, caution ought to be used since no modeling approach, whether MAR or MNAR, can recover the lack of information due to incompleteness of the data.

Table 1. *Overview of number of patients and post baseline visits per study*

	Number of patients	Post-baseline visits
Study 1	167	4–11
Study 2	342	4–8
Study 3	713	3–8

First, if MAR can be guaranteed to hold, a standard analysis would follow. However, only rarely is such an assumption known to hold (Murray and Findlay, 1988). Nevertheless, ignorable analyses may provide reasonably stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of the unseen data to be similar to that of the observed data (Mallinckrodt *et al.*, 2001a,b). A discussion of this phenomenon in the survey context has been given in Rubin *et al.* (1995). These authors argue that, in rigidly controlled experiments (some surveys and many clinical trials), the assumption of MAR is often reasonable. Second, and very importantly for confirmatory trials, an MAR analysis can be specified *a priori* without additional work relative to a situation with complete data. Third, while MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on untestable and often implicit assumptions regarding the distribution of the unobserved measurements given the observed measurements. The quality of the fit to the observed data need not reflect at all the appropriateness of the implied structure governing the unobserved data. This point is irrespective of the MNAR route taken, whether a parametric model of the type of Diggle and Kenward (1994) is chosen, or a semiparametric approach such as in Robins *et al.* (1998). Hence, in incomplete-data settings, a definitive MNAR analysis does not exist. We therefore argue that clinical trial practice should shift away from the *ad hoc* methods and focus on likelihood-based ignorable analyses instead. The cost involved in having to specify a model will likely be small to moderate in realistic clinical trial settings. To explore the impact of deviations from the MAR assumption on the conclusions, one should ideally conduct a sensitivity analysis, within which MNAR models and pattern-mixture models can play a major role (Verbeke and Molenberghs, 2000, Chapter 18–20).

A three-trial case study is introduced in Section 2. The general data setting is introduced in Section 3, as well as a formal framework for incomplete longitudinal data. A discussion on the problems associated with simple methods is presented in Section 4. In Section 5, using algebraic derivations, we explore the origins of the asymptotic bias in LOCF, complete-case and likelihood-based ignorable analyses. The case study is analyzed in Section 6. A perspective on sensitivity analysis is sketched in Section 7.

2. CASE STUDIES

The ideas developed in this paper are motivated from, and applied to, data from three clinical trials of anti-depressants. The three trials contained 167, 342, and 713 patients with post-baseline data, respectively (Mallinckrodt *et al.*, 2003b). The Hamilton Depression Rating Scale ($HAMD_{17}$) was used to measure the depression status of the patients. For each patient, a baseline assessment was available. Post-baseline visits differ by study (Table 1).

For blinding purposes, therapies are recoded as A1 for primary dose of experimental drug, A2 for secondary dose of experimental drug, and B and C for non-experimental drugs. The treatment arms across the three studies are as follows: A1, B, and C for study 1; A1, A2, B, and C for study 2; A1 and B for study 3. The primary contrast is between A1 and C for studies 1 and 2, whereas in study 3 one is interested in A *versus* B.

In this case study, emphasis is on the difference between the treatment arms in mean change of the

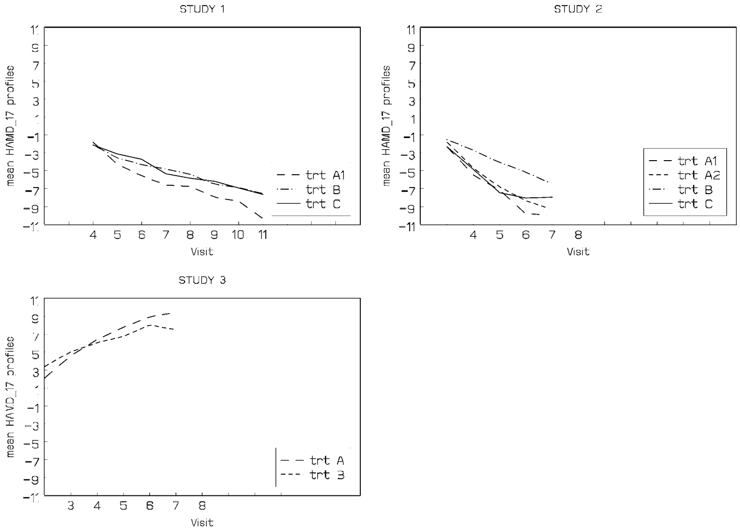


Fig. 1. Mean profiles for each of the three studies.

$HAMD_{17}$ score at the endpoint. For each study, mean profiles within each treatment arm are given in Figure 1. However, as time evolves, more and more patients drop out, resulting in fewer observations for later visits. Indeed, a graphical representation of dropout, per study and per arm, is given in Figure 2. Due to this fact, Figure 1 might be misleading if interpreted without acknowledging the diminishing basis of inference.

3. Data setting and modeling framework

Assume that for subject $i = 1, \ldots, N$ in the study a sequence of responses Y_{ij} is designed to be measured at occasions $j = 1, \ldots, n$. The outcomes are grouped into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$. In addition, define a dropout indicator D_i for the occasion at which dropout occurs and make the convention that $D_i = n + 1$ for a complete sequence. It is often necessary to split the vector $\boldsymbol{Y}_i$ into observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components respectively.

In principle, one would like to consider the density of the full data $f(\boldsymbol{y}_i, d_i|\boldsymbol{\theta}, \boldsymbol{\psi})$, where the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ describe the measurement and missingness processes, respectively. Covariates are assumed to be measured, but have been suppressed from notation for simplicity.

Most strategies used to analyze such data are, implicitly or explicitly, based on two choices.

Model for measurements. A choice has to be made regarding the modeling approach to the measurements. Several views are possible.

View 1. One can choose to analyze the entire longitudinal profile, irrespective of whether interest focuses on the entire profile (e.g. difference in slope between groups) or on a specific time point (e.g.

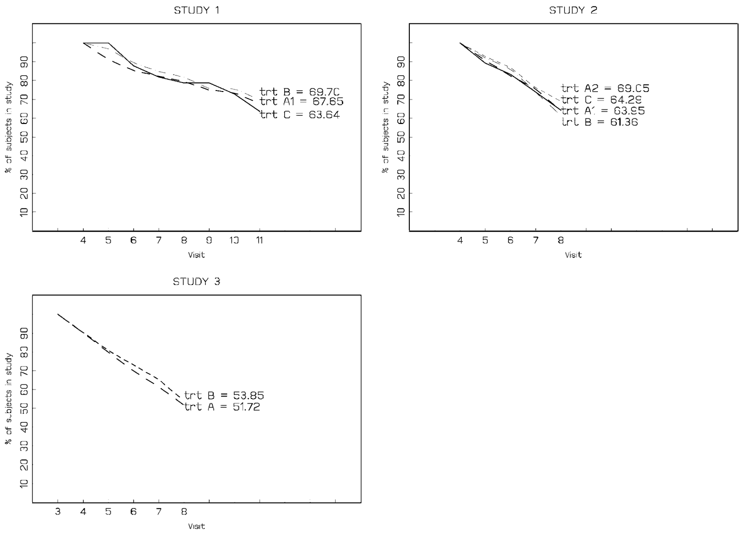


Fig. 2. Evolution of dropout per study and per treatment arm. Treatment arms of primary interest, are shown in bolder typeface.

the last planned occasion). In the latter case, one would make inferences about such an occasion using the posited model.

View 2. One states the scientific question in terms of the outcome at a well-defined point in time. Several choices are possible:

View 2a. The scientific question is defined in terms of the *last planned occasion*. In this case, one can either accept the dropout as it is or use one or other strategy (e.g. imputation) to incorporate the missing outcomes.

View 2b. One can choose to define the question and the corresponding analysis in terms of the *last observed measurement*.

While Views 1 and 2a necessitate reflection on the missing data mechanism, View 2b avoids the missing data problem because the question is couched completely in terms of observed measurements. Thus, under View 2b, an LOCF analysis might be acceptable, provided it matched the scientific goals, but is then better described as a Last Observation analysis because nothing is carried forward. Such an analysis should properly be combined with an analysis of time to dropout, perhaps in a survival analysis framework. Of course, an investigator should reflect very carefully on whether View 2b represents a relevant and meaningful scientific question (see also Shih and Quan, 1997).

Method for handling missingness. A choice has to be made regarding the modeling approach for the missingness process. Under certain assumptions this process can be ignored (e.g. a likelihood-based ignorable analysis). Some simple methods, such as a complete case analysis and LOCF, do not explicitly address the missingness process either.

We first describe the measurement and missingness models in turn, then formally introduce and comment on ignorability.

The measurement model will depend on whether or not a full longitudinal analysis is done. When the focus is on the last observed measurement or on the last measurement occasion only, one typically opts for classical two- or multi-group comparisons (t test, Wilcoxon, etc.). When a longitudinal analysis is deemed necessary, the choice depends on the nature of the outcome. For continuous outcomes, such as in our case studies, one typically assumes a linear mixed-effects model, perhaps with serial correlation:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{W}_i + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

(Verbeke and Molenberghs, 2000) where $\boldsymbol{Y}_i$ is the n-dimensional response vector for subject i, $1 \leqslant i \leqslant N$,$N$ is the number of subjects, X_i and Z_i are $(n \times p)$ and $(n \times q)$ known design matrices, $\boldsymbol{\beta}$ is the p-dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\mathbf{0}, D)$ is the q-dimensional vector containing the random effects, $\varepsilon_i \sim N(\mathbf{0}, \sigma^2 I_{n_i})$ is a n-dimensional vector of measurement error components, and $\boldsymbol{b}_1, \dots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \dots, \varepsilon_N$ are assumed to be independent. Serial correlation is captured by the realization of a Gaussian stochastic process, $\boldsymbol{W}_i$, which is assumed to follow a $N(\mathbf{0}, \tau^2 H_i)$ law. The serial covariance matrix H_i only depends on i through the number n of observations and through the time points t_{ij} at which measurements are taken. The structure of the matrix H_i is determined through the autocorrelation function $\rho(t_{ij} - t_{ik})$. This function decreases such that $\rho(0) = 1$ and $\rho(u) \to 0$ as $u \to \infty$. Finally, D is a general $(q \times q)$ covariance matrix with (i, j) element $d_{ij} = d_{ji}$. Inference is based on the marginal distribution of the response $\boldsymbol{Y}_i$ which, after integrating over random effects, can be expressed as

$$\boldsymbol{Y}_i \sim N(X_i\boldsymbol{\beta}, Z_i D Z_i' + \Sigma_i). \tag{3.2}$$

Here, $\Sigma_i = \sigma^2 I_{n_i} + \tau^2 H_i$ is a $(n \times n)$ covariance matrix combining the measurement error and serial components.

Assume that incompleteness is due to dropout only, and that the first measurement Y_{i1} is obtained for everyone. A possible model for the dropout process is a logistic regression for the probability of dropout at occasion j, given that the subject is still in the study. We denote this probability by $g(\boldsymbol{h}_{ij}, y_{ij})$ in which $\boldsymbol{h}_{ij}$ is a vector containing all responses observed up to but not including occasion j, as well as relevant covariates. We then assume that $g(\boldsymbol{h}_{ij}, y_{ij})$ satisfies

$$\operatorname{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = \operatorname{logit}\left[\operatorname{pr}(D_i = j | D_i \geqslant j, \boldsymbol{y}_i)\right] = \boldsymbol{h}_{ij}\boldsymbol{\psi} + \omega y_{ij}, \qquad i = 1, \dots, N, \tag{3.3}$$

(Diggle and Kenward, 1994). When ω equals zero, the dropout model is MAR, and all parameters can be estimated using standard software since the measurement model, for which we use a linear mixed model, and the dropout model, assumed to follow a logistic regression, can then be fitted separately. If $\omega \neq 0$, the posited dropout process is MNAR. Model (3.3) provides the building blocks for the dropout process $f(d_i | \boldsymbol{y}_i, \boldsymbol{\psi})$.

Rubin (1976) and Little and Rubin (1987) have shown that, under MAR and the condition that parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are functionally independent, likelihood-based inference remains valid when the missing data mechanism is ignored (see also Verbeke and Molenberghs, 2000). Practically speaking, the likelihood of interest is then based upon the factor $f(\boldsymbol{y}_i^o | \boldsymbol{\theta})$. This is called *ignorability*. The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects, manipulates the correct likelihood, providing valid parameter estimates and likelihood ratio values. Note that the estimands are the parameters of (3.2), which is a model for complete data, corresponding to what one would expect to see in the absence of dropouts.

A few cautionary remarks are warranted. First, when at least part of the scientific interest is directed towards the nonresponse process, obviously both processes need to be considered. Under MAR, both processes can be modeled and parameters estimated separately. Second, likelihood inference is often surrounded with references to the sampling distribution (e.g. to construct measures of precision for estimators and for statistical hypothesis tests; Kenward and Molenberghs, 1998). However, the practical implication is that standard errors and associated tests, when based on the observed rather than the expected information matrix and given that the parametric assumptions are correct, are valid. Thirdly, it may be hard to rule out the operation of an MNAR mechanism. This point was brought up in the introduction and will be discussed further in Section 7. Fourthly, such an analysis can proceed only under View 1, i.e. a full longitudinal analysis is necessary, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion. In the latter case, the fitted model can be used as the basis for inference at the last occasion. A common criticism is that a model needs to be considered, with the risk of model misspecification. However, it should be noted that in many clinical trial settings the repeated measures are balanced in the sense that a common (and often limited) set of measurement times is considered for all subjects, allowing the a priori specification of a saturated model (e.g. full group by time interaction model for the fixed effects and unstructured variance–covariance matrix). Such an ignorable linear mixed model specification is termed MMRM (mixed-model random missingness) by Mallinckrodt *et al.* (2001a,b). Thus, MMRM is a particular form of a linear mixed model, fitting within the ignorable likelihood paradigm. Such an approach is a promising alternative to the often used simple methods such as complete-case analysis or LOCF. These will be described in the next section and further studied in subsequent sections.

4. Simple methods

We will briefly review a number of relatively simple methods that still are commonly used. For the validity of many of these methods, MCAR is required. For others, such as LOCF, MCAR is necessary but not sufficient. The focus will be on the complete case method, for which data are removed, and on imputation strategies, where data are filled in. Regarding imputation, one distinguishes between single and multiple imputation. In the first case, a single value is substituted for every ‘hole’ in the data set and the resulting data set is analyzed as if it represented the true complete data. Multiple imputation acknowledges the uncertainty stemming from filling in missing values rather than observing them (Rubin, 1987; Schafer, 1997). LOCF will be discussed within the context of imputation strategies, although LOCF can be placed in other frameworks as well.

A *complete case analysis* includes only those cases for which all measurements were recorded. This method has obvious advantages. It is simple to describe and almost any software can be used since there are no missing data. Unfortunately, the method suffers from severe drawbacks. Firstly, there is nearly always a substantial loss of information. For example, suppose there are 20 measurements, with 10% of missing data on each measurement. Suppose, further, that missingness on the different measurements is independent; then, the estimated percentage of incomplete observations is as high as 87%. The impact on precision and power may be dramatic. Even though the reduction of the number of complete cases will be less severe in settings where the missingness indicators are correlated, this loss of information will usually militate against a CC analysis. Secondly, severe bias can result when the missingness mechanism is MAR but not MCAR. Indeed, should an estimator be consistent in the complete data problem, then the derived complete case analysis is consistent only if the missingness process is MCAR. A CC analysis can be conducted when Views 1 and 2 of Section 3 are adopted. It obviously is not a reasonable choice with View 2b.

An alternative way to obtain a data set on which complete data methods can be used is to fill in rather than delete (Little and Rubin, 1987). Concern has been raised regarding imputation strategies. Dempster and Rubin (1983) write: 'The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.' For example, Little and Rubin (1987) show that the application of imputation could be considered acceptable in a linear model with one fixed effect and one error term, but that it is generally not acceptable for hierarchical models, split-plot designs, repeated measures with a complicated error structure, random-effects, and mixed-effects models.

Thus, the user of imputation strategies faces several dangers. First, the imputation model could be wrong and, hence, the point estimates biased. Second, even for a correct imputation model, the uncertainty resulting from missingness is ignored. Indeed, even when one is reasonably sure about the mean value the unknown observation *would have had*, the actual stochastic realization, depending on both the mean and error structures, is still unknown. In addition, most methods require the MCAR assumption to hold while some even require additional and often unrealistically strong assumptions.

A method that has received considerable attention (Siddiqui and Ali, 1998; Mallinckrodt *et al.*, 2003a,b) is *last observation carried forward* (LOCF). In the LOCF method, whenever a value is missing, the last observed value is substituted. The technique can be applied to both monotone and nonmonotonic missing data. It is typically applied in settings where incompleteness is due to attrition.

LOCF can, but should not necessarily, be regarded as an imputation strategy, depending on which of the views of Section 3 is taken. The choice of viewpoint has a number of consequences. First, when the problem is approached from a missing data standpoint, one has to think it plausible that subjects' measurements do not change from the moment of dropout onwards (or during the period they are unobserved in the case of intermittent missingness). In a clinical trial setting, one might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Secondly, LOCF shares with other single imputation methods that it artificially increases the amount of information in the data, by treating imputed and actually observed values on an equal footing. This is especially true if a longitudinal view is taken. Verbeke and Molenberghs (1997, Chapter 5) have shown that all features of a linear mixed model (group difference, evolution over time, variance structure, correlation structure, random effects structure, ...) can be affected. A similar conclusion, based on the case study, is reached in Section 6.

Thus, scientific questions with which LOCF is compatible will be those that are phrased in terms of the last obtained measurement (View 2b). Whether or not such questions are sensible should be the subject of scientific debate, which is quite different from a *post hoc* rationale behind the use of LOCF. Likewise, it can be of interest to model the complete cases separately and to make inferences about them. In such cases, a CC analysis is of course the only reasonable way forward. This is fundamentally different from treating a CC analysis as one that can answer questions about the randomized population as a whole.

We will briefly describe two other imputation methods. The idea behind *unconditional mean imputation* (Little and Rubin, 1987) is to replace a missing value with the average of the observed values on the same variable over the other subjects. Thus, the term *unconditional* refers to the fact that one does not use (i.e. condition on) information on the subject for which an imputation is generated. Since values are imputed that are unrelated to a subject's other measurements, all aspects of a model, such as a linear mixed model, are typically distorted (Verbeke and Molenberghs, 1997). In this sense, unconditional mean imputation can be as damaging as LOCF.

Buck's method or *conditional mean imputation* (Buck, 1960; Little and Rubin, 1987) is similar in complexity to mean imputation. Consider, for example, a single multivariate normal sample. The first step is to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ from the complete cases, assuming that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \Sigma)$. For a subject with missing components, the regression of the missing components ($\boldsymbol{Y}_i^m$)

on the observed ones ($\boldsymbol{y}_i^o$) is

$$\boldsymbol{Y}_i^m|\boldsymbol{y}_i^o \sim N(\boldsymbol{\mu}^m + \Sigma^{mo}(\Sigma^{oo})^{-1}(\boldsymbol{y}_i^o - \boldsymbol{\mu}_i^o), \Sigma^{mm} - \Sigma^{mo}(\Sigma^{oo})^{-1}\Sigma^{om}). \tag{4.1}$$

The second step calculates the conditional mean from the regression of the missing components on the observed components, and substitutes the conditional mean for the corresponding missing values. In this way, 'vertical' information (estimates for $\boldsymbol{\mu}$ and Σ) is combined with 'horizontal' information ($\boldsymbol{y}_i^o$). Buck (1960) showed that under mild conditions, the method is valid under MCAR mechanisms. Little and Rubin (1987) added that the method is also valid under certain MAR mechanisms. Even though the distribution of the observed components is allowed to differ between complete and incomplete observations, it is very important that the regression of the missing components on the observed ones is constant across missingness patterns. Again, this method shares with other single imputation strategies that, although point estimation may be consistent, the precision will be overestimated. There is a connection between *the concept* of conditional mean imputation and a likelihood-based ignorable analysis, in the sense that the latter analysis produces expectations for the missing observations that are formally equal to those obtained under a conditional mean imputation. However, in likelihood-based ignorable analyses, no explicit imputation takes place, hence the amount of information in the data is not overestimated and important model elements, such as mean structure and variance components, are not distorted.

Historically, an important motivation behind the simpler methods was their simplicity. Currently, with the availability of commercial software tools such as, for example, the SAS procedures MIXED and NLMIXED and the SPlus and R nlme libraries, this motivation no longer applies. Arguably, a MAR analysis is the preferred choice. Of course, the correctness of a MAR analysis rests upon the truth of the MAR assumption, which is, in turn, never completely verifiable. Purely resorting to MNAR analyses is not satisfactory either since important sensitivity issues then arise. These and related issues are briefly discussed in the next section (see also Verbeke and Molenberghs, 2000).

It is often quoted that LOCF or CC, while problematic for parameter estimation, produce randomization-valid hypothesis testing, but this is questionable. First, in a CC analysis partially observed data are selected out, with probabilities that that may depend on post-randomization outcomes, thereby undermining any randomization justification. Second, if the focus is on one particular time point, e.g. the last one scheduled, then LOCF plugs in data. Such imputations, apart from artificially inflating the information content, may deviate in complicated ways from the underlying data (see next section). In contrast, a likelihood-based MAR analysis uses all available data, with the need for neither deletion nor imputation, which suggests that a likelihood-based MAR analysis would usually be the preferred one for testing as well. Third, although the size of a randomization based LOCF test may reach its nominal size under the null hypothesis of no difference in treatment profiles, there will be other regions of the alternative space where the power of the LOCF test procedure is equal to its size, which is completely unacceptable.

5. Bias in LOCF, CC, and ignorable likelihood methods

Using the simple but insightful setting of two repeated follow-up measures, the first of which is always observed while the second can be missing, we establish some properties of the LOCF and CC estimation procedures under different missing data mechanisms, against the background of a MAR process operating. In this way, we bring LOCF and CC within a general framework that makes clear their relationships with more formal modeling approaches, enabling us to make a coherent comparison among the different approaches. The use of a moderate amount of algebra leads to some interesting conclusions.

Let us assume each subject i is to be measured on two occasions $t_i = 0, 1$. Subjects are randomized to one of two treatment arms: $T_i = 0$ for the standard arm and 1 for the experimental arm. The probability of an observation being observed on the second occasion ($D_i = 2$) is p_0 and p_1 for treatment groups 0 and 1, respectively. We can write the means of the observations in the two dropout groups as follows:

$$\text{dropouts } D_i = 1 : \beta_0 + \beta_1 T_i + \beta_2 t_i + \beta_3 T_i t_i, \tag{5.1}$$

$$\text{completers } D_i = 2 : \gamma_0 + \gamma_1 T_i + \gamma_2 t_i + \gamma_3 T_i t_i. \tag{5.2}$$

The true underlying population treatment difference at time $t_i = 1$, as determined from (5.1)–(5.2), is equal to

$$\begin{aligned}\Delta_{\text{true}} = {} & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1 + \beta_2 + \beta_3)\\ & -[p_0(\gamma_0 + \gamma_2) + (1 - p_0)(\beta_0 + \beta_2)].\end{aligned} \tag{5.3}$$

If we use LOCF as the estimation procedure, the expectation of the corresponding estimator equals

$$\begin{aligned}\Delta_{\text{LOCF}} = {} & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1)\\ & -[p_0(\gamma_0 + \gamma_2) + (1 - p_0)\beta_0].\end{aligned} \tag{5.4}$$

Alternatively, if we use CC, the above expression changes to

$$\Delta_{\text{CC}} = \gamma_1 + \gamma_3. \tag{5.5}$$

In general, these are both biased estimators.

We will now consider the special but important cases where the true missing data mechanisms are MCAR and MAR, respectively. Each of these will impose particular constraints on the β and γ parameters in model (5.1)–(5.2). Under MCAR, the β parameters are equal to their γ counterparts and (5.3) simplifies to

$$\Delta_{\text{MCAR,true}} = \beta_1 + \beta_3 \equiv \gamma_1 + \gamma_3. \tag{5.6}$$

Suppose we apply the LOCF procedure in this setting, the expectation of the resulting estimator then simplifies to

$$\Delta_{\text{MCAR,LOCF}} = \beta_1 + (p_1 - p_0)\beta_2 + p_1\beta_3. \tag{5.7}$$

The bias is given by the difference between (5.6) and (5.7):

$$B_{\text{MCAR,LOCF}} = (p_1 - p_0)\beta_2 - (1 - p_1)\beta_3. \tag{5.8}$$

While of a simple form, we can learn several things from this expression by focusing on each of the terms in turn. First, suppose $\beta_3 = 0$ and $\beta_2 \neq 0$, implying that there is no differential treatment effect between the two measurement occasions but there is an overall time trend. Then, the bias can go in either direction depending on the sign of $p_1 - p_0$ and the sign of β_2. Note that $p_1 = p_0$ only in the special case that the dropout rate is the same in both treatment arms. Whether or not this is the case has no impact on the status of the dropout mechanism (it is MCAR in either case, even though in the second case dropout is treatment-arm dependent), but is potentially very important for the bias implied by LOCF. Second, suppose $\beta_3 \neq 0$ and $\beta_2 = 0$. Again, the bias can go in either direction depending on the sign of β_3, i.e. depending on whether the treatment effect at the second occasion is larger or smaller than the treatment effect at the first occasion. In conclusion, even under the strong assumption of MCAR, we see that the bias in the LOCF estimator typically does not vanish and, even more importantly, the bias can be positive or negative and can even induce an apparent treatment effect when one does not exist.

In contrast, as can be seen from (5.5) and (5.6), the CC analysis is unbiased.

Let us now turn to the MAR case. In this setting, the constraint implied by the MAR structure of the dropout mechanism is that the conditional distribution of the second observation given the first is the same in both dropout groups (Molenberghs *et al.*, 1998). Based on this result, the expectation of the second observation in the standard arm of the dropout group is

$$E(Y_{i2}|D_i = 1, T_i = 0) = \gamma_0 + \gamma_2 + \sigma(\beta_0 - \gamma_0) \tag{5.9}$$

where $\sigma = \sigma_{21}\sigma_{11}^{-1}$, σ_{11} is the variance of the first observation in the fully observed group and σ_{12} is the corresponding covariance between the pair of observations. Similarly, in the experimental group we obtain

$$E(Y_{i2}|D_i = 1, T_i = 1) = \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 + \sigma(\beta_0 + \beta_1 - \gamma_0 - \gamma_1). \tag{5.10}$$

The true underlying population treatment difference (5.3) then becomes

$$\Delta_{\text{MAR,true}} = \gamma_1 + \gamma_3 + \sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \tag{5.11}$$

In this case, the bias in the LOCF estimator can be written as

$$\begin{aligned} B_{\text{MAR,LOCF}} = {} & p_1(\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3) + (1 - p_1)(\beta_0 + \beta_1) \\ & -p_0(\gamma_0 + \gamma_2) - (1 - p_0)\beta_0 - \gamma_1 - \gamma_3 \\ & -\sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \end{aligned} \tag{5.12}$$

Again, although involving more complicated relationships, it is clear that the bias can go in either direction, thus contradicting the claim often put forward that the bias in LOCF leads to conservative conclusions. Further, it is far from clear what conditions need to be imposed in this setting for the corresponding estimator to be either unbiased or conservative.

The bias in the CC estimator case takes the form

$$B_{\text{MAR,CC}} = -\sigma[(1 - p_1)(\beta_0 + \beta_1 - \gamma_0 - \gamma_1) - (1 - p_0)(\beta_0 - \gamma_0)]. \tag{5.13}$$

Even though this expression is simpler than in the LOCF case, it is still true that the bias can operate in either direction.

Thus, in all cases, LOCF typically produces bias of which the direction and magnitude depend on the true but unknown treatment effects. Hence, caution is needed when using this method. In contrast, an ignorable likelihood based analysis, as outlined in Section 4, provides a consistent estimator of the true treatment difference at the second occasion under both MCAR and MAR. While this is an assumption, it is rather a mild one in contrast to the stringent conditions required to justify the LOCF method, even when the qualitative features of the bias are considered more important than the quantitative ones. Note that the LOCF method is not valid even under the strong MCAR condition, whereas the CC approach is valid under MCAR.

6. Analysis of case studies

We now analyze the three clinical trials, introduced in Section 2. The primary null hypothesis (zero difference between the treatment and placebo in mean change of the HAMD17 total score at endpoint) is tested using a model of the type (3.1). The model includes the fixed categorical effects of treatment, investigator, time, and treatment by time interaction, as well as the continuous, fixed covariates of baseline score and baseline score-by-time interaction. In line with the protocol design, we use the heterogeneous compound symmetric covariance structure. Satterthwaite's approximation will be used to estimate denominator degrees of freedom. The significance of differences in least-square means is based on Type III tests. These examine the significance of each partial effect, that is, the significance of an effect with all the other effects in the model. Analyses are implemented using the SAS procedure MIXED.

Given this description, the effect of simple approaches, such as LOCF and CC, *versus* MAR, can be studied in terms of their impact on various linear mixed model aspects (fixed effects, variance structure, correlation structure). It will be shown that the impact of the simplifications can be noticeable. This is the subject of Section 6.1, dedicated to View 1. Section 6.2 focuses on Views 2a and 2b, where the last planned occasion and the last measurement obtained are of interest, respectively. In addition, we consider the issues arising when switching from a two-treatment arm to an all-treatment arm comparison.

6.1 *View 1: longitudinal analysis*

For each study in this longitudinal analysis, we will only consider the treatments that are of direct interest. This means we estimate the main difference between these treatments (treatment main effect) as well as the difference between both over time (treatment by time interaction). Treatment main effect estimates and standard errors, p values for treatment main effect and treatment by time interaction, and estimates for the within-patient correlation are reported in Table 2. When comparing LOCF, CC, and MAR, there is little difference between the three methods, in either the treatment main effect or the treatment by time interaction. Nevertheless, some important differences will be established between the strategies in terms of other model aspects. These will be seen to be in line with the reports in Verbeke and Molenberghs (1997, 2000).

Two specific features of the mean structure are the time trends and the treatment effects (over time). We discuss these in turn. The placebo time trends as well as the treatment effects (i.e. differences between the active arms and the placebo arms) are displayed in Figure 3. Both LOCF and CC are different from MAR, with a larger difference for CC. The effect is strongest in the third study. It is striking that different studies lead to different conclusions in terms of relative differences between the approaches. While there is a relatively small difference between the three methods in Study 2 and a mild one for Study 1, for Study 3 there is a strong separation between LOCF and CC on the one hand, and MAR on the other hand. Importantly, the *average* effect is smaller for MAR than for LOCF and CC. This result is in agreement with the proofs in Section 5, which showed that the direction of the bias on LOCF is in fact hard to anticipate.

The variance–covariance structure employed is heterogeneous compound symmetry (CSH), i.e. a common correlation and a variance specific to each measurement occasion. The latter feature allows us to plot the fitted variance function over time. This is done in Figure 4. It is very noticeable that MAR and CC produce a relatively similar variance structure, which tends to rise only mildly. LOCF on the other hand, deviates from both and points towards a (linear) increase in variance. If further modeling is done, MAR and CC produce homogeneous or classical compound symmetry (CS) and hence a random-intercept structure. LOCF on the other hand, suggests a random-slope model. The reason for this discrepancy is that an incomplete profile is completed by means of a flat profile. Within a pool of linearly increasing or decreasing profiles, this leads to a progressively wider spread as study time elapses. Noting that the fitted variance function has implications for the computation of mean-model standard errors, the potential for misleading inferences is clear.

The fitted correlations are given in Table 2. Clearly, CC and MAR produce virtually the same correlation. However, the correlation coefficient estimated under LOCF is much stronger. This is entirely due to the fact that after dropout, a constant value is imputed for the remainder of the study period, thereby

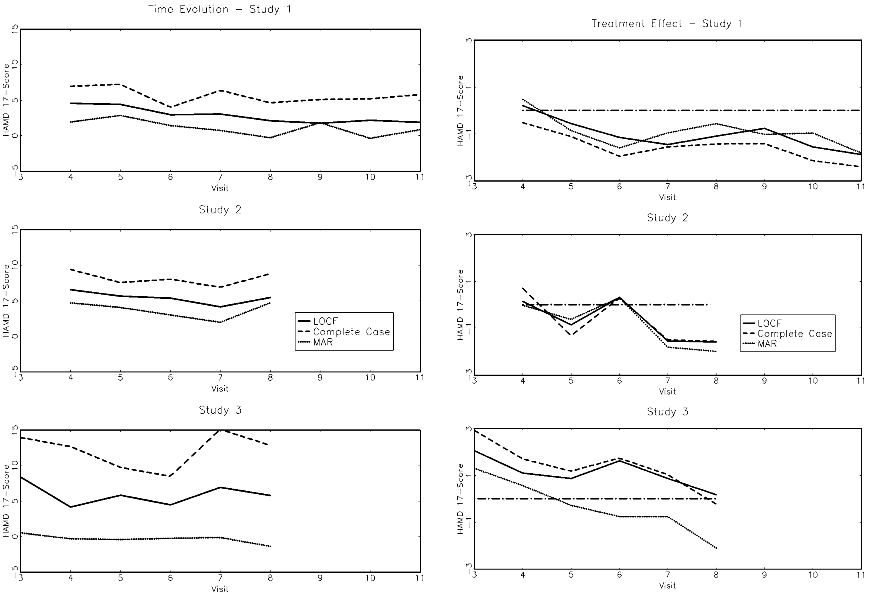


Fig. 3. Summary of all placebo time evolutions (left hand panels) and all treatment effects (right hand panels).

increasing the correlation between the repeated measurements. Of course, the problem is even more severe than shows from this analysis since, under LOCF, a constant correlation structure can be changed into one which progressively strengthens as time elapses. It should be noted that the correlation structure has an impact on all longitudinal aspects of the mean structure. For example, estimates and standard errors of time trends and estimated interactions of time with covariates can all be affected. In particular, if the estimated correlation is too high, the time trend can be ascribed a precision which is too high, implying the potential for a *liberal* error.

In conclusion, all aspects of the linear mixed models (mean structure, variance structure, correlation structure) may be influenced by the method of analysis. This is in line with results reported in Verbeke and Molenberghs (1997, 2000). It is important to note that, generally, the direction of the errors (conservative or liberal) is not clear *a priori*, since different distortions (in mean, variance, or correlation structure) may counteract each other. We will now study a number of additional analyses that are extremely relevant from a clinical trial point of view.

6.2 *Views 2a and 2b and all-* versus *two-treatment arms*

When emphasis is on the last measurement occasion, LOCF and CC are straightforward to use. When the last observed measurement is of interest, the corresponding analysis is not different from the one obtained under LOCF. In these cases, a *t* test will be used. Note that it is still possible to obtain inferences from a full linear mixed-effects model in this context. While this seems less sensible, since one obviously would

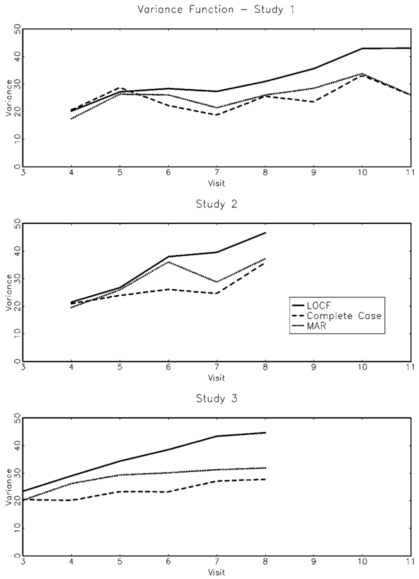


Fig. 4. Variance functions per study and per method.

Table 2. *Analysis of case study. View 1. Treatment effects (standard errors), p values for treatment main effect and for treatment by time interaction, and within-patient correlation coefficients*

Study	Method	Treatment effect (s.e.)	p value (effect, interaction)	Within-patient correlation
1	LOCF	−1.60(1.40)	(0.421, 0.565)	0.65
	CC	−1.96(1.38)	(0.322, 0.684)	0.57
	MAR	−1.81(1.24)	(0.288, 0.510)	0.53
2	LOCF	−1.61(1.05)	(0.406, 0.231)	0.54
	CC	−1.97(1.16)	(0.254, 0.399)	0.37
	MAR	−2.00(1.12)	(0.191, 0.138)	0.39
3	LOCF	1.12(0.71)	(0.964,<0.001)	0.74
	CC	1.75(0.77)	(0.918,<0.001)	0.57
	MAR	2.10(0.69)	(0.476,<0.001)	0.60

get distorted estimates of such longitudinal characteristics as time evolution, etc., we nevertheless add these for the sake of comparison. However, it should be understood that the t test analysis is more in line with clinical trial practice.

For MAR, by its very nature, one is drawn to consider the incomplete profiles, to use the information contained in these for the correct estimation of effects at later times, where there may be some missingness.

Table 3. *Analysis of case study. Views 2a and 2b. p values are reported. ('mixed' refers to the assessment of treatment at the last visit based on a linear mixed model)*

Method	Model	Data used	Study 1	Study 2	Study 3
CC	mixed	All treatments	0.076	0.055	0.001
		Two treatments	0.070	0.088	0.001
CC	*t* test	All treatments	0.092	0.156	0.017
		Two treatments	0.092	0.156	0.017
LOCF	mixed	All treatments	0.053	0.052	0.001
		Two treatments	0.056	0.082	0.001
	t test	All treatments	0.246	0.172	0.120
		Two treatments	0.246	0.172	0.120
MAR	mixed	All treatments	0.052	0.048	0.001
		Two treatments	0.047	0.077	0.001

Thus, one has to consider the full linear mixed model. To this end, the MMRM approach has been developed (Mallinckrodt *et al.*, 2001a,b).

An important issue that occurs whenever there are more than two treatment arms is whether one uses all treatments or only the two of interest. This choice has an effect on the p value in the linear mixed model case. Consider, for example, the covariance structure. Model-based smoothing of the covariance structure takes place either on two arms or on all arms. Hence, due to correlations between model parameters, the estimated treatment effects and also the resulting p values might change. Generally, one might argue that efficiency can be gained by using all treatment arms, but this comes at the cost of an increased risk of misspecification. This risk can be avoided by assuming a treatment-arm specific covariance matrix in conjunction with a treatment-arm specific mean evolution. For the t tests, however, there is no change. Of course, one might entertain the possibility of correcting for multiple comparisons when more than two arms are involved, but this is not the purpose of the current paper and does not substantially affect our conclusions.

Table 3 summarizes results in terms of p values. In study 3, which has a relatively large sample size, all p values indicate a significant difference with, very importantly, the sole exception of the t tests under LOCF. This re-emphasizes the problems with the LOCF method as discussed in Section 6.1. In studies 1 and 2, more subtle differences are observed.

For study 1, we have the following conclusions. All mixed models lead to borderline differences: LOCF and CC are not significant, MAR is borderline (depending on the number of treatments included). An endpoint analysis (i.e. using the last available measurement) leads to a completely different picture, with clearly non-significant results. For study 2, the mixed models lead to small differences, with a noticeable shift towards borderline significance for MAR with all treatments. An endpoint analysis shows, again, results that are notably different (non-significant) from the mixed models.

If the t tests under LOCF and CC are compared with the mixed analysis of MAR, studies 1 and 2 show dramatic differences. Such a comparison is not contrived since the t tests for LOCF and CC are well in line with common data-analytic practice and under MAR only the mixed analysis makes sense.

These results, in conjunction with those of Section 6.1, underscore the limitations of LOCF and CC. By selecting a subset (CC), a different type of patient might be retained in the treated versus the untreated arm. This can be explained by a difference in therapeutic effect, a difference in side effects or a combination thereof. As with CC, the difference of complete versus incomplete observations can cause distortions within an LOCF analysis. In addition to differences in sets to which the techniques are applied, there are further distortions which take place, in the mean structure, the variance structure and the correlation structure. These effects may counteract and/or strengthen each other, depending on the situation.

Table 4. *Analysis of case study. Fitted MAR and MNAR models to the case study data. Columns MAR and MNAR report twice the negative likelihood. The resulting likelihood ratio is given in the column labeled* χ^2

	MAR	MNAR		
Study	−2 likelihood		χ^2	p
1	2005.89	2004.99	0.90	0.32
2	2330.06	2320.41	9.65	0.0019
3	10234.53	10199.05	35.48	<0.0001
	Treat. effect (s.e.)			
1	−1.58(1.14)	−1.55(1.10)		
2	−1.84(1.07)	−1.64(1.07)		
3	1.98(0.65)	2.04(0.64)		

In conclusion, use of likelihood-based ignorable methods is more justifiable than LOCF and CC.

7. SENSITIVITY ANALYSIS

Although the assumption of likelihood ignorability encompasses both MAR and the more stringent and often implausible MCAR mechanisms, it is difficult to exclude the option of a more general nonrandom dropout mechanism. One solution is to fit an MNAR model as proposed by Diggle and Kenward (1994) who fitted models to the full data using the simplex algorithm (Nelder and Mead, 1965). The result of fitting these models to studies 1–3, using GAUSS code developed by the authors, is presented in Table 4. The effects of treatment, time, the interaction between time and treatment, and baseline value were all included in the model. The model for dropout is based on (3.3) and includes the effect of the previous outcome (MAR), with in addition the effect for current, possibly unobserved outcome in the MNAR case.

Note that the results are not directly comparable to those reported in Table 3, where inference is based on the last measurement, but rather to the treatment main effect results reported in Table 2. The model considered here is somewhat simpler than the model considered in Section 6.1, since fitting such a complicated model in the MNAR case may become computationally prohibitive. Note that studies 1–3 show a dramatically different picture in terms of evidence for MNAR, with apparently no, fairly strong, and very strong evidence for MNAR, respectively. However, as pointed out in the introduction and by several authors (discussion to Diggle and Kenward, 1994; Verbeke and Molenberghs, 2000, Chapter 18), one has to be extremely careful with interpreting evidence for or against MNAR using only the data under analysis.

A sensible compromise between blindly shifting to MNAR models or ignoring them altogether, is to make them a component of a sensitivity analysis. In that sense, it is important to consider the effect on key parameters such as treatment effect. Here, in line with several other observations (Molenberghs *et al.*, 2001; Verbeke *et al.*, 2001) we see that the impact on the treatment effect parameter is extremely small, providing additional support for the use of likelihood-based ignorable models. One such route for sensitivity analysis is to consider pattern-mixture models as a complement to selection models (Thijs *et al.*, 2002; Michiels *et al.*, 2002). Further routes to explore sensitivity are based on global and local influence methods (Verbeke *et al.*, 2001). A more extensive case study on the advantages and problems related to several sensitivity analysis is a topic of ongoing research.

The same considerations can be made when compliance data are available. In such a case, arguably a definitive analysis would not be possible and it might be sensible to resort to sensitivity analysis ideas (Cowles *et al.*, 1996).

8. Discussion

In this paper, we have used both formal derivations and case studies to show that there is little justification for analyzing incomplete data from longitudinal clinical trials by means of such simple methods as LOCF and CC. This is true even if a single point in time (e.g. the last measurement occasion) is of primary interest. It is more sensible to use linear mixed models in combination with the assumption of MAR. Such an approach, tailored to the needs of clinical trials, has been proposed by Mallinckrodt *et al.* (2001a,b). This type of analysis is stable and provides sensible assessments of important aspects such as treatment effect and time evolution, even if the assumption of MAR is violated in favor of MNAR. This is in line with analyses conducted by Diggle and Kenward (1994), Molenberghs *et al.* (1997, 2001) and Verbeke *et al.* (2001). Moreover, such analyses can be conducted routinely using standard statistical software such as the SAS procedures MIXED and NLMIXED.

A related and, for the regulatory clinical trial context, very important set of assertions is the following: (1) an ignorable likelihood analysis can be specified a priori in a protocol without any difficulty; (2) an ignorable likelihood analysi is consistent with the intention to treat (ITT) principle, even when only the measurement at the last occasion is of interest; (3) the difference between an LOCF and an ignorable likelihood analysis can be both liberal and conservative. The first is easy to see since, given ignorability, formulating a linear mixed model for either complete or incomplete data involves exactly the same steps. Let us expand on the second issue. It is often believed that when the last measurement is of interest a test for the treatment effect at the last occasion neglects sequences with dropout, even when such sequences contain post-randomization outcomes. As a result, it is often asserted that to be consistent with ITT some form of imputation, based on an incomplete patient's data, e.g. using LOCF, is necessary. However, as Little and Rubin (1987, Chapter 6) showed, likelihood based estimation of means in an incomplete multivariate setting involves adjustment in terms of the conditional expectation of the unobserved measurements given the observed ones. Thus, a likelihood based ignorable analysis (such as MMRM) should be seen as a proper way to accommodate information on a patient with post-randomization outcomes, even when such a patient's profile is incomplete. This fact, in conjunction with the use of treatment allocation as randomized rather than as received, shows that MMRM is fully consistent with ITT. Regarding the third issue, the case study produced smaller p values under MAR than under LOCF (Table 3). Conversely, consider a situation where the treatment difference increases over time, reaches a maximum around the middle of the study period, with a decline thereafter until complete disappearance at the end of the study. Suppose further that the bulk of dropout occurs around the middle of the study. Then, an endpoint analysis based on MAR will produce the correct nominal level, whereas LOCF might reject the null hypothesis too often. When considering LOCF, we often have in mind examples in which the disease shows progressive improvement over time. However, when the goal of a treatment is maintenance of condition in a progressively worsening disease state, LOCF can exaggerate the treatment benefit. For example, in Alzheimer's disease the goal is to prevent the patient from worsening. Thus, in a one-year trial where a patient on active treatment drops out after one week, carrying the last value forward implicitly assumes no further worsening. This is obviously not conservative.

Note that the inadequacy of LOCF, especially when conceived as a single imputation method, will vary across types of disease. LOCF is particularly inappropriate if either the effect of treatment is expected to change over time or there are secular trends. Thus, it would do slightly better in diseases where the treatment induces a steady but reversible response, such as asthma or rheumatism (Senn *et al.*, 2000).

When there is residual doubt about the plausibility of MAR, one can conduct a sensitivity analysis. This is a very active area of research. Obviously, a number of MNAR models can be fitted, provided one is prepared to approach formal aspects of model comparison with due caution. Such analyses can be complemented with appropriate (global and/or local) influence analyses. Another route is to construct pattern-mixture models and to compare the conclusions with those obtained from the selection model framework. Alternative frameworks for sensitivity analyses are provided by Robins *et al.* (1998) and Forster and Smith (1998), who present a Bayesian sensitivity analysis, and Raab and Donnelly (1999).

ACKNOWLEDGEMENTS

The first four authors gratefully acknowledge support from *Fonds Wetenschappelijk Onderzoek-Vlaanderen* Research Project G.0002.98 ‘Sensitivity Analysis for Incomplete and Coarse Data’ and from Belgian IUAP/PAI network ‘Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data’. Raymond Carroll’s research was supported by a grant from the National Cancer Institute (CA–57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30–ES09106). We are thankful to Eli Lilly for kind permission to use their data.

REFERENCES

AFIFI, A. AND ELASHOFF, R. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association* **61**, 595–604.

BUCK, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B* **22**, 302–306.

COWLES, M. K., CARLIN, B. P. AND CONNETT, J. E. (1996). Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association* **91**, 86–98.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

DEMPSTER, A. P. AND RUBIN, D. B. (1983). Overview. In Madow, W. G., Olkin, I. and Rubin, D. B. (eds), *Incomplete Data in Sample Surveys*, Vol. II: Theory and Annotated Bibliography. New York: Academic, pp. 3–10.

DIGGLE, P. J. AND KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* **43**, 49–93.

FORSTER, J. J. AND SMITH, P. W. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society, Series B* **60**, 57–70.

HARTLEY, H. O. AND HOCKING, R. (1971). The analysis of incomplete data. *Biometrics* **27**, 7783–7808.

HEYTING, A., TOLBOOM, J. AND ESSERS, J. (1992). Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine* **11**, 2043–2061.

KENWARD, M. G. AND MOLENBERGHS, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science* **12**, 236–247.

LAVORI, P. W., DAWSON, R. AND SHERA, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine* **14**, 1913–1925.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.

LITTLE, R. J. A. AND RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

MALLINCKRODT, C. H., CLARK, W. S. AND STACY, R. D. (2001a). Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal* **35**, 1215–1225.

MALLINCKRODT, C. H., CLARK, W. S. AND STACY, R. D. (2001b). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics* **11**, 9–21.

MALLINCKRODT, C. H., CLARK, W. S., CARROLL, R. J. AND MOLENBERGHS, G. (2003a). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics* **13**, 179–190.

MALLINCKRODT, C. H., SANGER, T. M., DUBE, S., DEBROTA, D. J., MOLENBERGHS, G., CARROLL, R. J., ZEIGLER POTTER, W. M. AND TOLLEFSON, G. D. (2003b). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry* **53**, 754–760.

MICHIELS, B., MOLENBERGHS, G., BIJNENS, L., VANGENEUGDEN, T. AND THIJS, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine* **21**, 1023–1041.

MOLENBERGHS, G., KENWARD, M. G. AND LESAFFRE, E. (1997). The analysis of longitudinal ordinal data with non-random dropout. *Biometrika* **84**, 33–44.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. G. AND DIGGLE, P. J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica* **52**, 153–161.

MOLENBERGHS, G., VERBEKE, G., THIJS, H., LESAFFRE, E. AND KENWARD, M. G. (2001). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis* **37**, 93–113.

MURRAY, G. D. AND FINDLAY, J. G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statististics in Medicine* **7**, 941–946.

NELDER, J. A. AND MEAD, R. (1965). A simplex method for function minimisation. *The Computer Journal* **7**, 303–313.

RAAB, G. M. AND DONNELLY, C. A. (1999). Information on sexual behaviour when some data are missing. *Applied Statistics* **48**, 117–133.

ROBINS, J. M., ROTNITZKY, A. AND SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* **93**, 1321–1339.

ROBINS, J. M., ROTNITZKY, A. AND ZHAO, L.‘P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

RUBIN, D. B., STERN, H. S. AND VEHOVAR, V. (1995). Handling “don’t know” survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association* **90**, 822–828.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

SENN, S. J., STEVENS, L. AND CHATURVEDI, N. (2000). Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Statistics in Medicine* **19**, 861–877.

SHIH, W. J. AND QUAN, H. (1997). Testing for treatment differences with dropouts present in clinical trials—A composite approach. *Statistics in Medicine* **16**, 1225–1239.

SIDDIQUI, O. AND ALI, M. W. (1998). A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics* **8**, 545–563.

THIJS, H., MOLENBERGHS, G., MICHIELS, B., VERBEKE, G. AND CURRAN, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3**, 245–265.

VERBEKE, G. AND MOLENBERGHS, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*, Lecture Notes in Statistics, 126. New York: Springer.

VERBEKE, G. AND MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

VERBEKE, G., MOLENBERGHS, G., THIJS, H., LESAFFRE, E. AND KENWARD, M. G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics* **57**, 7–14.

[*Received April 25, 2003; first revision October 30, 2003; second revision December 23, 2003; accepted for publication January 19, 2004*]

Bibliography of Raymond J. Carroll

The following is a list of all books, journal articles, and other articles authored or co-authored by Raymond J. Carroll, starting with his first in 1975 to the Fall of 2012.

Books

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. Boca Raton, Florida: Chapman and Hall/CRC Press.

Liang, F., Liu, C., and Carroll, R. J. (2010). *Advanced Markov Chain Monte Carlo: Learning from Past Samples*. New York: Wiley.

Publications

Johnson, N. L., Wegman, E. J., and Carroll, R. J. (1975). Report on a research study to determine the effects of class openness and the effects of kindergarten experience on selected student measures. Submitted as a public document to the North Carolina State Board of Education.

Carroll, R. J. (1975). Density estimation at unknown points and tail orderings. *Communications in Statistics*, 4, 565–574.

Carroll, R. J., Gupta, S. S., and Huang, D. Y. (1975). Selection procedures for the t-best populations. *Communications in Statistics*, 4, 987–1008.

Carroll, R. J. (1976). On sequential density estimation. *Zeitschrift für Wahrscheinlichkeithstheorie und verwandte Gebiete*, 36, 137–151.

Wegman, E. J. and Carroll, R. J. (1976). Final Report: General description of the sample for the North Carolina assessment of educational progress of ninth grade students. Submitted as a public document to the North Carolina State Department of Public Instruction.

Carroll, R. J. and Gupta, S. S. (1977). On the probabilities of rankings of k populations with applications. *Journal of Statistical Computation and Simulation*, 5, 145–157.

Hawkins, D., Carroll, R. J., and Wegman, E. J. (1977). Final Report: The 1976–77 North Carolina assessment of educational progress of third grade students. Submitted as a public document to the North Carolina State Department of Public Instruction.

Carroll, R. J. (1977). On the asymptotic normality of stopping times based on robust estimates. *Sankhya, Series A*, 355–377.

Wegman, E. J. and Carroll, R. J. (1977). A Monte-Carlo study of robust estimators of location. *Communications in Statistics*, 6, 795–812.

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6,
© Springer International Publishing Switzerland 2014

Carroll, R. J. (1977). A comparison of two approaches to fixed-width confidence interval estimators. *Journal of the American Statistical Association*, 72, 901–907.

Carroll, R. J. (1977). On the uniformity of sequential procedures. *Annals of Statistics*, 5, 1039–1046.

Carroll, R. J. (1978). On almost sure expansion for M-estimates. *Annals of Statistics*, 6, 314–318.

Carroll, R. J. (1978). Sequential confidence intervals for the mean of a subpopulation of a finite population. *Journal of the American Statistical Association*, 73, 408–413.

Carroll, R. J. (1978). On the asymptotic distribution of multivariate M-estimates. *Journal of Multivariate Analysis*, 8, 361–371.

Carroll, R. J. (1979). On sequential elimination procedures. *Sankhya, Series B*, 41, 226–238.

Carroll, R. J. (1979). Estimating variances of robust estimators when the errors are asymmetric. *Journal of the American Statistical Association*, 74, 674–679.

Carroll, R. J. (1979). On sequential estimation of the largest normal mean. *Sankhya, Series A*, 40, 294–302.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society, Series B*, 42, 71–78.

Ruppert D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 77, 828–838.

Holt, R. N. and Carroll, R. J. (1980). Classification of commercial bank loans through policy capturing. *Accounting, Organizations and Society*, 5, 285–296.

Carroll, R. J. (1980). Robust methods for factorial designs with outliers. *Applied Statistics*, 29, 246–251.

Carroll, R. J. and Ruppert, D. (1981). On robust tests for heteroscedasticity. *Annals of Statistics*, 9, 206–210.

Carroll, R. J. and Ruppert, D. (1981). Prediction and the power transformation family. *Biometrika*, 68, 609–616.

Hooton, T. M., Haley, R. W., Culver, D. H., White, J. W., Morgan, W., and Carroll, R. J. (1981). The joint associations of multiple risk factors with the occurrence of nosocomial infections. *American Journal of Medicine*, 70, 960–790.

Briles, D. G. and Carroll, R. J. (1981). A simple method for estimating the number of different antibodies by examining the repeat frequencies of the sequences of isoelectric focusing patterns. *Molecular Immunology*, 18, 29–38.

Carroll, R. J. (1982). Two examples of transformations when there are possible outliers. *Applied Statistics*, 31, 149–152.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics*, 10, 429–441.

Carroll, R. J. and Ruppert, D. (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, 77, 878–882.

Carroll, R. J. (1982). Robust estimation in certain heteroscedastic linear models when there are many parameters. *Journal of Statistical Planning and Inference*, 7, 1–12.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Annals of Statistics*, 10, 1224–1233.

Carroll, R. J. (1982). Power transformations when the choice of power is restricted to a finite set. *Journal of the American Statistical Association*, 77, 908–915.

Carroll, R. J., Ruppert, D., and Holt, R. N. (1982). Some aspects of estimation in heteroscedastic linear models. In *Statistical Decision Theory and Related Topics III, Volume I*, S. S. Gupta and J. O. Berger, eds. New York: Academic Press.

Carroll, R. J. and Gallo, P. P. (1982). Some aspects of robustness in functional errors-in-variables regression models. *Communications in Statistics, Series A*, 11, 2573–2585.

Carroll, R. J. and Ruppert, D. (1982). Weak convergence of bounded influence regression estimates with applications to repeated significance tests in clinical trials. *Journal of Statistical Planning and Inference*, 7, 117–129.

Carroll, R. J. (1983). Tests for regression parameters in power transformation models. *Scandinavian Journal of Statistics*, 9, 217–222.

Carroll, R. J. (1983). Discussion of "Minimax aspects of bounded influence regression," by P. J. Huber. *Journal of the American Statistical Association*, 78, 78–79.

Holt, R. N., Scarpello, V., and Carroll, R. J. (1983). Towards understanding the contents of the "Black Box" for predicting complex decision making outcomes. *Decision Sciences*, 14, 1253–1269.

Carroll, R. J. and Ruppert, D. (1983). Robust estimation in random coefficient regression models. In *Contributions to Statistics: Essays in Honour of Norman L. Johnson*, P. K. Sen, ed. Amsterdam: North Holland.

Oberpriller, J. O., Ferans, V. J., and Carroll, R. J. (1983). Changes in DNA content, number of nuclei and cellular dimensions of young rat atrial myocytes in response to left coronary artery ligation. *Journal of Molecular and Cellular Cardiology*, 14, 31–42.

Ruppert, D., Reish, R. L., Deriso, R. B., and Carroll, R. J. (1984). Monte-Carlo optimization by stochastic approximation, with application to harvesting of Atlantic menhaden. *Biometrics*, 40, 353–546.

Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79, 321–328.

Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71, 19–26.

Abbott, R. D. and Carroll, R. J. (1984). Interpreting multiple logistic regression coefficients in prospective observational studies. *American Journal of Epidemiology*, 119, 830–836.

Carroll, R. J. and Ruppert, D. (1984). Discussion of "The analysis of transformed data," by D. V. Hinkley and G. Runger. *Journal of the American Statistical Association*, 79, 312–313.

Carroll, R. J. and Gallo, P. P. (1984). Comparisons between maximum likelihood and method of moments in a linear errors-in-variables regression model. In *Design of Experiments: Ranking and Selection*, T. J. Santner and A. C. Tamhane, eds. New York: Marcel Dekker.

Oberpriller, J. O., Ferrans, V. J., and Carroll, R. J. (1984). DNA synthesis in rat atrial myocytes as a response to left ventrical infarction. *Journal of Molecular and Cellular Cardiology*, 16, 1119–1126.

Carroll, R. J. and Ruppert, D. (1985). Transformations: a robust analysis. *Technometrics*, 27, 1–12.

Reish, R. L., Deriso, R. B., Ruppert, D., and Carroll, R. J. (1985). An investigation of the population dynamics of Atlantic menhaden (Brevoortia tyrannus). *Canadian Journal of Fisheries and Aquatic Sciences*, 42, 147–157.

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics*, 13, 1335–1351.

Carroll, R. J. and Lombard, F. (1985). A note on N-estimators for the binomial distribution. *Journal of the American Statistical Association*, 80, 423–426.

Carroll, R. J. and Schneider, H. (1985). A note on Levene's test for heteroscedasticity. *Statistics and Probability Letters*, 3, 191–194.

Ruppert, D., Reish, R. L., Deriso, R. B., and Carroll, R. J. (1985). A stochastic model for managing the Atlantic menhaden fishery and assessing managerial risks. *Canadian Journal of Fisheries and Aquatic Sciences*, 42, 1371–1379.

Carroll, R. J., Gallo, P. P., and Gleser, L. J. (1985). Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association*, 80, 929–932.

Hollister, R. M., Carroll, R. J., and the Panel on Youth Employment (1985). *Youth Employment and Training Programs: The YEPDA Years.* Washington, DC: National Academy of Sciences Press

Ruppert, D. and Carroll, R. J. (1985). Data transformations in regression analysis with applications to stock recruitment relationships. In *Resource Management: Lecture Notes in Biomathematics* 61, M. Mangel, ed. New York: Springer Verlag.

Abbott, R. D. and Carroll, R. J. (1986). Conditional regression models for transient state survival analysis. *American Journal of Epidemiology*, 121, 278–735.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models, with applications to logistic regression. *Biometrika*, 73, 413–425.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986). Some new methods for weighted regression when there are possible outliers. *Technometrics*, 28, 219–230.

Carroll, R. J. and Spiegelman, C. H. (1986). The effect of small measurement error on precision instrument calibration. *Journal of Quality Technology*, 18, 170–173.

Carroll, R. J. and Ruppert, D. (1986). Discussion of Wu's paper "Jackknife, bootstrap and other resampling plans." *Annals of Statistics* 14, 1298–1301.

Gleser, L. J., Carroll, R. J., and Gallo, P. P. (1987). The limiting distribution of least squares in an errors-in-variables linear regression model. *Annals of Statistics*, 15, 220–233.

Simpson, D. G., Carroll, R. J., and Ruppert, D. (1987). M-estimation for discrete data: Asymptotic distribution theory and implications. *Annals of Statistics*, 15, 657–669.

Carroll, R. J. and Ruppert, D. (1987). Diagnostics and robustness for the transform-both-sides approach to nonlinear regression. *Technometrics*, 29, 287–299.

Watters, R. L., Carroll, R. J., and Spiegelman, C. H. (1987). Error modeling and confidence interval estimation for inductively coupled plasma calibration curves. *Analytical Chemistry* 59, 1639–1643.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1092.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74, 703–716.

Carroll, R. J. (1988). The effects of variance function estimation on prediction and calibration: an example. In *Statistical Decision Theory and Related Topics IV, Volume 2*, S. S. Gupta and J. O. Berger, eds. New York: Springer Verlag, New York.

Carroll, R. J. and Cline, D. B. H. (1988). An asymptotic theory for weighted least squares with weights estimated by replication. *Biometrika*, 75, 35–43.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175–188.

Davidian, M. and Carroll, R. J. (1988). A note on extended quasilikelihood estimation. *Journal of the Royal Statistical Society, Series B*, 50, 74–82.

Carroll, R. J., Sacks, J., and Spiegelman, C. H. (1988). A new, easy to use multiple calibration curve procedure. *Technometrics*, 30, 137–142.

Street, J. O., Ruppert, D., and Carroll, R. J. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *American Statistician*, 42, 152–154.

Davidian, M., Carroll, R. J., and Smith, W. (1988). Variance functions and the minimum detectable concentration in assays. *Biometrika*, 75, 549–556.

Carroll, R. J., Wu, C. F. J., and Ruppert, D. (1988). The effect of estimating weights in linear regression. *Journal of the American Statistical Association*, 83, 1045–1054.

Carroll, R. J., and Ruppert, D. (1988). Discussion of "Signal-to-Noise ratios, performance criteria, and transformations," by G. Box. *Technometrics*, 30, 30–31.

Carroll, R. J. and Härdle, W. (1988). Symmetrized nearest neighbor estimates. *Statistics and Probability Letters*, 7, 315–318.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184–1186.

Altschul, S. F., Carroll, R. J., and Lipman, D. J. (1989). Weights for data related by a tree. *Journal of Molecular Biology*, 207, 647–651.

Carroll, R. J. and Härdle, W. (1989). Second order effects in semiparametric weighted least squares regression. *Statistics*, 20, 179–186.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, 51, 3–14.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8, 1075–1093.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84, 460–466.

Ruppert, D., Cressie, N., and Carroll, R. J. (1989). A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics*, 45, 637–656.

Rowe, D. E., Carroll, R. J., and Day, C. L. (1989). Long term recurrence rates in previously untreated (primary) basal cell carcinoma: implications for patient followup. *Journal of Dermatologic Surgery and Oncology*, 15, 315–327.

Rowe, D. E., Carroll, R. J., and Day, C. L. (1989). Mohs surgery is the treatment of choice for recurrent (previously treated) basal cell carcinoma. *Journal of Dermatologic Surgery and Oncology*, 15, 424–431.

Carroll, R. J. (1989). Redescending M-estimates. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson, eds. New York: Wiley.

Carroll, R. J. and Welsh, A. H. (1989). A note on asymmetry and robustness in linear regression. *The American Statistician*, 42, 285–287.

Carroll, R. J. and Hall, P. (1990). Nonparametric estimation of optimal performance criteria in quality engineering. *Annals of Statistics*, 18, 281–302.

Stefanski, L. A. and Carroll, R. J. (1990). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society, Series B*, 52, 345–359.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 165–184.

Härdle, W. and Carroll, R. J. (1990). Biased crossvalidation for a kernel regression estimator and its derivatives. *Österreichische Zeitschrift für Statistik und Informatik*, 20, 53–64.

Yin, Y. and Carroll, R. J. (1990). A simple robust diagnostic for heteroscedasticity based on the Spearman rank correlation. *Statistics and Probability Letters*, 10, 69–76.

Stefanski, L. A. and Carroll, R. J. (1990). Structural logistic regression measurement error models. In *Proceedings of the Conference on Measurement Error Models*, P. J. Brown and W. A. Fuller, eds. New York: Wiley.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652–663.

Carroll, R. J. (1990). Review of Nonlinear regression, functional relations and robust methods, by H. Bunke and O. Bunke, eds. *Biometrics*, 46, 877–878.

Stefanski, L. A. and Carroll, R. J. (1991). Deconvolution based score tests in measurement error models. *Annals of Statistics*, 19, 249–259.

Carroll, R. J. and Ruppert, D. (1991). Prediction intervals and quantile estimation in nonlinear regression with transformation and/or weighting. *Technometrics*, 33, 197–210.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, 53, 573–585.

Ruppert, D., Cressie, N., and Carroll, R. J. (1991). Response to "Generalized linear models for enzyme-kinetic data" by J. A. Nelder. *Biometrics*, 47, 1610–1612.

Freedman, L. S., Carroll, R. J., and Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology*, 134, 510–520.

Hsing, T. and Carroll, R. J. (1992). Asymptotic properties of sliced inverse regression. *Annals of Statistics*, 20, 1040–1061.

Carroll, R. J., van Rooij, A., and Ruymgaart, F. (1992). Theoretical aspects of ill-posed problems in statistics. *Acta Applicandae Mathematicae*, 24, 113–140.

Carroll, R. J. (1992). Approaches to estimation with errors in predictors. In *Advances in GLIM and Statistical Modeling*, Lecture Notes in Statistics #78, L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz, eds. New York: Springer Verlag.

Simpson, D. G., Ruppert, D., and Carroll, R. J. (1992). One-step GM-estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, 87, 439–450.

Carroll, R. J. and Li, K. C. (1992). Errors in variables for nonlinear regression: dimension reduction and data visualization. *Journal of the American Statistical Association*, 87, 1040–1050.

Carroll, R. J. and Spiegelman, C. H. (1992). Diagnostics for nonlinearity and heteroscedasticity in errors in variables regression. *Technometrics*, 34, 186–196.

Rosenberg, P. S., Gail, M. H., and Carroll, R. J. (1992). Projecting AIDS incidence in the presence of therapeutic effects using backcalculation and a health care access model. *Statistics in Medicine*, 11, 1633–1655.

Carroll, R. J., Gail, M. H., and Lubin, J. H. (1993). Case-control studies with errors in predictors. *Journal of the American Statistical Association*, 88, 185–199.

Sepanski, J. H. and Carroll, R. J. (1993). Semiparametric quasilikelihood and variance function estimation in measurement error models. *Journal of Econometrics*, 58, 226–253.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, 55, 693–706.

Carroll, R. J. and Hall, P. G. (1993). Semiparametric comparison of regression curves via normal likelihoods. *Australian Journal of Statistics*, 34, 471–487.

Carroll, R. J., Eltinge, J. L. , and Ruppert, D. (1993). Robust linear regression in replicated measurement error models. *Statistics and Probability Letters*, 16, 169–175.

Wang, C. Y. and Carroll, R. J. (1993). On robust estimation in logistic case-control studies. *Biometrika*, 80, 237–241.

Wang, C. Y. and Carroll, R. J. (1993). Robust estimation in case-control studies with errors in predictors. In *Statistical Decision Theory and Related Topics, V*, J. O. Berger and S. S. Gupta, eds. New York: Springer Verlag.

Carroll, R. J. (1993). Comment on "Report of the Ad Hoc Committee on Double-Blind Refereeing," by D. Cox, L. Gleser, M. Perlman, N. Reid, and K. Roeder. *Statistical Science*, 5, 323.

Welsh, A. H., Carroll, R. J., and Ruppert, D. (1994). Fitting heteroscedastic regression models. *Journal of the American Statistical Association*, 89, 100–116.

Carroll, R. J., Hall, P. G., and Ruppert, D. (1994). Estimation of lag in misregistration problems for averaged signals. *Journal of the American Statistical Association*, 89, 219–229.

Wacholder, S., Carroll, R. J., Pee, D. Y., and Gail, M. H. (1994). The partial questionnaire design for case-control studies. *Statistics in Medicine*, 13, 623–634.

Carroll, R. J. and Stefanski, L. A. (1994). Meta-analysis, measurement error and corrections for attenuation. *Statistics in Medicine*, 13, 1265–1282.

Sepanski, J. H., Knickerbocker, R., and Carroll, R. J. (1994). A semiparametric correction for attenuation. *Journal of the American Statistical Association*, 89, 1366–1373.

Wang, C. Y. and Carroll, R. J. (1995). Robust estimation in case-control studies with weights depending on the response. *Journal of Statistical Planning and Inference*, 331–340.

Landin, R., Carroll, R. J., and Freedman, L. S. (1995). Adjusting for time trends when estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *Biometrics*, 51, 169–181.

Carroll, R. J., Wang, C. Y., and Wang, S. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90, 157–169.

Carroll, R. J. and Li, K. C. (1995). Binary regressors in dimension reduction models: a new look at treatment comparisons. *Statistica Sinica*, 5, 667–688.

Carroll, R. J., Knickerbocker, R. K., and Wang, C. Y. (1995). Dimension reduction in semiparametric measurement error models. *Annals of Statistics*, 23, 161–181.

Kim, M. Y., Pasternack, B. S., Carroll, R. J., Koenig, K. L., and Toniolo, P. G. (1995). Estimating the reliability of an exposure variable in the presence of confounders. *Statistics in Medicine*, 14, 1437–1446.

Gutierrez, R. G., Carroll, R. J., Wang, N., Taylor, B., and Hee, G. (1995). Analysis of tomato root initiation using a mixture normal distribution. *Biometrics*, 51, 1461–1468.

Wang, N., Carroll, R. J., and Liang, K. Y. (1996). Quasilikelihood and variance functions in measurement error models with replicates. *Biometrics*, 52, 401–411.

Carroll, R. J., Lombard, F., Küchenhoff, H., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, 91, 242–250.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91, 722–732.

Carroll, R. J., Freedman, L., and Hartman, A. (1996). The use of semiquantitative food frequency questionnaires to estimate the distribution of usual intake. *American Journal of Epidemiology*, 143, 392–404.

Gail, M. H., Mark, S., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15, 1069–1092.

Carroll, R. J. and Ruppert, D. (1996). The use and misuse of orthogonal regression estimation in linear errors-in-variables models. *American Statistician*, 50, 1–6.

Carroll, R. J. (1996). Review of *Measurement, Regression and Calibration*, by P. J. Brown. *Statistics in Medicine*, 15, 561–562.

Simpson, D. G., Guth. D., Zhou, H., and Carroll, R. J. (1996). Interval censoring and marginal analysis in ordinal regression. *Journal of Agricultural, Biological and Environmental Statistics*, 1, 354–376.

Carroll, R. J. (1997). Discussion of "Optimal estimating functions, quasilikelihood and statistical modeling," by A. F. Despond. *Journal of Statistical Planning and Inference*, 60, 104–106.

Carroll, R. J., Chen, R., Li, T. H., Newton, H. J., Schmiediche, H., Wang, N., and George, E. I. (1997). Modeling ozone exposure in Harris County, Texas (with discussion). *Journal of the American Statistical Association*, 92, 392–413.

Wang, C. Y., Wang, S., and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*, 77, 65–86.

Küchenhoff, H. and Carroll, R. J. (1997). Segmented regression with errors in predictors. *Statistics in Medicine*, 16, 169–188.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92, 477–489.

Carroll, R. J. and Stefanski, L. A. (1997). Asymptotic theory for the SIMEX estimator in measurement error models. In *Advances in Statistical Decision Theory and Methodology*, N. Balakrishnan and S. Panchapekesan, eds. Berlin: Birkholder.

Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics*, 53, 262–272.

Carroll, R. J., Pee, D., Freedman, L. S., and Brown, C. C. (1997). Design of calibration studies when selection is at random. *American Journal of Clinical Nutrition*, 65, 1187–1189.

Carroll, R. J., Iturria, S. J., and Gutierrez, R. G. (1997). Estimating covariance matrices using estimating functions in nonparametric and semiparametric regression. In *Selected Proceedings of the Symposium on Estimating Functions*, I. Basawa, V. P. Godambe, and R. L. Taylor, eds. IMS Lecture Notes-Monograph Series, Hayward, California: Institute of Mathematical Statistics.

Xie, M., Simpson, D. G., and Carroll, R. J. (1997). Scaled link functions heterogeneous ordinal response data. In *Modeling Longitudinal and Spatially Correlated Data*, T. Gregoire, ed. New York: Springer Verlag.

Carroll, R. J., Lin, X., and Wang, N. (1997). Generalized Linear Mixed Measurement Error Models. In *Modeling Longitudinal and Spatially Correlated Data*, T. Gregoire, ed. New York: Springer Verlag.

Guth, D. J., Carroll, R. J., Simpson, D. G., and Zhou, H. (1997). Categorical regression analysis of acute inhalation exposure to tetrachloroethylene. *Risk Analysis*, 17, 321–332. .

Borkowf, C., Gail, M. H., Carroll, R. J., and Gill, R. D. (1997). Analyzing bivariate continuous data that have been grouped into categories defined by sample quantiles of the marginal distribution. *Biometrics*, 53, 690–699.

Carroll, R. J., Freedman, L. S., and Pee, D. (1997). Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics*, 53, 1440–1451.

Carroll, R. J. (1997). Surprising effects of measurement error on an aggregate data estimator. *Biometrika*, 84, 231–234.

Carroll, R. J. (1998). Measurement error in epidemiologic studies. In *Encyclopedia of Biostatistics*, P. Armitage and T. Colton, eds. New York: Wiley.

Carroll, R. J., Ruppert, D., and Welsh, A. (1998). Local estimating equations. *Journal of the American Statistical Association*, 93, 214–227.

Carroll, R. J., Freedman, L. S., Kipnis, V., and Li, L. (1998). A new class of measurement error models, with applications to estimating the distribution of usual intake. *Canadian Journal of Statistics*, 26, 467–477.

Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998). Generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93, 249–261.

Kauermann, G., Müller, M., and Carroll, R. J. (1998). The efficiency of bias-corrected estimators for nonparametric kernel estimation based on local estimating equations. *Statistics and Probability Letters*, 37, 41–47.

Carroll, R. J. and Galindo, C. D. (1998). Measurement error, biases and the validation of complex models. *Environmental Health Perspectives*, 106 (Supplement 6), 1535–1539.

Gail, M. H., Pee, D., Benichou, J., and Carroll, R. J. (1998). Designing studies to estimate the penetrance of an identified autosomal dominant mutation. *Genetic Epidemiology*, 16, 15–39.

Wang, C.Y., Wang, S., Gutierrez, R., and Carroll, R. J. (1998). Local linear regression for generalized linear models with missing data. *Annals of Statistics*, 26, 1028–1050.

Carroll, R. J., Freedman, L. S., and Kipnis, V. (1998). Measurement error and dietary intake. In *Mathematical Modeling in Experimental Nutrition*, A. J. Clifford and H. G. Müller, eds. New York: Plenum. pages 139–146.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1998). Nonparametric regression splines for generalized linear measurement error models. In *Econometrics in Theory and Practice: Festschrift in The Honour of Hans Schneeweiss*, R. Galata and H. Küchenhoff, eds. Heidelberg: Physica Verlag.

Carroll, R. J., Roeder, K., and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics*, 55, 44–54.

Potischman, N., Carroll, R. J., Iturria, S., Mittl, B., Curtin, J., Thompson, F., and Brinton, L. (1999). Comparison of the 60- and 100-item NCI-Block questionnaires with validation data. *Nutrition and Cancer*, 34, 70–75.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1999). Comment on "Regression depth," by P. J. Rousseeuw and M. Hubert. *Journal of the American Statistical Association*, 94, 410–411.

Schafer, D. W., Stefanski, L. A., and Carroll, R. J. (1999). Consideration of measurement errors in the international radiation study of cervical cancer. In *Uncertainties in Radiation Dosimetry and Their Impact on Dose Response Analysis*, E. Ron and F. O. Hoffman, eds. National Institutes of Health Publication 99-4541.

Carroll, R. J. (1999). Risk assessment with subjectively derived doses. In *Uncertainties in Radiation Dosimetry and Their Impact on Dose Response Analysis*, E. Ron and F. O. Hoffman, eds. Bethesda, Maryland: National Cancer Institute Press.

Wang, S. and Carroll, R. J. (1999). High-order asymptotics for retrospective sampling problems. *Biometrika*, 84, 881–897.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression with errors in covariates. *Biometrika*, 86, 541–554.

Kipnis, V., Carroll, R. J., Freedman, L. S., and Li, L. (1999). A new dietary measurement error model and its application to the estimation of relative risk: application to four validation studies. *American Journal of Epidemiology*, 150, 642–651.

Iturria, S., Carroll, R. J., and Firth, D. (1999). Multiplicative measurement error estimation: estimating equations. *Journal of the Royal Statistical Society, Series B*, 61, 547–562.

Lin, X. and Carroll, R. J. (1999). SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics*, 55, 613–619.

Liang, H., Härdle, W., and Carroll, R. J. (1999). Large sample theory in a semiparametric partially linear errors in variables model. *Annals of Statistics*, 27, 1519–1535.

Hong, M. Y., Chapkin, R. S., Wild, C. P., Morris, J. S., Wang, N., Carroll, R. J., Turner, N. D., and Lupton, J. R. (1999). Relationship between DNA adduct levels, repair enzyme and apoptosis as a function of DNA methylation by Azoxymethane. *Cell Growth and Differentiation*, 10, 749–758.

Gail, M. H., Pee, D., Carroll, R. J., and Wacholder, S. W. (1999). Kin-cohort designs for gene characterization. *Journal of the National Cancer Institute*, 26, 55–60.

Xie, M., Simpson, D. G., and Carroll, R. J. (2000). Random effects in interval-censored ordinal regression: latent structure and Bayesian approach. *Biometrics*, 56, 376–383.

Ruppert, D. and Carroll, R. J. (2000). Spatially adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, 42, 205–223.

Ruckstuhl, A., Welsh, A. H., and Carroll, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica*, 10, 51–71.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, 95, 520–534.

Carroll, R. J., Gail, M. H., Pee, D., and Benichou, J. (2000). Score tests for familial correlation in genotyped proband designs. *Genetic Epidemiology*, 18, 293–306.

Satten, G. A. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics*, 56, 384–400.

Mick, R., Crowley, J. J., and Carroll, R. J. (2000). Phase II clinical trial design for noncytotoxic anticancer agents for which time to disease progression is the primary endpoint. *Controlled Clinical Trials*, 21, 343–359.

Gail, M. H., Pfeiffer, R., van Houwelingen, H. C., and Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics*, 1, 231–246.

Davidson, L. A., Brown, R. E., Chang, W-C. L., Lupton, J. R., Morris, J. S., Wang, N., Carroll, R. J., Turner, N. D., and Chapkin, R. S. (2000). Morphodensitometric analysis of protein kinase C β_{II} expression in the rat colon: modulation by diet and relation to in situ cell proliferation and apoptosis. *Carcinogenesis*, 21(8), 1513–1519.

Hong, M. Y., Lupton, J. R., Morris, J. S., Wang, N., Carroll, R. J., Davidson, L. A., Elder, R. H., and Chapkin, R. S. (2000). Dietary fish oil reduces O^6-methylguanine DNA adduct levels in the rat colon in part by increasing apoptosis during tumor initiation. *Cancer Epidemiology, Biomarkers and Prevention*, 9, 819–826.

Spiegelman, D., Carroll, R. J., and Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine*, 20, 139–160.

Galindo, C. D., Kauermann, G., Liang, H., and Carroll, R. J. (2001). Bootstrap confidence intervals for local likelihood, local estimating equations and varying coefficient models. *Statistica Sinica*, 11, 121–134.

Lin, X. and Carroll, R. J. (2001). Comment on "Semiparametric and nonparametric regression analysis of longitudinal data," by D. Y. Lin and Z. Ying. *Journal of the American Statistical Association*, 96, 114–116.

Gail, M. H., Pee, D., and Carroll, R. J. (2001). Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *Journal of Statistical Planning and Inference*, 96, 121–129.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D. Hong, M. Y., and Carroll, R. J. (2001). Understanding the relationship between carcinogen-induced DNA adduct levels in distal and proximal parts of the colon. In *Mathematical Modeling and Nutrition in the Health Sciences*, J. A. Novotny, M. J. Green, and R. C. Boston, eds. New York: Kluwer Academic/Plenum Publishing.

Carroll, R. J. (2001). Review times in Statistics: tilting at windmills? *Biometrics*, 57, 1–6.

McShane, L., Midthune, D. N., Dorgan, J. F., Freedman, L. S., and Carroll, R. J. (2001). Covariate measurement error adjustment for matched case-control studies. *Biometrics*, 57, 62–73.

Strauss, W. J., Carroll, R. J., Bortnick, S. M., Menkedick, J. R., and Schulz, B. D. (2001). Combining datasets to predict the effects of regulation of environmental lead exposure in housing stock. *Biometrics*, 57, 203–210.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D. Hong, M. Y., and Carroll, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen-induced DNA adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association*, 96, 816–826.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96, 1045–1056.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data with a nonparametric cluster-level component. *Biometrika*, 88, 1179–1185.

Kipnis V., Midthune D., Freedman L.S., Bingham S., Schatzkin A., Subar A., and Carroll R. J. (2001). Empirical evidence of correlated biases in dietary assessment instruments and its implications. *American Journal of Epidemiology*, 153, 394–403.

Jiang, W., Kipnis, V., Midthune, D., and Carroll, R. J. (2001). Parameterization and inference for nonparametric regression problems, with applications to dietary intake instruments. *Journal of the Royal Statistical Society, Series B*, 63, 583–591.

Hong, M. Y., Chapkin, R. S. , Morris, J. S., Wang, N., Carroll, R. J., Turner, N. D., Chang, W. C. L., Davidson, L. A., and Lupton, J. R. (2001). Anatomical site-specific response to DNA damage is related to later tumor development in the rat AOM colon carcinogenesis model. *Carcinogenesis*, 22, 1831–1835.

Schafer, D.W., Lubin, J. H., Ron, E., Stovall, M., and Carroll, R. J. (2001). Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Biometrics*, 57, 689–697.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96, 1387–1396.

Welsh, A. H., Lin, X., and Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods *Journal of the American Statistical Association*, 97, 482–493.

Carroll, R. J., Härdle, W., and Mammen, E. (2002). Estimation in an additive model when components are linked parametrically. *Econometric Theory*, 18, 886–912.

Berry, S. A., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97, 160–169.

Mallick, B., Hoffman, F. O., and Carroll, R. J. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. *Biometrics*, 58, 13–20.

Sarkar, S., Watts, S., Ohashi, Y., and Carroll, R. J. (2001). Bridging data between two ethnic populations: a new application of matched case control methodology. *Drug Information Journal*.

Berry, S. A., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing for measurement error problems. In *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, S. van Huffel and P. Lemmerling, eds. Dordrecht: Kluwer Academic Publishers.

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D. Hong, M. Y., and Carroll, R. J. (2002). A Bayesian analysis of colonic crypt structure and coordinated response incorporating missing crypts. *Biostatistics*, 3, 529–546.

Kim, I., Cohen, N. D., and Carroll, R. J. (2002). A method for graphical representation of effect heterogeneity by a matched covariate in matched case-control studies exemplified using data from a study of colic in horses. *American Journal of Epidemiology*, 156, 463–470

Chapkin, R. S., Carroll, R. J., Apanaosovich, T. A., and McMurray, D. M. (2002). Dietary ω-3 PUFA affect TcR-mediated activation of purified murine T cells and accessory cell function in co-cultures. *Clinical and Experimental Immunology*, 130, 12–18.

Nguyen, D., Arpat, A. B., Wang, N., and Carroll, R. J. (2002). DNA microarray experiments: biological and technological issues. *Biometrics*, 58, 701–717.

Potischman, N., Coates, R. J., Swanson, C. A., Carroll, R. J., Daling, J. R., Brogan, D. R., Gammon, M. D., Midthune, D., Curtin, J., and Brinton, L. A. (2002). Increased risk of early stage breast cancer related to consumption of sweet foods among women less than age 45. *Cancer Causes and Control*, 13, 937–46.

Hong, M. Y., Chapkin, R. S., Barhoumi, R., Burghardt, R. C., Turner, N. D., Henderson, C. E., Sanders, L. M., Fan, Y. Y., Davidson, L. A., Murphy, M. E., Spinka, C. M., Carroll, R. J., and Lupton, J. R. (2002). Fish oil increases mitochondrial phospholipid unsaturation, upregulating reactive oxygen species and apoptosis in rat colonocytes. *Carcinogenesis*, 23, 1919–1925.

Rathouz, P. J., Satten, G. A., and Carroll, R. J. (2002). Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika*, 89, 905–916.

Liang, H., Wu, H., and Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics*, 4, 297–312.

Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. (2003). The structure of dietary measurement error: results of the OPEN biomarker study. *American Journal of Epidemiology*, 158, 14–21.

Linton, O. B., Mammen, E., Lin, X., and Carroll, R. J. (2004). Correlation and marginal longitudinal kernel nonparametric regression. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, D. Y. Lin and P. J. Heagerty, eds. New York: Springer.

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, 98, 573–597.

Mallinckrodt, C. H., Sanger, T. M., Dube, S., Debrota, D. J., Molenberghs, G., Carroll, R. J., Potter, W. M., and Tollefson, G. D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biological Psychiatry*, 53, 754–760.

Mallinckrodt, C. H., Clark, W. S., Carroll, R. J., and Molenberghs, G. (2003). Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *Journal of Biopharmaceutical Statistics*, 13, 179–190.

Kipnis, V., Midthune, D., Freedman, L. S., Bingham, S., Day, N. E., Riboli, E., and Carroll, R. J. (2003). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition*, 5, 915–923.

Carroll, R. J. (2003). Variances are not always nuisance parameters: The 2002 R. A. Fisher Lecture. *Biometrics*, 59, 211–220.

Schatzkin, A., Kipnis, V., Subar, A. F., Midthune, D., Carroll, R. J., Bingham, S., Schoeller, D. A., Troiano, R., and Freedman, L. S. (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based OPEN study. *International Journal of Epidemiology*, 32, 1054–1062.

Bancroft, L. K., Lupton, J. R., Taddeo, S. S., Davidson, L. A., Murphy, M. E., Carroll, R. J., and Chapkin, R. S. (2003). Dietary fish oil reduces oxidating DNA damage in rat colonocytes. *Free Radical Biology and Medicine*, 35, 149–159.

Johnson, C. D., Tadesse, M., Carroll, R. J., Dougherty, E. R., and Ramos, K. S. (2003). Genomic profiles and predictive biological networks in oxidant-induced atherogenesis. *Physiological Genomics*, 13, 263–275.

Apanasovich, T. V., Sheather, S., Lupton, J. R., Popovic, N., Turner, N. D., Chapkin, R. S., and Carroll, R. J. (2003). Testing for spatial correlation in nonstationary binary data with application to aberrant crypt foci in colon carcinogenesis. *Biometrics*, 59, 752–761.

Kim, I., Carroll, R. J., and Cohen, N. D. (2003). Semiparametric regression splines in matched case-control studies. *Biometrics*, 59, 1160–1169.

Hong, M. Y., Chapkin, R. S., Davidson, L. A., Turner, N. D., Morris, J. S., Carroll, R. J., and Lupton, J. R. (2003). Fish oil enhances targeted apoptosis during colon tumor initiation in part by down regulating BCL-2. *Nutrition and Cancer*, 46, 44–51.

Xiao, Z., Linton, O. B., Carroll, R. J., and Mammen, E. (2003). More efficient kernel estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association*, 98, 980–992.

Carroll, R. J. and Hall, P. (2004). Low-order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society, Series B*, 66, 31–46.

Lin, X., Wang, N., Welsh, A. H., and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for longitudinal/clustered data. *Biometrika*, 91, 177–194.

Freedman, L. S., Feinberg, V., Kipnis, V., Midthune, D., and Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60, 171–181.

Nguyen, D., Wang, N., and Carroll, R. J. (2004). Missing value estimation for cancer microarray gene expression data. *Journal of Data Science*, 2, 347–370.

Braga-Neto, U., Hashimoto, R., Dougherty, E. R., Nguyen, D. V., and Carroll, R. J. (2004). Is crossvalidation better than resubstitution for ranking genes? *Bioinformatics*, 20, 243–258.

Lubin, J. H., Schafer, D. W., Ron, E., Stovall, M., and Carroll, R. J. (2004). A reanalysis of thyroid neoplasms in the Israeli tinea capitis study accounting for dose uncertainties. *Radiation Research*, 161, 359–368.

Hu, Z., Wang, N., and Carroll, R. J. (2004). Profile-kernel versus backfitting in the partially linear model for longitudinal/clustered data. *Biometrika*, 91, 251–262.

Carroll, R. J., Hall, P., Apanasovich, T. V., and Lin, X. (2004). Histospline method in nonparametric regression models with application to clustered/longitudinal data. *Statistica Sinica*, 14, 633–658.

Balagurnathan, Y., Wang, N., Dougherty, E. R., Nguyen, D., Chen, Y., Bittner, M. L., and Carroll, R. J. (2004). Noise factor analysis for cDNA microarrays. *Journal of Biomedical Optics*, 9, 663–678.

Mohlenberghs, G., Thijs, H., Kenward, M. G., Carroll, R. J., Mallinckrodt, C., Jansen, I., and Beunckens, C. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5, 445–464.

Liang, H., Wang, S., Robins, J., and Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*, 99, 357–367.

Freedman, L. S., Midthune, D., Carroll, R. J., Krebs-Smith, S., Subar, A., Troiano, R. P., Dodd, K., Schatzkin, A., Ferrari, P., and Kipnis, V. (2004). Adjustments to improve the estimation of usual dietary intake distributions in the population. *Journal of Nutrition*, 134, 1836–1843.

Carroll, R. J., Ruppert, D., Tosteson, T. D., Crainiceanu, C., and Karagas, M. R. (2004). Nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, 99, 736–750.

Mallinckrodt, C. H., Kaiser, C. J., Watkin, J. G., Molenberghs, G., and Carroll, R. J. (2004). The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. *Clinical Trials*, 1, 477–489.

Davidson, L. A., Nguyen, D. V., Hokanson, R. M., Callaway, E. S., Isett, R. B., Turner, N. D., Dougherty, E. R., Lupton, J. R., Carroll, R. J., and Chapkin, R. S. (2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Research*, 64, 6797–6804.

Sanders, L. M., Henderson, C., Hong, M. H. Wang, N., Spinka, C. M., Carroll, R. J., Turner, N. D., Chapkin, R. S., and Lupton, J. R. (2004). An increase in reactive oxygen species by dietary fish oil coupled with the attenuation of antioxidant defenses by dietary pectin enhances rat colonocyte apoptosis. *Journal of Nutrition*, 134, 3233–3238.

Carroll, R. J. (2005). Measurement error in epidemiologic studies. In *Encyclopedia of Biostatistics, Second Editon*, P. Armitage and T. Colton, eds. New York: Wiley.

Durban, M., Harelezk, J., Wand, M. P., and Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24, 1153–1167.

Fu, W., Dougherty, E. R., Mallick, B. K., and Carroll, R. J. (2005). How many samples are needed to build a classifier: a general sequential approach. *Bioinformatics*, 21, 63–70.

Wang, N., Carroll, R. J., and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100, 147–157.

Baladandayuthapani, V., Mallick, B. K., and Carroll, R. J. (2005). Spatially adaptive Bayesian regression splines. *Journal of Computational and Graphical Statistics*, 14, 378–394.

Chatterjee, N., Kalaylioglu, Z., and Carroll, R. J. (2005). A new paradigm of conditional-likelihoods for exploiting gene-environment independence in family based case-control studies. *Genetic Epidemiology*, 28, 138–156.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K., and Carroll, R. J. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association*, 100, 591–601.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, 92, 399–418.

Fu, W., Haynes, T., Kohli, R., Carroll, R. J., Meininger, C. J., and Wu, G. (2005). L-Arginine is a novel anti-obesity nutrient for Zucker diabetic fatty rats. *Journal of Nutrition*, 135, 714–721.

Fu, W., Carroll, R. J., and Wang, S. (2005). Estimating misclassification error with small samples via bootstrap crossvalidation. *Bioinformatics*, 21, 1979–1986.

Leyk, M., Nguyen, D. V., Attoor, S. N., Dougherty, E. R., Turner, N. D., Bancroft, L. K., Chapkin, R. S., Lupton, J. R., and Carroll, R. J. (2005). Comparing automatic and manual image processing in FLARE assay analysis for colon carcinogenesis. *Statistical Applications in Genetics and Molecular Biology*, 4 (electronic).

Hong, M. Y., Turner, N., Chapkin, R., Carroll, R. J., and Lupton, J. R. (2005). Differential response to oxidative DNA damage may explain aspects of the cancer susceptibility between small and large intestine. *Experimental Biology and Medicine*, 230, 464–471.

Spinka, C., Carroll, R. J., and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology*, 29, 108–127.

Hong, M. Y., Bancroft, L. K., Turner, N. D., Davidson, L. A., Murphy, M. E., Carroll, R. J., Chapkin, R. S., and Lupton, J. R. (2005). Fish oil reduces oxidative DNA damage by enhancing apoptosis in rat colon. *Nutrition and Cancer*, 52, 166–175.

Carroll, R. J. (2005). Comment on “Statistical Issues Arising in the Women’s Health Initiative,” by R. L. Prentice, M. Pettinger, and G. L. Anderson. *Biometrics*, 61, 911.

Sherman, M., Apanasovich, T. V., and Carroll, R. J. (2006). On estimation in binary autologistic spatial models. *Journal of Statistical Computation and Simulation*, 76, 167–179.

Carroll, R. J., Midthune, D., Freedman, L. S., and Kipnis, V. (2006). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, 62, 75–84.

Cantwell, M. M., Millen, A. E., Carroll, R. J., Mittl, B. L., Hermansen, S., Brinton, L. A., and Postichman, N. (2006). Does a debriefing session with a nutritionist improve dietary assessment using food diaries? *Journal of Nutrition*, 136, 440–445.

Mallinckrodt, C. H., Detke, M. J., Kaiser, C. J., Watkin, J. G., Mohlenberghs, G., and Carroll, R. J. (2006). Comparing onset of antidepressant action using a repeated measures approach and a traditional assessment. *Statistics in Medicine*, 25, 2384–2397.

Chatterjee, N., Chen, J., Spinka, C., and Carroll, R. J. (2006). Comment on “Likelihood based inference on haplotype effects in genetic association studies,” by D. Y. Lind, and D. Zeng. *Journal of the American Statistical Association*, 101, 108–110.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, 68, 68–88.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68, 179–199.

Carroll, R. J. and Ruppert, D. (2006). Comment on “Conditional growth charts,” by Y. Wei and X. He. *Annals of Statistics*, 34, 2098–2104.

Fu, W., Hi, J., Spenser, T., Carroll, R. J., and Wu, G. (2006). Statistical models in assessing fold change of gene expression in Real-Time RT-PCR Experiments. *Computational Biology and Chemistry*, 30, 21–26.

Sun, N., Carroll, R. J., and Zhao, H. (2006). Bayesian error analysis model (BEAM) for reconstructing transcriptional regulatory networks. *Proceedings of the National Academy of Sciences*, 103, 7988–7993.

Baladandayuthapani, V., Holmes, C. C., Mallick, B. K., and Carroll, R. J. (2006). Modeling nonlinear gene interactions using Bayesian MARS. In *Bayesian Inference for Gene Expression and Proteomics*, K. Do, P. Müeller, and M. Vannucci, eds. Cambridge: Cambridge University Press.

Freedman, L. S., Potischman, N., Kipnis, V., Midthune, D., Schatzkin, A., Thompson, F., Troiano, R., Prentice, R., Patterson, R., Carroll, R. J., and Subar, A. (2006). A comparison of two dietary instruments for evaluating the fat - breast cancer relationship. *International Journal of Epidemiology*, 35, 1011–1021.

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J., and Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically-consumed foods with application to their distribution. *Journal of the American Dietetic Association*, 106, 1575–1587.

Lyon, J. L., Alder, S. C., Stone, M. B., Scholl, A., Reading, J. C. Holubkov, R., Sheng, X. White, G. L., Hegmann, K. T., Anspaugh, L., Hoffman, F. O., Simon, S. L., Thomas, B., Carroll, R. J., and Meikle, A. W. (2006). Thyroid disease associated with exposure to the Nevada Test Site radiation: a reevaluation based on corrected dosimetry and examination data. *Epidemiology*, 17, 604–614.

Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, 101, 1465–1474.

Hoffman, F. O., Ruttenber, J., Greenland, S., and Carroll, R. J. (2006). Radiation exposure and thyroid cancer: Letter to the editor. *Journal of the American Medical Association*, 296, 513.

Ruppert, D. and Carroll, R. J. (2007). Comment on "Does the effect of micronutrient supplementation on neonatal survival vary with respect to the percentiles of the birth weight distribution?" by F. Dominici, S. L. Zeger, G. Parmigiani, J. Katz, and P. Christian. *Bayesian Analysis*, 2, 37–42.

Liang, H., Wang, S., and Carroll, R. J. (2007). Partially linear models with missing response variables and error-prone covariates. *Biometrika*, 94, 185–198.

Hoffman, F. O., Ruttenber, A. J., Apostoaei, A. I., Carroll, R. J., and Greenland, S. (2007). The Hanford Thyroid Disease Study: an alternative view of the findings. *Health Physics*, 92, 99–111.

Thiebaut, A., Freedman, L. S., Kipnis, V., and Carroll, R. J. (2007). Is it necessary to correct for measurement error in nutritional epidemiology? *Annals of Internal Medicine*, 146, 65–68.

Van Keilegom, I. and Carroll, R. J. (2007). Backfitting versus profiling in general criterion functions. *Statistica Sinica*, 17, 797–816.

Liang, F., Liu, C., and Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, 102, 305–320.

Claeskens, G. and Carroll, R. J. (2007). Post-model selection inference in semiparametric models. *Biometrika*, 94, 249–265.

Maity, A., Ma, Y., and Carroll, R. J. (2007). Efficient estimation of population-level summaries in general semiparametric regression models with missing response. *Journal of the American Statistical Association*, 102, 123–139.

Crainiceanu, C., Carroll, R. J., and Ruppert, D. (2007). Spatially adaptive Bayesian P-splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16, 265–288.

Maity, A., Apanasovich, T. V., and Carroll, R. J. (2007). Estimation of population-level summaries in general semiparametric repeated measures regression models. In *Beyond Parametrics in Interdisciplinary Research, Fetschrift to P.K. Sen*, N. Balakrishnan, E. Pena, and M. J. Silvapulle, eds. IMS Lecture Notes-Monograph Series, Hayward, California: Institute of Mathematical Statistics.

Li, Y., Wang, N., Hong, M., Turner, N. D., Lupton, J. R., and Carroll, R. J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *Annals of Statistics*, 35, 1608–1643.

Li, Y., Guolo, A., Hoffman, F. O., and Carroll, R. J. (2007). Shared uncertainty in measurement error problems, with application to Nevada Test Site Fallout data. *Biometrics*, 63, 1226–1236.

Thompson, F. E., Kipnis, V., Midthune, D., Carroll, R. J., Freedman, L. S., Subar, A. F., Mouw, T., Leitzmann, M., and Schatzkin, A. (2007). Performance of a food frequency questionnaire in the National Institutes of Health-AARP Diet and Health Study. *Public Health Nutrition*, 11, 183–195.

Tadesse, M., He, Q., Johnson, C. D., Carroll, R. J., and Ramos, K. S. (2007). Comparison of high-density short-oligonucleotide microarray platforms. *Current Bioinformatics*, 2: 203–213.

Carroll, R. J. and Maity, A. (2007). Discussion of "Nonparametric inference with generalized likelihood ratio tests," by J. Fan and J. Jiang. *TEST*, 16, 456–458.

Baladandayuthapani, V., Hong, M. Y., Mallick, B. K., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64, 64–73.

Midthune, D., Kipnis, V., Freedman, L. S., and Carroll, R. J. (2008). Binary regression in truncated samples, with application to comparing dietary instruments in a large prospective study. *Biometrics*, 64, 289–298.

Chen, Y.-H., Carroll, R. J., and Chatterjee, N. (2008). Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*, 9, 81–99.

Pfeiffer, R. M., Carroll, R. J., Wheeler, B., Whitby, D., and Mbulaiteye, S. (2008). Combining assays for estimating prevalence of human herpesvirus 8 infection using multivariate mixture models. *Biostatistics*, 9, 137–151.

Carroll, R. J., Delaigle, A., and Hall, P. (2008). Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *Journal of the Royal Statistical Society, Series B*, 69, 859–878.

Apanasovich, T. V., Ruppert, D., Lupton, J. R., Popovic, N., Turner, N. D., Chapkin, R. S., and Carroll, R. J. (2008). Semiparametric longitudinal-spatial binary regression, with application to colon carcinogenesis. *Biometrics*, 64, 490–500.

Yanetz, R., Kipnis, V., Carroll, R. J., Dodd, K. W., Subar, A. F., Schatzkin, A., and Freedman, L. S. (2008). Using biomarker data to adjust estimates of the distribution of usual intakes for misreporting: application to energy intake in the US population. *Journal of the American Dietetic Association*, 108, 455–464.

Lobach, I., Carroll, R. J., Spinka, C., Gail, M. H., and Chatterjee, N. (2008). Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, 64, 673–684.

Vanamala1, J., Glagolenko, A., Yang, P., Carroll, R. J., Murphy, M. E., Newman, R. A., Ford, J. R., Braby, L. A., Chapkin, R. S., Turner, N. D., and Lupton, J. R. (2008). Dietary fish oil and pectin enhance colonocyte apoptosis in part through suppression of PPAR /PGE2 and elevation of PGE3. *Carcinogenesis*, 29, 790–796.

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modeling of paired sparse functional data using principal components. *Biometrika*, 95, 601–619.

Xie, M., Simpson, D. G., and Carroll, R. J. (2008). Semiparametric analysis of heterogeneous data using varying scale generalized linear models. *Journal of the American Statistical Association*, 103, 650–660.

Carroll, R. J. and Wang, Y. (2008). Nonparametric variance estimation in the analysis of microarray data: a measurement error approach. *Biometrika*, 95, 437–449.

Henderson, D. J., Carroll, R. J., and Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, 144, 257–275.

Warren, C. A., Paulhill, P. J., Davidson, L. A., Lupton, J. R., Taddeo, S. S., Hong, M. Y., Carroll, R. J., Chapkin, R. S., and Turner, N. D. (2009). Quercetin may suppress rat aberrant crypt foci formation by suppressing inflammatory mediators that influence proliferation and apoptosis. *Journal of Nutrition*, 139, 101–105.

Freedman, L. S., Midthune, D., Carroll, R. J., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 27, 5195–5216.

Senturk, D., Nguyen, D., and Carroll, R. J. (2008). Covariate-adjusted linear mixed effects model with an application to longitudinal data. *Journal of Nonparametric Statistics*, 20, 459–481.

Ferrari, P., Carroll, R. J., Gustafson, P., and Riboli, E. (2008). A Bayesian multi-level model for estimating the diet/disease relationship in a multicenter study with exposure measured with error: The EPIC study. *Statistics in Medicine*, 27, 6037–6054.

Maity, A., Carroll, R. J., Mammen, E., and Chatterjee, N. (2009). Testing in semiparametric models with interaction, with applications to gene-environment interactions. *Journal of the Royal Statistical Society, Series B*, 71, 75–96.

Senturk, D., Nguyen, D., Tassone, F., Hagerman, R., J., Carroll, R. J., and Hagerman, P. J. (2009). Covariate adjusted correlation analysis with application to FMR1 premutation female carrier data. *Biometrics*, 65, 781–792.

Carroll, R. J., Delaigle, A., and Hall, P. (2009). Nonparametric prediction in measurement error models. *Journal of the American Statistical Association*, 104, 993–1014. (Editor's Invited Paper for 2009).

Wang, Y., Ma, Y., and Carroll, R. J. (2009). Variance estimation in the analysis of microarray data. *Journal of the Royal Statistical Society, Series B*, 71, 425–445.

Lin, X. and Carroll, R. J. (2009). Nonparametric and semiparametric regression methods: Introduction and overview. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, eds. Boca Raton, Florida: Chapman and Hall/CRC Press.

Lin, X. and Carroll, R. J. (2009). Nonparametric and semiparametric regression methods for longitudinal data. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, eds. Boca Raton, Florida: Chapman and Hall/CRC Press.

Chen, Y.-H., Chatterjee, N., and Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*, 104, 220–233.

Delaigle, A., Fan, J., and Carroll, R. J. (2009). Design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*, 104, 348–359.

Carroll, R. J., Maity, A., Mammen, E., and Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika*, 96, 383–398.

Apanasovich, T. V., Carroll, R. J., and Maity, A. (2009). SIMEX and variance estimation in semiparametric measurement error models. *Electronic Journal of Statistics*, 3, 318–348.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J., and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65, 1003–1010.

Wei, Y. and Carroll, R. J. (2009). Quantile regression with measurement error. *Journal of the American Statistical Association*, 104, 1129–1143.

Carroll, R. J., Maity, A., Mammen, E., and Yu, K. (2009). Efficient semiparametric marginal estimation for the partially linear additive model for longitudinal/clustered data. *Statistics in Biosciences*, 1, 10–31.

Turner, N. D., Paulhill, K. J., Warren, C. A., Carroll, R. J., Wang, N., Davidson, L. A., Chapkin, R. S., and Lupton, J. R. (2009). Quercetin suppresses early colon carcinogenesis partly through inhibition of inflammatory mediators. *Acta Horticulturae*, 841, 237–241.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2009). Semiparametric regression during 2003–2008. *Electronic Journal of Statistics*, 3, 1193–1256.

Chatterjee, N., Chen, Y.-H., Luo, S., and Carroll, R. J. (2009). Analysis of case-control association studies: SNPs, imputation, and haplotypes. *Statistical Science*, 24, 489–502.

Chen, X., Hu, Y., and Carroll, R. J. (2010). Identification and inference in nonlinear models using two samples with nonclassical measurement errors (with discussion). *Journal of Nonparametric Statistics*, 22, 379–423.

Martinez, J. G., Huang, J. Z., Burghardt, R. C., Barhoumi, R., and Carroll, R. J. (2010). Use of multiple singular value decompositions to analyze complex intracellular calcium ion signals. *Annals of Applied Statistics*, 3, 1467–1492.

Sinha, S., Mallick, B. K., Kipnis, V., and Carroll, R. J. (2010). Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. *Biometrics*, 66, 444–454.

Sun, Y., Carroll, R. J., and Li, D. (2009). Semiparametric estimation of fixed effects panel data varying coefficient models. In *Nonparametric Econometric Methods*, Q. Li and J. S. Racine, eds. Bingley, United Kingdom: Emerald Group Publishing.

Wang, S., Qian, L., and Carroll, R. J. (2010). Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*, 97, 79–93.

Martinez, J. G., Liang, F., and Carroll, R. J. (2010). Functional principal component selection via Stochastic Approximation Monte Carlo (SAMC). *Canadian Journal of Statistics*, 38, 256–270.

Zhou, L., Huang, J., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association*, 105, 390–400.

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11, 177–194.

Al-Kadiri, M., Carroll, R. J., and Wand, M. P. (2010). Marginal longitudinal semiparametric regression via penalized splines. *Statistics and Probability Letters*, 80, 1242–1252.

Fu, W., Stromberg, A. J., Viele, K., Carroll, R. J., and Wu, G. (2010). Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology. *Journal of Nutritional Biochemistry*, 21, 561–572.

Dhavala, S., Datta, S., Mallick, B. K., Carroll, R. J., Khare, S., Lawhon, S. D., and Adams, L. G. (2010). Bayesian modeling of MPSS data: gene expression analysis of bovine salmonella infection. *Journal of the American Statistical Association*, 105, 956–967.

Martinez, J. G. , Carroll, R. J., Mueller, S., Sampson, J. N., and Chatterjee, N. (2010). A note on the effect on power of score tests via dimension reduction by penalized regression under the null. *International Journal of Biostatistics*, 6 : Issue 1, Article 12. DOI: 10.2202/1557-4679.1231.

Li, Y., Wang, N., and Carroll, R. J. (2010). Generalized functional latent feature models with single-index interactions. *Journal of the American Statistical Association*, 105, 621–633.

Chen, Y. A., Almeida, J. S., Richards, A. J., Müller, P., Carroll, R. J., and Rohrer, B. (2010). A nonparametric approach to detect nonlinear correlation in gene expression. *Journal of Computational and Graphical Statistics*, 19, 552–568.

Leonardi, T., Vanamala, J. Taddeo, S. S., Davidson, L. A., Murphy, M. E., Patil, B.S., Wang, N., Carroll, R. J., Chapkin, R. S., Lupton, J. R., and Turner, N. D. (2010). Apigenin and naringenin suppress colon carcinogenesis through the aberrant crypt stage in azoxymethane-treated rats. *Experimental Biology and Medicine*, 235, 710–717.

Lobach, I., Fan, R. and Carroll, R. J. (2010). Genotype-based association mapping of complex diseases: gene-environment interactions with multiple genetic markers and measurement error in environmental exposures. *Genetic Epidemiology*, 34, 792–802.

Sturino, J. M., Zorych, I., Mallick, B., Chang, Y.-Y., Carroll, R. J., and Bliznuyk, N. (2010). Statistical methods for comparative phenomics using high-throughput phenotype microarrays. *International Journal of Biostatistics*, 6, Issue 1, Article 29. DOI: 10.2202/1557-4679.1227.

Martinez, J. G., Huang, J. Z., and Carroll, R. J. (2010). A note on using multiple singular value decompositions to cluster complex intracellular calcium ion signals. In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, T. K. and G. Tutz, eds. Heidelberg, Germany: Physica-Verlag.

Tooze, J. A., Kipnis, V., Buckman, D. W., Carroll, R. J., Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., and Dodd, K. W. (2010). A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Statistics in Medicine*, 29, 2857–2868.

Kukush, A., Shklyar, S., Masiuk, S., Likhtarov, I., Kovgan, L., Carroll, R. J., and Bouville, A. (2011). Mixtures of classical and Berkson uncertainties in the analysis of the Chernobyl accident. *International Journal of Biostatistics*, 7 : Issue 1, Article 15. DOI: 10.2202/1557-4679.1281.

Zhang, S. Midthune, D., Pérez, A, Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W., Krebs-Smith, S. M., and Carroll, R. J. (2011). Fitting a bivariate measurement error model for episodically consumed dietary components. *International Journal of Biostatistics*, 7, Issue 1, Article 1. DOI: 10.2202/1557-4679.1267.

Carroll, R. J., Hart, J. D., and Ma, Y. (2011). Local and omnibus tests in classical measurement error models. *Journal of the Royal Statistical Society, Series B*, 73, 81–98.

Calderon, C. P., Martinez, J. G., Carroll, R. J., and Sorensen, D. C. (2011). P-splines using derivative information. *Multiscale Modeling and Simulation*, 8, 1562–1580.

Lobach, I., Mallick, B. K., and Carroll, R. J. (2011). Semiparametric Bayesian analysis of gene-environment interactions with error in measurement of environmental covariates and missing genetic data. *Statistics and its Interface*, 4, 305–315.

Carroll, R. J., Delaigle, A., and Hall, P. (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *Journal of the American Statistical Association*, 106, 191–202.

Wei, J., Carroll, R. J., and Maity, A. (2011). Testing for constant nonparametric effects in general semiparametric regression models with interactions. *Statistics and Probability Letters*, 81, 717–723.

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L. S., and Carroll, R. J. (2011). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Annals of Applied Statistics*, 5, 1456–1487.

Divers, J., Redden, D. T., Carroll, R. J., and Allison, D. B. (2011). How to estimate measurement error variance associated with ancestry proportion estimates. *Statistics and its Interface*, 4, 327–337.

Midthune, D., Schatzkin, A., Subar, A. F., Thompson, F. E., Freedman, L. S., Carroll, R. J., Shumakovich, M. A., and Kipnis, V. (2011). Validating a food frequency questionnaire for intake of episodically consumed foods: application to the National Institutes of Health-AARP Diet and Health Study. *Public Health Nutrition*, 14, 1212–1221.

Freedman, L. S., Midthune, D., Carroll, R. J., Tasevska, N., Schatzkin, A., Mares, J., Tinker, L., Potischman, N., and Kipnis, V. (2011). Using regression calibration equations that combine self-reported intake and biomarker measures to obtain unbiased estimates and more powerful tests of dietary associations. *American Journal of Epidemiology*, 174, 1238–1245.

Ma, Y., Hart, J. D., and Carroll, R. J. (2011). Density estimation in several populations with uncertain population membership. *Journal of the American Statistical Association*, 106, 1180–1192.

Xun, X., Mallick, B. K., Carroll, R. J., and Kuchment, P. (2011). A Bayesian approach to detection of small low emission sources. *Inverse Problems*, 27, 115009, doi:10.1088/0266-5611/27/11/115009.

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2012). Generalized additive partial linear models-polynomial spline smoothing estimation and variable selection procedures. *Annals of Statistics*, 39, 1827–1851.

Martinez, J. G., Carroll, R. J., Müller, S., Sampson, J. N., and Chatterjee, N. (2012). Empirical performance of crossvalidation with oracle methods in a genomics context. *American Statistician*, 65, 223–228.

Wei, J., Carroll, R. J., Harden, K. K., and Wu, G. (2012). Comparisons of treatment means when factors do not interact in two-factor studies. *Amino Acids*, 42, 2031–2035.

Collier, B. A., Groce, J. E., Morrison, M. L., Newnam, J. C., Campomizzi, A.J., Farrell, S. J., Mathewson, H. A., Snelgrove, R. T., Carroll, R. J., and Wilkins, R. N. (2012). Predicting patch occupancy in fragmented landscapes at the rangewide scale for endangered species: an example of an American warbler. *Diversity and Distributions*,18, 158–167.

Park, J.-H., Gail, M. H., Weinberg, C. Carroll, R. J., Chung, C., Wang, Z., Chanock, S., Fraumeni, J. F., and Chatterjee, N. (2012). Distribution of allele frequencies, effect-sizes and their interrelationships for common susceptibility variants. *Proceedings of the National Academy of Sciences*, 108, 18026–18031.

Ma, S., Yang, L., and Carroll, R. J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica*, 22, 95–122.

Pérez, A., Zhang, S., Kipnis, V., Freedman, L. S., and Carroll, R. J. (2012). Intake_epis_food(): An R function for fitting a bivariate measurement error model to estimate usual and energy intake for episodically consumed foods. *Journal of Statistical Software*, 46, 1–17.

Yi, G., Ma, Y., and Carroll, R. J. (2012). A robust, functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99, 151–165.

Bliznyuk, N., Carroll, R. J., Genton, M., and Wang, Y. (2012). Variogram estimation in the presence of trend. *Statistics and its Interface*, 5, 155–168.

Wei, Y., Ma, Y., and Carroll, R. J. (2012). Multiple imputation in quantile regression. *Biometrika*, 99, 423–438.

Carroll, R. J., Midthune, D., Subar, A. F., Shumakovich, M., Freedman, L. S., Thompson, F. E., and Kipnis, V. (2012). Taking advantage of the strengths of two different dietary assessment instruments to improve intake estimates for nutritional epidemiology. *American Journal of Epidemiology*, 175, 340–347.

Cho, Y. Kim, H., Turner, N. D., Mann, J. C., Wei, J., Taddeo, S. S., Davidson, L. A., Wang, N., Vannucci, M., Carroll, R. J., Chapkin, R. S., and Lupton, J. R. (2011). A chemoprotective fish oil and pectin-containing diet temporally alters gene expression profiles in exfoliated rat colonocytes throughout oncogenesis. *Journal of Nutrition*, 141, 1029–1035.

Kipnis, V., Midthune, D., Freedman, L. S., and Carroll, R. J. (2012). Regression calibration with more instruments than mismeasured variables. *Statistics in Medicine*, 31, 2713–2732.

Carroll, R. J., Delaigle, A., and Hall, P. (2012). Deconvolution when classifying noisy data involving transformations. *Journal of the American Statistical Association*, 107, 1166–1177.

Tekwe, C. D., Carroll, R. J., and Dabney, A. R. (2012). Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. *Bioinformatics*, 28, 1998–2003.