Monographs
on Statistics and
Applied Probability 63

Measurement Error in Nonlinear Models

R.J. Carroll
D. Ruppert
and
L.A. Stefanski

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

D.R. Cox, D.V. Hinkley, N. Keiding, N. Reid, D.B. Rubin and B.W. Silverman

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Bartlett* (1960)
2 Queues *D.R. Cox and W.L. Smith* (1961)
3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5 Population Genetics *W.J. Ewens* (1969)
6 Probability, Statistics and Time *M.S. Bartlett* (1975)
7 Statistical Inference *S.D. Silvey* (1975)
8 The Analysis of Contingency Tables *B.S. Everitt* (1977)
9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10 Stochastic Abundance Models *S. Engen* (1978)
11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12 Point Processes *D.R. Cox and V. Isham* (1980)
13 Identification of Outliers *D.M. Hawkins* (1980)
14 Optimal Design *S.D. Silvey* (1980)
15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16 Classification *A.D. Gordon* (1981)
17 Distribution-free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19 Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikainen and E. Pesonen* (1984)
21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25 The Statistical Analysis of Compositional Data *J. Aitchison* (1986)
26 Density Estimation for Statistics and Data Analysis *B.W. Silverman*
27 Regression Analysis with Applications *G.B. Wetherill* (1986)

28 Sequential Methods in Statistics, 3rd edition
G.B. Wetherill and K.D. Glazebrook (1986)
29 Tensor Methods in Statistics *P. McCullagh* (1987)
30 Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)
31 Asymptotic Techniques for Use in Statistics *O.E. Barndorff-Nielsen and D.R. Cox* (1989)
32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33 Analysis of Infectious Disease Data *N.G. Becker* (1989)
34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36 Symmetric Multivariate and Related Distributions *K.-T. Fang S. Kotz and K.W. Ng* (1990)
37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38 Cyclic and Computer Generated Designs, 2nd edition *J.A. John and E.R. Williams* (1995)
39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40 Subset Selection in Regression *A.J. Miller* (1990)
41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44 Inspection Errors for Attributes in Quality Control *N.L. Johnson, S. Kotz and X. Wu* (1991)
45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)
46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1993)
47 Longitudinal Data with Serial Correlation: A State-space Approach *R.H. Jones* (1993)
48 Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49 Markov Models and Optimization *M.H.A. Davis* (1993)
50 Networks and Chaos – Statistical and Probabilistic Aspects *O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall* (1993)
51 Number-theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)
52 Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)
53 Practical Risk Theory for Actuaries *C.D. Daykin, T. Pentikäinen and M. Pesonen* (1994)

54 Biplots *J.C. Gower and D.J. Hand* (1996)
55 Predictive Inference: An Introduction *S. Geisser* (1993)
56 Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)
57 An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)
58 Nonparametric Regression and Generalized Linear Models *P.J. Green and B.W. Silverman* (1994)
59 Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)
60 Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)
61 Statistics for Long Memory Processes *J. Beran* (1995)
62 Nonlinear Models for Repeated Measurement Data *M. Davidian and D.M. Giltinan* (1995)
63 Measurement Error in Nonlinear Models *R.J. Carroll, D. Ruppert and L.A. Stefanski* (1995)

(Full details concerning this series are available from the Publishers).

Measurement Error in Nonlinear Models

R.J. CARROLL
Professor of Statistics
Texas A&M University, USA

D. RUPPERT
Professor of Operations Research and Industrial Engineering
Cornell University, USA

and

L.A. STEFANSKI
Professor, Department of Statistics
North Carolina State University, USA

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

First edition 1995

Originally published by Chapman & Hall, in 1995
Softcover reprint of the hardcover 1st edition 1995

ISBN 978-0-412-04721-3 ISBN 978-1-4899-4477-1 (eBook)
DOI 10.1007/978-1-4899-4477-1

A catalogue record for this book is available from the British Library

Library of Congress Catalog Card Number: 95-69067

Printed on permanent acid-free text paper, manufactured in accordance with ANSI/NISO Z39.48-1992 and ANSI/NISO Z39.48-1984 (Permanence of Paper)

To

Marcia

Matthew

Donna, Nick, and Doug

Contents

Preface xvii

Guide to Notation xxiii

1 INTRODUCTION 1
1.1 Measurement Error Examples 1
1.1.1 Nutrition Studies 1
1.1.2 Nurses' Health Study 3
1.1.3 Bioassay in a Herbicide Study 3
1.1.4 Lung Function in Children 4
1.1.5 Coronary Heart Disease and Blood Pressure 4
1.1.6 A-Bomb Survivor Data 5
1.1.7 Blood Pressure and Urinary Sodium Chloride 5
1.2 Functional and Structural Models 6
1.3 Models for Measurement Error 7
1.3.1 General Approaches 7
1.3.2 Transportability of Models 10
1.3.3 Potential Dangers of Transporting Models 11
1.4 Sources of Data 12
1.5 Is There an "Exact" Predictor? 13
1.6 Differential and Nondifferential Error 16
1.7 True and Approximate Replicates 17
1.8 Measurement Error as a Missing Data Problem 18
1.9 Prediction 18
1.10 A Brief Tour 19

2 REGRESSION AND ATTENUATION 21
2.1 Introduction 21

2.2 Bias Caused by Measurement Error 21
2.2.1 Simple Linear Regression with Additive Error 22
2.2.2 Simple Linear Regression, More Complex Error Structure 23
2.2.3 Multiple Regression: Single Covariate Measured with Error 25
2.2.4 Multiple Covariates Measured with Error 26
2.3 Correcting for Bias 27
2.3.1 Method of Moments 27
2.3.2 Orthogonal Regression 28
2.4 Bias Versus Variance 32
2.5 Attenuation in General Problems 34
2.5.1 An Illustration of Nondifferential Measurement Error 36
2.6 Other References 37
2.7 Appendix 38

3 REGRESSION CALIBRATION 40
3.1 Overview 40
3.2 The Regression Calibration Algorithm 41
3.2.1 Correction for Attenuation 42
3.3 NHANES Example 42
3.4 Estimating the Calibration Function Parameters 46
3.4.1 Overview and First Methods 46
3.4.2 Best Linear Approximations Using Replicate Data 47
3.4.3 Nonlinear Calibration Function Models 48
3.4.4 Alternatives When Using Partial Replicates 50
3.4.5 James-Stein Calibration 50
3.5 Standard Errors 50
3.6 Expanded Regression Calibration Models 51
3.6.1 The Expanded Approximation Defined 52
3.6.2 Implementation 54
3.6.3 Models Without Severe Curvature 55
3.7 Bioassay Data 55
3.8 Heuristics and Accuracy of the Approximations 61
3.9 Examples of the Approximations 62
3.9.1 Linear Regression 63
3.9.2 Logistic Regression 63
3.9.3 Loglinear Mean Models 66

3.10 Theoretical Examples 67
3.10.1 Homoscedastic Regression 67
3.10.2 Quadratic Regression with Homoscedastic Regression Calibration 68
3.10.3 Loglinear Mean Model 68
3.10.4 Small Curvature, Heteroscedastic Calibration 69
3.11 Other References 69
3.12 Appendix 69
3.12.1 Error Variance Estimation in the CSFII 69
3.12.2 Standard Errors and Replication 72
3.12.3 Quadratic Regression: Details of The Expanded Calibration Model 78

4 SIMULATION EXTRAPOLATION 79
4.1 Overview 79
4.2 Simulation Extrapolation Heuristics 80
4.3 The SIMEX Algorithm 82
4.3.1 The Simulation and Extrapolation Steps 82
4.3.2 Modifications of the Simulation Step 83
4.3.3 Estimating the Measurement Error Variance 83
4.3.4 Extrapolant Function Considerations 84
4.3.5 Inference and Standard Errors 86
4.3.6 Relation to the Jackknife 86
4.4 Nonadditive Measurement Error 87
4.5 Framingham Heart Study 88
4.5.1 Full Replication 89
4.5.2 Partial Replication 92
4.6 SIMEX in Some Important Special Cases 94
4.6.1 Multiple Linear Regression 95
4.6.2 Loglinear Mean Models 95
4.6.3 Quadratic Mean Models 96
4.6.4 Segmented Linear Regression Mean Models 97
4.7 Theory and Variance Estimation 97
4.7.1 Simulation Extrapolation Variance Estimation 99
4.7.2 Estimating Equation Approach to Variance Estimation 101

5 INSTRUMENTAL VARIABLES 107
5.1 Overview 107
5.2 Approximate Instrumental Variable Estimation 108

5.2.1 First Regression Calibration Instrumental Variable Algorithm 109
5.2.2 Second Regression Calibration Instrumental Variable Algorithm 110
5.3 An Example 111
5.4 Derivation of the Estimators 112
5.4.1 First Regression Calibration Instrumental Variable Algorithm 113
5.4.2 Second Regression Calibration Instrumental Variable Algorithm 114
5.5 Asymptotic Distribution Approximations 116
5.5.1 Two-Stage Estimation 117
5.5.2 Computing Estimates and Standard Errors 120

6 FUNCTIONAL METHODS 122
6.1 Overview 122
6.2 Linear, Logistic and Gamma-Loglinear Models 123
6.3 Framingham Data 125
6.4 Unbiased Score Functions via Conditioning 125
6.4.1 Linear and Logistic Regression 128
6.4.2 Other Canonical Models 129
6.4.3 Computation 129
6.4.4 Inference 130
6.5 Exact Corrected Estimating Equations 131
6.5.1 Likelihoods With Exponentials and Powers 132
6.5.2 Asymptotic Distribution Approximation 133
6.6 Estimated Σ_{uu} 133
6.7 Infinite Series Corrected Estimating Equations 134
6.7.1 Rare-Event Logistic Regression 135
6.7.2 Extensions to Mean and Variance Function Models 136
6.8 Comparison of Methods 137
6.9 Appendix 139
6.9.1 Technical Complements to Conditional Score Theory 139
6.9.2 Technical Complements to Distribution Theory for Estimated Σ_{uu} 139

7 LIKELIHOOD AND QUASILIKELIHOOD 141
7.1 Introduction 141

7.1.1 Identifiable Models 143
7.2 Measurement Error Models and Missing Data 144
7.3 Likelihood Methods when **X** is Unobserved 146
7.3.1 Error Models 147
7.3.2 Likelihood and External Second Measures 149
7.3.3 The Berkson Model 150
7.3.4 Error Model Choice 151
7.4 Likelihood When **X** is Partly Observed 152
7.5 Numerical Computation of Likelihoods 153
7.6 Framingham Data 154
7.7 Bronchitis Example 156
7.8 Quasilikelihood and Variance Function Models 160
7.9 Appendix 161
7.9.1 Monte-Carlo Computation of Integrals 161
7.9.2 Linear, Probit and Logistic Regression 162

8 BAYESIAN METHODS 165
8.1 Overview 165
8.2 The Gibbs Sampler 168
8.2.1 Direct Sampling without Measurement Error 168
8.2.2 The Weighted Bootstrap 169
8.2.3 Forming Complete Data 170
8.3 Importance Sampling 171
8.4 Cervical Cancer 173
8.5 Framingham Data 175
8.5.1 Details of the Gibbs Sampler and Weighted Bootstrap 178

9 SEMIPARAMETRIC METHODS 182
9.1 Using Only Complete Data 183
9.2 Special Two-Stage Designs for Binary Responses 184
9.3 Pseudolikelihood 185
9.4 Mean Score Method 187
9.5 General Unbiased Estimating Functions 188
9.5.1 Using Polynomials 190
9.5.2 Optimal Moment-Based Estimators 191
9.5.3 Mean Based Moment-Based Estimators 191
9.6 Semiparametric Regression Calibration 192
9.7 Comparison of the Methods 193
9.8 Appendix 194

9.8.1 Use of Complete Data Only 194
9.8.2 Theory for Complete Data Only 196
9.8.3 Theory of Moment-Estimating Functions 197

10 UNKNOWN LINK FUNCTIONS 199
10.1 Overview 199
10.1.1 Constants of Proportionality 200
10.2 Estimation Methods 201
10.2.1 Some Basic Facts 201
10.2.2 Least Squares and Sliced Inverse Regression 201
10.2.3 Details of Implementation 202
10.3 Framingham Heart Study 203
10.4 Appendix 203
10.4.1 Basic Theory 203

11 HYPOTHESIS TESTING 206
11.1 Overview 206
11.2 The Regression Calibration Approximation 207
11.2.1 Testing $H_0 : \beta_x = 0$ 208
11.2.2 Testing $H_0 : \beta_z = 0$ 208
11.2.3 Testing $H_0 : (\beta_x^t, \beta_z^t)^t = 0$ 208
11.3 Hypotheses about Subvectors of β_x and β_z 209
11.4 Efficient Score Tests of $H_0 : \beta_x = 0$ 210
11.4.1 Generalized Score Tests 211

12 DENSITY ESTIMATION AND NONPARAMETRIC REGRESSION 215
12.1 Deconvolution 215
12.1.1 Parametric Deconvolution via Moments 219
12.1.2 Estimating Distribution Functions 219
12.1.3 Optimal Score Tests 220
12.1.4 Framingham Data 220
12.1.5 NHANES Data 222
12.2 Nonparametric Regression 223
12.2.1 SIMEX 224
12.2.2 Regression Calibration 225
12.2.3 QVF and Likelihood Models 226
12.2.4 Framingham Data 226
12.2.5 Other Methods 228

13 RESPONSE VARIABLE ERROR **229**
13.1 Additive/Multiplicative Error and QVF Models 230
13.1.1 Unbiased Measures of True Response 230
13.1.2 Recommendations 233
13.1.3 Biased Responses 233
13.1.4 Calibration 233
13.2 Likelihood Methods 235
13.2.1 General Likelihood Theory and Surrogates 235
13.2.2 Use of Complete Data Only 237
13.2.3 Other Methods 238
13.3 Semiparametric Methods 238
13.3.1 Pseudolikelihood—Simple Random Subsampling 238
13.3.2 Modified Pseudolikelihood—Other Types of Subsampling 239
13.4 Example 240

14 OTHER TOPICS **243**
14.1 Logistic Case-Control Studies 243
14.1.1 The Case that $\mathbf{X}$ is Observed 243
14.1.2 Measurement Error 244
14.1.3 Normal Discriminant Model 245
14.2 Differential Measurement Error 245
14.2.1 Likelihood Formulation 245
14.2.2 Functional Methods in Two-Stage Studies 246
14.2.3 Comparison of Functional and Likelihood Approaches 247
14.3 Mixture Methods as Functional Modeling 247
14.3.1 Overview 247
14.3.2 Nonparametric Mixture Likelihoods 248
14.3.3 A Cholesterol Example 250
14.3.4 Covariates Measured Without Error 251
14.4 Design of Two-Stage Validation and Replication Studies 251
14.5 Misclassification 253
14.6 Survival Analysis 254
14.6.1 General Considerations 254
14.6.2 Rare Events 254
14.6.3 Risk Set Calibration 255

A FITTING METHODS AND MODELS **257**
A.1 Overview 257
A.2 Likelihood Methods 257
A.2.1 Notation 257
A.2.2 Maximum likelihood Estimation 258
A.2.3 Likelihood Ratio Tests 259
A.2.4 Profile Likelihood and Likelihood Ratio Confidence Intervals 259
A.2.5 Efficient Score Tests 260
A.3 Unbiased Estimating Equations 261
A.3.1 Introduction and Basic Large Sample Theory 261
A.3.2 Sandwich Formula Example: Linear Regression Without Measurement Error 264
A.3.3 Sandwich Method and Likelihood-type Inference 265
A.3.4 Unbiased, But Conditionally Biased, Estimating Equations 266
A.3.5 Biased Estimating Equations 267
A.3.6 Stacking Estimating Equations: Using Prior Estimates of Some Parameters 267
A.4 Quasilikelihood and Variance Function (QVF) Models 269
A.4.1 General Ideas 269
A.4.2 Estimation and Inference for QVF Models 270
A.5 Generalized Linear Models 273
A.6 Bootstrap Methods 273
A.6.1 Introduction 273
A.6.2 Nonlinear Regression Without Measurement Error 274
A.6.3 Bootstrapping Heteroscedastic Regression Models 276
A.6.4 Bootstrapping Logistic Regression Models 277
A.6.5 Bootstrapping Measurement Error Models 277
A.6.6 Bootstrap Confidence Intervals 278

References **280**

Author index **298**

Subject index **301**

Preface

This monograph is about analysis strategies for regression problems in which predictors are measured with error. These problems are commonly known as *measurement error modeling* or *errors-in-variables.* There is an enormous literature on this topic in linear regression, as summarized by Fuller (1987). Our interest lies almost exclusively in the analysis of nonlinear regression models, defined generally enough to include generalized linear models, transform-both-sides models, and quasilikelihood and variance function problems.

The effects of measurement error are well-known, and we basically assume that the reader understands that measurement error in predictors causes biases in estimated regression coefficients, and hence that the field is largely about correcting for such effects. Chapter 2 summarizes much of what is known about the consequences of measurement error for estimating linear regression parameters, although the material is not exhaustive.

Nonlinear errors-in-variables modeling began in earnest in the early 1980s with the publication of a series of papers on diverse topics: Prentice (1982) on survival analysis, Carroll, Spiegelman, Lan, Bailey & Abbott (1984) and Stefanski & Carroll (1985) on binary regression, Armstrong (1985) on generalized linear models, Amemiya (1985) on instrumental variables and Stefanski (1985) on estimating equations. David Byar and Mitchell Gail organized a workshop on the topic in 1987 at the National Institutes of Health, which in 1989 was published as a special issue of *Statistics in Medicine.* Since these early papers, the field has grown dramatically, as evidenced by the bibliography at the end of this book. Unlike the early 1980s, the literature is now so large that it is difficult to understand the main ideas from individual papers. Indeed,

a first draft of this book, completed in late 1990, consisted only of the material in four of the first five chapters. Essentially all the rest of the material has been developed since 1990. In a field as rapidly evolving as this one, and with the entrance of many new researchers into the area, we can present but a snapshot of the current state of knowledge.

This book can be divided broadly into four main parts: Chapters 1–2, 3–6, 7–8, and 9–14. In addition, there is Appendix A, a review of relevant fitting methods and statistical models.

The first part is introductory. Chapter 1 gives a number of applications where measurement error is of concern, and defines basic terminology of error structure, data sources and the distinction between functional and structural models. Chapter 2 gives an overview of the important ideas from linear regression, particularly the biases caused by measurement error and some estimation techniques.

The second part gives the basic ideas and techniques of what we call *functional modeling*, where the distribution of the true predictor is *not* modeled parametrically. In addition, in these chapters it is assumed that the true predictor is never observable. The focus is on the additive measurement error model, although periodically we describe modifications for the multiplicative error model. Chapters 3 and 4 discuss two broadly applicable functional methods, regression calibration and simulation-extrapolation (SIMEX), which can be thought of as the default approaches. Chapter 5 discusses a broadly based approach to the use of instrumental variables. All three of these chapters focus on estimators which are easily computed but yield only approximately consistent estimates. Chapter 6 is still based on the assumption that the true predictor is never observable, but here we provide functional techniques which are fully and not just approximately consistent. This material is somewhat more daunting in (algebraic) appearance than the approximate techniques, but even so the methods themselves are often easily programmed. Throughout this part of the book, we use examples of binary regression modeling.

The third part of the book concerns *structural modeling*, meaning that the distribution of the true predictor is parametrically modeled. Chapter 7 describes the likelihood approach to estimation and inference in measurement error models, while Chapter 8 briefly covers Bayesian modeling. Here we become more focused on

the distinction between functional and structural modeling, and also describe the measurement error problem as a missing data problem. We also allow for the possibility that the true predictor can be measured in a subset of the study population. The discussion is fully general, and applies to categorical data as well as to the additive and multiplicative measurement error models. While at this point the use of structural modeling in measurement error models is not very popular, we believe it will become more so in the very near future.

The fourth part of the book is devoted to more specialized topics. Chapter 9 takes up the study of functional techniques which are applicable when the predictor can be observed in a subset of the study. Chapter 10 discusses functional estimation in models with generalized linear structure and an unknown link function. Chapter 11 describes the effects that measurement error has on hypothesis testing. Nonparametric regression and density function estimation are addressed in Chapter 12. Errors in the response rather than in predictors are described in Chapter 13. In Chapter 14, a variety of topics are addressed briefly: case-control studies, differential measurement error, functional mixture methods, design of two-stage studies and survival analysis.

We have tried to design the text so that it can be read at two levels. Many readers will be interested only in the background material and in the definition of the specific methods that can be employed. These readers will find that the chapters in the middle two parts of the text (functional and structural modeling) begin with preliminary discussion, move into the definition of the methods, and are then followed by a worked numerical example. The end of the example serves as a flag that the material is about to become more detailed, with justifications of the methods, derivations of estimated standard errors, etc. Those readers who are not interested in such details should skip the material following the examples at first (and perhaps last) reading.

It is our intention that the part of the book on functional models (Chapters 3–6) can be understood at an overview level without an extensive background in theoretical statistics, at least through the numerical examples. The structural modeling approach requires that one knows about likelihood and Bayesian methods, but with this exception the material is not particularly specialized. The fourth part of the book (Chapters 9–14) is more technical, and we

suggest that those interested mainly in an overview simply read the first section of each of those chapters.

A full appreciation of the text, especially its details, requires a strong background in likelihood methods, estimating equations and quasilikelihood and variance function models. For inference, we typically provide estimated standard errors, as well as suggest use of "the" bootstrap. These topics are all covered in Appendix A, albeit briefly. For more background on the models used in this monograph, we highly recommend reading Chapter 1 of Fuller (1987) for an introduction to linear measurement error models and the first four chapters of McCullagh and Nelder (1989) for further discussion of generalized linear models, including logistic regression.

This is a book about general ideas and strategies of estimation and inference, and not a book about a specific problem. Our interest in the field started with logistic regression, and many of our examples are based upon this problem. However, our philosophy is that measurement error occurs in many fields, and in a variety of guises, and what is needed is an outline of strategies for handling progressively more difficult problems. While logistic regression may well be the most important nonlinear measurement error model, the strategies here are applied to a hard core nonlinear regression bioassay problem (Chapter 3), a changepoint problem (Chapter 7), and a 2×2 table with misclassification (Chapter 8). Our hope is that the strategies will be sufficiently clear that they can be applied to new problems as they arise.

We have tried to represent the main themes in the field, and to reference as many research papers as possible. Obviously, as in any monograph, the selection of topics and material to be emphasized reflects our own interests. We apologize in advance to those workers whose work we have neglected to cite, or whose work should have been better advertised.

Carroll's research and the writing of this book were supported by grants from the National Cancer Institute (CA–57030 and CA–61067). After January 1, 1996, Splus and SAS computer programs (on SPARC architecture SunOS versions 4 and 5 and for Windows on PC's) which implement (for major generalized linear models) many of the functional methods described in this book can be obtained by sending a message to qvf@stat.tamu.edu. The body of the text should contain only a valid return email address. This will generate an automatic reply with instructions on how to get the

software.

Much of Stefanski's research on measurement error problems has been supported by grants from the National Science Foundation (DMS–8613681 and DMS–9200915) and by funding from the Environmental Monitoring and Assessment Program, U.S. Environmental Protection Agency.

We want to thank Jim Calvin, Bobby Gutierrez, Stephen Eckert, Joey Lin, C. Y. Wang and Naisyin Wang for helpful general comments, Donna Spiegelman for a detailed reading of the manuscript, Jeff Buzas, John Cook, Tony Olsen and Scott Overton for ideas and comments related to our research, and Viswanath Devanarayan for computing assistance and comments. Rob Abbott stimulated our initial interest in the field in 1981 with a question concerning the effects of measurement error in the Framingham Heart Study; this example appears throughout our discussion. Larry Freedman and Mitch Gail have commented on much of our work and have been instrumental in guiding us to interesting problems. Nancy Potischman introduced us to the world of nutritional epidemiology, where measurement error is of fundamental concern. Our friend Leon Gleser has been a source of support and inspiration for many years, and has been a great influence on our thinking.

This book uses data supplied by the National Heart, Lung, and Blood Institute, NIH, DHHS from the Framingham Heart Study. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the National Heart, Lung, and Blood Institute or of the Framingham Study.

Guide to Notation

In this section we give brief explanations and representative examples of the notation used in this monograph. For precise definitions, see the text.

$\widehat{A}_n$, $\widehat{B}_n$	components of the sandwich formula
α_0	intercept in model for $E(\mathbf{X}\|\mathbf{Z},\mathbf{W})$
α_x	coefficient of $\mathbf{X}$ in model for $E(\mathbf{X}\|\mathbf{Z},\mathbf{W})$
α_z	coefficient of $\mathbf{Z}$ in model for $E(\mathbf{X}\|\mathbf{Z},\mathbf{W})$
β_0	intercept in a model for $E(\mathbf{Y}\|\mathbf{X},\mathbf{Z})$
β_x	coefficient of $\mathbf{X}$ in model for $E(\mathbf{Y}\|\mathbf{X},\mathbf{Z})$
β_z	coefficient of $\mathbf{Z}$ in model for $E(\mathbf{Y}\|\mathbf{X},\mathbf{Z})$
$\beta_{\underline{1}ZX}$	coefficient of $\mathbf{1}$ in generalized linear regresssion
Δ	indicator of validation data, e.g., where $\mathbf{X}$ is observed
$\dim(\beta)$	dimension of the vector β
$f(\mathbf{Z},\mathbf{X},\beta)$	$E(\mathbf{Y}\|\mathbf{Z},\mathbf{X})$ in QVF (quasilikelihood variance function) model
f_x	$(\partial/\partial x)f$
f_{xx}	$(\partial^2/\partial x^2)f$
f_X	density of $\mathbf{X}$
$f_{Y,W,T\|Z}$	density of $(\mathbf{Y},\mathbf{W},\mathbf{T})$ given $\mathbf{Z}$
$\mathcal{F}(\cdot)$	unknown link function
$\sigma^2 g(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)$	$\mathrm{var}(\mathbf{Y}\|\mathbf{Z},\mathbf{X})$ in QVF model
$\mathcal{G}$	extrapolant function in SIMEX
$\mathcal{G}_Q$	quadratic extrapolant function
$\mathcal{G}_{RL}$	rational linear extrapolant function
$\gamma_{0,\mathrm{cm}}$	intercept in a regression calibration model
$\gamma^t_{z,\mathrm{cm}}$	coefficient of $\mathbf{Z}$ in a regression calibration model

$\gamma_{w,\mathrm{cm}}^{t}$	coefficient of $\mathbf{W}$ in a regression calibration model
$\gamma_{0,\mathrm{em}}$	intercept in an error model
$\gamma_{x,\mathrm{em}}^{t}$	coefficient of $\mathbf{X}$ in an error model
$\gamma_{0,\mathrm{em}}$	coefficient of $\mathbf{W}$ in an error model
$H(v)$	$(1+\exp(-v))^{-1}$, e.g., the logistic function
h	bandwidth in nonparametric regression or density estimation
$I_n(\Theta)$	Fisher information
$K(\cdot)$	kernel used in nonparametric regression or density estimation
κ_{cm}	$\sigma_{\mathrm{cm}}^2/\sigma^2$
$\Lambda(\cdot)$	likelihood ratio
$\mathcal{L}(\cdot)$	generalized score function
$m(\mathbf{Z},\mathbf{W},\gamma_{\mathrm{cm}})$	$E(\mathbf{X},\mathbf{Z},\mathbf{W})$
$\pi(\mathbf{Y},\mathbf{Z},\mathbf{W},\alpha)$	probability of selection into a validation study
Ψ, ψ	estimating functions
$\mathbf{S}$	$\mathbf{Y}$ measured with error ($\mathbf{S}=\mathbf{Y}+\mathbf{V}$)
$s_i(y\|\Theta)$	score function
σ_U^2	variance of $\mathbf{U}$
$\sigma_{X\|Z}^2$	conditional variance of $\mathbf{X}$ given $\mathbf{Z}$
Σ_{ZX}	covariance matrix between the random vectors $\mathbf{Z}$ and $\mathbf{X}$
$\mathbf{T}$	observation related to $\mathbf{X}$
$\Theta_b(\lambda)$	simulated estimator used in SIMEX
$\Theta(\lambda)$	average of the $\Theta_b(\lambda)$s
$\mathbf{U}$	observation error in an error model
$\mathbf{U}_{b,k}$	pseudo-error in SIMEX
$\mathbf{V}$	measurement error in the response
$\mathbf{W}$	observation related to $\mathbf{X}$
$\mathbf{X}$	covariates measured with error
$\mathbf{Y}$	response
$\mathbf{Z}$	covariates measured without error
$\overline{\mathbf{Y}}_{i\cdot}$	average of Y_{ij} over j
$[\tilde{\mathbf{Y}}\|\tilde{\mathbf{Z}},\tilde{\mathbf{X}},\mathcal{B}]$	density of $\tilde{\mathbf{Y}}$ given $(\tilde{\mathbf{Z}},\tilde{\mathbf{X}},\mathcal{B})$ (Bayesian notation)

CHAPTER 1

INTRODUCTION

1.1 Measurement Error Examples

Nonlinear measurement error models commonly begin with an underlying nonlinear model for the response $\mathbf{Y}$ in terms of the predictors. We distinguish between two kinds of predictors: $\mathbf{Z}$ represents those predictors which for all practical purposes are measured without error, and $\mathbf{X}$ those which cannot be observed exactly for all study subjects. The distinguishing feature of a measurement error problem is that we can observe a variable $\mathbf{W}$ which is related to $\mathbf{X}$. The parameters in the model relating $\mathbf{Y}$ and $(\mathbf{Z}, \mathbf{X})$ cannot, of course, be estimated directly by fitting $\mathbf{Y}$ to $(\mathbf{Z}, \mathbf{X})$. The goal of measurement error modeling is to obtain nearly unbiased estimates of these parameters indirectly by fitting a model for $\mathbf{Y}$ in terms of $(\mathbf{Z}, \mathbf{W})$. Attainment of this goal requires careful analysis. Substituting $\mathbf{W}$ for $\mathbf{X}$, but making no adjustments in the usual fitting methods for this substitution, leads to estimates that are biased, sometimes seriously.

In assessing measurement error, careful attention needs to be given to the type and nature of the error, and the sources of data which allow modeling of this error. The following examples illustrate some of the different types of problems that are considered in this book.

1.1.1 Nutrition Studies

The NHANES-I Epidemiologic Study Cohort data set (Jones, et al., 1987), is a cohort study originally consisting of 8,596 women, who were interviewed about their nutrition habits and then later examined for evidence of cancer. We restrict attention to a subcohort of 3,145 women aged 25–50 who have no missing data on

the variables of interest.

The response $\mathbf{Y}$ indicates the presence of breast cancer. The predictor variables $\mathbf{Z}$ assumed to be measured without significant error include the following: age, poverty index ratio, body mass index, alcohol (Yes-No), family history of breast cancer, age at menarche, and menopausal status. We are primarily interested in the effects of nutrition variables $\mathbf{X}$ that are known to be imprecisely measured, e.g., "long-term" saturated fat intake.

If all these underlying variables were observable, then a standard logistic regression analysis would be performed. However, it is both difficult and expensive to measure long-term diet in a large cohort. In the NHANES data, instead of observing $\mathbf{X}$, the measured $\mathbf{W}$ was a 24 hour recall, i.e., each participant's diet in the previous 24 hours was recalled and nutrition variables computed. That the measurement error is large in 24-hour recalls has been documented previously (Beaton, et al., 1979; Wu, et al., 1986). Indeed, there is evidence to support the conclusion that more than half of the variability in the observed data is due to measurement error.

There are several sources of the measurement error. First, there is the error in the ascertainment of food consumption in the previous 24-hours, especially amounts. Some of this type of error is purely random, while another part is due to systematic bias, e.g., some people resist giving an accurate description of their consumption of snacks. The size of potential systematic bias can be determined in some instances (Freedman, et al., 1991), but in the present study we have available only the 24-hour recall information, and any systematic bias is unidentifiable.

The major source of "error" is the fact that a single day's diet does not serve as an adequate measure of the previous year's diet. There *are* seasonable differences in diet, as well as day-to-day variations. This points out the fact that measurement error is much more than simple recording or instrument error: it encompasses many different sources of variability.

There is insufficient information in the NHANES data to model measurement error directly. Instead, the measurement error structure was modeled using an *external* data set, the CSFII (Continuing Survey of Food Intakes by Individuals) data (Thompson, et al., 1992). The CSFII data contain the 24-hour recall measures $\mathbf{W}$, as well as 3 additional 24-hour recall phone interviews. Using external data, rather than assessing measurement error on an internal

subset of the primary study, entails certain risks that we will be discussing later in this chapter. The basic problem is that parameters in the external study may differ from parameters in the primary study, leading to bias when external estimates are transported to the primary study.

1.1.2 Nurses' Health Study

A second nutrition and breast cancer study has been considered by Rosner, Willett & Spiegelman (1989) and Rosner, Spiegelman & Willett (1990), namely the Nurses' Health Study. The study is much larger than the NHANES study, with over 80,000 participants and over 500 breast cancer cases. The variables are much the same, with the exceptions that (1) alcohol is assessed differently; and (2) a food frequency questionnaire was used instead of 24-hour recall interviews. The size of the measurement error in the nutrition variables is still quite large. Here, $\mathbf{X}$ = (alcohol intake, nutrient intake).

The Nurses' Health Study was designed so that a direct assessment of measurement error is possible. Specifically, 173 nurses recorded alcohol and nutrient intakes in diary form for four different weeks over the course of a year. The average, $\mathbf{T}$, of these diary entries is taken to be an unbiased estimate of $\mathbf{X}$. We will call $\mathbf{T}$ a *second measure* of $\mathbf{X}$. Thus, in contrast to NHANES, measurement error was assessed on data internal to the primary study.

1.1.3 Bioassay in a Herbicide Study

Rudemo, et al. (1989) consider a bioassay experiment with plants, in which eight herbicides were applied. For each of these eight combinations, six (common) nonzero doses were applied and the dry weight $\mathbf{Y}$ of five plants grown in the same pot was measured. In this instance, the predictor variable $\mathbf{X}$ of interest is the amount of the herbicide actually absorbed by the plant, a quantity which cannot be measured. Here the response is continuous, and if $\mathbf{X}$ were observable then a nonlinear regression model would have been fit, probably by nonlinear least squares. The four-parameter logistic model (not to be confused with logistic regression where the response is binary) is commonly used.

However, $\mathbf{X}$ is not observable, but instead we know only the

nominal concentration $\mathbf{W}$ of herbicide applied to the plant. The sources of error include not only the error in diluting to the nominal concentration, but also the fact that two plants receiving the same amount of herbicide may absorb different amounts.

In this example, the measurement error was not assessed directly. Instead, the authors assumed that the true amount $\mathbf{X}$ was linearly related to the nominal amount $\mathbf{W}$ with nonconstant variance. This error model, combined with the approach discussed in Chapter 3, was used to construct a new model for the observed data.

1.1.4 Lung Function in Children

Tosteson, et al. (1989) describe an example in which the response was the presence ($\mathbf{Y} = 1$) or absence ($\mathbf{Y} = 0$) of wheeze in children, which is an indicator of lung dysfunction. The predictor variable of interest is $\mathbf{X}$ = personal exposure to NO_2. Since $\mathbf{Y}$ is a binary variable, if $\mathbf{X}$ were observable, the authors would have used logistic or probit regression to model the relationship of $\mathbf{Y}$ and $\mathbf{X}$. However, $\mathbf{X}$ was not available in their study. Instead, the investigators were able to measure a bivariate variable $\mathbf{W}$ consisting of observed kitchen and bedroom concentrations of NO_2 in the child's home. School age children spend only a portion of their time in their homes, and only a portion of that time in their kitchens and bedrooms. Thus, it is clear that the true NO_2 concentration is not fully explained by what happens in the kitchen and bedroom.

While $\mathbf{X}$ was not measured in the primary data set, two independent, i.e., external, studies were available in which both $\mathbf{X}$ and $\mathbf{W}$ were observed. We will describe this example in more detail later in this chapter.

1.1.5 Coronary Heart Disease and Blood Pressure

The Framingham study (Kannel, et al., 1986) is a large cohort study following individuals for the development $\mathbf{Y}$ of coronary heart disease. The main predictor of interest in the study is systolic blood pressure, but other variables include age at first exam, body mass, serum cholesterol and whether the person is a smoker or not. In principle, at least, $\mathbf{Z}$ consists only of age, body mass and smoking status, while the variables $\mathbf{X}$ measured with error are serum cholesterol and systolic blood pressure. It should be noted that in

a related analysis MacMahon, et al. (1990) consider only the last as a variable measured with error. We will follow this convention in our discussion.

Again, it is impossible to measure long-term systolic blood pressure $\mathbf{X}$. Instead, what is available is the blood pressure $\mathbf{W}$ observed during a clinic visit. The reason that the long-term $\mathbf{X}$ and the single-visit $\mathbf{W}$ differ is that blood pressure has major daily, as well as seasonal, variation.

In this experiment, we have an extra measurement of blood pressure $\mathbf{T}$ from a clinic visit taken 4 years before $\mathbf{W}$ was observed. Hence, unlike any of the other studies we have discussed, in the Framingham study, we have information on measurement error for each individual. One can look at $\mathbf{T}$ as simply a replicate of $\mathbf{W}$. However, $\mathbf{T}$ may be a biased measure of $\mathbf{X}$ because of secular changes in the distribution of blood pressure in the population. Both ways of looking at the data are useful, and lead to different methods of analysis.

1.1.6 A-Bomb Survivor Data

Pierce, et al. (1992) consider analysis of A-bomb survivor data from the Hiroshima explosion. They discuss various responses $\mathbf{Y}$, including the number of chromosomal aberrations. The true radiation dose $\mathbf{X}$ cannot be measured, and instead estimates $\mathbf{W}$ are available. They adopt a fully parametric approach assuming that $\mathbf{W}$ is lognormally distributed with median $\mathbf{X}$ and coefficient of variation of 30%. They assumed that if $\mathbf{X}$ is positive, it has a Weibull distribution.

1.1.7 Blood Pressure and Urinary Sodium Chloride

Liu & Liang (1992) describe a problem of logistic regression where the response $\mathbf{Y}$ is the presence of high systolic blood pressure (greater than 140). In principle the fact that systolic blood pressure is measured with error should cause the response to be measured with error, i.e., the binary response should be subject to misclassification. However, in this particular study blood pressure was measured many times and the average recorded, so that the amount of measurement error in the average systolic blood pressure is reasonably small. The predictors $\mathbf{Z}$ measured without error are age and

body mass index. The predictor $\mathbf{X}$ subject to measurement error is urinary sodium chloride, which is subject to error because of intra-individual variation over time and also possibly due to measurement error in the chemical analyses. In order to understand the effects of measurement error, 24-hour urinary sodium chloride was measured on 6 consecutive days.

1.2 Functional and Structural Models

Historically, the taxonomy of measurement error models has been based upon two major defining characteristics. The first characteristic includes both the structure of the error model relating $\mathbf{W}$ to $\mathbf{X}$ and the type and amount of additional data available to assess the important features of this error model, e.g., replicate measurements as in the Framingham data, or second measurements as in the NHANES study. These two factors (error structure and data structure) are clearly related since more sophisticated error models can be entertained only if sufficient data are available for estimation. We take up the issue of error models in detail in section 1.3, although this is a recurrent theme throughout the book.

The second defining characteristic is determined by properties of the unobserved true values $\mathbf{X}_i$, $i = 1, \ldots, n$. The literature traditionally makes the distinction between *classical functional* models, in which the $\mathbf{X}$'s are regarded as a sequence of unknown fixed constants, and *classical structural* models, in which the $\mathbf{X}$'s are regarded as random variables. We believe that it is more fruitful to make a distinction between *functional modeling*, where the $\mathbf{X}$'s may be either fixed or random, but in the latter case no or at least only minimal assumptions are made about the distribution of the $\mathbf{X}$'s, and *structural modeling*, where models, usually parametric, are placed on the distribution of the random $\mathbf{X}$'s. We discuss this issue in more detail in section 7.2, along with the connection of measurement error modeling to missing data problems. Here we give a brief overview of some of the important issues in functional and structural models and modeling.

Since we believe that the key distinction is between functional modeling and structural modeling, we will use that terminology throughout.

Consider modeling the relationship between aquatic species diversity $\mathbf{Y}$, and acid neutralizing capacity $\mathbf{X}$, given data consisting

of measures on $(\mathbf{Y}, \mathbf{X})$ from each of n lakes. If the only lakes of interest are those represented in the sample, then it is appropriate to treat $\mathbf{X}_i$, $i = 1, \ldots, n$ as unknown constants. The $\mathbf{X}$'s are not a random sample from any population, and one might reasonably be reluctant to hypothesis a parametric model for their "distribution". A functional modeling approach would either treat the $\mathbf{X}$'s as fixed unknown constants to be estimated, or would attempt to avoid their consideration entirely.

Alternatively, if the lakes represented in the data set are a random sample from a large population of lakes, then it is appropriate to treat $\mathbf{X}_i$, $i = 1, \ldots, n$ as independent and identically distributed random variables. We might still adopt a functional modeling approach, which is attractive because of the lack of assumptions that are made. Here, however, we have the alternative of hypothesizing a parametric model for the $\mathbf{X}$'s, and thus adopt a structural modeling approach.

An important fact to keep in mind is that if an estimator can be found that is consistent under a functional modeling approach, then it is distributional-robust, i.e., it may be used without making any assumptions about the distribution of the $\mathbf{X}$'s. Functional modeling is at the heart of the first part of this book, especially in Chapters 3, 4, 6 and in the more advanced (and less developed) literature discussed in Chapter 9. The key point is that even when the $\mathbf{X}$'s form a random sample from a population, functional modeling is useful because it leads to estimation procedures that are robust to misspecification of the distribution of $\mathbf{X}$. As described in Chapter 7, structural modeling has an important role to play (see also Chapter 8) in applications, but a major concern must be the appropriateness of any assumptions made about the distribution of $\mathbf{X}$.

Throughout, we will treat $\mathbf{Z}_1, ..., \mathbf{Z}_n$ as fixed constants, and our analyses will be conditioned on their values. The practice of conditioning on known covariates is standard in regression analysis.

1.3 Models for Measurement Error

1.3.1 General Approaches

A fundamental prerequisite for analyzing a measurement error problem is specification of a model for the measurement error pro-

cess. There are two general types:

- Error Models, including Classical Measurement Error models and Error Calibration models, where the conditional distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ is modeled;
- Regression Calibration models, which are also known as controlled-variable or Berkson error models, where the conditional distribution of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is modeled.

The *classical error model*, in its simplest form, is appropriate when an attempt is made to determine $\mathbf{X}$ directly, but one is unable to do so because of various errors in measurement. For example, consider systolic blood pressure (SBP), which is known to have strong daily and seasonal variations. In trying to measure SBP, the various sources of error include simple machine recording error, administration error, time of day, season of the year, etc. In such a circumstance, it sometimes makes sense to hypothesize an unbiased additive error model, which we write as

$$\mathbf{W} = \mathbf{X} + \mathbf{U}. \tag{1.1}$$

In this model, stating that $\mathbf{W}$ is an unbiased measure of $\mathbf{X}$ says that $\mathbf{W}$ has conditional mean, given both $\mathbf{X}$ and any covariates measured without error, equal to $\mathbf{X}$, i.e., in symbols, $E(\mathbf{U}|\mathbf{X}, \mathbf{Z}) = 0$. The error structure of $\mathbf{U}$ could be homoscedastic (constant variance) or heteroscedastic.

A slightly more general model allows for systematic biases. For example, it is common to measure long-term food intake via a food frequency questionnaire. There is some evidence in the literature that these questionnaires have systematic biases (Freedman, et al., 1991); in particular, it might be the case that those with the largest amounts of intake of something like saturated fat underreport their true intake more than someone with a standard diet. This phenomenon can often be modeled by a regression relationship,

$$\mathbf{W} = \gamma_{0,\mathrm{em}} + \gamma_{x,\mathrm{em}}^t \mathbf{X} + \gamma_{z,\mathrm{em}}^t \mathbf{Z} + \mathbf{U}, \quad E(\mathbf{U}|\mathbf{X}, \mathbf{Z}) = 0. \tag{1.2}$$

We use the designation "em" to stand for "error model". In either case, the basic idea is that we observe truth contaminated by error.

To distinguish the classical additive error model (1.1) from (1.2), we will call the latter an *error calibration* model. The term calibration means that $\mathbf{W}$ is biased for $\mathbf{X}$ and has to be calibrated to make it unbiased, e.g., by using $\left(\gamma_{x,\mathrm{em}}^t\right)^{-1} (\mathbf{W} - \gamma_{0,\mathrm{em}} - \gamma_{z,\mathrm{em}}^t \mathbf{Z})$.

By a *regression calibration model* we mean one which focuses on the distribution of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$. We use the term "regression calibration" as opposed to "error calibration" to make it clear that while the latter adjusts to a classical additive error model, the former involves more complex modeling. The *controlled variable model* as a form of a regression calibration is especially applicable to laboratory studies. As an example, consider the herbicide study of Rudemo, et al. (1989) (section 1.1.3). In that study, a nominal measured amount $\mathbf{W}$ of herbicide was applied to a plant. However, the actual amount $\mathbf{X}$ absorbed by the plant differed from $\mathbf{W}$, both because of potential errors in application (the nominal amount was not applied) and because of the absorption process itself. In this case, the true response should be modeled as a function of $\mathbf{W}$, e.g.,

$$\mathbf{X} = \gamma_{0,\mathrm{cm}} + \gamma_{1,\mathrm{cm}}^t \mathbf{W} + \gamma_{2,\mathrm{cm}}^t \mathbf{Z} + \mathbf{U}_*, \quad E(\mathbf{U}_* | \mathbf{Z}, \mathbf{W}) = 0. \quad (1.3)$$

Here we use the designation "cm" to denote a "regression calibration model". If true $\mathbf{X}$ is unbiased for nominal $\mathbf{W}$, so that $\gamma_{0,\mathrm{cm}} = \gamma_{2,\mathrm{cm}} = 0$ and $\gamma_{1,\mathrm{cm}} = 1$. Model (1.3) is usually called the *Berkson model.*

Determining an appropriate error model to use in the data analysis depends upon the circumstances and the available data. For example, in the herbicide study, the measured concentration $\mathbf{W}$ is fixed by design and the true concentration $\mathbf{X}$ varies due to error, so that model (1.3) is appropriate. On the other hand, in the measurement of long-term systolic blood pressure, it is the true long-term blood pressure which is fixed for an individual, and the measured value which is perturbed by error, so model (1.2) should be used. Estimation and inference procedures have been developed both for error and calibration models. While working through this monograph, the reader will observe that we provide methods for both cases.

Sometimes it is not obvious whether an error calibration or a regression calibration model is most realistic, and in these cases the choice between them necessarily is made on the basis of convenience. Empirical considerations obviously should determine the form of the model. For example, consider the lung function study of Tosteson, et al. (1989). In this study, interest was in the relationship of long-term true NO_2 intake $\mathbf{X}$ in children on the eventual development of lung disease. In their study, $\mathbf{X}$ was not available. The vector $\mathbf{W}$ consists of bedroom and kitchen NO_2 levels as mea-

sured by in situ or stationary recording devices. Certainly, $\mathbf{X}$ and $\mathbf{W}$ are related, but children are exposed to other sources of NO_2, e.g., in other parts of the house, school, etc.

The available data consisted of the primary study in which $\mathbf{Y}$ and $\mathbf{W}$ were observed, and two external studies (from different locations, study populations and investigators) in which $(\mathbf{X}, \mathbf{W})$ were observed. In this problem, the regression calibration model (1.3) seems physically reasonable, because a child's total exposure $\mathbf{X}$ can be thought of as a sum of in-home exposure and other uncontrolled factors ($\mathbf{U}_*$). Tosteson, et al. (1989) fit (1.3) to each of the external studies, found remarkable similarities in the estimated γ's, and concluded that the assumption of a common model for all three studies was a reasonable working assumption.

The error calibration model (1.2) could also have been fit. However, $\mathbf{W}$ here is bivariate, $\mathbf{X}$ is univariate, and implementation of estimates and inferences is simply less convenient here that it is for a regression calibration model.

1.3.2 Transportability of Models

In some studies, the measurement error process is not assessed directly, but instead data from other independent studies (called external data sets) are used. In this section, we discuss the appropriateness of using information from independent studies and the manner in which this information should be used.

We say that parameters of a model can be transported from one study to another if the model holds with the same parameter values in both studies. Typically, in applications only a subset of the model parameters need be transportable. Transportability means that not only the model but also the relevant parameter estimates can be transported without bias.

In many instances, approximately the same classical error model holds across different populations. For example, consider systolic blood pressure at two different clinical centers. Assuming similar levels of training for technicians making the measurements and a similar protocol, e.g., sitting after a resting period, it is reasonable to expect that the distribution of the error in the recorded measure $\mathbf{W}$ depends only on $(\mathbf{Z}, \mathbf{X})$ and not on the clinical center one enters, or on the technician making the measurement, or on the value of $\mathbf{X}$ being measured (except possibly for heteroscedasticity). Thus,

in classical error models it is often reasonable to assume that the error distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ is the same across different populations.

Similarly, the same regression calibration or controlled-variable model can sometimes be assumed to hold across different studies. For example, consider the NO_2 study described earlier. If we have two populations of suburban children, then it may be reasonable to assume that the sources of NO_2 exposure other than the bedroom and kitchen will be approximately the same, and the error models transportable. However, if one study consists of suburban children living in a nonindustrial area, and the second study consists of children living in an inner-city near an industrialized area, the assumption of transportable error models would be tenuous at best.

1.3.3 Potential Dangers of Transporting Models

The use of independent-study data to assess error model structure carries with it the danger of introducing estimation bias into the primary study analysis.

First consider the controlled variable model for NO_2 intake. The primary data set of Tosteson, et al. (1989) (section 1.1.4) is a sample from Watertown, Massachusetts. Two independent data sets were used to fit the parameters in (1.3), one from the Netherlands and one from Portage, Wisconsin. The parameter estimates for this model in the two external data sets were essentially the same. Tosteson, et al. used this evidence suggesting that the regression relationship from the Dutch and Portage studies was appropriate for the Watertown study. However, as these authors note in some detail, it is important to remember that this is an *assumption*, plausible in this instance, but still one not to be made lightly. If Watertown were to have a much different pattern of NO_2 exposure than Portage or the Netherlands, then the parameters to (1.3) fit by the latter two studies, while similar, might be biased for the Watertown study, and the results for Watertown hence incorrect.

The issue of transporting results for error models is critical in the classical measurement error model as well. Consider the MRFIT study (Kannel, et al., 1986), in which $\mathbf{X}$ is long-term systolic blood pressure. The external data set is the Framingham data (MacMahon, et al., 1990). Carroll & Stefanski (1994) discuss these studies in detail, but here we use the studies only to illustrate the potential

pitfalls of extrapolating across studies. It is reasonable to assume that model (1.1) holds with the same measurement error variance for both studies, which reduces to stating that the distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ is the same in the two studies. However, the distribution of $\mathbf{X}$ appears to differ substantially in the two studies, with the MRFIT study having smaller variance. Under these circumstances, while the error model is probably transportable a regression calibration model formed from Framingham would not be transportable to MRFIT. The problem is that (by Bayes' theorem) the distribution of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ depends both on the distribution of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ and on the distribution of $\mathbf{X}$ given $\mathbf{Z}$, and the later is not transportable.

1.4 Sources of Data

In order to perform a measurement error analysis, as seen in (1.2)-(1.3), one needs information about either $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Z})$ (*classical measurement error or error calibration*) or about $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ (*regression calibration*).

In this section, we will discuss various data sources that allow estimation of the critical distributions. These data sources can be broken up into two main categories:

- *Internal* subsets of the primary data;
- *External* or independent studies. Within each of these broad categories, there are three types of data, all of which we assume to be available in a random subsample of the data set in question:
- *Validation* data in which $\mathbf{X}$ is observable directly.
- *Replication* data, in which replicates of $\mathbf{W}$ are available.
- *Instrumental* data, in which another variable $\mathbf{T}$ is observable in addition to $\mathbf{W}$.

An internal validation data set is the ideal, because it can be used with all known techniques, permits direct examination of the error structure, and typically leads to much greater precision of estimation and inference. We cannot express too forcefully that if it is possible to construct an internal validation data set, one should strive to do so. External validation data can be used to assess any of the models (1.1)-(1.3) in the external data, but one is always making an assumption when transporting such models to the primary data.

Usually, one would make replicate measurements if there were good reason to believe that the replicated mean is a better estimate of $\mathbf{X}$ than a single observation, i.e., the classical error model is the target. Such data cannot be used to test whether $\mathbf{W}$ is unbiased for $\mathbf{X}$ as in (1.1) or biased as in (1.2). However, if one is willing to assume (1.1), then replication data can be used to estimate the variance of the measurement error, $\mathbf{U}$.

Internal instrumental data sets containing a second measure $\mathbf{T}$ are useful for instrumental variable analysis (Chapter 5). If external, they are only useful if $\mathbf{T}$ is unbiased for $\mathbf{X}$, in which case they can be used to estimate the γ's in (1.3); regression calibration (Chapter 3) is one technique which can be applied in this case.

1.5 Is There an "Exact" Predictor?

We have based our discussion on the existence of an exact predictor $\mathbf{X}$, and measurement error models which provide information about this predictor. However, in practice, it is often the case that the definition of "exact" needs to be carefully defined prior to discussion of error models.

For example, consider the NHANES study in which long-term intake of saturated fat is of interest. Ideally, one wishes to measure the actual long-term average of saturated fat intake, but even here we have a definitional problem. If this is long-term average intake over a subject's entire life, it is clearly never measurable. Even if we define "long-term" as the average intake of saturated fat within a year of entry into the study, we still cannot measure this variable without error. The problem is that, at the present time, saturated fat intake can only be assessed in practice through the use of a dietary measurement such as a food record, 24-hour recall or a food-frequency questionnaire. Such instruments measure *actual* food intake with error.

In almost all cases, one has to take an operational definition for the exact predictor. In the measurement error literature the term "gold standard" is often used for the operationally defined exact predictor, though sometimes this term is used for an exact predictor that cannot be operationally defined, In the NHANES study the operational definition is the average saturated food intake over a year-long period *as measured by the average of 24-hour recall instruments.* One can think of this as the best measure of

exposure that could possibly be determined in practice, and even here it is extremely difficult to measure this quantity. Having made this operational definition for $\mathbf{X}$, we are in a position to undertake an analysis, for clearly the observed measure $\mathbf{W}$ is unbiased for $\mathbf{X}$ when measured on a randomly selected day. In this case, the measurement error model (1.1) is reasonable. However, in order to ascertain the distributional properties of the measurement error, one requires a replication experiment, and even modeling the replicates is somewhat subtle. The simplest way to take replicates is to perform 24-hour recalls on a few consecutive days (see also section 1.1.7), but the problem here is that such replicates are probably not conditionally independent given the long-term average. This type of replication does not measure the true error, which is highly influenced by intra-individual variation in diet. Hence, with replicates on consecutive days, estimating the variance of the measurement error by components-of-variance techniques will underestimate the measurement error.

The same problem may occur in the urinary sodium chloride example (section 1.1.7), because the replicates were recorded on consecutive days. The authors suggest that intra-individual variation is an important component of variability, and the design is not ideal for measuring this variation.

If one wants to estimate the measurement error variance consistently, it is much simpler if replicates can be taken far enough apart in time that the errors can reasonably be considered independent (see Chapter 3 for details). Otherwise, assumptions must be made about the form of the correlation structure, see Wang, Carroll & Liang (1995) and also the analysis of the CSFII component of the NHANES study in section 3.12.1. In the CSFII component of the NHANES study, measurements were taken at least two months apart, but there was still some small correlation between errors. In the Nurses Health Study (section 1.1.2), the exact predictor is the long-term average intake as measured by food records. Replicated food records were taken at four different points during the year, thus properly accounting for intra-individual variation.

Using an operational definition for an "exact" predictor is often reasonable and justifiable on the grounds that it is the best one could ever possibly hope to accomplish. However, such definitions may be controversial. For example, consider the breast cancer and fat controversy. One way to determine whether changing one's fat

intake lowers the risk of developing breast cancer is to do a clinical trial, where the treatment group is actively encouraged to change their dietary behavior. Even this is controversial, because noncompliance can occur in either the treatment or the control arm. If instead one uses prospective data, as in the NHANES study, along with an operational definition of long-term intake, one should be aware that the results of a measurement error analysis could be invalid if true long-term intake and operational long-term intake differ in subtle ways. Suppose that the operational definition of fat and calories could be measured, and call these $(\text{Fat}_O, \text{Calories}_O)$, while the actual long-term intake is $(\text{Fat}_A, \text{Calories}_A)$. If breast cancer risk is associated with age and fat intake through the logistic regression model

$$\begin{aligned}&\Pr(\mathbf{Y} = 1|\text{Fat}_A, \text{Calories}_A, \text{Age})\\ &\quad = H\left(\beta_0 + \beta_1 \text{Age} + \beta_2 \text{Calories}_A + \beta_3 \text{Fat}_A\right),\end{aligned}$$

then the important parameter is β_3, with $\beta_3 > 0$ corresponding to the conclusion that increased fat intake at a given level of calories leads to increased cancer risk.

However, suppose that the observed fat and calories are actually biased measures of the long-term average:

$$\begin{aligned}\text{Fat}_O &= \gamma_{1,\text{em}} \text{Fat}_A + \gamma_{2,\text{em}} \text{Calories}_A;\\ \text{Calories}_O &= \gamma_{3,\text{em}} \text{Fat}_A + \gamma_{4,\text{em}} \text{Calories}_A.\end{aligned}$$

Then a little algebra shows that the regression of disease on the operationally defined measures has a slope for operationally defined fat of

$$\left(\gamma_{4,\text{em}}\beta_3 - \gamma_{3,\text{em}}\beta_2\right) / \left(\gamma_{1,\text{em}}\gamma_{4,\text{em}} - \gamma_{2,\text{em}}\gamma_{3,\text{em}}\right).$$

Depending on the parameter configurations, this can take on a sign different from β_3. For example, suppose that $\beta_3 = 0$ and there really is no fat effect. Using the operational definition, a measurement error analysis would lead to a fat effect of $-\gamma_{3,\text{em}}\beta_2/(\gamma_{1,\text{em}}\gamma_{4,\text{em}} - \gamma_{2,\text{em}}\gamma_{3,\text{em}})$, which may be nonzero. Hence, in this instance, there really is no fat effect, but our operational definition would lead us to find one.

1.6 Differential and Nondifferential Error

It is important to make a distinction between *differential* and *nondifferential* measurement error. Nondifferential measurement error occurs when $\mathbf{W}$ has no information about $\mathbf{Y}$ other than what is available in $\mathbf{X}$ and $\mathbf{Z}$. The technical definition is that measurement error is nondifferential if the distribution of $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ depends only on $(\mathbf{X}, \mathbf{Z})$. In this case $\mathbf{W}$ is said to be a *surrogate*. In other words, $\mathbf{W}$ is a surrogate if it is *conditionally independent* of the response given the true covariates; measurement error is *differential* otherwise.

For instance, consider the Framingham example of section 1.1.5. The predictor of major interest is long-term systolic blood pressure ($\mathbf{X}$), but we can only observe blood pressure on a single day ($\mathbf{W}$). It seems plausible that a single day's blood pressure contributes essentially no information over and above that given by true long-term blood pressure, and hence that measurement error is nondifferential. The same remarks apply to the nutrition examples in sections 1.1.1 and 1.1.2: measuring diet on a single day should not contribute information not already available in long-term diet.

Many problems can plausibly be classified as having nondifferential measurement error, especially when the true and observed covariates occur at a fixed point in time, and the response is measured at a later time.

There are two exceptions that need to be kept in mind. First, in case-control or choice-based sampling studies (section 14.1), the response is obtained first and then subsequent follow-up ascertains the covariates. In nutrition studies, this ordering of measurement typically causes differential measurement error. For instance, here the true predictor would be long-term diet before diagnosis, but the nature of case-control studies is that reported diet is obtainable only after diagnosis. A woman who develops breast cancer may well change her diet, so the reported diet as measured after diagnosis is clearly still correlated with cancer outcomes, even after taking into account long-term diet before diagnosis.

A second setting for differential measurement error occurs when $\mathbf{W}$ is not merely a mismeasured version of $\mathbf{X}$, but is a separate variable acting as a type of proxy for $\mathbf{X}$. For example, Satten & Kupper (1993) use an example for estimating the risk of coronary heart disease where $\mathbf{X}$ is an indicator of elevated LDL (low density

lippoprotein cholesterol level), taking the values 1 and 0 according as the LDL does or does not exceed 160. For their value $\mathbf{W}$ they use total cholesterol. In their particular data set, both $\mathbf{X}$ and $\mathbf{W}$ are available, and it transpires that the relationship between $\mathbf{W}$ and $\mathbf{Y}$ is differential, i.e., there is still a relationship between the two even after accounting for $\mathbf{X}$. While the example is somewhat forced, one should be aware that problems in which $\mathbf{W}$ is not merely a mismeasured version of $\mathbf{X}$ may well have differential measurement error.

The reason why nondifferential measurement error is important is that, as we will show in subsequent chapters, one can typically estimate parameters in models for responses given true covariates even when the true covariates ($\mathbf{X}$) are not observable. With differential measurement error, this is not the case: one must observe the true covariate on some study subjects. Most of this book focuses on nondifferential measurement error models, although some work for differential measurement error is described in section 14.2.

Here is a little technical argument illustrating why nondifferential measurement error is so useful. With nondifferential measurement error the relationship between $\mathbf{Y}$ and $\mathbf{W}$ is greatly simplified relative to the case of differential measurement error. In simple linear regression, for example, it means that the regression in the observed data is a linear regression of $\mathbf{Y}$ on $E(\mathbf{X}|\mathbf{W})$, because

$$\begin{aligned} E(\mathbf{Y}|\mathbf{W}) &= E\{E(\mathbf{Y}|\mathbf{X},\mathbf{W})|\mathbf{W}\} \\ &= E\{E(\mathbf{Y}|\mathbf{X})|\mathbf{W}\} \\ &= E(\beta_0 + \beta_x\mathbf{X}|\mathbf{W}) \\ &= \beta_0 + \beta_x E(\mathbf{X}|\mathbf{W}). \end{aligned}$$

The assumption of nondifferential measurement error is used to justify the second equality above. This argument forms the heart of the method of regression calibration, see Chapter 3.

1.7 True and Approximate Replicates

In the classical homoscedastic additive error model (1.1), to estimate the measurement error variance it is typical to take replicates, i.e., observe $\mathbf{W}$ twice on some of the study participants. Estimating the measurement error variance when there are replicates is discussed in detail in section 3.4.2, but here we point out a vexing

practical problem with replicates which should be kept in mind for any error analysis.

What often happens is that the replicates are subject to *drift*, so that for example the second time $\mathbf{W}$ is observed on a study participant, there is a tendency for the mean to decrease (or increase). This is a well-known phenomenon in nutrition, where individuals tend to report steadily decreasing total calories in diet the more they are interviewed. The simplest way to handle such a drift is to add a constant to the second measurements so that their sample mean equals the sample mean of the first measurements. This method is very simple, and it often works amazingly well (Landin, et al., 1995).

1.8 Measurement Error as a Missing Data Problem

In section 7.2, we discuss in detail the relationship between measurement error modeling and the vast literature on missing data (Little & Rubin, 1987). We leave the discussion until then, but here provide a very brief overview.

From one perspective, measurement error models are special kinds of missing data problems, because the $\mathbf{X}$'s, being mostly and often entirely unobservable, are obviously missing as well. Readers who are already comfortable with linear measurement error models and functional modeling will be struck by the fact that most of the recent missing data literature has pursued likelihood and Bayesian methods, i.e., structural modeling approaches. Readers familiar with missing data analysis will also be interested to know that, in large part, the measurement error model literature has pursued functional modeling approaches. We feel that both functional and structural modeling approaches are useful in the measurement error context, and this book will pursue both strategies.

1.9 Prediction

In Chapter 2 we discuss the biases caused by measurement error for estimating regression parameters, and the effects on hypothesis testing are described in Chapter 11. Much of the rest of the book is taken up with methods for removing the biases caused by measurement error, with brief descriptions of inference at each step.

Prediction of a response is, however, another matter. If a predic-

tor $\mathbf{X}$ is measured with error, and one wants to predict a response *based on the error-prone version* $\mathbf{W}$ *of* $\mathbf{X}$, then except for a special case discussed below, it rarely makes any sense to worry about measurement error. The reason for this is quite simple. If one has an original set of data $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$, one can fit a convenient model to $\mathbf{Y}$ as a function of $(\mathbf{Z}, \mathbf{W})$. Predicting $\mathbf{Y}$ from $(\mathbf{Z}, \mathbf{W})$ is merely a matter of using this model for prediction. There is no need then for measurement error to play a role in the problem.

The one situation requiring that we model the measurement error occurs when we develop a prediction model using data from one population but we wish to predict in another population. A naive prediction model that ignores measurement error may not be transportable.

1.10 A Brief Tour

As noted in the preface, this monograph is structured into four parts: background material, functional modeling, structural modeling, and specialized topics. Here we provide another brief overview of where we are going.

It is commonly thought that the effect of measurement error is "bias towards the null", and hence that one can ignore measurement error for the purpose of testing whether a predictor is "statistically significant". This lovely and appealing folklore is sometimes true but unfortunately often wrong. The reader may find Chapters 2 (especially section 2.5) and 11 instructive, for it is in these chapters that we describe in detail the effects of ignoring measurement error.

With continuously measured variables, the classical error model (1.1) is often assumed. The question of how one checks this assumption has not been discussed in the literature. Section 4.4 suggests one such method, namely plotting the intra-individual standard deviation against the mean, which should show no structure if (1.1) holds. This and a simple graphical device to check for normality of the errors are described in section 7.6. Often, the measured value of $\mathbf{W}$ is replicated, and the usual assumption is that the replicates are independent. Methods to check this assumption are described in section 3.12.1.

Having specified an error model, one can either use functional modeling methods (Chapters 3-6 and 10) or structural modeling

methods (Chapters 7-8). If $\mathbf{X}$ is observable for a subset of the study, then other functional methods are applicable (Chapter 9). Density estimation, nonparametric regression, response error and other topics are discussed in Chapters 12-14.

CHAPTER 2

REGRESSION AND ATTENUATION

2.1 Introduction

This chapter summarizes some of the known results about the effects of measurement error in linear regression, and describes some of the statistical methods used to correct for those effects. Our discussion of the linear model is intended only to set the stage for our main topic, nonlinear measurement error models, and is far from complete. A comprehensive account of linear measurement error models can be found in Fuller (1987).

In addition to the background material on linear models, the problem of *attenuation* in nonlinear models is discussed.

2.2 Bias Caused by Measurement Error

Many textbooks contain a brief description of measurement error in linear regression, usually focusing on simple linear regression and arriving at the conclusion that the effect of measurement error is to bias the slope estimate in the direction of 0. Bias of this nature is commonly referred to as *attenuation* or *attenuation to the null.*

In fact though, even this simple conclusion has to be qualified, because it depends on the relationship between the measurement, $\mathbf{W}$, and the true predictor, $\mathbf{X}$, and possibly other variables in the regression model as well. In particular, the effect of measurement error depends upon the model under consideration and on the joint distribution of the measurement error and the other variables. In multiple linear regression, the effects of measurement error vary depending on: (i) the regression model, be it simple or multiple regression; (ii) whether or not the predictor measured with error

is univariate or multivariate; and (iii) the presence of bias in the measurement. The effects can range from the simple attenuation described above to situations where: (i) real effects are hidden; (ii) observed data exhibit relationships that are not present in the error-free data; and (iii) even the signs of estimated coefficients are reversed relative to the case with no measurement error.

The key point is that the measurement error distribution determines the effects of measurement error, and thus appropriate methods for correcting for the effects of measurement error depend on the measurement error distribution.

2.2.1 Simple Linear Regression with Additive Error

We start with the simple linear regression model $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, where $\mathbf{X}$ has mean μ_x and variance σ_x^2, and ϵ is independent of $\mathbf{X}$, has mean zero and variance σ_ϵ^2. The predictor $\mathbf{X}$ cannot be observed, and instead one observes $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $\mathbf{X}$, has mean zero, and variance σ_u^2. This is the classical additive measurement error model, where it is well-known that an ordinary least squares regression of $\mathbf{Y}$ on $\mathbf{W}$ is a consistent estimate not of β_x, but instead of $\beta_{x*} = \lambda\beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1. \tag{2.1}$$

Thus ordinary least squares regression of $\mathbf{Y}$ on $\mathbf{W}$ produces an estimator that is attenuated to 0. The attenuating factor, λ, is called the *reliability ratio* (Fuller, 1987).

One would expect that because $\mathbf{W}$ is an error-prone predictor, it has a weaker relationship with the response than does $\mathbf{X}$. This can be seen both by the attenuation, and also by the fact that the residual variance of this regression of $\mathbf{Y}$ on $\mathbf{W}$ is

$$\text{var}(\mathbf{Y}|\mathbf{W}) = \sigma_\epsilon^2 + \frac{\beta_x^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2}.$$

This facet of the problem is often ignored, but it is important. Measurement error causes a double-whammy: not only is the slope attenuated, but the data are more noisy, with an increased error about the line.

To illustrate the attenuation associated with the classical additive measurement error, the results of a small simulation are dis-

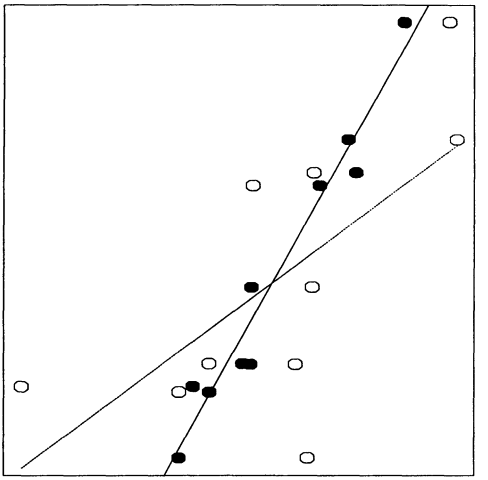

Figure 2.1. *Illustration of additive measurement error model. The filled circles are the true* $(\mathbf{Y}, \mathbf{X})$ *data and the steeper line is the least squares fit to these data. The empty circles and attenuated line are the observed* $(\mathbf{Y}, \mathbf{W})$ *data and the associated least squares regression line. For these data* $\sigma_x^2 = \sigma_u^2 = 1$, $(\beta_0, \beta_x) = (0, 1)$ *and* $\sigma_\epsilon^2 = .25$.

played in Figure 2.1.

Ten observations were generated with $\sigma_x^2 = \sigma_u^2 = 1$, $(\beta_0, \beta_x) = (0, 1)$ and $\sigma_\epsilon^2 = .25$. The filled circles and steeper line depict the true but unobservable data $(\mathbf{Y}, \mathbf{X})$ and the regression line of $\mathbf{Y}$ on $\mathbf{X}$. The empty circles and attenuated line depict the observed $(\mathbf{Y}, \mathbf{W})$ data and the linear regression of $\mathbf{Y}$ on $\mathbf{W}$.

2.2.2 Simple Linear Regression, More Complex Error Structure

Despite admonitions of Fuller (1987) and others to the contrary, it is a common perception that the effect of measurement error is always to attenuate the line, but in fact attenuation depends critically on the classical additive measurement error model. In this section, we discuss two deviations from the classical additive

error model that do not lead to attenuation.

We continue with the simple linear regression model, but now we make the error structure more complex in two ways. First, we will no longer insist that $\mathbf{W}$ be unbiased for $\mathbf{X}$. The intent of studying this departure from the classical additive error model is to study what happens when one pretends that one has an unbiased surrogate, but in fact the surrogate is biased.

A second departure from the additive model is to allow the errors in the linear regression model to be correlated with the errors in the predictors. One example where this problem arises naturally is in dietary calibration studies (Freedman, Carroll & Wax, 1991). In a typical dietary calibration study, one is interested in the relationship between a self-administered food frequency questionnaire (FFQ, the value of $\mathbf{Y}$) and usual (or long-term) dietary intake (the value of $\mathbf{X}$) as measures of, for example, the percentage of calories from fat in a person's diet. FFQ's are thought to be biased for usual intake, and in a calibration study researchers will obtain a second measure (the value of $\mathbf{W}$), typically either from a food diary or from an interview where the study subject reports their diet in the previous 24-hours. In this context, it is often assumed that the diary or recall is unbiased for usual intake. In principle, then, we have simple linear regression with an additive measurement error model, but in practice a complication can arise. It is often the case that the FFQ and the diary/recall are given very nearly contemporaneously in time, as in the Women's Health Trial Vanguard Study (Henderson, et al., 1990). In this case, it makes little sense to pretend that the error in the relationship between the FFQ ($\mathbf{Y}$) and usual intake ($\mathbf{X}$) is uncorrelated with the error in the relationship between a diary/recall ($\mathbf{W}$) and usual intake. This correlation has been demonstrated (Freedman, et al., 1991), and in this section we will discuss its effects.

To express the possibility of bias in $\mathbf{W}$, we write the model as $\mathbf{W} = \gamma_{0,\mathrm{em}} + \gamma_{1,\mathrm{em}}\mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $\mathbf{X}$ and has mean zero and variance σ_u^2. To express the possibility of correlated errors, we will write the correlation between ϵ and $\mathbf{U}$ as $\rho_{\epsilon u}$. The classical additive measurement error model sets $\gamma_{0,\mathrm{em}} = 0$, $\rho_{\epsilon u} = 0$ and $\gamma_{1,\mathrm{em}} = 1$, so that $\mathbf{W} = \mathbf{X} + \mathbf{U}$.

If $(\mathbf{X}, \epsilon, \mathbf{U})$ are jointly normally distributed, then the regression

of $\mathbf{Y}$ on $\mathbf{W}$ is linear with intercept

$$\beta_{0*} = \beta_0 + \beta_x\mu_x - \beta_{x*}(\gamma_{0,\text{em}} + \gamma_{1,\text{em}}\mu_x)$$

and slope

$$\beta_{x*} = \frac{\beta_x\gamma_{1,\text{em}}\sigma_x^2 + \rho_{\epsilon u}\sqrt{\sigma_\epsilon^2\sigma_u^2}}{\gamma_{1,\text{em}}^2\sigma_x^2 + \sigma_u^2}. \tag{2.2}$$

Examination of (2.2), shows that if $\mathbf{W}$ is biased ($\gamma_{1,\text{em}} \neq 1$) or if there is significant correlation between the measurement error and the error about the true line ($|\rho_{\epsilon u}| > 0$), it is possible for $|\beta_{x*}| > |\beta_x|$, an effect exactly the opposite of attenuation. Thus, correction for bias induced by measurement error clearly depends on the nature, as well as the extent of the measurement error.

For purposes of completeness, we note that the residual variance of the linear regression of $\mathbf{Y}$ on $\mathbf{W}$ is

$$\text{var}(\mathbf{Y}|\mathbf{W}) = \sigma_\epsilon^2 + \frac{\beta_x^2\sigma_u^2\sigma_x^2 - \rho_{\epsilon u}^2\sigma_\epsilon^2\sigma_u^2 - 2\beta_x\gamma_{1,\text{em}}\sigma_x^2\rho_{\epsilon u}\sqrt{\sigma_\epsilon^2\sigma_u^2}}{\gamma_{1,\text{em}}^2\sigma_x^2 + \sigma_u^2}.$$

2.2.3 Multiple Regression: Single Covariate Measured with Error

In multiple linear regression the effects of measurement error are more complicated, even for the classical additive error model.

We now consider the case where $\mathbf{X}$ is scalar, but there are additional covariates $\mathbf{Z}$ measured without error. The linear model is now

$$\mathbf{Y} = \beta_0 + \beta_x\mathbf{X} + \beta_z^t\mathbf{Z} + \epsilon, \tag{2.3}$$

where $\mathbf{Z}$ and β_z are column vectors, and β_z^t is a row vector. In the appendix it is shown that if $\mathbf{W}$ is unbiased for $\mathbf{X}$, and the measurement error $\mathbf{U}$ is independent of $\mathbf{X}$, $\mathbf{Z}$ and ϵ, then the least squares regression estimator of the coefficient of $\mathbf{W}$ consistently estimates $\lambda_1\beta_x$, where

$$\lambda_1 = \frac{\sigma_{x|z}^2}{\sigma_{w|z}^2} = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}, \tag{2.4}$$

and $\sigma_{w|z}^2$ and $\sigma_{x|z}^2$ are the (residual) variances of the regressions of $\mathbf{W}$ on $\mathbf{Z}$ and $\mathbf{X}$ on $\mathbf{Z}$, respectively. Note that λ_1 is equal to the simple linear regression attenuation $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ only when $\mathbf{X}$ and $\mathbf{Z}$ are uncorrelated.

The problem of measurement error-induced bias is not restricted to the regression coefficient of $\mathbf{X}$. The coefficient of $\mathbf{Z}$ is also biased in general, unless $\mathbf{Z}$ is independent of $\mathbf{X}$ (Carroll, et al., 1985; Gleser, et al., 1987). In the appendix it is shown that for the model (2.3), the naive ordinary least squares estimates not β_z but rather

$$\beta_{z*} = \beta_z + \beta_x(1 - \lambda_1)\Gamma_z, \tag{2.5}$$

where Γ_z^t is the coefficient of $\mathbf{Z}$ in the regression of $\mathbf{X}$ on $\mathbf{Z}$, i.e., $E(\mathbf{X} \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z^t\mathbf{Z}$.

This result has important consequences when interest centers on the effects of covariates measured without error. Carroll, et al. (1985) and Carroll (1989) show that in the two-group analysis of covariance where $\mathbf{Z}$ is a treatment assignment variable, naive linear regression produces a consistent estimate of the treatment effect only if the design is balanced, i.e., $\mathbf{X}$ has the same mean in both groups and is independent of treatment. With considerable imbalance, the naive analysis may lead to the conclusion that: (i) there is a treatment effect when none actually exists; and (ii) the effects are negative when they are actually positive, and vice-versa.

2.2.4 Multiple Covariates Measured with Error

Now suppose that there are covariates $\mathbf{Z}$ measured without error, that $\mathbf{W}$ is unbiased for $\mathbf{X}$ which may consist of multiple predictors, and that the linear regression model is $\mathbf{Y} = \beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z} + \epsilon$. If we write Σ_{ab} to be the covariance matrix between random variables $\mathbf{A}$ and $\mathbf{B}$, then naive ordinary linear regression consistently estimates not (β_x, β_z) but rather

$$\begin{aligned}
\begin{pmatrix} \beta_{x*} \\ \beta_{z*} \end{pmatrix} &= \begin{pmatrix} \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \\
&\qquad \left\{ \begin{pmatrix} \Sigma_{xy} \\ \Sigma_{zy} \end{pmatrix} + \begin{pmatrix} \Sigma_{u\epsilon} \\ 0 \end{pmatrix} \right\} \qquad (2.6) \\
&= \begin{pmatrix} \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \\
&\qquad \left\{ \begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \beta_x \\ \beta_z \end{pmatrix} + \begin{pmatrix} \Sigma_{u\epsilon} \\ 0 \end{pmatrix} \right\}.
\end{aligned}$$

Thus, ordinary linear regression is biased. We turn next to bias correction.

2.3 Correcting for Bias

As we have just seen, the ordinary least squares estimator is typically biased under measurement error, and the direction and magnitude of the bias depends on the regression model and the measurement error distribution. In this section, we describe two methods for eliminating bias that are commonly used.

2.3.1 Method of Moments

In simple linear regression with the classical additive error model, we have seen in (2.1) that ordinary least squares is an estimate of $\lambda\beta_x$; recall that λ is called the reliability ratio. If the reliability ratio were known, then one could obtain a proper estimate of β_x simply by dividing the ordinary least squares slope β_{x*} by the reliability ratio.

Of course, the reliability ratio is rarely known in practice, and one has to estimate it. If $\widehat{\sigma}_u^2$ is an estimate of the measurement error variance (this is discussed in section 3.4), and if $\widehat{\sigma}_w^2$ is the sample variance of the $\mathbf{W}$'s, then a consistent estimate of the reliability ratio is $\widehat{\lambda} = (\widehat{\sigma}_w^2 - \widehat{\sigma}_u^2)/\widehat{\sigma}_w^2$. The resulting estimate is $\beta_{x*}/\widehat{\lambda}$.

In small samples the sampling distribution of $\widehat{\beta}_{x*}/\widehat{\lambda}$ is highly skewed, and in such cases a modified version of the method-of-moments estimator is recommended (Fuller, 1987).

The algorithm described above is called the *method-of-moments* estimator. The terminology is apt, because ordinary least squares and the reliability ratio depend only on moments of the observed data.

The method-of-moments estimator can be constructed for the general linear model, and not just for simple linear regression. Suppose that $\mathbf{W}$ is unbiased for $\mathbf{X}$, and consider the general linear regression model with $\mathbf{Y} = \beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z} + \epsilon$. The ordinary least squares estimator is biased even in large samples because it estimates (2.6).

When Σ_{uu} and $\Sigma_{u\epsilon}$ are known or can be estimated, (2.6) can be used to construct a simple method-of-moments estimator that is commonly used to correct for the bias. Let S_{ab} be the sample covariance between random variables $\mathbf{A}$ and $\mathbf{B}$. The method-of-moments estimator that corrects for the bias in the case that Σ_{uu}

and $\Sigma_{u\epsilon}$ are known is

$$\begin{pmatrix} S_{ww} - \Sigma_{uu} & S_{wz} \\ S_{zw} & S_{zz} \end{pmatrix}^{-1} \begin{pmatrix} S_{wy} - \Sigma_{u\epsilon} \\ S_{zy} \end{pmatrix}, \tag{2.7}$$

In the case that Σ_{uu} and $\Sigma_{u\epsilon}$ are estimated, the estimates replace the known values in (2.7). It is often reasonable to assume that $\Sigma_{u\epsilon} = 0$. in which case (2.7) simplifies accordingly.

In the event that $\mathbf{W}$ is biased for $\mathbf{X}$, i.e., $\mathbf{W} = \gamma_{0,\mathrm{em}} + \gamma_{x,\mathrm{em}}\mathbf{X} + \mathbf{U}$, i.e., the error calibration model, the method-of-moments estimator can still be used provided estimates of $(\gamma_{0,\mathrm{em}}, \gamma_{x,\mathrm{em}})$ are available. The strategy is to calculate the estimators above using the error-calibrated variate $\mathbf{W}_* = \widehat{\gamma}_{x,\mathrm{em}}^{-1}(\mathbf{W} - \widehat{\gamma}_{0,\mathrm{em}})$.

2.3.2 Orthogonal Regression

Another well publicized method for linear regression in the presence of measurement error is *orthogonal regression*; see Fuller (1987, section 1.3.3). However, for reasons given below, we are skeptical about the general utility of orthogonal regression, in large part because it is so easily misused. Although not fundamental to understanding later material on nonlinear models, we take the opportunity to discuss orthogonal regression at length here in order to emphasize the potential pitfalls associated with it. This section can be skipped by those who are interested only in estimation for nonlinear models.

Let $\mathbf{Y} = \beta_0 + \beta_x\mathbf{X} + \epsilon$ and $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where ϵ and $\mathbf{U}$ are uncorrelated. Whereas the method-of-moments estimator (section 2.3) requires knowledge or estimability of the measurement error variance σ_u^2, orthogonal regression requires the same for the ratio $\eta = \sigma_\epsilon^2/\sigma_u^2$.

The orthogonal regression estimator minimizes the orthogonal distance of $(\mathbf{Y}, \mathbf{W})$ to the line $\beta_0 + \beta_x\mathbf{X}$, weighted by η, i.e., it minimizes

$$\sum_{i=1}^{n} \left\{ (\mathbf{Y}_i - \beta_0 - \beta_x x_i)^2 + \eta\,(\mathbf{W}_i - x_i)^2 \right\} \tag{2.8}$$

in the unknown parameters $(\beta_0, \beta_x, x_1, ..., x_n)$.

In fact (2.8) is the sum of squared orthogonal distances between the points $(\mathbf{Y}_i, \mathbf{W}_i)_1^n$, and the line $y = \beta_0 + \beta_x x$, only in the special case that $\eta = 1$. However, the term orthogonal regression is used

$\mathbf{W}_i$	$\mathbf{Y}_{i1}$	$\mathbf{Y}_{i2}$
−1.8007	−0.5558	−0.9089
−0.7717	0.2076	0.6499
−0.4287	−1.7365	−1.8542
−0.0857	−0.9018	0.2040
0.2572	−0.2312	−0.3097
0.6002	0.2967	0.5072
0.9432	0.5928	1.5381
1.2862	1.2420	1.2599

Table 2.1. *Orthogonal regression example with replicated response.*

to describe the method regardless of the value of $\eta < \infty$.

The orthogonal regression estimator is the functional maximum likelihood estimator (sections 1.2 and 6.1) assuming that $(\mathbf{X}_1, ..., \mathbf{X}_n)$ are unknown fixed constants, and that the errors $(\epsilon, \mathbf{U})$ are independent and normally distributed.

Orthogonal regression has the appearance of greater applicability than method-of-moments estimation in that only the ratio, η, of the error variances need be known or estimated. However, it is our experience that in the majority of problems η cannot be specified or estimated correctly, and use of orthogonal regression with an improperly specified value of η often results in an unacceptably large *over correction* for attenuation due to measurement error.

We illustrate the problem with some data from a consulting problem (Table 2.1). The data include two measurements of a response variable, $\mathbf{Y}_{i1}$ and $\mathbf{Y}_{i2}$, and one predictor variable, $\mathbf{X}_i$, $i = 1, \ldots, 8$. The data are proprietary and we cannot disclose the nature of the application. Accordingly, all of the variables have been standardized to have sample means and variances 0 and 1 respectively.

We take as the response variable to be used in the regression analysis, $\mathbf{Y}_i = (\mathbf{Y}_{i1} + \mathbf{Y}_{i2})/2$, the average of the two response measurements.

From an independent experiment it had been estimated that

$\sigma_u^2 \approx 0.0424$, also after standardization. Because the sample standard deviation of $\mathbf{W}$ is 1.0, measurement error induces very little bias here. The estimated reliability ratio is $\widehat{\lambda} = 1/1.0424 \approx 0.96$ and so attenuation is only about 4%. The ordinary least squares estimated slope from regressing the average of the responses on $\mathbf{W}$ is 0.65, while the method-of-moments slope estimate is $\widehat{\lambda}^{-1}0.65 \approx 0.68$.

In a first analysis of these data, our client thought that orthogonal regression was an appropriate method for these data. A components-of-variance analysis resulted in the estimate 0.0683 for the response measurement error variance. If η is estimated by $\widehat{\eta} = 0.0683/0.0424 \approx 1.6118$, then the resulting orthogonal regression slope estimate is 0.88.

The difference in these two estimates, $|0.88 - 0.68|$, is larger than would be expected from random variation alone. Clearly something is amiss. The method-of-moments correction for attenuation is only $\widehat{\lambda}^{-1} \approx 1.04$, whereas orthogonal regression in effect, produces a correction for attenuation of approximately $1.35 \approx 0.88/0.65$.

The problem lies in the nature of the regression model error ϵ, that is typically the sum of two components: (i) ϵ_M, the measurement error in determination of the response; and (ii) ϵ_L, the *equation error*, i.e., the variation about the regression line of the true response in the absence of measurement error.

If we have replicated measurements, $\mathbf{Y}_{ij}$, of the true response, then $\mathbf{Y}_{ij} = \beta_0 + \beta_x \mathbf{X}_i + \epsilon_{L,i} + \epsilon_{M,ij}$, and of course their average is $\overline{\mathbf{Y}}_{i\cdot} = \beta_0 + \beta_x \mathbf{X}_i + \epsilon_{L,i} + \bar{\epsilon}_{M,i\cdot}$. Here and throughout the book a subscript "dot" and over bar means averaging. For example, with k replicates,

$$\overline{\mathbf{Y}}_{i\cdot} = k^{-1} \sum_{j=1}^{k} \mathbf{Y}_{ij}; \quad \bar{\epsilon}_{M,i\cdot} = k^{-1} \sum_{j=1}^{k} \epsilon_{M,ij}.$$

The components of variance analysis estimates *only* the variance of the average measurement error $\bar{\epsilon}_{M,i\cdot}$ in the responses, but completely ignores the variability, $\epsilon_{L,i}$, about the line. The net effect is to under estimate η and thus overstate the correction required of the ordinary least squares estimate, because $\text{var}(\bar{\epsilon}_{M,i\cdot})/\sigma_u^2$ is used as the estimate of η instead of the larger, appropriate value $\{\text{var}(\bar{\epsilon}_{M,i\cdot}) + \text{var}(\epsilon_{L,i})\}/\sigma_u^2$.

The naive use of orthogonal regression on the data in Table

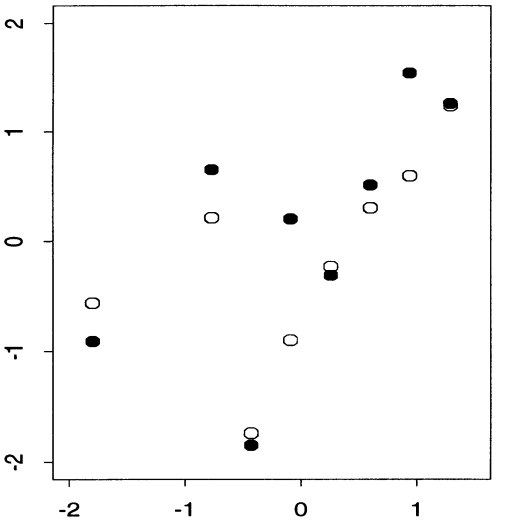

Figure 2.2. *Illustration of the dangers of orthogonal regression. The filled and empty circles represent replicated values of the response. Note the evidence of equation error.*

2.1 has assumed that there is no additional variability about the line in addition to that due to measurement error in the response, i.e., $\epsilon_{L,i} = 0$. To check this, refer to Figure 2.2. Each replicated response is indicated by a solid and filled circle. Remember that there is little measurement error in $\mathbf{W}$. In addition, the replication analysis suggested that the standard deviation of the replicates was less than 10% of the variability of the responses. Thus, in the absence of equation error we would expect to see the replicated pairs falling along a clearly delineated straight line. This is far from the case, suggesting that the equation error $\epsilon_{L,i}$ is a large part of the variability of the responses. Indeed, while the replication analysis suggests that $\text{var}(\bar{\epsilon}_{M,i\cdot}) \approx 0.0683$, a method-of-moments analysis suggests $\text{var}(\epsilon_{L,i}) \approx 0.4860$.

Fuller (1987) was one of the first to emphasize the importance of equation error. In our experience, outside of some special laboratory validation studies, equation error is almost always important in

linear regression. In this majority of cases, orthogonal regression is an inappropriate technique, unless estimation of both the response measurement error and the equation error is possible.

In some cases, $\mathbf{Y}$ and $\mathbf{W}$ are measured in the same way, e.g., if they are both blood pressure measurements. Here, it is often entirely reasonable to *assume* that the variance of ϵ_M equals σ_u^2, and then there is a temptation to ignore equation error and hence set $\eta = 1$. This temptation is especially acute when replicates are absent, so that σ_u^2 cannot be estimated and the method-of-moments estimator cannot be used.

2.4 Bias Versus Variance

Estimates which do not account for measurement error are typically biased. Correcting for this bias entails what is often referred to as a *bias versus variance* tradeoff. What this means is that in most problems, the very nature of correcting for bias is that the resulting corrected estimator will be more variable than the biased estimator. Of course, when an estimator is more variable, the confidence intervals associated with it become longer.

We will discuss this in detail for linear regression, but the bias versus variance tradeoff occurs far more generally. For example, Rosner, Willett & Spiegelman (1989) describe a problem in logistic regression, where the response is the development of breast cancer, and the predictor measured with error is daily saturated fat intake (adjusted for caloric intake). Ignoring measurement error, they obtained an estimated odds ratio (the exponential function of the logistic regression slope) for saturated fat of 0.92, with a 95% confidence interval from 0.80 to 1.05. Having corrected for measurement error, the estimated odds ratio becomes 0.83 with a confidence interval from 0.61 to 1.12. Note the key point: by correcting for error, the length of the confidence interval was increased (by inference, this means that the corrected estimator is more variable).

In this section, we will illustrate the bias versus variance tradeoff theoretically in simple linear regression. This material is somewhat technical, and readers may skip it without any loss of understanding of the main points of measurement error models.

Consider the simple linear regression model, $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, with additive independent measurement error, $\mathbf{W} = \mathbf{X} + \mathbf{U}$,

under the simplifying assumption of joint normality of $\mathbf{X}$, $\mathbf{U}$ and ϵ. Further, suppose that the reliability ratio λ in (2.1) is known. We make this assumption only to simplify the discussion in this section. Generally in applications it is seldom the case that this parameter is known, although there are exceptions (Fuller, 1987).

Let $\widehat{\beta}_{x*}$ denote the least squares estimate of slope from the regression of $\mathbf{Y}$ on $\mathbf{W}$. We know that its mean is $E(\widehat{\beta}_{x*}) = \lambda\beta_x$. Denote its variance by σ_*^2.

The method-of-moments estimator of β_x, is $\widehat{\beta}_{x,mm} = \lambda^{-1}\widehat{\beta}_{x*}$ and has mean $E(\widehat{\beta}_{x,mm}) = \beta_x$, and variance $\mathrm{Var}(\widehat{\beta}_{x,mm}) = \lambda^{-2}\sigma_*^2$.

Because $\lambda < 1$ it is clear that while the correction-for-attenuation in $\widehat{\beta}_{x,mm}$ reduces its bias to 0, there is an increase in variability relative to the variance of the biased estimator $\widehat{\beta}_{x*}$.

The price for reduced bias is increased variance. This phenomenon is not restricted to the simple model and estimator in this section, but occurs with almost universal generality in the analysis of measurement error models. In cases where the absence of bias is of paramount importance, then there is usually no escaping the increase in variance. In cases where some bias can be tolerated then consideration of mean squared error is necessary.

In the following material, we indicate that there are compromise estimators which may outperform both uncorrected and corrected estimators, at least in small samples. Surprisingly, outside of the work detailed in Fuller (1987), such compromise estimators have not been much investigated, especially for nonlinear models.

Remember that mean squared error (MSE) is the sum of the variance plus the square of the bias. This is an interesting criterion to use, because uncorrected estimators have more bias but smaller variance than corrected estimators, and the bias versus variance tradeoff is transparent. Note that

$$\begin{aligned} \mathrm{MSE}\left(\widehat{\beta}_{x*}\right) &= \sigma_*^2 + (1-\lambda)^2\beta_x^2; \text{ and} \\ \mathrm{MSE}\left(\widehat{\beta}_{x,mm}\right) &= \lambda^{-2}\sigma_*^2. \end{aligned} \tag{2.9}$$

It follows that

$$\mathrm{MSE}\left(\widehat{\beta}_{x,mm}\right) < \mathrm{MSE}\left(\widehat{\beta}_{x*}\right)$$

if and only if

$$\sigma_*^2 < \frac{\lambda^2(1-\lambda)\beta_x^2}{1+\lambda}.$$

Because σ_*^2 decreases with increasing sample size we can conclude that in sufficiently large samples it is always beneficial, in terms of mean squared error, to correct for attenuation due to measurement error.

Consider now the alternative estimator $\widehat{\beta}_{x,a} = a\beta_{x*}$ for a fixed constant a. The mean squared error of this estimator is $a^2\sigma_*^2 + (a\lambda - 1)^2\beta_x^2$, which is minimized when $a = a_* = \lambda\beta_x^2/(\sigma_*^2 + \lambda^2\beta_x^2)$. Ignoring the fact that a_* depends on unknown parameters we consider the "estimator" $\widehat{\beta}_{x,*} = a_*\beta_{x*}$, which has smaller mean squared error than either $\widehat{\beta}_{x,mm}$ or $\widehat{\beta}_{x*}$. Note that as $\sigma_*^2 \to 0$, $a_* \to \lambda^{-1}$.

The estimator $\widehat{\beta}_{x,*}$ achieves its mean-squared-error superiority by making a partial correction for attenuation in the sense that $a_* < \lambda^{-1}$. This simple exercise illustrates that estimators that make only partial corrections for attenuation can have good mean-squared-error performance.

We make one final use of the simple model and estimator in this section. Note that for testing the null hypothesis $H_0 : \beta_x = 0$, the test statistic obtained by dividing the parameter estimate by its standard error is exactly the same regardless of which estimator, $\widehat{\beta}_{x*}$ or $\widehat{\beta}_{x,mm}$, is used. In other words, the correction for attenuation has no effect on the power to detect the presence of a linear relationship.

Although we have used a simple model and a somewhat artificial estimator to facilitate the discussion of bias and variance, all of the conclusions made above hold, at least to a very good approximation, in general for both linear and nonlinear regression measurement error models.

2.5 Attenuation in General Problems

We have already seen that with multiple covariates, even in linear regression the effects of measurement error are complex, and not easily described. In this section, we provide a brief overview of what happens in nonlinear models.

Consider a scalar covariate $\mathbf{X}$ measured with error, and suppose that there are no other covariates. In the classical error model

for simple linear regression we have seen that the bias caused by measurement error is always in the form of attenuation, so that ordinary least squares preserves the sign of the regression coefficient asymptotically, but is biased towards zero. Attenuation is a consequence then of (i) the simple linear regression model; and (ii) the classical additive error model. Without (i)–(ii), the effects of measurement error are more complex; we have already seen that attenuation may not hold if (ii) is violated.

In logistic regression when $\mathbf{X}$ is measured with additive error, attenuation does not always occur (Stefanski & Carroll, 1985), but it is typical. More generally, in most problems with a scalar $\mathbf{X}$ and no covariates $\mathbf{Z}$, the underlying *trend* between $\mathbf{Y}$ and $\mathbf{X}$ is preserved under nondifferential measurement error, in the sense that the correlation between $\mathbf{Y}$ and $\mathbf{W}$ is positive whenever both $E(\mathbf{Y}|\mathbf{X})$ and $E(\mathbf{W}|\mathbf{X})$ are increasing functions of $\mathbf{X}$ (Weinberg, et al., 1993). Technically, this follows because with nondifferential measurement error, $\mathbf{Y}$ and $\mathbf{W}$ are uncorrelated given $\mathbf{X}$, and hence the covariance between $\mathbf{Y}$ and $\mathbf{W}$ is just the covariance between $E(\mathbf{Y}|\mathbf{X})$ and $E(\mathbf{W}|\mathbf{X})$.

Positively, this result says that for the very simplest of problems (scalar $\mathbf{X}$, no covariates $\mathbf{Z}$ measured without error) the general trend in the data is typically unaffected by nondifferential measurement error. However, the result illustrates only part of a complex picture, because it describes only the *correlation* between $\mathbf{Y}$ and $\mathbf{W}$, and says nothing about the structure of this relationship.

For example, one might expect that if the regression $E(\mathbf{Y}|\mathbf{X})$ of $\mathbf{Y}$ on $\mathbf{X}$ is nondecreasing in $\mathbf{X}$, and if $\mathbf{W} = \mathbf{X} + \mathbf{U}$ where $\mathbf{U}$ is independent of $\mathbf{X}$ and $\mathbf{Y}$, then the regression of $\mathbf{Y}$ on $\mathbf{W}$ would also be nondecreasing. But Hwang & Stefanski (1994) have shown that this need not be the case, although it is true in linear regression normally distributed measurement error. However, these results show that the problem of making inferences about details in the relationship of $\mathbf{Y}$ and $\mathbf{X}$, based on the observed relationship between $\mathbf{Y}$ and $\mathbf{W}$, is a difficult problem in general.

There are other practical reasons why ignoring measurement error is not acceptable. First, estimating the direction of the relationship between $\mathbf{Y}$ and $\mathbf{X}$ correctly is nice, but as emphasized by MacMahon, et al. (1990) we can be misled if we severely underestimate its magnitude. Secondly, the result does not apply to multiple covariates. Indeed, we have already seen that in multiple linear re-

gression under the additive measurement error model, the observed and underlying trends may be entirely different. Finally, it is also the case (section 11.1) that especially with multiple covariates one can use error modeling to improve the power of inferences. In large classes of problems then, there is simply no alternative to careful consideration of the measurement error structure.

2.5.1 An Illustration of Nondifferential Measurement Error

To show that trends are not always preserved under nondifferential measurement error, we consider the following theoretical example (Dosemeci, Wacholder & Lubin, 1990). Suppose that the 924 subjects are exposed at no ($\mathbf{X} = 0$), low ($\mathbf{X} = 1$) and high ($\mathbf{X} = 2$) levels to a harmful substance. Suppose the chance of an adverse outcome is 1/2, 2/3 and 6/7 for no, low and high exposures, while the chances of the exposures themselves are .0059347, .8902077 and .1038576, respectively. If true exposure could be ascertained, the expected outcomes would be as in Table 2.2 in the section labeled **TRUE**. If we were to do a regression of $\mathbf{Y}$ on the dummy variables $\mathbf{X}_1$ indicating low exposure ($\mathbf{X}_1 = 1$), and $\mathbf{X}_2$ indicating high exposure ($\mathbf{X}_2 = 1$), then the true logistic regression parameters for $\mathbf{X}_1$ and $\mathbf{X}_2$ are $\log(2) = .69$ and $\log(6) = 1.79$, respectively, indicating that the two higher exposure levels have response rates higher than the response rate associated with the no-exposure level. The true data clearly indicate a harmful effect due to exposure.

Now suppose, however, that measurement error (in this case misclassification) occurs, so that 40% of those truly at high exposure are misclassified into the no exposure group, and 40% of those truly at low exposure are misclassified into the high exposure group. Let $\mathbf{W}$ be the resulting variable taking on the three observed levels of exposure, with corresponding dummy variables $\mathbf{W}_1$ and $\mathbf{W}_2$. This is a theoretical example, of course, and one can criticize it for not being particularly realistic, but it is an example of nondifferential measurement error. The observed data we expect to see using the surrogates $\mathbf{W}_1$ and $\mathbf{W}_2$ are also given in Table 2.2.

The observed logistic regression parameters for $\mathbf{W}_1$ and $\mathbf{W}_2$ are $\log(.46) = -.78$ and $\log(.53) = -.63$, respectively, indicating that the two higher exposure levels have response rates lower than the response rate associated with the no-exposure level. The observed data suggest a beneficial effect due to exposure!

Disease Status	Exposure = None	Exposure = Low	Exposure = High
True			
$\mathbf{Y} = 1$	4	800	120
$\mathbf{Y} = 0$	4	400	20
Observed			
$\mathbf{Y} = 1$	52	480	392
$\mathbf{Y} = 0$	12	240	172

Table 2.2. *A hypothetical logistic regression example with nondifferential measurement error. The entries are the expected counts. The true logistic parameters for dummy variables low and high exposure are log(2) and log(6), respectively, while the observed coefficients for the error prone data are log(.46) and log(.53), respectively.*

For this example the sufficient condition of Weinberg, et al. (1993) is violated. We have that $E(\mathbf{W}|\mathbf{X} = 0) = 0$, $E(\mathbf{W}|\mathbf{X} = 1) = 1.4$, and $E(\mathbf{W}|\mathbf{X} = 2) = 1.2$, which is not increasing in $\mathbf{X}$. The results is that the trend in the true data is obscured by the nondifferential misclassification.

2.6 Other References

The linear regression problem has a long history and continues to be the subject of research. Excellent historic background can be found in the papers by Lindley (1953), Lord (1960), Cochran (1968) and Madansky (1969). Further more technical analyses are given by Fuller (1980), Carroll & Gallo (1982, 1984), Carroll, Gallo & Gleser (1985). Diagnostics are discussed by Carroll & Spiegelman (1986, 1992) and Cheng & Tsai (1992). Robustness is discussed by Ketellapper & Ronner (1984), Zamar (1988, 1992), Cheng & van Ness (1988) and Carroll, Eltinge & Ruppert (1993). Ganse, Amemiya & Fuller (1983) discuss an interesting prediction problem. Hwang (1986) and Hasenabeldy, Fuller & Ware (1989) discuss problems with unusual error structure. Boggs, et al. (1988) discusses com-

putational aspects of orthogonal regression in nonlinear models.

2.7 Appendix

Here we establish (2.4) and (2.5) under the assumption of multivariate normality.

Taking expectations of both sides of (2.3) conditional on $(\mathbf{X}, \mathbf{Z})$ leads to the identity

$$E(\mathbf{Y} \mid \mathbf{W}, \mathbf{Z}) = \beta_0 + \beta_x E(\mathbf{X} \mid \mathbf{W}, \mathbf{Z}) + \beta_z^t \mathbf{Z}. \tag{2.10}$$

Under joint normality the regression of $\mathbf{X}$ on $(\mathbf{W}, \mathbf{Z})$ is linear. To facilitate the derivation we parameterize this as

$$\begin{aligned} E(\mathbf{X} \mid \mathbf{W}, \mathbf{Z}) = \gamma_0 + \gamma_w \{\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})\} \\ + \gamma_z^t \{\mathbf{Z} - E(\mathbf{Z})\}. \end{aligned} \tag{2.11}$$

Because of the orthogonalization in (2.11) it is immediate that

$$\begin{aligned} \gamma_w &= \frac{E\left(E\left[\mathbf{X}\{\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})\} \mid \mathbf{Z}\right]\right)}{E\left(E\left[\{\mathbf{W} - E(\mathbf{W} \mid \mathbf{Z})\}^2 \mid \mathbf{Z}\right]\right)} \\ &= \frac{E\{E(\mathbf{XW} \mid \mathbf{Z}) - E(\mathbf{X} \mid \mathbf{Z})E(\mathbf{W} \mid \mathbf{Z})\}}{\sigma^2_{w|z}}, \end{aligned} \tag{2.12}$$

where $\sigma^2_{w|z} = \mathrm{var}(\mathbf{W} \mid \mathbf{Z})$.

Now because $\mathbf{U}$ is independent of $\mathbf{Z}$, $E(\mathbf{W} \mid \mathbf{Z}) = E(\mathbf{X} \mid \mathbf{Z})$, $E(\mathbf{XW} \mid \mathbf{Z}) = E(\mathbf{X}^2 \mid \mathbf{Z})$, and the numerator in (2.12) is just $\sigma^2_{x|z}$. Independence of $\mathbf{U}$ and $\mathbf{Z}$ also implies that $\sigma^2_{w|z} = \sigma^2_{x|z} + \sigma^2_u$. It follows that

$$\gamma_w = \frac{\sigma^2_{x|z}}{\sigma^2_{w|z}} = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u} \tag{2.13}$$

as claimed.

Suppose now that $E(\mathbf{X} \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z^t \mathbf{Z}$. As noted previously $E(\mathbf{W} \mid \mathbf{Z}) = E(\mathbf{X} \mid \mathbf{Z})$, and thus $E(\mathbf{W} \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z^t \mathbf{Z}$ also.

Again because of the orthogonalization in (2.11) it is immediate that $\gamma_z = \Gamma_z$.

If we now replace $E(\mathbf{W} \mid \mathbf{Z})$ with $\Gamma_0 + \Gamma_z^t \mathbf{Z}$ in (2.11), and substitute the right hand side of (2.11) into (2.10), and then collect coefficients of $\mathbf{Z}$ using the definition of (2.13), we find that the

coefficient of $\mathbf{Z}$ in (2.10) is

$$\beta_{z*}^t = \beta_z^t + \beta_x(1 - \lambda_1)\Gamma_z^t. \tag{2.14}$$

CHAPTER 3

REGRESSION CALIBRATION

3.1 Overview

In this monograph we will describe two simple, generally applicable approaches to measurement error analysis, regression calibration in this chapter and simulation extrapolation (SIMEX) in Chapter 4.

The basis of regression calibration is the replacement of $\mathbf{X}$ by the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$. After this approximation, one performs a standard analysis. This *regression calibration* algorithm was suggested as a general approach by Carroll & Stefanski (1990) and Gleser (1990). Prentice (1982) pioneered the idea for the proportional hazard model, where it is still the default option, and a modification of it has been suggested for this topic by Clayton (1991); see section 14.6. Armstrong (1985) suggests regression calibration for generalized linear models, and Fuller (1987, pp 261–262) briefly mentions the idea. Rosner, Willett & Spiegelman (1989, 1990) have developed the idea for logistic regression into a workable and popular methodology, complete with a good computer program. In some special cases, regression calibration is equivalent to the classical method of moments bias correction; see section 3.4.2.

Regression calibration is simple and potentially applicable to any regression model, provided the approximation is sufficiently accurate. SIMEX shares these advantages but is more computationally intensive. The simplicity of regression calibration is somewhat mitigated by the need to develop and fit a calibration model to the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$. Calibration modeling is discussed in section 3.4.

There are two justifications of the regression calibration approximation:

- For some models, e.g., loglinear mean models (section 3.9.3)

and linear regression when the variance of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is constant (section 3.9.1), the regression calibration approximation is exact except for a change in the intercept parameter. For logistic regression, in many cases the approximation is almost exact (section 3.9.2).

- The approximation can be developed using a Taylor series expansion, assuming that the measurement error variance is small. By taking extra terms in the Taylor series, refined approximations, called expanded regression calibration models, are possible. See sections 3.8 and 3.6.

In section 3.2 the basic algorithm is given. We give a first example, the NHANES data, in section 3.3. Basic to the algorithm is a model for $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ and methods of fitting such models are discussed in section 3.4. Section 3.5 provides details of calculating standard errors. The expanded regression calibration approximation in section 3.6 attempts to improve the basic regression calibration approximation; the following section includes a second example, the bioassay data. Sections 3.8, 3.9 and 3.10 are devoted to theoretical justification of regression calibration and expanded regression calibration. Technical details are relegated to the appendix, section 3.12.

3.2 The Regression Calibration Algorithm

The regression calibration algorithm is as follows:

- Using replication, validation or instrumental data, estimate the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$, $m(\mathbf{Z}, \mathbf{W}, \gamma_{\rm cm})$, depending on parameters $\gamma_{\rm cm}$ which are estimated by $\widehat{\gamma}_{\rm cm}$. Here "cm" stands for "regression calibration model".
- Replace the unobserved $\mathbf{X}$ by its estimate $m(\mathbf{Z}, \mathbf{W}, \widehat{\gamma}_{\rm cm})$, and then run a standard analysis to obtain parameter estimates.
- Adjust the resulting standard errors to account for the estimation of $\gamma_{\rm cm}$, using either the bootstrap or sandwich method, consult Appendix A for the definition of these techniques.

Suppose for example that the mean of $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z})$ can be described by $f(\mathbf{Z}, \mathbf{X}, \mathcal{B})$ for some unknown parameter $\mathcal{B}$. The replacement of $\mathbf{X}$ by its estimated value in effect proposes a modified model for the observed data, namely

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) \approx f\{\mathbf{Z}, m(\mathbf{Z}, \mathbf{W}, \gamma_{\rm cm}), \mathcal{B}\}. \tag{3.1}$$

It is important to emphasize that the regression calibration model (3.1) is an approximate, working model for the observed data. It is not necessarily the same as the actual mean and variance function for the observed data, but in many cases is only modestly different. Even as an approximation, the regression calibration model can be improved, see section 3.6 for refinements.

3.2.1 Correction for Attenuation

The simplest form of regression calibration is the "correction for attenuation" used in linear regression. It is easiest to describe in the following situation:

(i) $\mathbf{X}$ is a scalar;
(ii) The measurement error is additive ($\mathbf{W} = \mathbf{X} + \mathbf{U}$) with error variance σ_u^2 estimated by $\widehat{\sigma}_u^2$ (section 3.4);
(iii) The covariates $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ are jointly normally distributed;
(iv) As in logistic regression and generalized linear models, the response is affected only by a linear combination of the predictors, namely $\beta_0 + \beta_x \mathbf{X} + \beta_z^t \mathbf{Z}$. This might be linear, logistic, probit or loglinear regression.

For estimating the effect of $\mathbf{X}$, namely β_x, the regression calibration estimator is formed by three steps:

- Let $\widehat{\beta}_x(\text{naive})$ be the naive estimator formed by ignoring measurement error;
- Let $\widehat{\sigma}_{w|z}^2$ be the regression mean squared error from a linear regression of $\mathbf{W}$ on $\mathbf{Z}$. This is the sample variance of the $\mathbf{W}$'s if there are no other covariates $\mathbf{Z}$;
- The regression calibration estimator is $\widehat{\beta}_x(\text{naive})\widehat{\sigma}_{w|z}^2/(\widehat{\sigma}_{w|z}^2 - \widehat{\sigma}_u^2)$.

In section 3.4, we discuss how to implement regression calibration when one wants to estimate β_z, when $\mathbf{X}$ is multivariate, for nonnormally distributed data, and when the measurement error is not additive.

3.3 NHANES Example

In this section, we consider the analysis of the NHANES–I Epidemiologic Study Cohort data set (Jones, et al., 1987). The predictor variables $\mathbf{Z}$ that are assumed to have been measured without appreciable error are age, poverty index ratio, body mass index, use

of alcohol (yes–no), family history of breast cancer, age at menarche (a dummy variable taking on the value 1 if the age is ≤ 12) and menopausal status (pre or post). The variable measured with error, $\mathbf{X}$, is daily intake of saturated fat (in grams). The response is breast cancer incidence. The analysis in this section is restricted to 3,145 women aged 25–50 with complete data on all the variables listed above; 59 had breast cancer. In general, logistic regression analyses with a small number of disease cases are very sensitive to misclassification, case deletion, etc.

Saturated fat was measured via a 24-hour recall, i.e., a participant's diet in the previous 24 hours was recalled and nutrition variables computed. It is measured with considerable error (Beaton, et al., 1979; Wu, et al., 1986), leading to considerable controversy as regards their use to assess breast cancer risk (Prentice, et al., 1989; Willett, et al., 1987).

Our analysis concerns the effect of saturated fat on risk of breast cancer, adjusted for the other variables. To give a first indication of the effects, we considered the marginal effect of saturated fat. Specifically, we considered the variable log(5+saturated fat) and computed kernel density estimates (Silverman, 1986) of this variable for the breast cancer cases and for the noncases. The transformation was chosen for illustrative purposes and because it makes the observed values nearly normally distributed. The results are given in Figure 3.1. Note that this figure indicates a small marginal *but protective* effect due to higher levels of saturated fat in the diet, which is in opposition to one popular hypothesis. Thus we should expect the logistic regression coefficient of saturated fat to be negative (hence, the higher the levels of fat, the lower the estimated risk of breast cancer).

In Table 3.1 we list the result of ignoring measurement error. This analysis suggests that transformed saturated fat is a highly significant predictor of risk with a negative logistic regression coefficient. From Chapter 11, the p-value is asymptotically valid because there are no other covariates measured with error.

There are at least two problems with these data that suggest that the results should be treated with extreme caution.

The first reason is simple sensitivity analysis, and has nothing to do with measurement error in the predictors. If we change the three individuals with the highest levels of fat from non-breast cancer cases to breast cancer cases, the logistic regression estimates

changed from the original -0.97 to -0.53 with nominal p-values becoming 0.061. Thus changing only three observations, less than 0.1% of the total data (alternatively, increasing by only 5% the number of breast cancer cases) nearly halved the logistic regression parameter estimate and changed the p-value from highly statistically significant to nonsignificant at the 0.05 level. The point here is that misclassification of breast cancer cases, or loss to follow-up of breast cancer cases with high fat intakes, can affect the final analysis enormously.

By using data from the Continuing Survey of Food Intake by Individuals (CSFII, see Thompson, et al., 1992), we estimate that over 75% of the variance of a single 24-hour recall is made up of measurement error (this analysis is fairly involved and is discussed in the appendix, section 3.12.1). In other words, there is more noise than signal in a single 24-hour recall. There seems to us to be almost no wisdom of putting much trust in such an unexplained outcome as a negative coefficient for saturated fat when the observed predictor is mostly noise.

In fact, the observed sample variance of $\mathbf{W}$ is 0.233, and for the additive measurement error model, the measurement error variance is estimated as $\widehat{\sigma}_u^2 = 0.171$. The mean squared error from the linear regression of $\mathbf{W}$ on $\mathbf{Z}$ is $\widehat{\sigma}_{w|z}^2 = 0.217$. The correction for attenuation estimate for the effect of transformed fat is thus

$$\widehat{\beta}_x = \frac{\widehat{\sigma}_{w|z}^2 \widehat{\beta}_x(\text{naive})}{\widehat{\sigma}_{w|z}^2 - \widehat{\sigma}_u^2} = \frac{0.217 \times (-0.97)}{0.217 - 0.171} = -4.67.$$

The "resampling pairs" bootstrap (section A.6.2) gave estimated standard error of 2.26, with a percentile 95% confidence interval from -10.37 to -1.38.

As would be typical in a dietary intake analysis, we have also examined the effect of adding the variable *total caloric intake* into the regression along with saturated fat. In this case, the predictors measured with error are (log-transformed) total caloric intake and (transformed) saturated fat. The usual tests of the hypothesis of *no* treatment effect, i.e., that neither caloric intake nor saturated fat affect risk, are valid in large samples, and in this case the effects are highly statistically significantly different from zero. The two nutrient intake measures have Pearson correlation of 0.80, and as expected with such multicollinearity in the observed data, along

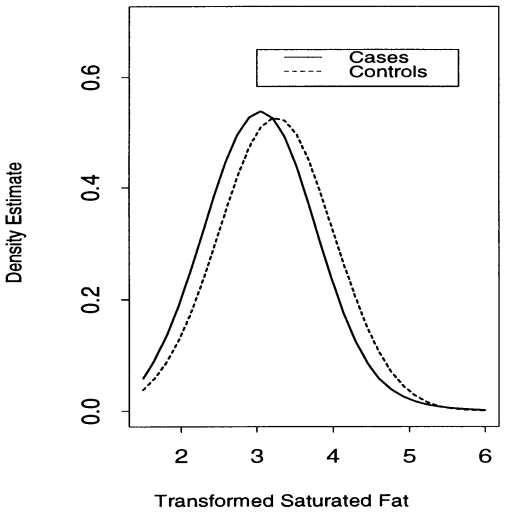

Figure 3.1. *Density estimates of transformed saturated fat for cases and controls: NHANES data.*

Variable	Estimate	Std. Error	p-value
Age /25	2.09	.53	$< .001$
Poverty Index	.13	.08	.10
Body Mass Index / 100	-1.67	2.55	.51
Alcohol	.42	.29	.14
Family History	.63	.44	.16
Age at Menarche	-0.19	.27	.48
Pre-menopausal?	.85	.43	.05
Race	.19	.38	.62
log(5 + Saturated Fat)	-0.97	.29	$< .001$

Table 3.1. *Logistic Regression in the NHANES data.*

with the small number of breast cancer cases, it was impossible to distinguish between caloric intake and saturated fat as a statistically significant predictor of risk.

3.4 Estimating the Calibration Function Parameters

3.4.1 Overview and First Methods

The basic point of using the regression calibration approximation is that one runs a favorite analysis with $\mathbf{X}$ replaced by the mean of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ as an approximation. In this section we will discuss methods for estimating this conditional mean.

With *internal validation data*, the simplest approach is to regress $\mathbf{X}$ on the other covariates $(\mathbf{Z}, \mathbf{W})$ in the validation data. While linear regression will be typical, it is not required. Internal validation data admit many other structural and functional modeling approaches, see Chapters 7 and 8 for the former and Chapters 9 and 14 for the latter.

In some problems, an *unbiased instrument* $\mathbf{T}$ is available for a subset of the study participants, see section 1.4. Here, by definition of "unbiased instrument," the regression of $\mathbf{T}$ on $(\mathbf{Z}, \mathbf{W})$ is an unbiased estimate of $m(\mathbf{Z}, \mathbf{W}, \gamma_{\rm cm})$ since $E(\mathbf{T}|\mathbf{Z}, \mathbf{W}) = E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$. This is the method used by Rosner, Spiegelman & Willett (1990) in their analysis of the Nurses' Health Study.

With validation data or an unbiased instrument, models for $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ can be checked by ordinary regression diagnostics such as residual plots.

When one has internal validation, one will of course want to use the validation data to improve the estimates of (β, θ); after all, one has $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ for these data and it makes sense to use them directly. We have already referenced work in other chapters which addresses this problem, but regression calibration can be used as a simple fallback. One method is to simply run the analysis with $\mathbf{X}$ estimated in the unvalidated data; in generalized linear models we suggest that this procedure include the addition into the regression of a dummy variable indicating whether $\mathbf{X}$ is or is not observed.

3.4.2 Best Linear Approximations Using Replicate Data

Here we consider the additive error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$ where conditional on $(\mathbf{Z}, \mathbf{X})$ the errors have mean zero and constant covariance matrix Σ_{uu}. We describe an algorithm yielding a linear approximation to the regression calibration function. The algorithm is applicable when Σ_{uu} is estimated via external data or via internal replicates. The method was derived independently by Carroll & Stefanski (1990) and Gleser (1990), and used by Liu & Liang (1992) and Wang, et al. (1995).

In this subsection, we will discuss using replicates of $\mathbf{X}$. We repeat here the warning made in section 1.7 about the difference between a true and approximate replicate. As described there, when necessary, the convention made in this book is to adjust the replicates a priori so that they have the same sample means.

Suppose there are k_i replicate measurements of $\mathbf{X}_i$, and $\overline{\mathbf{W}}_{i\cdot}$ is their mean. Replication enables us to estimate the measurement error covariance matrix Σ_{uu} by the usual components of variance analysis, as follows:

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t}{\sum_{i=1}^{n}(k_i - 1)}. \tag{3.2}$$

In (3.2), remember that we are using the "dot and overline" notation to mean averaging over the indicated subscript.

Write Σ_{ab} as the covariance matrix between two random variables and let μ_a be the mean of a random variable. The best linear approximant to $\mathbf{X}$ given $(\mathbf{Z}, \overline{\mathbf{W}})$ is

$$\begin{aligned} E(\mathbf{X}|\mathbf{Z}, \overline{\mathbf{W}}) &\approx \mu_x \\ &+ \begin{pmatrix} \Sigma_{xx} \\ \Sigma_{zx} \end{pmatrix} \begin{bmatrix} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{\mathbf{W}} - \mu_w \\ \mathbf{Z} - \mu_z \end{pmatrix}. \end{aligned} \tag{3.3}$$

Here is how one can operationalize (3.3) based on observations $(\mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot})$, replicate sample sizes k_i and estimated error covariance matrix $\widehat{\Sigma}_{uu}$. We use analysis of variance formulae. Let

$$\widehat{\mu}_x = \widehat{\mu}_w = \sum_{i=1}^{n} k_i \overline{\mathbf{W}}_{i\cdot} / \sum_{i=1}^{n} k_i; \quad \widehat{\mu}_z = \overline{\mathbf{Z}}_{\cdot};$$

$$\nu = \sum_{i=1}^{n} k_i - \sum_{i=1}^{n} k_i^2 / \sum_{i=1}^{n} k_i;$$

$$\widehat{\Sigma}_{zz} = (n-1)^{-1} \sum_{i=1}^{n} \left(\mathbf{Z}_i - \overline{\mathbf{Z}}_\cdot\right)\left(\mathbf{Z}_i - \overline{\mathbf{Z}}_\cdot\right)^t;$$

$$\widehat{\Sigma}_{xz} = \sum_{i=1}^{n} k_i \left(\overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w\right)\left(\mathbf{Z}_i - \overline{\mathbf{Z}}_\cdot\right)^t / \nu;$$

$$\widehat{\Sigma}_{xx} = \left[\left\{\sum_{i=1}^{n} k_i \left(\overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w\right)\left(\overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w\right)^t\right\} - (n-1)\widehat{\Sigma}_{uu}\right] / \nu.$$

The resulting estimated calibration function is

$$E(\mathbf{X}_i|\mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot}) \approx \widehat{\mu}_w \qquad (3.4)$$
$$+(\widehat{\Sigma}_{xx}, \widehat{\Sigma}_{xz}) \begin{bmatrix} \widehat{\Sigma}_{xx} + \widehat{\Sigma}_{uu}/k_i & \widehat{\Sigma}_{xz} \\ \widehat{\Sigma}_{xz}^t & \widehat{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w \\ \mathbf{Z}_i - \overline{\mathbf{Z}}_\cdot \end{pmatrix}.$$

In linear regression, if there are no replicates ($k_i \equiv 1$) but an external estimate $\widehat{\Sigma}_{uu}$ is available, or if there are exactly two replicates ($k_i \equiv 2$) in which case $\widehat{\Sigma}_{uu}$ is half the sample covariance matrix of the differences $\mathbf{W}_{i1} - \mathbf{W}_{i2}$, regression calibration reproduces the classical method of moments estimates, i.e., the estimators of section 2.3 with Σ_{uu} estimated from replicates and $\Sigma_{\epsilon u}$ assumed to be 0.

When the number of replicates is not constant, the algorithm can be shown to produce consistent estimates in linear regression, and (approximately!) to logistic regression. For loglinear mean models, one should add a dummy variable to the regression indicating whether or not an observation is replicated.

3.4.3 Nonlinear Calibration Function Models

Schafer (1992) describes ways to approximate $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ via nonlinear models when $\mathbf{X}$ and $\mathbf{W}$ are scalar variables. If we add a quadratic term in $\mathbf{W}$, the model is

$$E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) = (1, \mathbf{Z}^t, \mathbf{W}, \mathbf{W}^2)\gamma_{\text{cm}}; \qquad (3.5)$$

If validation data are available, then estimating (3.5) is of course a standard regression problem. Partial replicates can be handled in a similar way, see section 3.4.4 below. If we let $\mathbf{R} = (1, \mathbf{Z}^t, \mathbf{W}, \mathbf{W}^2)^t$, then γ_{cm} is defined by the linear regression formula

$$\gamma_{\text{cm}} = \left\{E(\mathbf{R}\mathbf{R}^t)\right\}^{-1} E(\mathbf{R}\mathbf{X}).$$

The first of these terms ($E\mathbf{R}\mathbf{R}^t$) is estimated by the sample version $\sum_1^n \mathbf{R}_i\mathbf{R}_i^t/n$. For the second term, note that

$$E(\mathbf{RX}) = E(\mathbf{RW}) - \left\{0, 0^t, \sigma_U^2, 2\sigma_U^2 E(\mathbf{W}) + E(\mathbf{U}^3)\right\}^t. \quad (3.6)$$

The first of these terms is estimated by $\sum_1^n \mathbf{R}_i\mathbf{W}_i/n$, while in the second term $E(\mathbf{W})$ is estimated the sample mean of the $\mathbf{W}$'s. Estimation in this case requires an estimate of the third moment $E(\mathbf{U}^3)$ of the errors. If one is reasonably confident that the measurement errors are symmetrically distributed, e.g., under the assumption of normal errors, then $E(\mathbf{U}^3) = 0$. Otherwise, the third moment can be estimated from replication data as follows. Let $\kappa_{3,j}$ be the third central sample moment of the jth replicate and let κ_3 be the third central moment of the mean of the two replicates. Then a consistent estimate of $E(\mathbf{U}^3)$ is $\kappa_{31} + \kappa_{32} - 4\kappa_3$.

Schafer proposed the following method for the multiplicative error model $\mathbf{W} = \mathbf{X}\mathbf{U}_*$. For the linear model, he noted that

$$E(\mathbf{RX}) = \begin{pmatrix} 1/E(\mathbf{U}_*) \\ E(\mathbf{ZW})/E(\mathbf{U}_*) \\ E(\mathbf{W}^2)E(\mathbf{U}_*)/E(\mathbf{U}_*^2) \end{pmatrix}.$$

If one knows or has estimates of the first two moments of $\mathbf{U}_*$, then one estimates $E(\mathbf{RX})$ by replacing $E(\mathbf{ZW})$ and $E(\mathbf{W}^2)$ by their sample averages.

For a quadratic model (3.5) and with multiplicative error,

$$E(\mathbf{RX}) = \begin{pmatrix} 1/E(\mathbf{U}_*) \\ E(\mathbf{ZW})/E(\mathbf{U}_*) \\ E(\mathbf{W}^2)E(\mathbf{U}_*)/E(\mathbf{U}_*^2) \\ E(\mathbf{W}^3)E(\mathbf{U}_*^2)/E(\mathbf{U}_*^3) \end{pmatrix}. \quad (3.7)$$

Just as before, implementation of this quadratic model requires knowledge of the first three moments of $\mathbf{U}_*$. In both cases, because the errors are multiplicative, it makes sense as a working hypothesis to assume that $\mathbf{U}_*$ has a lognormal distribution.

There can be some value in the quadratic approximations. Schafer presents a simulation in which $\mathbf{Y}$ given $\mathbf{X}$ is logistic, $\mathbf{U}_*$ is lognormal and the moments of $\mathbf{U}_*$ are known. In his simulation, the quadratic approximation had about 30% smaller mean squared error for estimating the logistic slope than did the linear approximation.

One potential difficulty with the multiplicative model is that the

efficiencies will likely deteriorate when the moments of $\mathbf{U}_*$ are estimated, as they must be in practice. To the best of our knowledge this has not been investigated.

3.4.4 Alternatives When Using Partial Replicates

The linear and quadratic approximations defined above are only approximations, but they can be checked by using the replicates themselves. As is typical, if only a partial subset of the study has an internal replicate ($k_i = 2$), while most of the data are unreplicated ($k_i = 1$), the partial replicates can be used to check the best linear and quadratic approximations to $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ defined above, by fitting models to the regression of $\mathbf{W}_{i2}$ on $(\mathbf{Z}_i, \mathbf{W}_{i1})$. If necessary, the partial replication data can be used in this way to estimate $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$.

3.4.5 James-Stein Calibration

Whittemore (1989) also proposed regression calibration in the case that $\mathbf{X}$ is scalar, there is no $\mathbf{Z}$, and the additive error model applies. If σ_u^2 is unknown and there are k replicates at each observation, then instead of the method of moments estimate (3.4) of $E(\mathbf{X}|\mathbf{W})$ she suggested use of the James-Stein estimate, namely

$$\overline{\mathbf{W}}_{\cdot\cdot} + \left\{1 - \frac{n-1}{n-3}\frac{n(k-1)}{n(k-1)+2}\frac{\widehat{\sigma}_u^2/k}{\widehat{\sigma}_{\overline{w}}^2}\right\}(\overline{\mathbf{W}}_{i\cdot} - \overline{\mathbf{W}}_{\cdot\cdot}),$$

where $\widehat{\sigma}_u^2$ is the usual components of variance estimate of σ_u^2 defined in (3.2) and $\widehat{\sigma}_{\overline{w}}^2$ is the sample variance of the terms $(\overline{\mathbf{W}}_{i\cdot})$. Typically, the James-Stein and moments estimates are nearly the same.

3.5 Standard Errors

It is possible to provide asymptotic formulae for standard errors (Carroll & Stefanski, 1990), but doing so is extremely tedious because of the multiplicity of special cases. Some explicit formulae are given in the appendix (section 3.12.2) for the case of generalized linear models, and models in which one specifies only the mean and variance of the response given the predictors.

The bootstrap (section A.6) requires less programming (and mathematics!!) but takes more computer time. This can be a real

issue, because as Donna Spiegelman has pointed out, it is not realistic to think that in applications investigators will repeatedly use the bootstrap while building models for their data.

In its simplest form, the bootstrap can be used to form standard error estimates and then t-statistics can be constructed using the bootstrap standard errors. The bootstrap percentile method can be used for confidence intervals. Approximate bootstrap pivots can be formed by ignoring the variability in the estimation of the calibration function.

3.6 Expanded Regression Calibration Models

The main purpose of regression calibration is to derive an approximate model for the observed $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ data in terms of the fundamental model parameters. The regression calibration method is one means to this end: merely replace $\mathbf{X}$ by an estimate of $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$.

Although these techniques apply in general, it is convenient for our purposes to cast the problems in the form of what are called mean and variance models (often called quasilikelihood and variance function models), which are described in more generality and detail in (A.21)–(A.22). Readers unfamiliar with the ideas of quasilikelihood may wish to skip this material at first reading, and continue into later chapters.

Mean and variance models specify the mean and variance of a response $\mathbf{Y}$ as functions of covariates $(\mathbf{X}, \mathbf{Z})$ and unknown parameters. For example, in linear regression, the mean is a linear function of the covariates, and the variance is constant. In logistic regression, the "mean" of a binary response $\mathbf{Y}$ is just the probability that the event occurs, which is described by the logistic function evaluated at a linear combination of predictors.

We write these models in general as

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = f(\mathbf{Z}, \mathbf{X}, \mathcal{B}) \tag{3.8}$$

$$\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta). \tag{3.9}$$

The replacement of $\mathbf{X}$ by its estimated value in effect proposes a modified model for the observed data, namely

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) \approx f\{\mathbf{Z}, m(\mathbf{Z}, \mathbf{W}, \gamma_{\text{cm}}), \mathcal{B}\}; \tag{3.10}$$

$$\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) \approx \sigma^2 g^2\{\mathbf{Z}, m(\mathbf{Z}, \mathbf{W}, \gamma_{\text{cm}}), \mathcal{B}, \theta\}. \tag{3.11}$$

We have emphasized that this is a *model* for the data which can be

checked via residual plots. In some cases, the model can be modified to improve the fit, see section 3.7 for a striking data application.

An example will help explain the possible need for refined approximations. Consider the simple linear homoscedastic regression model $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_x\mathbf{X}$ and $\text{var}(\mathbf{Y}|\mathbf{X}) = \sigma^2$. Suppose the measurement process induces a heteroscedastic Berkson model where $E(\mathbf{X}|\mathbf{W}) = \mathbf{W}$ and $\text{var}(\mathbf{X}|\mathbf{W}) = \sigma^2_{\text{cm}}\mathbf{W}^{2\gamma}$, where "cm" stands for "calibration model". The regression calibration approximate model states that the observed data follow a simple linear homoscedastic regression model with $\mathbf{X}$ replaced by $E(\mathbf{X}|\mathbf{W}) = \mathbf{W}$. However, while this gives a correct mean function, the actual variance function for the observed data is heteroscedastic: $\text{var}(\mathbf{Y}|\mathbf{W}) = \sigma^2 + \sigma^2_{\text{cm}}\beta_x^2\mathbf{W}^{2\gamma}$. Hence the regression calibration model gives a consistent estimate of the slope and intercept, but the estimate is inefficient because weighted least squares should have been used. If important enough to effect the efficiency of the estimates, the heteroscedasticity should show up in residual plots.

The preceding example shows that a refined approximation can improve efficiency of estimation, while the next describes a simple situation where bias can also be corrected; another example is discussed in the loglinear mean model case in section 3.9.3. Consider ordinary homoscedastic quadratic regression with $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_{x,1}\mathbf{X} + \beta_{x,2}\mathbf{X}^2$. Use the same heteroscedastic Berkson model as before. Then the regression calibration approximation suggests a homoscedastic model with $\mathbf{X}$ replaced by $\mathbf{W}$, while in fact the observed data have mean $\beta_0 + \beta_{x,1}\mathbf{W} + \beta_{x,2}(\mathbf{W}^2 + \sigma^2_{\text{cm}}\mathbf{W}^{2\gamma})$. If the Berkson error model is heteroscedastic, the regression calibration approximation will lead to a biased estimate of the regression parameters.

It is important to stress that these examples do not invalidate regression calibration as a method, because the heteroscedasticity in the Berkson error model has to be fairly severe before much effect will be noticed. However, there clearly is a need for refined approximations which take over when the regression calibration approximation breaks down.

3.6.1 The Expanded Approximation Defined

We will consider the QVF models (3.8)–(3.9). We will focus entirely on the case that $\mathbf{X}$ is a scalar. Although the general theory (Carroll

& Stefanski, 1990) does allow multiple predictors, the algebraic details are unusually complex. The simplest approximate models are based upon the mean and variance models

$$E(\mathbf{X}|\mathbf{Z},\mathbf{W}) = m(\mathbf{Z},\mathbf{W},\gamma_{\rm cm}); \quad (3.12)$$

$$\mathrm{var}(\mathbf{X}|\mathbf{Z},\mathbf{W}) = \sigma^2_{\rm cm}V^2(\mathbf{Z},\mathbf{W},\gamma_{\rm cm}). \quad (3.13)$$

We wish to construct approximations to the mean and variance function of the observed regression of $\mathbf{Y}$ on $(\mathbf{Z},\mathbf{W})$. Carroll & Stefanski (1990) base such approximations on *pretending* that $\sigma^2_{\rm cm}$ is "small"; if it equals zero, the resulting approximate model is the regression calibration model.

Here is how the approximation works. Let f_x and f_{xx} be the first and second derivatives of $f(z,x,\mathcal{B})$ with respect to x, and let $s_x(z,w,\mathcal{B},\theta,\gamma_{\rm cm})$ and $s_{xx}(\cdot)$ be the first and second derivatives of $g^2(z,x,\mathcal{B},\theta)$ with respect to x and evaluated at $x = E(\mathbf{X}|\mathbf{Z} = z, \mathbf{W} = w)$. Defining $m(\cdot) = m(\mathbf{Z},\mathbf{W},\gamma_{\rm cm})$ and $V(\cdot) = V(\mathbf{Z},\mathbf{W},\gamma_{\rm cm})$, simple Taylor series expansions in section 3.8 with $\sigma^2_{\rm cm} \to 0$ yield the following approximate model, which we call the *expanded regression calibration model*,

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \approx f\{\mathbf{Z},m(\cdot),\mathcal{B}\} + (1/2)\sigma^2_{\rm cm}V^2(\cdot)f_{xx}\{\mathbf{Z},m(\cdot),\mathcal{B}\}; \quad (3.14)$$

$$\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \approx \sigma^2 g^2\{\mathbf{Z},m(\cdot),\mathcal{B},\theta\} + \sigma^2_{\rm cm}V^2(\cdot)\left\{f_x^2(\cdot) + (1/2)\sigma^2 s_{xx}(\cdot)\right\}. \quad (3.15)$$

There are important points to note about the approximate model (3.14)–(3.15):

- By setting $\sigma^2_{\rm cm} = 0$, it reduces to the regression calibration model, in which we need only estimate $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$.
- It is an approximate model which serves as a guide to final model construction in individual cases. We are not assuming that the measurement error is small, only pretending that it is in order to derive a plausible model for the observed data in terms of the regression parameters of interest. In some instances terms can be dropped or combined with others to form even simpler useful models for the observed data.
- It is a mean and variance models for the observed data. Hence, the techniques of model fitting and model exploration discussed

in Carroll & Ruppert (1988) can be applied to nonlinear measurement error model data.

One potential problem with the model (3.14)–(3.15) is that it might not be range preserving. For example, because of the term $s_{xx}(\cdot)$, the variance function (3.15) need not necessarily be positive. If the original function $f(\cdot)$ is positive, the new approximate mean function (3.14) need not be positive because of the term $f_{xx}(\cdot)$. A range preserving expanded regression calibration model for the observed data is

$$\begin{aligned} & E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \\ & \approx f\left[\mathbf{Z},\ m(\cdot)+\frac{1}{2}\sigma^2_{\rm cm}\frac{V^2(\cdot)f_{xx}(\cdot)}{f_x(\cdot)},\mathcal{B}\right]; \end{aligned} \tag{3.16}$$

$$\begin{aligned} \mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \ \approx\ & \sigma^2_{\rm cm}f^2_X\{\mathbf{Z},m(\cdot),\mathcal{B}\}V^2(\cdot) \\ & +\sigma^2 g^2\left[\mathbf{Z},m(\cdot)+\frac{1}{2}\sigma^2_{\rm cm}\frac{V^2(\cdot)s_{xx}(\cdot)}{s_x(\cdot)},\mathcal{B},\theta\right]. \end{aligned} \tag{3.17}$$

3.6.2 Implementation

The approximations (3.14)–(3.15) require specification of the mean and variance functions. In the Berkson model, the former is just $\mathbf{W}$ and a flexible model for the latter is $\sigma^2_{\rm cm}\mathbf{W}^{2\gamma}$, with $\gamma = 0$ indicating homoscedasticity. We will see later in a variety of examples that for this Berkson class the model parameters $(\mathcal{B},\theta)$ are often estimable via QVF techniques using the approximate models, without the need for any validation data. The Berkson framework thus serves as an ideal environment for expanded regression calibration models.

Outside the Berkson class, validation, replication or instrumental data are typically required. We have already discussed in Chapter 3 methods for estimating the conditional mean of $\mathbf{X}$. If possible, one should use such data to estimate the conditional variance function. For example, if there are k unbiased replicates in an additive measurement error model, then the natural counterpart to the best linear estimate of the mean function is the usual formula for the variance in a regression, namely $\mathrm{var}(\mathbf{X}|\mathbf{Z},\mathbf{W}) = \sigma^2_{\rm cm}$, where if σ^2_x

is the variance of $\mathbf{X}$ and σ_u^2 is the measurement error variance,

$$\sigma_{\text{cm}}^2 = \sigma_x^2 - \left(\sigma_x^2, \Sigma_{xz}\right) \begin{bmatrix} \sigma_x^2 + \sigma_u^2/k & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{bmatrix}^{-1} \left(\sigma_x^2, \Sigma_{xz}\right)^t .$$

This can be estimated using the formulae of section 3.4.2. For validation data, one would specify a model for the regression calibration mean and variance functions and estimate the parameters using likelihood or QVF techniques.

3.6.3 Models Without Severe Curvature

When the models for the mean and variance are are not severely curved, f_{xx} and s_{xx} are small relative to $f(\cdot)$ and $g^2(\cdot)$, respectively. In this case, setting $\kappa_{\text{cm}} = \sigma_{\text{cm}}^2/\sigma^2$, the mean and variance functions of the observed data greatly simplify to

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx f\{\mathbf{Z}, m(\cdot), \mathcal{B}\} \\ \text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &\approx \sigma^2 \left[g^2\{\mathbf{Z}, m(\cdot), \mathcal{B}, \theta\} + \kappa_{\text{cm}} V^2(\cdot) f_x^2(\cdot)\right] . \end{aligned}$$

Having estimated the mean function $m(\cdot)$, this is just a QVF model in the parameters $(\mathcal{B}, \theta_*)$, where θ_* consists of θ, κ_{cm} and the other parameters in the function $V^2(\cdot)$. In principle, the QVF fitting methods of Chapter A can be used.

3.7 Bioassay Data

Rudemo, et al. (1989) describe a bioassay problem following a heteroscedastic Berkson error model. In this experiment, four herbicides were applied either as technical grades or as commercial formulations; thus there are eight herbicides, four pairs of two herbicides each. The herbicides were applied at the six different nonzero doses 2^{j-5} for $j = 0, 1, ..., 5$. There were also two zero dose observations. The response $\mathbf{Y}$ was the dry weight of five plants grown in the same pot. There were three complete replicates of this experiment done at three different time periods, so that the replicates are a blocking factor. The data are listed in Table 3.2.

Let $\mathbf{Z}_1$ be a vector of size eight with a single nonzero element indicating which herbicide was applied, and let $\mathbf{Z}_2$ be a vector of size four indicating the herbicide pair. Let $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$. For zero doses, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ may be defined arbitrarily as any nonzero value. In the absence of measurement error for doses, and if there were

no random variation, the relationship between response and dose, $\mathbf{X}$, is expected to be

$$\mathbf{Y} \approx f(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = \beta_0 + \frac{\beta_x - \beta_0}{1 + \left\{ \frac{\mathbf{X}}{\beta_{z,1}^t \mathbf{Z}_1} \right\}^{\beta_{z,2}^t \mathbf{Z}_2}}. \tag{3.18}$$

Model (3.18) is typically referred to as the *four–parameter logistic* model. Physically, the parameters β_0 and β_x should be nonnegative, since they are the approximate dry weight at infinite and zero doses, respectively.

An initial ordinary nonlinear least squares fit to the data with a fixed block effect had a negative estimate of β_0. Figure 3.2 displays a plot of absolute residuals versus predicted means. Also displayed are box plots of the residuals formed by splitting the data into six equal-sized groups ordered on the basis of predicted values. Both figures show that the residuals are clearly heteroscedastic, with the response variance an increasing function of the predicted value.

This problem is exactly of the type amenable to analysis by the *transform-both-sides* (TBS) methodology of Carroll & Ruppert (1988), see also Ruppert, et al. (1989). Specifically, model (3.18) is a theoretical model for the data in the absence of any randomness, which when fit shows a pattern of heteroscedasticity. The TBS methodology suggests controlling for the heteroscedasticity by transforming both sides of the equation:

$$h(\mathbf{Y}, \lambda) \approx h\left\{ f(\mathbf{Z}, \mathbf{X}, \mathcal{B}), \lambda \right\}, \tag{3.19}$$

where the transformation family can be arbitrary but is taken here as the power transformation family:

$$\begin{aligned} h(v, \lambda) &= (v^\lambda - 1)/\lambda \text{ if } \lambda \neq 0; \\ &= \log(v) \text{ if } \lambda = 0. \end{aligned}$$

Of course, the actual dose applied $\mathbf{X}$ may be different from the nominal dose applied $\mathbf{W}$. It seems reasonable in this context to consider the Berkson error model with mean $\mathbf{W}$ and variance $\sigma_{\mathrm{cm}}^2 \mathbf{W}^{2\gamma}$, the heteroscedasticity basically indicating the perfectly plausible assumption that the size of the error made depends on the nominal dose applied. With this specification, the regression calibration approximation replaces $\mathbf{X}$ by $\mathbf{W}$. Letting $\mathbf{Y}_{ij}$ be the jth replicate at the ith herbicide/dose combination, the TBS–

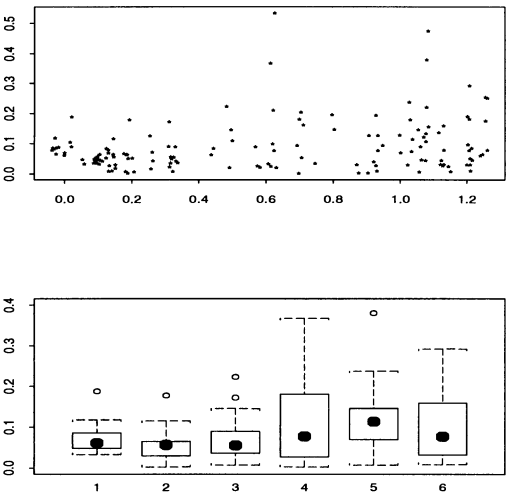

Figure 3.2. *Bioassay data. Absolute residual analysis for an ordinary nonlinear least squares fit. Note the increasing variability for larger predicted values.*

regression calibration model incorporating randomness is

$$h(\mathbf{Y}_{ij}, \lambda) = h\{f(\mathbf{Z}_i, \mathbf{W}_i, \mathcal{B}), \lambda\} + \eta_j + \epsilon_{ij}, \tag{3.20}$$

where ϵ_{ij} is the homoscedastic random effect with variance σ^2, and η_j is the fixed block effect. The parameters were fit using maximum likelihood assuming that the errors are normally distributed, as described by Carroll & Ruppert (1988, Chapter 4). This involves maximizing the loglikelihood

$$-\frac{1}{2}\sum_{i,j}\left(\frac{[h(\mathbf{Y}_{ij}, \lambda) - h\{f(\mathbf{Z}_i, \mathbf{W}_i, \mathcal{B}), \lambda\} - \eta_j]^2}{\sigma^2}\right.$$
$$\left. + \log(\sigma^2) - 2(\lambda - 1)\log(\mathbf{Y}_{ij})\right).$$

The estimated transformation, $\widehat{\lambda} = 0.117$, is very near the log transformation. The residual plots are given in Figure 3.3, where

H	W	Y	H	W	Y	H	W	Y	H	W	Y
0	0	1.51	0	0	1.43	1	1	0.05	1	2	0.06
1	4	0.15	1	8	0.40	1	16	0.76	1	32	0.95
2	1	0.04	2	2	0.07	2	4	0.13	2	8	0.52
2	16	0.79	2	32	1.17	3	1	0.05	3	2	0.26
3	4	0.28	3	8	0.70	3	16	1.05	3	32	1.30
4	1	0.11	4	2	0.42	4	4	0.59	4	8	0.90
4	16	1.08	4	32	1.24	5	1	0.04	5	2	0.06
5	4	0.19	5	8	0.50	5	16	0.84	5	32	1.17
6	1	0.04	6	2	0.04	6	4	0.24	6	8	0.70
6	16	1.21	6	32	1.01	7	1	0.05	7	2	0.08
7	4	0.14	7	8	0.60	7	16	1.20	7	32	1.30
8	1	0.38	8	2	0.64	8	4	0.88	8	8	1.09
8	16	1.50	8	32	1.30						
0	0	1.01	0	0	1.34	1	1	0.05	1	2	0.07
1	4	0.09	1	8	0.26	1	16	0.55	1	32	1.21
2	1	0.04	2	2	0.06	2	4	0.19	2	8	1.16
2	16	0.96	2	32	1.13	3	1	0.04	3	2	0.17
3	4	0.33	3	8	0.50	3	16	1.11	3	32	1.20
4	1	0.12	4	2	0.30	4	4	0.41	4	8	1.06
4	16	1.29	4	32	1.17	5	1	0.04	5	2	0.07
5	4	0.19	5	8	0.36	5	16	0.88	5	32	1.16
6	1	0.04	6	2	0.05	6	4	0.22	6	8	0.61
6	16	1.15	6	32	1.39	7	1	0.04	7	2	0.18
7	4	0.27	7	8	0.88	7	16	0.97	7	32	1.26
8	1	0.29	8	2	0.98	8	4	1.12	8	8	1.10
8	16	1.13	8	32	1.31						

Table 3.2. *The Bioassay Data. Here* **Y** *is the response and* **W** *is the nominal dose time 32. The herbicides* **H** *are listed as 1–8, and* $\mathbf{H} = 0$ *means a zero dose. The replicates* **R** *are separated by horizontal lines. The herbicide pairs are (1,5), (2,6), (3,7) and (4,8).* *Continued on next page.*

H	W	Y	H	W	Y	H	W	Y	H	W	Y
0	0	1.21	0	0	1.10	1	1	0.04	1	2	0.09
1	4	0.12	1	8	0.25	1	16	0.56	1	32	1.04
2	1	0.05	2	2	0.06	2	4	0.14	2	8	0.35
2	16	0.90	2	32	1.12	3	1	0.06	3	2	0.21
3	4	0.37	3	8	0.60	3	16	1.01	3	32	0.70
4	1	0.10	4	2	0.20	4	4	0.47	4	8	0.95
4	16	1.07	4	32	0.93	5	1	0.05	5	2	0.07
5	4	0.09	5	8	0.29	5	16	0.78	5	32	1.05
6	1	0.05	6	2	0.07	6	4	0.16	6	8	0.39
6	16	0.78	6	32	0.97	7	1	0.04	7	2	0.11
7	4	0.24	7	8	0.48	7	16	0.94	7	32	1.30
8	1	0.15	8	2	0.26	8	4	0.60	8	8	0.87
8	16	0.61	8	32	0.98						

Table 3.2 continued.

we still see some unexplained structure to the variability, since the extremes of the predicted means have smaller variability than the centers (even after accounting for leverage).

To account for the unexplained variability, we now consider higher order approximate models. Denoting the left hand side of (3.19) by $\mathbf{Y}_*$ and the right hand side by $f_*(\cdot)$, and noting that the four–parameter logistic model is one in which f_{xx}/f is typically small, the approximate model (3.15) says that $\mathbf{Y}_*$ has mean $h\{f(\mathbf{Z},\mathbf{W},\mathcal{B})\}$ and variance $\sigma^2 + \sigma^2_{\text{cm}}\mathbf{W}^{2\gamma}\left\{f^{\lambda-1}(\mathbf{Z},\mathbf{W},\mathcal{B})f_x(\mathbf{Z},\mathbf{W},\mathcal{B})\right\}^2$. If we define $\kappa = \sigma^2_{\text{cm}}/\sigma^2$, in contrast to (3.20) an approximate model for the data is

$$\begin{aligned} h(\mathbf{Y}_{ij},\lambda) &= h\{f(\cdot),\lambda\} + \eta_j \\ &\quad + \epsilon_{ij}\left[1 + \kappa\mathbf{W}_i^{2\gamma}\left\{f^{\lambda-1}(\cdot)f_x(\cdot)\right\}^2\right]^{1/2}, \end{aligned} \tag{3.21}$$

where as before ϵ_{ij} has variance σ^2. This is a heteroscedastic TBS model, all of whose parameters are identifiable and hence estimable from the observed data. The identifiability of parameters in the Berkson model is a general phenomenon, and it taken up in more

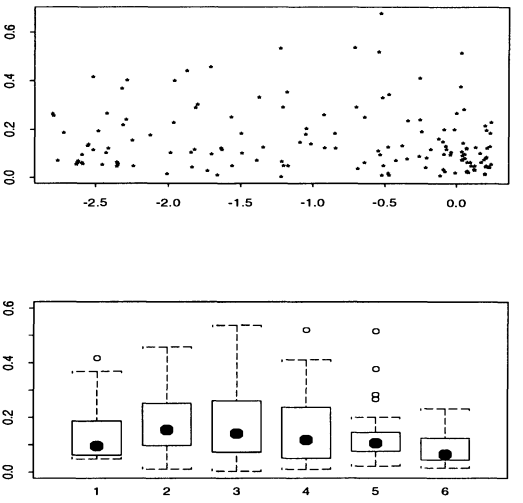

Figure 3.3. *Bioassay data. Absolute residual analysis for an ordinary transform-both-sides fit. Note the unexplained structure of the variability.*

detail in section 3.10. The likelihood of (3.21) is the same as before but with σ^2 replaced by

$$\sigma^2 \left[1 + \kappa \mathbf{W}_i^{2\gamma} \left\{ f^{\lambda-1}(\cdot) f_x(\cdot) \right\}^2 \right].$$

This model was fit to the data, and $\widehat{\lambda} \approx -1/3$ with an approximate standard error of 0.12. The corresponding residual plots are given in Figure 3.4. Here we see no real hint of unexplained variability. As a further check, we can contrast the models (3.21) and (3.20) by means of a likelihood ratio test, the two extra parameters being $(\sigma^2_{\mathrm{cm}}, \kappa)$. The likelihood ratio test for the hypothesis that these two parameters equal zero had a chisquared value of over 30, indicating a large improvement in the fit due to allowing for possible heteroscedasticity in the Berkson error model. On the other hand, after further calculation one finds that adding a correction term to $f_*(\cdot)$ as in (3.14) or (3.16) offers only a negligible improvement.

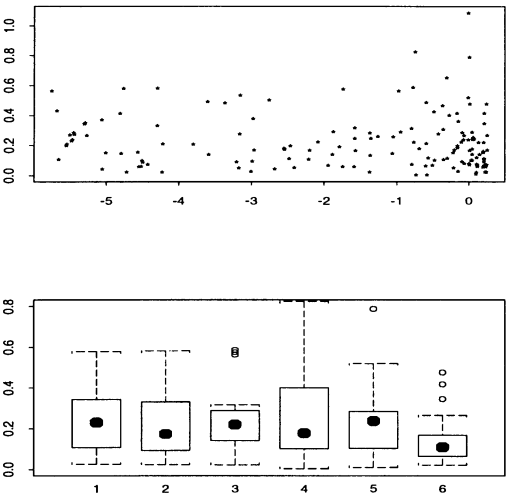

Figure 3.4. *Bioassay data. Absolute residual analysis for a second order approximate transform-both-sides fit.*

3.8 Heuristics and Accuracy of the Approximations

The essential step in regression calibration is the replacement of $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{W},\mathbf{Z}) = m(\mathbf{Z},\mathbf{W},\gamma_{\rm cm})$ in (3.8) and (3.9) leading to the model (3.10)–(3.11). This model can be justified by a "small-σ" argument, i.e., by assuming that the measurement error is small. The basic idea is that under small measurement error, $\mathbf{X}$ will be close to its expectation. However, even with small measurement error, $\mathbf{X}$ may not be close to $\mathbf{W}$, so naively replacing $\mathbf{X}$ by $\mathbf{W}$ may lead to large bias, hence the need for calibration. For simplicity, assume that $\mathbf{X}$ is univariate. Let $\mathbf{X} = E(\mathbf{X}|\mathbf{Z},\mathbf{W}) + \mathbf{V}$, where $E(\mathbf{V}|\mathbf{Z},\mathbf{W}) = 0$ and $\mathrm{var}(\mathbf{V}|\mathbf{Z},\mathbf{W}) = \sigma^2_{X|Z,W}$. Let $m(\cdot) = m(\mathbf{Z},\mathbf{W},\gamma_{\rm cm})$. Let f_x and f_{xx} be the first and second partial derivatives of $f(z,x,\mathcal{B})$ with respect to x. Assuming that $\sigma^2_{X|Z,W}$ is small and hence that $\mathbf{V}$ is small with high probability, we have the Taylor approximation:

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = E\Big\{E(\mathbf{Y}|\mathbf{Z},\mathbf{W},\mathbf{X})\Big|\mathbf{Z},\mathbf{W}\Big\}$$

$$\begin{aligned}
&\approx E\Big\{ f(\mathbf{Z}, m(\cdot), \mathcal{B}) + f_x(\mathbf{Z}, m(\cdot), \mathcal{B})\, \mathbf{V} \\
&\quad + (1/2) f_{xx}(\mathbf{Z}, m(\cdot), \mathcal{B})\, \mathbf{V}^2 \Big| \mathbf{Z}, \mathbf{W} \Big\} \\
&= f\{\mathbf{Z}, m(\cdot), \mathcal{B}\} + (1/2) f_{xx}\{\mathbf{Z}, m(\cdot), \mathcal{B}\}\, \sigma^2_{X|Z,W}.
\end{aligned}$$

Model (3.10) results from dropping the term involving $\sigma^2_{X|Z,W}$, which can be justified by the small-σ assumption. This term is retained in the expanded regression calibration model developed in section 3.6.

To derive (3.11), note that

$$\begin{aligned}
\mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) \;=\; & \mathrm{var}\Big\{ E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \mathbf{X}) \Big| \mathbf{Z}, \mathbf{W} \Big\} \\
& + E\Big\{ \mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \mathbf{X}) \Big| \mathbf{Z}, \mathbf{W} \Big\}.
\end{aligned} \tag{3.22}$$

The first term on the right hand side of (3.22) is

$$\begin{aligned}
\mathrm{var}\{f(\mathbf{Z}, \mathbf{X}, \mathcal{B})|\mathbf{Z}, \mathbf{W}\} &\approx \mathrm{var}\{f_x(\mathbf{Z}, m(\cdot), \mathcal{B})\mathbf{V}|\mathbf{Z}, \mathbf{W}\} \\
&= f_x^2\{\mathbf{Z}, m(\cdot), \mathcal{B}\}\, \sigma^2_{X|Z,W},
\end{aligned}$$

which represents variability in $\mathbf{Y}$ due to measurement error and is set equal to 0 in the regression calibration approximation, but is used in the expanded regression calibration approximation of section 3.6. Let s_x and s_{xx} be the first and second partial derivatives of $g^2(z, x, \mathcal{B}, \theta)$ with respect to x. The second term on the right hand side of (3.22) is

$$\begin{aligned}
E\{\sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta)|\mathbf{Z}, \mathbf{W}\} &\approx \sigma^2 g^2(\mathbf{Z}, m(\cdot), \mathcal{B}, \theta) \\
&\quad + \frac{1}{2} s_{xx}(\mathbf{z}, m(\cdot), \mathcal{B}, \theta) \sigma^2_{X|Z,W}.
\end{aligned}$$

Setting the term involving $\sigma^2_{X|Z,W}$ equal to 0 gives the regression calibration approximation, while both terms are used in expanded regression calibration.

3.9 Examples of the Approximations

In this section, we investigate the appropriateness of the regression calibration algorithm in a variety of settings, paying particular attention to the variance function regression models of section A.4.

3.9.1 Linear Regression

Consider linear regression when the variance of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ is constant, so that the mean and variance of $\mathbf{Y}$ when given $(\mathbf{Z}, \mathbf{X})$ are $\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}$ and σ^2, respectively. As an approximation, the regression calibration model says that the observed data also have constant variance but have regression function given by $E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x^t m(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}}) + \beta_z^t \mathbf{Z}$. Because we assume nondifferential measurement error (section 1.6), the regression calibration model accurately reproduces the regression function, but the observed data have a different variance, namely

$$\mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \sigma^2 + \beta_x^t \mathrm{var}(\mathbf{X}|\mathbf{Z}, \mathbf{W})\beta_x.$$

Note the difference here: the regression calibration model is a working model for the observed data, which may differ somewhat from the actual or true model for the observed data. In this case, the regression calibration approximation gives the correct mean function, and the variance function is also correct and constant if $\mathbf{X}$ has a constant covariance matrix given $(\mathbf{Z}, \mathbf{W})$.

If, however, $\mathbf{X}$ has nonconstant conditional variance, the regression calibration approximation would suggest the homoscedastic linear model when the variances are heteroscedastic. In this case, while the least squares estimates would be consistent, the usual standard errors are incorrect. There are two options: (i) use least squares but employ the resampling-vectors form of the bootstrap (section A.6.2) or the sandwich method for constructing standard errors (section A.3); and (ii) expand the model using the methods of section 3.6.

3.9.2 Logistic Regression

Regression calibration is also well established in logistic regression, at least as long as the effects of the variable $\mathbf{X}$ measured with error is not "too large" (Rosner, et al., 1989, 1990; Whittemore, 1989). Let the binary response $\mathbf{Y}$ follow the logistic model $\Pr(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{X}) = H(\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z})$, where $H(v) = \{1 + \exp(-v)\}^{-1}$ is the logistic distribution function. The key problem is computing the probability of a response $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$. For example, suppose that $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is normally distributed with mean $m(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})$ and (co)variance function $V(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})$. Let p be

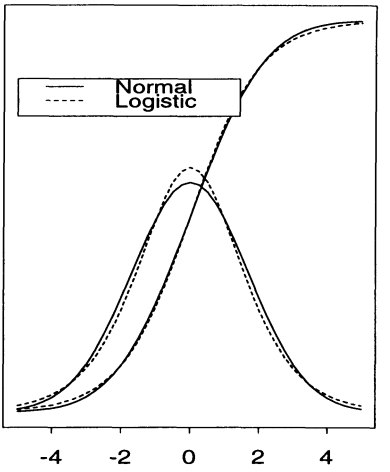

Figure 3.5. *The standard logistic distribution and density functions compared to the normal distribution and density functions with standard deviation 1.70.*

the number of components of $\mathbf{X}$. As described in more detail in Chapter 7, the probability that $\mathbf{Y} = 1$ for values of $(\mathbf{Z}, \mathbf{W})$ is

$$\frac{\int H(\cdot)\exp\left[-(1/2)\left\{x - m(\cdot)\right\}^t V^{-1}(\cdot)\left\{x - m(\cdot)\right\}\right] dx}{(2\pi)^{p/2}|V(\cdot)|^{1/2}}, \quad (3.23)$$

where $H(\cdot) = H(\beta_0 + \beta_x^t x + \beta_z^t \mathbf{Z})$. Formulae (3.23) does not have a closed-form solution; Crouch & Spiegelman (1990) develop a fast algorithm which they have implemented in FORTRAN. Monahan & Stefanski (1991) describe a different method easily applicable to all standard computer packages. However, a simple technique often works just as well, namely to approximate the logistic by the probit. For $c \approx 1.70$, it is well–known that $H(v) \approx \Phi(v/c)$, where $\Phi(\cdot)$ is the standard normal distribution function (Johnson & Kotz, 1970; Liang & Liu, 1991; Monahan & Stefanski, 1991).

In Figure 3.5 we plot the density and distribution functions of the logistic and normal distributions, and the reader will note that the

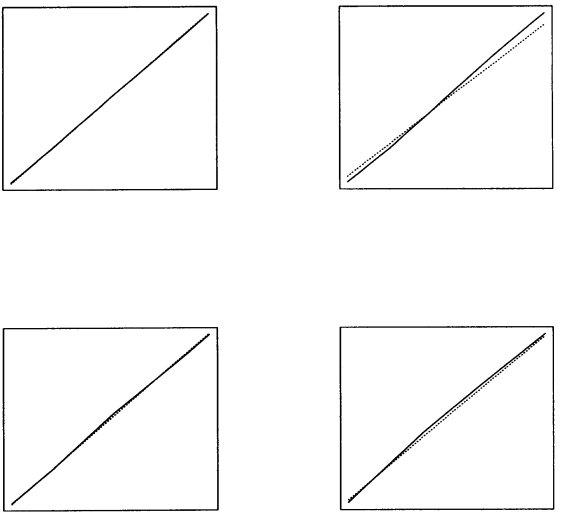

Figure 3.6. *Values of* $\mathrm{pr}(\mathbf{Y} = 1|\mathbf{W})$ *are plotted against* $\mathbf{W}$ *in the solid line, while the regression calibration approximation is the dotted line. The measurement error is additive on the first row and multiplicative on the second row. The fact that the lines are nearly indistinguishable is the whole point. See text for more details.*

logistic and normal are very similar. With some standard algebra (Carroll, et al. (1984)), one can approximate (3.23) by

$$\Pr(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{W}) \approx H\left[\frac{\beta_0 + \beta_x^t m(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}}) + \beta_z^t \mathbf{Z}}{\{1 + \beta_x^t V(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})\beta_x / c^2\}^{1/2}}\right]. \quad (3.24)$$

In most cases, the denominator in (3.24) is very nearly one, and regression calibration is a good approximation; the exception is for "large" $\beta_x^t V(\cdot)\beta_x$.

The approximation (3.24) is often remarkably good, even when the true predictor $\mathbf{X}$ is rather far from normally distributed. To test this, we dropped $\mathbf{Z}$ and computed the approximations and ex-

act forms of $\mathrm{pr}(\mathbf{Y} = 1|\mathbf{W})$ under the following scenario. For the distribution of $\mathbf{X}$, we chose either a standard normal distribution or the chi-squared distribution with one degree of freedom. The logistic intercept β_0 and slope β_x were chosen so that there was a 10% positive response rate ($\mathbf{Y} = 1$) on average, and so that $\exp\{\beta_x(q_{90} - q_{10})\} = 3$, where q_a is the ath percentile of the distribution of $\mathbf{X}$. In the terminology of epidemiology, this means that the "relative risk" is 3.0 in moving from the 10th to the 90th percentile of the distribution of $\mathbf{X}$, a representative situation.

In Figure 3.6 we plot values of $\mathrm{pr}(\mathbf{Y} = 1|\mathbf{W})$ against $\mathbf{W}$ in the solid line, for the range from the 5th to the 95th percentile of the distribution of $\mathbf{W}$. The regression calibration approximation is the dotted line. The measurement error is additive on the first row and multiplicative on the second row. The top left plot has $\mathbf{W} = \mathbf{X} + \mathbf{U}$ where $(\mathbf{X}, \mathbf{U})$ follow a bivariate standard normal distribution, while the top right plot differs in that both follow a chi-squared distribution with one degree of freedom. The bottom row has $\mathbf{W} = \mathbf{XU}$, where $\mathbf{U}$ follows a chi-squared distribution with one degree of freedom; on the left, $\mathbf{X}$ is standard normal, while on the right, $\mathbf{X}$ is chi-squared. Note that the solid and dashed lines very nearly overlap. In all of these cases, the measurement error is *very* large, so in some sense we are displaying a worst case scenario. For these four very different situations, the regression calibration approximation works very well indeed.

3.9.3 Loglinear Mean Models

As might occur for gamma or lognormal data, suppose $E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \exp(\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z})$ and $\mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \sigma^2\{E(\mathbf{Y}|\mathbf{Z}, \mathbf{X})\}^2$. Suppose that the calibration of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$ has mean $m(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})$, and denote the moment generating function of the calibration distribution by

$$E\{\exp(a^t\mathbf{X})|\mathbf{Z}, \mathbf{W}\} = \exp\{a^t m(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}}) + v(a, \mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})\},$$

where $v(\cdot)$ is a general function which differs from distribution to distribution. If $(\cdot) = (\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})$, the observed data then follow the model

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) &= \exp\{\beta_0 + \beta_x^t m(\cdot) + \beta_z^t\mathbf{Z} + v(\beta_x, \cdot)\}; \\ \mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) &= \exp\{2\beta_0 + 2\beta_x^t m(\cdot) + 2\beta_z^t\mathbf{Z} + v(2\beta_x, \cdot)\} \end{aligned}$$

$$\times \left[\sigma^2 + 1 - \exp\left\{2v(\beta_x, \cdot) - v(2\beta_x, \cdot)\right\}\right].$$

If the calibration distribution for $\mathbf{X}$ is normally distributed with constant covariance matrix Σ_{xx}, then $v(a, \cdot) = (1/2)a^t\Sigma_{xx}a$. Remarkably, for $\beta_{0*} = \beta_0 + (1/2)\beta_x^t\Sigma_{xx|z,w}\beta_x$, the observed data *also* follow the loglinear mean model with intercept β_{0*} and a new variance parameter σ_*^2. Thus, the regression calibration approximation is exactly correct for the slope parameters (β_x, β_z)! The conclusion holds more generally, requiring only that $\mathbf{X} - m(\mathbf{Z}, \mathbf{W}, \gamma_{\mathrm{cm}})$ have distribution independent of $(\mathbf{Z}, \mathbf{W})$.

In some instances, the intercept itself is of interest, and the regression calibration approximation must be modified. In loglinear mean models, the regression calibration approximation breaks down if the calibration is "badly" heteroscedastic. Both problems can be handled by the methods described in section 3.6.

3.10 Theoretical Examples

3.10.1 Homoscedastic Regression

The simple homoscedastic linear regression model is $f(z, x, \mathcal{B}) = \beta_0 + \beta_x x + \beta_z z$ with $g^2(\cdot) = V^2(\cdot) = 1$. If the variance function (3.13) is homoscedastic, then the approximate model (3.14)–(3.15) is exact in this case with $E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x m(\cdot) + \beta_z \mathbf{Z}$ and $\mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \sigma^2 + \sigma_{\mathrm{cm}}^2\beta_x^2$, i.e., a homoscedastic regression model. One sees clearly that the effect of measurement error is to inflate the error about the observed line.

In simple linear regression satisfying a Berkson error model with possibly heteroscedastic calibration variances $\sigma_{\mathrm{cm}}^2\mathbf{W}^{2\gamma}$, the approximations are again exact: $E(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x\mathbf{W} + \beta_z\mathbf{Z}$ and $\mathrm{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) = \sigma^2\left\{1 + \beta_x^2(\sigma_{\mathrm{cm}}^2/\sigma_2)\mathbf{W}^{2\gamma}\right\}$. The reader will recognize this as a QVF model, where the parameter $\theta = (\gamma, \kappa = \sigma_{\mathrm{cm}}^2/\sigma^2)$. As long as $\gamma \neq 0$, all the parameters are estimable by standard QVF techniques, without recourse to validation or replication data.

This problem is an example of a remarkable fact, namely that in Berkson error problems, the approximations (3.14)–(3.15) often lead to an identifiable model, so that the parameters can all be estimated without recourse to validation data. Of course, if one does indeed have validation data, then it can be used to improve upon the approximate QVF estimators.

3.10.2 Quadratic Regression with Homoscedastic Regression Calibration

Ordinary quadratic regression has mean function $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_{x,1}\mathbf{X} + \beta_{x,2}\mathbf{X}^2$. With homoscedastic regression calibration, the observed data have mean function

$$\begin{aligned} E(\mathbf{Y}|\mathbf{W}) &= (\beta_0 + \beta_z\sigma^2_{\mathrm{cm}}) + \beta_x m(\mathbf{W}) + \beta_z m^2(\mathbf{W}) \\ &= \beta_0^* + \beta_x m(\mathbf{W}) + \beta_z m^2(\mathbf{W}). \end{aligned}$$

As we remarked in Chapter 3, the regression calibration model accurately reflects the observed data in terms of the slope parameters, but it is off by a constant, since its intercept β_0^* differs from β_0. Here, however, the approximate expanded mean model (3.14) is exact, and β_0 can be estimated as long as one has available an estimate of the calibration variance σ^2_{cm}, see the previous section.

If the error of $\mathbf{X}$ about its conditional mean is homoscedastic and symmetrically distributed, e.g., normally distributed, then the expanded regression calibration model accurately reflects the form of the variance function for the observed data. Details are given in the appendix, section 3.12.3. If the error is asymmetric, then the expanded model (3.15) misses a term involving the third error moment.

3.10.3 Loglinear Mean Model

The loglinear mean model of section 3.9.3 has $E(\mathbf{Y}|\mathbf{X}) = \exp(\beta_0 + \beta_x\mathbf{X})$, and variance proportional to the square of the mean with constant of proportionality σ^2. If calibration is homoscedastic and normally distributed, the actual mean function for the observed data is $E(\mathbf{Y}|\mathbf{W}) = \exp\left\{\beta_0 + (1/2)\beta_x^2\sigma^2_{\mathrm{cm}} + \beta_x m(\mathbf{W})\right\}$. The mean model of regression calibration is $\exp\{\beta_0 + \beta_x m(\mathbf{W})\}$. As discussed in Chapter 3, regression calibration yields a consistent estimate of the slope β_x but not of the intercept.

In this problem, the range–preserving expanded regression calibration model (3.16) correctly captures the mean of the observed data. Interestingly, it also captures the essential feature of the variance function, since both the actual and approximate variance functions (3.17) are a constant times $\exp\{2\beta_0 + 2\beta_x m(\mathbf{W})\}$.

3.10.4 Small Curvature, Heteroscedastic Calibration

In many nonlinear problems, the second derivative f_{xx} is small relative to f, so that $E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) \approx f\{\mathbf{Z}, m(\cdot), \mathcal{B}\}$ and $\text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = \sigma^2 + \sigma^2_{\text{cm}} f_x^2(\cdot) V^2(\cdot)$. In the Berkson case, this again typically reduces to an identified QVF model. Examples of such models in bioassay experiments are described by Rudemo, et al. (1989) and Racine–Poon, et al. (1991).

3.11 Other References

There is a long history of approximately consistent estimates in nonlinear problems, of which regression calibration and the SIMEX method (Chapter 4) are the most recent such methods. Readers should also consult Stefanski & Carroll (1985), Stefanski (1985), Amemiya & Fuller (1988), Amemiya (1985, 1990a, 1990b), and Whittemore & Keller (1988) for other approaches.

3.12 Appendix

3.12.1 Error Variance Estimation in the CSFII

We now turn to an analysis of measurement error in the NHANES study. There is no internal validation or replication of the NHANES-I data set, so we use the CSFII (Continuing Survey of Food Intakes by Individuals) data set, a data set collected in 1985-86 by the USDA (Thompson, et al., 1992). The portion of the CSFII data set we used was restricted to the 1,722 women aged 25-50, the same age range as the group we are investigating.

The method of replication was as follows. First, a 24-hour recall was administered in person by an interviewer, yielding the fallible covariate $\mathbf{W}$; this is essentially the same method used in NHANES-I. Then over the course of a year, the women were reinterviewed by phone five additional times, although the computer file available contains only three of these follow-up interviews, the saturated fat levels from which we call $(\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3)$. The ordering is in time, namely $\mathbf{T}_1$ is the first measurement recorded, and $\mathbf{T}_3$ the last.

The means and standard deviations of $(\mathbf{W}, \mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3)$ in the data are (–1.31,–1.43,–1.45,–1.44) and (.481,.505,.515,.522) respectively. The $\mathbf{T}$'s are very nearly identically distributed. To understand the difference between $\mathbf{W}$ and the $\mathbf{T}$'s, note that the mean

of $\mathbf{W} - \mathbf{T}_3$ is 0.127 with a standard deviation of 0.619, which indicates a small but statistically significant difference in the means $\mathbf{W}$ and the $\mathbf{T}$'s.

A reasonable model for these data is

$$\begin{aligned} \mathbf{W} &= \mathbf{X} + \mathbf{U}; \\ \mathbf{T}_j &= \gamma_{0,\mathrm{em}} + \gamma_{1,\mathrm{em}}\mathbf{X} + \mathbf{V}_j, \end{aligned}$$

where the errors $\mathbf{U}$ and $\mathbf{V}_j$ are independent of $(\mathbf{X}, \mathbf{Z})$ and have mean zero and variances σ_u^2 and σ_v^2, and where "em" stands for "error model". If we knew the latter variance, and if we assume that the errors in $\mathbf{W}$ and $\mathbf{T}_3$ are independent (see below), then we can easily estimate the unknown parameters. There are two possibilities. The first is to assume that all random variables are normally distributed and compute the maximum likelihood estimate of the parameters. This method has the drawback that it requires specification of the correlation between $\mathbf{W}$ and the $\mathbf{T}$'s.

A second method avoids this issue, as long as $\mathbf{W}$ and $\mathbf{T}_3$ are uncorrelated (or practically so). If we define $a = \mathrm{var}(\mathbf{W})$, $b = \mathrm{var}(\mathbf{T}_3) - \sigma_v^2$ and $c = \mathrm{Var}(\mathbf{T}_3 - \mathbf{W})$, then

$$\gamma_{1,\mathrm{em}} = \left\{2b/(b + a - c + \sigma_v^2)\right\}; \gamma_{0,\mathrm{em}} = E(\mathbf{T}_3) - \gamma_{1,\mathrm{em}}E(\mathbf{W});$$
$$\sigma_u^2 = \mathrm{var}(\mathbf{W}) - \left\{\mathrm{var}(\mathbf{T}_3) - \sigma_v^2\right\}/\gamma_{1,\mathrm{em}}^2 = \mathrm{var}(\mathbf{W}) - b/\gamma_{1,\mathrm{em}}.$$

The terms $\mathrm{var}(\mathbf{W})$, $\mathrm{var}(\mathbf{T}_3)$ and $\mathrm{var}(\mathbf{T}_3 - \mathbf{W})$ can be estimated by the corresponding sample variances of $\mathbf{W}$, $\mathbf{T}_3$ and $\mathbf{T}_3 - \mathbf{W}$, respectively, while the population means $E(\mathbf{W})$ and $E(\mathbf{T}_3)$ can be estimated by their sample means. It thus remains to estimate the intra-individual variance σ_v^2 of the errors in the $\mathbf{T}$'s. If these errors were independent of one another, we would simply use a components of variance method, see (3.2). However, there is some concern in these data that the $\mathbf{V}$'s, which we will call instrumental errors, might not be independent. Suppose that the instrumental errors have the following covariance matrix:

$$\begin{pmatrix} \theta_0 & \theta_1 & \theta_2 \\ \theta_1 & \theta_0 & \theta_1 \\ \theta_3 & \theta_1 & \theta_0 \end{pmatrix}.$$

The term θ_0 is the variance of the instrumental errors, while the correlation between any two adjacent pairs is θ_1/θ_0, and the correlation between the first and last instrument is θ_2/θ_0. It is impossible

to estimate all of $(\theta_0, \theta_1, \theta_2)$ without hypothesizing a more specific model. To be more precise, with three replicates only two parameters in the covariance matrix of the errors can be estimated. We are thus forced to lower the number of parameters by hypothesizing reasonable correlation models.

A natural model in this context is the AR(1) model. If $\rho = \theta_1/\theta_0$ is the correlation between the errors of two adjacent replicates, $\theta_2 = \rho^2\theta_0$. Recall that b is the variance of $\gamma_{0,\mathrm{em}} + \gamma_{1,\mathrm{em}}\mathbf{X}$. Let $e_{3,3}$ be a 3×3 matrix of all ones, and let $e_{3,1}$ be the 3×1 vector of all ones. If we let d be the unknown mean of any $\mathbf{T}$, then the vector $\widetilde{\mathbf{T}} = (\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3)^t$ has mean $de_{3,1}$ and covariance matrix

$$\Sigma(b, \rho, \theta_0) = be_{3,3} + \theta_0 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}.$$

The unknown parameters (d, b, θ_0, ρ) can be estimated by the method of moments. Specifically, we can pretend that all random variables are normally distributed, and even without this assumption obtain consistent estimates by maximizing

$$-(n/2)\log\left[\det\left\{\Sigma(b, \rho, \theta_0)\right\}\right]$$
$$-(1/2)\sum_{i=1}^{n}\left(\widetilde{\mathbf{T}}_i - de_{3,1}\right)^t \left\{\Sigma(b, \rho, \theta_0)\right\}^{-1}\left(\widetilde{\mathbf{T}}_i - de_{3,1}\right).$$

Since the Gaussian MLE's are also method of moments estimates, they are consistent when the random variables are non-Gaussian. For other applications, see Wang, et al. (1995).

When we apply the AR(1) error model to the CSFII data, we find that the estimated variance of an instrumental error is $\widehat{\sigma}_v^2 = 0.188$, $\sigma_u^2 = 0.171$, and the correlation between any two adjacent measurements is 0.07, with the correlation between the first and the last instrumental error being less than 0.01. Various significance tests, including the bootstrap and the methods of Hotelling (1940) and Wolfe (1976), suggested a marginally statistically significant correlation. In addition, if the intra-individual errors are independent or nearly so, then the plot of $\mathbf{T}_2 - \overline{\mathbf{T}}$ against $\mathbf{T}_1 - \overline{\mathbf{T}}$ and the plot of $\mathbf{T}_3 - \overline{\mathbf{T}}$ against $\mathbf{T}_1 - \overline{\mathbf{T}}$ should have approximately the same negative slope, a finding basically confirmed in Figure 3.7 (the slight discrepancy in the slopes is reflective of the small amount of autocorrelation). In other words, in light of the large sample size

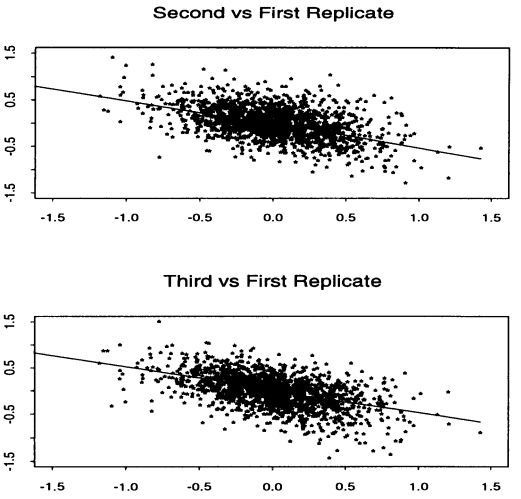

Figure 3.7. *CSFII Data. The plot of* $\mathbf{T}_2 - \overline{\mathbf{T}}$ *against* $\mathbf{T}_1 - \overline{\mathbf{T}}$ *and the plot of* $\mathbf{T}_3 - \overline{\mathbf{T}}$ *against* $\mathbf{T}_1 - \overline{\mathbf{T}}$*, with lowess lines. These should have the same negative slope if the intraindividual errors are independent.*

(1,722 women), there is some but not overwhelming evidence of an intra-individual correlation between adjacent recalls.

One implication of this analysis is that since the intraindividual correlation between *adjacent* recalls is modest, the amount of correlation between the original interview recall (the surrogate $\mathbf{W}$) and the last telephone recall (the instrument $\mathbf{T}_3$) is likely to be negligible. Hence, in the analysis we assumed that these two recalls are independent.

3.12.2 Standard Errors and Replication

As promised in section 3.5, here we provide formulae for asymptotic standard errors for generalized linear models, wherein

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) &= f(\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}); \\ \text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) &= \sigma^2 g^2(\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}). \end{aligned}$$

Let $f^{(1)}(\cdot)$ be the derivative of the function $f(\cdot)$, and let $\mathcal{B} = (\beta_0, \beta_x^t, \beta_z^t)^t$.

We will use here the best linear approximations of section 3.4.2. Let n be the size of the main data set, and $N - n$ the size of any independent data set giving information about the measurement error variance Σ_{uu}. Let $\Delta = 1$ mean that the main data set is used, and $\Delta = 0$ otherwise. Remember that there are k_i replicates for the ith individual and that $\nu = \sum_{i=1}^n k_i - \sum_{i=1}^n k_i^2 / \sum_{i=1}^n k_i$.

Make the definitions $\alpha = (n-1)/\nu$, $\widehat{\Sigma}_{wz} = \widehat{\Sigma}_{xz}$, $\widehat{\Sigma}_{zw} = \widehat{\Sigma}_{wz}^t$, $\widehat{\Sigma}_{ww} = \widehat{\Sigma}_{xx} + \alpha\widehat{\Sigma}_{uu}$, $r_{wi} = (\overline{\mathbf{W}}_{i\cdot} - \mu_w)$, $r_{zi} = (\mathbf{Z}_i - \mu_z)$ and

$$\widehat{\mu}_w = \sum_{i=1}^N \Delta_i k_i \overline{\mathbf{W}}_{i\cdot} / \sum_{i=1}^N \Delta_i k_i; \quad \widehat{\mu}_z = n^{-1} \sum_{i=1}^N \Delta_i \mathbf{Z}_i; \tag{3.25}$$

$$\Psi_{1i*} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & (nk_i/\nu) r_{wi} r_{wi}^t & (nk_i/\nu) r_{wi} r_{zi}^t \\ 0 & (nk_i/\nu) r_{zi} r_{wi}^t & \{n/(n-1)\} r_{zi} r_{zi}^t \end{bmatrix};$$

$$\Psi_{1i} = \Psi_{1i*} - V_i;$$

$$V_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & b_{i1} & b_{i2} \\ 0 & b_{i2}^t & b_{i3} \end{bmatrix};$$

$$b_{i1} = \Sigma_{xx} \frac{nk_i}{\nu} \left\{ 1 - 2k_i / \sum_j \Delta_j k_j + \sum_j \Delta_j k_j^2 / (\sum_j \Delta_j k_j)^2 \right\}$$
$$+ \Sigma_{uu} (n/\nu)(1 - k_i / \sum_j \Delta_j k_j);$$

$$b_{i2} = \Sigma_{xz} (n/\nu)(k_i - k_i^2 / \sum_j \Delta_j k_j); \quad b_{i3} = \Sigma_{zz}.$$

In what follows, except where explicitly noted, we assume that the data have been centered, so that $\widehat{\mu}_w = 0$ and $\widehat{\mu}_z = 0$. This is accomplished by subtracting the original values of the quantities (3.25) from the $\mathbf{W}$'s and $\mathbf{Z}$'s, and has an effect only on the intercept. Reestimating the intercept after "uncentering" is described at the end of this section.

The analysis requires an estimate of Σ_{uu}. For this we only assume

that for some random variables Ψ_{2i} and Ψ_{3i}, if

$$\widehat{S} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \widehat{\Sigma}_{uu} & 0 \\ 0 & 0 & 0 \end{pmatrix}; \qquad S = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Sigma_{uu} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

then

$$\begin{aligned} \widehat{S} - S &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \widehat{\Sigma}_{uu} - \Sigma_{uu} & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &\approx n^{-1} \sum_{i=1}^{N} \{\Delta_i \Psi_{2i} + (1 - \Delta_i)\Psi_{3i}\}. \end{aligned} \tag{3.26}$$

For example, if the estimator comes from an independent data set of size $N - n$, then $\Psi_{2i} = 0$ and

$$\Psi_{3i} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \psi_{3i} & 0 \\ 0 & 0 & 0 \end{pmatrix}; \text{ where}$$

$$\psi_{3i} = \frac{\sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t - (k_i - 1)\Sigma_{uu}}{n^{-1} \sum_{l=1}^{N} (1 - \Delta_l)(k_l - 1)}.$$

If the estimate of Σ_{uu} comes from internal data, then $\Psi_{3i} = 0$ and

$$\Psi_{2i} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \psi_{2i} & 0 \\ 0 & 0 & 0 \end{pmatrix}; \text{ where}$$

$$\psi_{2i} = \frac{\sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t - (k_i - 1)\Sigma_{uu}}{n^{-1} \sum_{l=1}^{N} \Delta_l (k_l - 1)}.$$

Now make the further definitions

$$\widehat{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \widehat{\Sigma}_{ww} & \widehat{\Sigma}_{wz} \\ 0 & \widehat{\Sigma}_{zw} & \widehat{\Sigma}_{zz} \end{bmatrix};$$

$$\widehat{c}_i = \left\{\widehat{D} - (\alpha - k_i^{-1})\widehat{S}\right\}^{-1}.$$

Let D and S be the limiting values of $\widehat{D}$ and $\widehat{S}$. Let I be the identity matrix of the same dimension as $\mathcal{B}$. Define $R_i = (1, \overline{\mathbf{W}}_{i\cdot}^t, \mathbf{Z}_i^t)^t$

and $\widehat{Q}_i = (\widehat{D} - \alpha\widehat{S})\widehat{c}_i R_i$. Using the fact that the data are centered, it is an easy but crucial calculation to show that $\widehat{Q}_i = (1, \widehat{E}(\mathbf{X}_i^t|\mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot}), \mathbf{Z}_i^t)^t$, i.e., it reproduces the regression calibration estimates. Now make the following series of definitions:

$$\widehat{s}_i = \left\{ f^{(1)}(\widehat{Q}_i^t\widehat{\mathcal{B}})/g(\widehat{Q}_i^t\widehat{\mathcal{B}}) \right\}^2;$$

$$\widehat{A}_{1n} = n^{-1}\sum_{i=1}^{N} \Delta_i \widehat{Q}_i \widehat{Q}_i^t \widehat{s}_i;$$

$$r_i = \left\{ Y_i - f(Q_i^t\mathcal{B}) \right\} f^{(1)}(Q_i^t\mathcal{B}) Q_i / g^2(Q_i^t\mathcal{B});$$

$$d_{in1} = n^{-1}\sum_{j=1}^{N} \Delta_j s_j Q_j R_j^t c_j \Psi_{1i} \left\{ I - c_j(D - \alpha S) \right\} \mathcal{B};$$

$$d_{in2} = n^{-1}\sum_{j=1}^{N} \Delta_j s_j Q_j R_j^t c_j \Psi_{2i} \left\{ (\alpha - k_j^{-1})(D - \alpha S)c_j - \alpha I \right\} \mathcal{B};$$

$$d_{in3} = n^{-1}\sum_{j=1}^{N} \Delta_j s_j Q_j R_j^t c_j \Psi_{3i} \left\{ (\alpha - k_j^{-1})(D - \alpha S)c_j - \alpha I \right\} \mathcal{B};$$

$$e_{in} = \Delta_i(r_i - d_{in1} - d_{in2}) - (1 - \Delta_i)d_{in3}.$$

Here and in what follows, s_i, Q_i, c_i, A_{1n}, etc. are obtained by removing the estimates in each of their terms. Similarly, $\widehat{r}_i$, $\widehat{d}_{in1}$, $\widehat{d}_{in2}$, $\widehat{e}_{in}$, etc. are obtained by replacing population quantities by their estimates.

We are going to show that

$$\widehat{\mathcal{B}} - \mathcal{B} \approx A_{1n}^{-1} n^{-1} \sum_{i=1}^{N} e_{in}, \tag{3.27}$$

and hence a consistent asymptotic covariance matrix estimate obtained by using the sandwich method is

$$n^{-1}\widehat{A}_{1n}^{-1}\widehat{A}_{2n}\widehat{A}_{1n}^{-1}, \quad \text{where} \tag{3.28}$$

$$\widehat{A}_{2n} = n^{-1}\sum_{i=1}^{N} \left\{ \Delta_i \left(\widehat{r}_i - \widehat{d}_{in1} - \widehat{d}_{in2} \right) \left(\widehat{r}_i - \widehat{d}_{in1} - \widehat{d}_{in2} \right)^t + (1 - \Delta_i)\widehat{d}_{in3}\widehat{d}_{in3}^t \right\}. \tag{3.29}$$

The information-type asymptotic covariance matrix uses

$$\widehat{A}_{2n,i} = \widehat{A}_{2n} + \widehat{A}_{1n} - n^{-1} \sum_{i=1}^{N} \Delta_i \widehat{r}_i \widehat{r}_i^t. \tag{3.30}$$

It is worth noting that deriving (3.28) and (3.30) takes considerable effort, and that programming it is not trivial. The bootstrap avoids both steps, at the cost of extra computer time.

To verify (3.27), note by the definition of the quasilikelihood estimator and by a Taylor series, we have the expansion

$$\begin{aligned} 0 &= n^{-1/2} \sum_{i=1}^{N} \Delta_i \left\{ Y_i - f(\widehat{Q}_i^t \widehat{\mathcal{B}}) \right\} f^{(1)}(\widehat{Q}_i^t \widehat{\mathcal{B}}) \widehat{Q}_i / g^2(\widehat{Q}_i^t \widehat{\mathcal{B}}) \\ &\approx n^{-1/2} \sum_{i=1}^{N} \Delta_i \left\{ r_i - s_i Q_i \left(\widehat{Q}_i^t \widehat{\mathcal{B}} - Q_i^t \mathcal{B} \right) \right\} \\ &\approx n^{-1/2} \sum_{i=1}^{N} \Delta_i \left\{ r_i - s_i Q_i \left(\widehat{Q}_i - Q_i \right)^t \mathcal{B} \right\} \\ &\qquad - A_{1n} n^{1/2} \left(\widehat{\mathcal{B}} - \mathcal{B} \right). \end{aligned} \tag{3.31}$$

However, by a standard linear expansion of matrices,

$$\begin{aligned} \widehat{Q}_i - Q_i &= \left\{ (\widehat{D} - \alpha \widehat{S}) \widehat{c}_i - (D - \alpha S) c_i \right\} R_i \\ &\approx \left\{ (\widehat{D} - D) - \alpha (\widehat{S} - S) \right\} c_i R_i \\ &\quad - (D - \alpha S) c_i \left\{ (\widehat{D} - D) - (\alpha - k_i^{-1}) (\widehat{S} - S) \right\} c_i R_i \\ &= \{ I - (D - \alpha S) c_i \} (\widehat{D} - D) c_i R_i \\ &\qquad + \left\{ (\alpha - k_i^{-1}) (D - \alpha S) c_i - \alpha I \right\} (\widehat{S} - S) c_i R_i. \end{aligned}$$

However, we have the linear expansion

$$n^{1/2} (\widehat{D} - D) \approx n^{-1/2} \sum_{i=1}^{N} \Delta_i \Psi_{1i},$$

and substituting this together with (3.26) means that

$$n^{-1/2} \sum_{i=1}^{N} \Delta_i \left\{ r_i - s_i Q_i \left(\widehat{Q}_i - Q_i \right)^t \mathcal{B} \right\}$$

$$
\begin{aligned}
&\approx n^{-1/2}\sum_{i=1}^{N}\Delta_i r_i \\
&-n^{-1/2}\sum_{i=1}^{N}\Delta_i s_i Q_i R_i^t c_i n^{-1}\sum_{j=1}^{N}\Delta_j \Psi_{1j}\left\{I-c_i(D-\alpha S)\right\}\mathcal{B} \\
&-n^{-1/2}\sum_{i=1}^{N}\Delta_i s_i Q_i R_i^t c_i n^{-1}\sum_{j=1}^{N}\left\{\Delta_j\Psi_{2j}+(1-\Delta_j)\Psi_{3j}\right\} \\
&\qquad\times\left\{(\alpha-k_i^{-1})(D-\alpha S)c_i-\alpha I\right\}\mathcal{B}.
\end{aligned}
$$

If we interchange the roles of i and j in the last expressions and inset into (3.31), we obtain (3.27).

While the standard error formulae have assumed centering, one can still make inference about the original intercept that would have been obtained had one not centered. Letting the original means of the $\mathbf{Z}_i$'s and $\overline{\mathbf{W}}_{i\cdot}$'s be $\widehat{\mu}_{z,o}$ and $\widehat{\mu}_{w,o}$, the original intercept is estimated by $\widehat{\beta}_0+\widehat{\beta}_x^t\widehat{\mu}_{w,o}+\widehat{\beta}_z^t\widehat{\mu}_{z,o}$. If one conditions on the observed values of $\widehat{\mu}_{z,o}$ and $\widehat{\mu}_{w,o}$, then this revised intercept is the linear combination $a^t\widehat{\mathcal{B}}=(1,\widehat{\mu}_{z,o}^t,\widehat{\mu}_{z,o}^t)\widehat{\mathcal{B}}$, and its variance is estimated by $n^{-1}a^t\widehat{A}_{1n}^{-1}\widehat{A}_{2n}\widehat{A}_{1n}^{-1}a$.

If Σ_{uu} is known, or if one is willing to ignore the variation in its estimate $\widehat{\Sigma}_{uu}$, set $d_{in2}=d_{in3}=0$. This may be relevant if $\widehat{\Sigma}_{uu}$ comes from a large, careful independent study, for which only summary statistics are available (a common occurrence).

In other cases, $\mathbf{W}$ is a scalar variable, Σ_{uu} cannot be treated as known and one must rely on an independent experiment which reports only an estimate of it. If that experiment reports an asymptotic variance $\widehat{\xi}/n$ based on a sample of size $N-n$, then Ψ_{3i} is a scalar and simplifications result which enable a valid asymptotic analysis. Define

$$
d_{n4}=n^{-1}\sum_{j=1}^{N}\Delta_j\widehat{s}_j\widehat{Q}_j R_j^t\widehat{c}_j\left\{(\alpha-k_j^{-1})(\widehat{D}-\alpha\widehat{S})\widehat{c}_j-\alpha I\right\}\widehat{\mathcal{B}}.
$$

Then, in (3.29) replace $n^{-1}\sum_i(1-\Delta_i)\widehat{d}_{in3}\widehat{d}_{in3}^{\,t}$ by $d_{n4}d_{n4}^t n\widehat{\xi}/(N-n)$.

3.12.3 Quadratic Regression: Details of The Expanded Calibration Model

Here we show that, as stated in section 3.10.2, in quadratic regression, if $\mathbf{X}$ given $\mathbf{W}$ is symmetrically distributed and homoscedastic, the expanded model (3.15) accurately summarizes the variance function. Let $\kappa = E\left\{(\mathbf{X} - m)^4 | \mathbf{W}\right\}$, which is constant because of the homoscedasticity. Then, if $r = \mathbf{X} - m$, the variance function is given by

$$\begin{aligned}
&\text{var}(\mathbf{Y}|\mathbf{W}) = \sigma^2 + \beta_{x,1}^2 \text{var}(\mathbf{X}|\mathbf{W}) + \beta_{x,2}^2 \text{var}(\mathbf{X}^2|\mathbf{W}) \\
&\quad + 2\beta_{x,1}\beta_{x,2} \text{cov}\left\{(\mathbf{X}, \mathbf{X}^2)|\mathbf{W}\right\} \\
&= \sigma^2 + \beta_{x,1}^2 \sigma_{\text{cm}}^2 + \beta_{x,2}^2 E\left\{\mathbf{X}^4 - (m^2 + \sigma_{\text{cm}}^2)^2 | \mathbf{W}\right\} \\
&\quad + 2\beta_{x,1}\beta_{x,2} E\left[r\left\{r^2 + 2mr - \sigma_{\text{cm}}^2\right\} | \mathbf{W}\right] \\
&= \sigma^2 + \beta_{x,1}^2 \sigma_{\text{cm}}^2 + \beta_{x,2}^2 (\kappa + 4m^2\sigma_{\text{cm}}^2 - \sigma_{\text{cm}}^4) + 4\beta_{x,1}\beta_{x,2} m \sigma_{\text{cm}}^2 \\
&= \sigma_*^2 + \sigma_{\text{cm}}^2 (\beta_{x,1} + 2\beta_{x,2} m)^2,
\end{aligned}$$

where $\sigma_*^2 = \sigma^2 + \beta_{x,2}^2 \kappa - \sigma_{\text{cm}}^4$. The approximation (3.15) is of exactly the same form. The only difference is that it replaces the correct σ_*^2 by σ^2, but this replacement is unimportant since both are constant.

CHAPTER 4

SIMULATION EXTRAPOLATION

4.1 Overview

Regression calibration (Chapter 3) is a simple, generally applicable approximate estimation method that is especially well suited to problems in which validation or replication data are available for modeling the calibration function $E(\mathbf{X} \mid \mathbf{W})$. We now describe a complementary approximate method that shares the simplicity of regression calibration and is well suited to problems with additive measurement error. Simulation extrapolation (SIMEX) is a simulation-based method of estimating and reducing bias due to measurement error. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error-induced bias versus the variance of the added measurement error, and extrapolating this trend back to the case of no measurement error. The technique was proposed by Cook & Stefanski (1995) and further developed by Carroll, Küchenhoff, Lombard & Stefanski (1996) and Stefanski & Cook (1996).

The fact that measurement error in a predictor variable induces bias in regression estimates is counter-intuitive to many people. An integral component of SIMEX is a self-contained simulation study resulting in graphical displays that illustrate the effect of measurement error on parameter estimates and the need for bias correction. The graphical displays are especially useful when it is necessary to motivate or explain a measurement error model analysis.

The key features of the SIMEX algorithm are described in the context of linear regression in the following section. A detailed description of the method is then given, followed by an example appli-

cation to data from the Framingham Heart Study. These sections will be sufficient for the reader to understand and implement the procedure. Following the example, theoretical aspects of SIMEX estimation are described in greater detail. Examples of linear, log-linear, quadratic and segmented regression are described in detail. The chapter ends with a section on asymptotic distribution theory and variance estimation for SIMEX estimators.

4.2 Simulation Extrapolation Heuristics

This section describes the basic idea of SIMEX, focusing on linear regression with additive measurement error. In section 4.4, we show how to extend SIMEX to nonadditive models. We assume that $\mathbf{Y} = \beta_1 + \beta_x \mathbf{X} + \epsilon$, with additive measurement error $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $(\mathbf{Y}, \mathbf{X})$ and has mean zero and variance σ_u^2. The ordinary least squares estimate of β_x, denoted $\widehat{\beta}_{x,\text{naive}}$, consistently estimates not β_x but rather $\beta_x \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ (Chapter 2). For this simple model the effect of measurement error on the least squares estimator is easily determined mathematically.

The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can also be determined experimentally via simulation. If we regard measurement error as a factor whose influence on an estimator is to be determined, we are naturally led to consider simulation experiments in which the level of the measurement error, i.e., its variance, is intentionally varied.

Suppose that in addition to the original data used to calculate $\widehat{\beta}_{x,\text{naive}}$, there are $M - 1$ additional data sets available, each with successively larger measurement error variances, say $(1 + \lambda_m)\sigma_u^2$, where $0 = \lambda_1 < \lambda_2 < \cdots, < \lambda_M$. The least squares estimate of slope from the mth data set, $\widehat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2/\{\sigma_x^2 + (1 + \lambda_m)\sigma_u^2\}$.

We can think of this problem as a nonlinear regression model, with dependent variable $\widehat{\beta}_{x,m}$ and independent variable λ_m, having a mean function of the form

$$\mathcal{G}(\lambda) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1 + \lambda)\sigma_u^2}, \qquad \lambda \geq 0.$$

The parameter of interest, β_x, is obtained from $\mathcal{G}(\lambda)$ by extrapolation to $\lambda = -1$. We describe the process schematically in Figure 4.1.

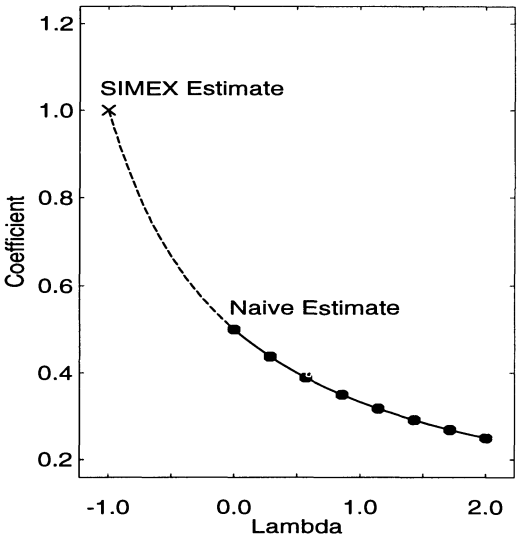

Figure 4.1. *A generic plot of the effect of measurement error of size* $(1+\lambda)\sigma_u^2$ *on parameter estimates. The value of* λ *is on the x-axis, while the value of the estimated coefficient is on the y-axis. The SIMEX estimate is an extrapolation to* $\lambda = -1$*. The naive estimate occurs at* $\lambda = 0$*.*

SIMEX imitates the procedure just described. In the *simulation step* additional independent measurement errors with variance $\lambda_m\sigma_u^2$ are generated and added to the original data, thereby creating data sets with successively larger measurement error variances. For the mth data set, the total measurement error variance is $\sigma_u^2+\lambda_m\sigma_u^2 = (1+\lambda_m)\sigma_u^2$. Next, estimates are obtained from each of the resulting contaminated data sets. The simulation and reestimation step is repeated a large number of times and the average value of the estimate for each level of contamination is calculated. These averages are plotted against the λ values and regression techniques are used to fit an extrapolant function to the averaged, error-contaminated estimates. Extrapolation back to the ideal case of no measurement error ($\lambda = -1$) yields the SIMEX estimate.

4.3 The SIMEX Algorithm

In this section, we describe the implementation of the SIMEX algorithm.

4.3.1 The Simulation and Extrapolation Steps

While SIMEX is a general methodology, it is easiest to understand when there is only a single, scalar predictor $\mathbf{X}$ subject to additive error, so that $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$, where $\mathbf{U}_i$ is a normal random variable with variance σ_u^2, and is independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{Y}_i$. Typically, the assumption of normality is not critical in practice. Initially we assume that the measurement error variance, σ_u^2, is known. Additivity of errors is not crucial, see section 4.4.

SIMEX, like regression calibration, is applicable to general estimation methods, e.g., least-squares, maximum likelihood, quasi-likelihood, etc. In this section, we will not distinguish among the methods, but instead will refer to "the estimator" to mean the chosen estimation method computed as if there were no measurement error. However, we do restrict attention to M-estimators. We let Θ denote the parameter of interest.

The first part of the algorithm is the simulation step. As described above, this involves using simulation to create additional datasets of increasingly large measurement error $(1+\lambda)\sigma_u^2$. For any $\lambda \geq 0$, define

$$\mathbf{W}_{b,i}(\lambda) = \mathbf{W}_i + \sqrt{\lambda}\,\mathbf{U}_{b,i}, \quad i = 1, \ldots, n, \quad b = 1, \ldots, B, \quad (4.1)$$

where the computer-generated *pseudo errors*, $\{\mathbf{U}_{b,i}\}_{i=1}^n$, are mutually independent, independent of all the observed data and identically distributed, normal random variables with mean 0 and variance σ_u^2.

Having generated the new predictors, we compute the resulting naive estimates. Define $\widehat{\Theta}_b(\lambda)$ to be the M-estimator when the $\{\mathbf{W}_{b,i}(\lambda)\}_1^n$ are used, and define the average of these estimators as

$$\widehat{\Theta}(\lambda) = B^{-1} \sum_{b=1}^{B} \widehat{\Theta}_b(\lambda). \quad (4.2)$$

By design, $\widehat{\Theta}(\lambda)$ is the sample mean of $\{\widehat{\Theta}_b(\lambda)\}_1^B$, and hence is the average of the estimates obtained from a large number of experiments with the same amount of measurement error. It is the points

$\{\widehat{\Theta}(\lambda_m), \lambda_m\}_1^M$ that are plotted as filled circles in Figure 4.1. This is the simulation component of SIMEX.

The extrapolation step of the proposal entails modeling each of the components of $\widehat{\Theta}(\lambda)$ as functions of λ for $\lambda \geq 0$, and extrapolating the fitted models back to $\lambda = -1$. The vector of extrapolated values yields the simulation extrapolation estimator denoted $\widehat{\Theta}_{\text{simex}}$. In Figure 4.1 the extrapolation is indicated by the dashed line and the SIMEX estimate is plotted as a cross.

4.3.2 Modifications of the Simulation Step

There is a simple modification to the simulation step that is sometimes useful. As described above the pseudo errors are generated independently of $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n$ as $\mathrm{N}(0, \sigma_u^2)$ random variables. The Monte Carlo variance in $\widehat{\Theta}(\lambda)$ can be reduced by the use of pseudo errors constrained so that for each fixed b, the sequence $(\mathbf{U}_{b,i})_{i=1}^n$ has mean zero, population variance σ_u^2, i.e., $\sum_{i=1}^n \mathbf{U}_{b,i}^2 = n\sigma_u^2$, and its sample correlations with $(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_i)_1^n$ are all zero. We call pseudo errors constrained in this manner, non-iid pseudo errors. In some simple models such as linear regression, the Monte Carlo variance is reduced to zero by the use of non-iid pseudo errors.

The non-iid pseudo errors are generated by first generating iid standard normal pseudo errors $(\mathbf{U}_{b,i}^*)_1^n$. Next fit a linear regression model of the iid pseudo errors on $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n$, including an intercept. The non-iid pseudo errors are obtained by multiplying the residuals from this regression by the constant

$$c = \left[n\sigma_u^2 / \{(n - p - 1)\ \text{MSE}\ \}\right]^{1/2},$$

where MSE is the usual linear regression mean squared error, and p is the dimension of $(\mathbf{Y}, \mathbf{Z}^t, \mathbf{W}^t)^t$.

The use of non-iid pseudo errors can be useful with small sample sizes. However, both in simulations (Cook & Stefanski, 1995) and theory (Carroll, Küchenhoff, Lombard & Stefanski, 1996) we have found that with large sample sizes the use of non-iid pseudo errors confers no significant advantage.

4.3.3 Estimating the Measurement Error Variance

When the measurement error variance σ_u^2 is unknown, it must be estimated with auxiliary data as described in Chapter 3, see espe-

cially (3.2). The estimate is then substituted for σ_u^2 in the SIMEX algorithm and standard errors are calculated as described in section 4.7.2.

4.3.4 Extrapolant Function Considerations

In multiple linear regression with non-iid pseudo errors, use of the extrapolant function,

$$\mathcal{G}_{\mathrm{RL}}(\lambda,\ \Gamma) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \lambda}. \tag{4.3}$$

reproduces the usual method-of-moments estimators, see Section 4.6.1. Since the function $\mathcal{G}_{\mathrm{RL}}(\lambda,\ \Gamma)$ may be represented as a ratio of two linear functions we call it the *rational linear extrapolant.*

SIMEX can be automated in the sense that $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ can be employed to the exclusion of other functional forms. However, this is not recommended, especially in new situations where the effects of measurement error are not reasonably well understood. SIMEX is a technique for studying the effects of measurement error in statistical models and approximating the bias due to measurement error. The extrapolation step should be approached as any other modeling problem, with attention paid to adequacy of the extrapolant based on theoretical considerations, residual analysis, and possibly the use of linearizing transformations. Of course, extrapolation is risky in general even when model diagnostics fail to indicate problems, and this should be kept in mind.

In many problems of interest the magnitude of the measurement error variance, σ_u^2, is such that the curvature in the best or "true" extrapolant function is slight and is adequately modeled by either $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ or the simple quadratic extrapolant,

$$\mathcal{G}_{\mathrm{Q}}(\lambda, \Gamma) = \gamma_1 + \gamma_2\lambda + \gamma_3\lambda^2. \tag{4.4}$$

An advantage of the quadratic extrapolant is that it is often numerically more stable than $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$. Instability of the rational linear extrapolant can occur when the effects of measurement error on a parameter are negligible and a constant, or nearly constant, extrapolant function is required. Such situations arise, for example, with the coefficient of an error-free covariate $\mathbf{Z}$ that is uncorrelated with $\mathbf{W}$. In this case $\gamma_2 \approx 0$ and γ_3 is nearly unidentifiable. In cases where $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ is used to model a nearly horizontal line,

$\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are well determined, but $\widehat{\gamma}_3$ is not. Problems arise when $0 < \widehat{\gamma}_3 < 1$, for then the fitted model has a singularity in the range of extrapolation $[-1, 0)$. The problem is easily solved by fitting $\mathcal{G}_{\mathrm{Q}}(\lambda, \Gamma)$ in these cases. The quadratic extrapolant typically results in conservative corrections for attenuation; however, the increase in bias is often offset by a reduction in variability.

The instability of $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ just described is fundamentally different from the problem that arises when there are manifest effects of measurement error, i.e., the plot of $\widehat{\Theta}(\lambda)$ versus λ is not close to horizontal, and yet $\widehat{\gamma}_3 < 1$. Here it simply may be that B is too small, in which case the solution is apparent. However, the problem can persist even for $B \to \infty$. For example, this occurs in the linear measurement error model when the sample variance of $(\mathbf{W}_i)_1^n$ is less than σ_u^2. As in the linear model, the likelihood that this problem will occur in large samples is small. Also, from a practical viewpoint the occurrence of this problem suggests reassessment of the measurement error model assumptions.

Simulation evidence and our experience with applications thus far suggest that the extrapolant be fit for λ in the range $[0, \lambda_{\max}]$ where $1 \leq \lambda_{\max} \leq 2$. We denote the grid of λ values employed by Λ, i.e., $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_M)$ where typically $\lambda_1 = 0$ and $\lambda_M = \lambda_{\max}$.

The quadratic extrapolant is a linear model and thus easily fits. The rational linear extrapolant generally requires a nonlinear least squares program to fit the model. However, it is possible to obtain exact analytic fits to three points and this provides a means of obtaining good starting values.

Let $\lambda_0^* < \lambda_1^* < \lambda_2^*$ and define $d_{ij} = a_i - a_j$, $0 \leq i < j \leq 2$. Then fitting $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ to the points $\{a_j,\ \lambda_j^*\}_0^2$ results in parameter estimates

$$\begin{aligned}
\widehat{\gamma}_3 &= \frac{d_{12}\lambda_2^*(\lambda_1^* - \lambda_0^*) - \lambda_0^* d_{01}(\lambda_2^* - \lambda_1^*)}{d_{01}(\lambda_2^* - \lambda_1^*) - d_{12}(\lambda_1^* - \lambda_0^*)} \\
\widehat{\gamma}_2 &= \frac{d_{12}(\widehat{\gamma}_3 + \lambda_1^*)(\widehat{\gamma}_3 + \lambda_2^*)}{\lambda_2^* - \lambda_1^*} \\
\widehat{\gamma}_1 &= a_0 - \frac{\widehat{\gamma}_2}{\widehat{\gamma}_3 + \lambda_0^*}
\end{aligned}$$

An algorithm we have employed successfully to obtain starting values for fitting $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ starts by fitting a quadratic model to

$\{\widehat{\theta}(\lambda_m),\ \lambda_m\}_1^M$ where the λ_m are equally spaced over $[0,\ \lambda_{\max}]$. Initial parameter estimates for fitting $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ are obtained from a three-point fit to $(\widehat{a}_j, \lambda_j^*)_0^2$ where $\lambda_0^* = 0$, $\lambda_1^* = \lambda_{\max}/2$, $\lambda_2^* = \lambda_{\max}$ and $\widehat{a}_j$ is the predicted value corresponding to λ_j^* from the fitted quadratic model. In our experience initial values obtained in this fashion are generally very good and frequently differ insignificantly from the fully iterated, nonlinear least squares parameter estimates.

4.3.5 Inference and Standard Errors

Inference for SIMEX estimators can be performed either via the bootstrap or the theory of M-estimators (Appendix A), in particular by means of the sandwich estimator. Because of the computational burden of the SIMEX estimator, the bootstrap requires considerably more computing time than do other methods. Without efficient implementation of the estimation scheme at each step, even with current computing resources the SIMEX bootstrap may take an inconveniently long (clock) time to compute. On our computing system for measurement error models, the implementation is efficient, and most bootstrap applications take place in a reasonable (clock) time.

Asymptotic covariance estimation methods based on the sandwich estimator are described in Section 4.7.2. This is easy to implement in specific applications. Since the formulae look forbidding, we leave their listing until later.

When σ_u^2 is known or nearly so, the SIMEX calculations themselves admit a simple standard error estimator. Let $\tau_b^2(\lambda)$ be any variance estimator attached to $\widehat{\Theta}_b(\lambda)$, e.g., the sandwich estimator or the inverse of the information matrix, and let $\tau^2(\lambda)$ be their average for $b = 1, \ldots, B$. Let $s_\Delta^2(\lambda)$ be the sample covariance matrix of the terms $\widehat{\Theta}_b(\lambda)$ for $b = 1, \ldots, B$. Then as shown in Section 4.7.1, variance estimates for the SIMEX estimator can be obtained by extrapolating the components of the differences, $\widehat{\tau}^2(\lambda) - s_\Delta^2(\lambda)$, to $\lambda = -1$.

4.3.6 Relation to the Jackknife

The SIMEX algorithm resamples pseudo errors from the measurement error distribution and thus is reminiscent of a parametric

bootstrap procedure. The extrapolation step is similar to that underlying Quenouille's jackknife estimator (Quenouille, 1956) for reducing finite-sample bias, as described in Efron (1982). The theoretical relationship between SIMEX and Quenouille's jackknife is such that SIMEX can also be motivated and derived as an adaptation of the jackknife to measurement error problems.

The connection between SIMEX and the jackknife is involved and not necessary to understand SIMEX, and we give no further details here. The relationship is studied by Stefanski & Cook (1996). However, we do note that just as the ordinary jackknife also provides a variance estimator (Tukey, 1958), so too does SIMEX. The variance estimator mentioned at the end of section 4.3.5 and described in section 4.7.1 is related in theory to the usual jackknife variance estimator.

4.4 Nonadditive Measurement Error

We have described the SIMEX algorithm in terms of the additive measurement error model. However, SIMEX applies far more generally, and is easily extended to other models.

For example, consider multiplicative error. Taking logarithms transforms the multiplicative model to the additive model. In regression calibration, multiplicative error is handled in special ways (section 3.4.3). SIMEX works somewhat more naturally, in that one performs the simulation step (4.1) on the logarithms of the $\mathbf{W}$'s, and not on the $\mathbf{W}$'s themselves. Thus,

$$\mathbf{W}_{b,i}(\lambda) = \exp\left\{\log(\mathbf{W}_i) + \sqrt{\lambda}\mathbf{U}_{b,i}\right\}.$$

In general, suppose we can transform $\mathbf{W}$ to an additive model by a transformation $\mathcal{H}$, so that $\mathcal{H}(\mathbf{W}) = \mathcal{H}(\mathbf{X}) + \mathbf{U}$. This is an example of the transform-both-sides model, see (3.19). If $\mathcal{H}$ has an inverse function $\mathcal{G}$, then the simulation step generates

$$\mathbf{W}_{b,i}(\lambda) = \mathcal{G}\left\{\mathcal{H}(\mathbf{W}_i) + \sqrt{\lambda}\mathbf{U}_{b,i}\right\}.$$

In the multiplicative model, $\mathcal{H} = \log$, and $\mathcal{G} = \exp$. A standard class of transformation models is the power family discussed in section 3.7.

With replicates, one can also investigate the appropriateness of different transformations. For example, after transformation the

standard deviation of the intra-individual replicates should be uncorrelated with their mean, and one can find the power transformation which makes the two uncorrelated.

4.5 Framingham Heart Study

We illustrate the methods using data from the Framingham Heart Study, correcting for bias due to measurement error in systolic blood pressure measurements. The Framingham study consists of a series of exams taken two years apart. We use Exam #3 as the baseline. There are 1,615 men aged 31–65 in this data set, with the outcome, $\mathbf{Y}$, indicating the occurrence of coronary heart disease (CHD) within an eight-year period following Exam #3; there were 128 such cases of CHD. Predictors employed in this example are the patient's age at Exam #2, smoking status at Exam #1 and serum cholesterol at Exam #3, 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. In addition to the measurement error in SBP measurements, there also is measurement error in the cholesterol measurements. However, for this example we ignore the latter source of measurement error and illustrate the methods under the assumption that only SBP is measured with error.

The covariates measured without error, $\mathbf{Z}$, are age, smoking status and serum cholesterol. For $\mathbf{W}$, we employ a modified version of a transformation originally due to Cornfield and discussed by Carroll, Spiegelman, Lan, Bailey & Abbott (1984), setting $\mathbf{W} = \log(\text{SBP} - 50)$. Implicitly, we are defining $\mathbf{X}$ as the long-term average of $\mathbf{W}$.

In addition to the variables discussed above, we also have SBP measured at Exam #2. The mean transformed SBP at Exams #2 and #3 are 4.37 and 4.35, respectively. Their difference has mean 0.02, and standard error 0.0040, so that the large-sample test of equality of means has p-value < 0.0001. Thus in fact, the measurement at Exam #2 is not *exactly* a replicate, but the difference in means from Exam #2 to Exam #3 is close to negligible for all practical purposes.

We present two sets of analyses. The first analysis employs the full complement of replicate measurements from Exam #2. In the second analysis we illustrate the procedures for the case when only

a single measurement is employed and σ_u^2 is estimated by a small replication data set, obtained in this example by randomly selecting a subset of the Exam #2 SBP measurements.

4.5.1 Full Replication

This analysis uses the replicate SBP measurements from Exams #2 and #3 for all study participants. The transformed data are $\mathbf{W}_{i,j}$, where i denotes the individual and $j = 1, 2$ refers to the transformed SBP at Exams #2 and #3, respectively. The overall surrogate is $\overline{\mathbf{W}}_{i,\cdot}$, the sample mean for each individual. The model is

$$\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j},$$

where the $\mathbf{U}_{i,j}$ have mean zero and variance σ_u^2. The components of variance estimator (3.2) is $\widehat{\sigma}_u^2 = 0.01259$.

We employ SIMEX using $\mathbf{W}_i^* = \overline{\mathbf{W}}_{i,\cdot}$ and $\mathbf{U}_i^* = \overline{\mathbf{U}}_{i,\cdot}$. The sample variance of $(\mathbf{W}_i^*)_1^n$ is $\widehat{\sigma}_{w,*} = 0.04543$, and the estimated measurement error variance is $\widehat{\sigma}_{u,*}^2 = \widehat{\sigma}_u^2/2 = 0.00630$. Thus the linear model correction for attenuation, i.e., inverse of the reliability ratio, for these data is 1.16. There are 1,614 degrees of freedom for estimating $\widehat{\sigma}_{u,*}^2$ and thus for practical purposes the measurement error variance is known.

In Table 4.1, we list the results of the naive analysis that ignores measurement error, the regression calibration analysis, and the SIMEX analysis. For the naive analysis, sandwich and information refer to the sandwich and information standard errors discussed in Appendix A; the latter is the output from standard statistical packages.

For the regression calibration analysis, the first set of sandwich and information standard errors are those obtained from a standard logistic regression analysis having substituted the calibration equation for $\mathbf{W}$, and ignoring the fact that the equation is estimated. The second set of sandwich standard errors are as described in Section 3.12, while the bootstrap analysis uses the methods of Appendix A.

For the SIMEX estimator, M-estimator refers to estimates derived from the theory of Section 4.7.2 for the case where σ_u^2 is estimated from the replicate measurements. Sandwich and Information refer to estimates defined in Section 4.3.5 (theory in Section

	Age	Smoke	Chol	LSBP
Naive	.055	.59	.0078	1.70
Sand.	.010	.24	.0019	.39
Info.	.011	.24	.0021	.41
Reg. Cal.	.053	.60	.0077	2.00
Sand.[1]	.010	.24	.0019	.46
Info.[1]	.011	.25	.0021	.49
Sand.[2]	.010	.24	.0019	.46
Bootstrap	.010	.25	.0019	.46
SIMEX	.053	.60	.0078	1.93
Simex, Sand.[3]	.010	.24	.0019	.43
Simex, Info.[3]	.011	.25	.0021	.47
M-est. [4]	.010	.24	.0019	.44

Table 4.1. *Estimates and standard errors from the Framingham data logistic regression analysis. This analysis assumes that all observations have replicated SBP. Sand., sandwich; Info., information;* [1]*, calibration function known;* [2]*, calibration function estimated;* [3]*,* σ_u^2 *known;* [4]*,* σ_u^2 *estimated. Here "Smoke" is smoking status, "Chol" is cholesterol and "LSBP" is log(SBP−50).*

4.7.1), with $\widehat{\tau}^2(\lambda)$ derived from the naive sandwich and naive information estimates, respectively. The M-estimation sandwich and SIMEX sandwich standard errors yield nearly identical standard errors because σ_u^2 is so well estimated.

Figure 4.2 contains plots of the logistic regression coefficients $\widehat{\Theta}(\lambda)$ for eight equally spaced values of λ spanning $[0, 2]$ (solid circles). For this example $B = 2000$. The points plotted at $\lambda = 0$ are the naive estimates $\widehat{\Theta}_{\text{naive}}$.

The nonlinear least-squares fits of $\mathcal{G}_{\text{RL}}(\lambda, \Gamma)$ to the components of $\{\widehat{\Theta}(\lambda_m), \lambda_m\}_1^8$ (solid curves) are extrapolated to $\lambda = -1$ (dashed curves) resulting in the SIMEX estimators (crosses). The open circles are the SIMEX estimators that result from fitting quadratic

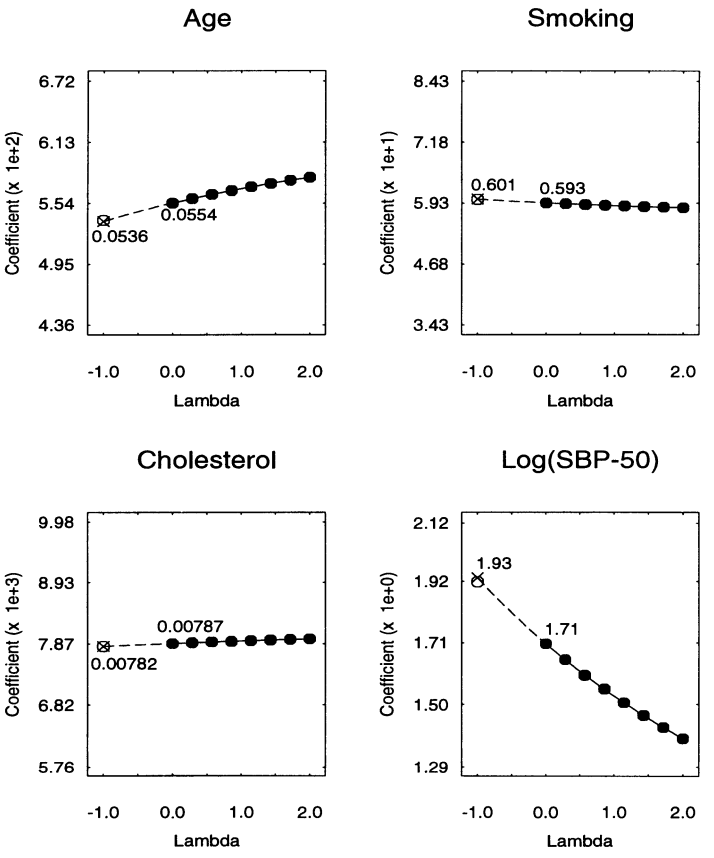

Figure 4.2. *Coefficient extrapolation functions for the Framingham logistic regression modeling. The simulated estimates* $\{\widehat{\Theta}(\lambda_m),\ \lambda_m\}_1^8$ *are plotted (solid circles) and the fitted rational linear extrapolant (solid line) is extrapolated to* $\lambda = -1$ *(dashed line) resulting in the SIMEX estimate (cross). Open circles indicate SIMEX estimates obtained with the quadratic extrapolant.*

extrapolants. To preserve clarity the quadratic extrapolants were not plotted. Note that the quadratic-extrapolant estimates are conservative relative to the rational linear-extrapolant estimates in the sense that they fall between the rational linear-extrapolant estimates and the naive estimates.

We have stated previously that the SIMEX plot displays the effect of measurement error on parameter estimates. This is especially noticeable in Figure 4.2. In each of the four graphs in Figure 4.2, the range of the ordinate corresponds to a one-standard error confidence interval for the naive estimate constructed using the information standard errors. Thus Figure 4.2 illustrates the effect of measurement error relative to the variability in the naive estimate. It is apparent that the effect of measurement error is of practical importance only on the coefficient of log(SBP − 50).

The SIMEX sandwich and the M-estimation (with σ_u^2 estimated) methods of variance estimation yield similar results in this example. The difference between the SIMEX sandwich and information methods is due to differences in the naive sandwich and information methods for these data.

Figure 4.3 displays the variance extrapolant functions fit to the components of $\widehat{\tau}^2(\lambda) - s_\Delta^2(\lambda)$ used to obtain the SIMEX information variances and standard errors. The figure is constructed using the same conventions used in the construction of Figure 4.2. For these plots the ranges of the ordinates are $(1/2)\widehat{\text{var}}(\text{naive})$ to $(4/3)\widehat{\text{var}}(\text{naive})$, where $\widehat{\text{var}}(\text{naive})$ is the information variance estimate of the naive estimator.

4.5.2 Partial Replication

We now illustrate the analyses for the case where σ_u^2 is estimated from a small replication data set. The measured predictor, $\mathbf{W}$, is the single SBP measurement from Exam #3. A randomly selected subset of 30 replicate measurements from Exams #2 and #3 were used to estimate σ_u^2. For these data the sample variance of $\mathbf{W}$ is .05252 and the estimate of σ_u^2 is .01306. The estimated linear model correction for attenuation, or inverse of the reliability ratio, is 1.33.

There are two major differences between this set of analyses and those from the previous section: (i) the measurement error variance is twice as large because we are using only Exam #3 and not its average with Exam #2, thus resulting in greater attenuation

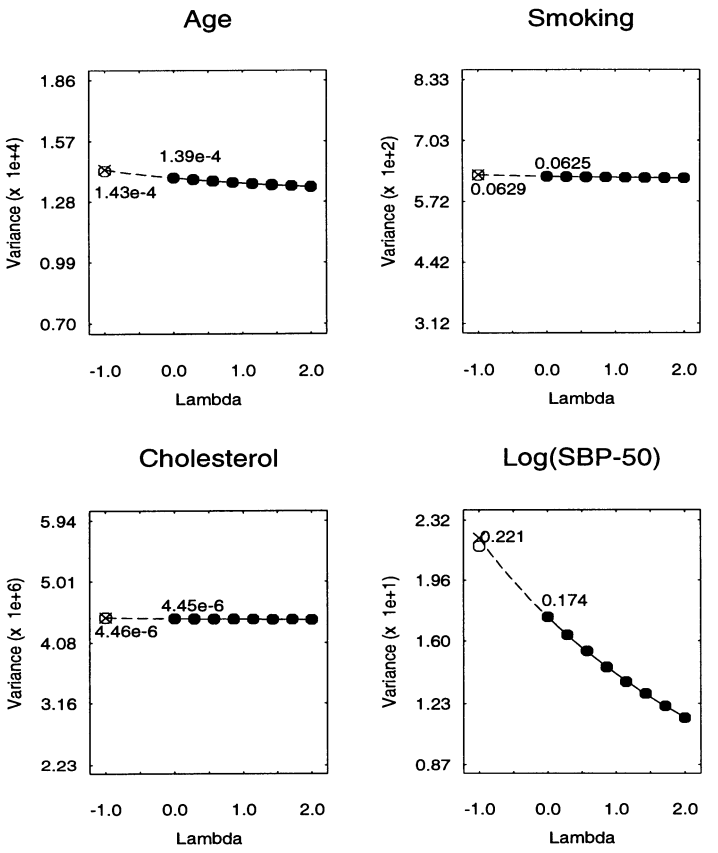

Figure 4.3. *Variance extrapolation functions for the Framingham logistic regression variance estimation. Values of* $\{(\widehat{\tau}^2(\lambda_m) - s^2_\Delta(\lambda_m)),\ \lambda_m\}_1^8$ *are plotted (solid circles) and the fitted rational linear extrapolant (solid line) is extrapolated to* $\lambda = -1$ *(dashed line) resulting in the SIMEX variance estimate (cross). Open circles indicate SIMEX variance estimates obtained with the quadratic extrapolant.*

	Age	Smoke	Chol	LSBP
Naive	.056	.573	.0078	1.52
Sand.	.010	.243	.0019	.36
Info.	.011	.249	.0021	.38
Reg. Cal.	.053	.582	.0075	2.07
Sand.[1]	.010	.243	.0020	.49
Infor.[1]	.011	.249	.0021	.52
Sand.[2]	.010	.243	.0020	.53
SIMEX	.053	.581	.0077	1.94
Simex, Sand.[3]	.011	.261	.0020	.46
Simex, Info.[3]	.012	.251	.0021	.49
M-est.[4]	.011	.245	.0020	.54

Table 4.2. *Estimates and standard errors from the Framingham data logistic regression analysis. This analysis is based upon a randomly selected replication data set of size 30. Sand., sandwich; Info., information;* [1]*, calibration function known;* [2]*, calibration function estimated;* [3]*,* σ_u^2 *known;* [4]*,* σ_u^2 *estimated.*

in the naive estimate; and (ii) the measurement error variance is estimated with far less precision, 29 degrees of freedom versus 1614, resulting in less precise corrected estimates.

Both of these differences are reflected in the results reported in Table 4.2. The standard errors are calculated as in Table 4.1 with the exception that bootstrap standard errors were not calculated for the regression calibration estimates.

4.6 SIMEX in Some Important Special Cases

This section describes the bias-correction properties of SIMEX in four important special cases.

4.6.1 Multiple Linear Regression

Consider the multiple linear regression model

$$\mathbf{Y}_i = \beta_1 + \beta_z^t \mathbf{Z}_i + \beta_x \mathbf{X}_i + \epsilon_i.$$

In the notation of section 4.3, $\Theta = (\beta_1, \beta_z^t, \beta_x)^t$. If non-iid pseudo errors are employed in the SIMEX simulation step, it is readily seen that

$$\widehat{\Theta}(\lambda) = \left\{ \sum_{i=1}^{n} \begin{pmatrix} 1 & \mathbf{Z}_i^t & \mathbf{W}_i \\ \mathbf{Z}_i & \mathbf{Z}_i\mathbf{Z}_i^t & \mathbf{Z}_i\mathbf{W}_i \\ \mathbf{W}_i & \mathbf{W}_i\mathbf{Z}_i^t & \mathbf{W}_i^2 + \lambda\sigma_u^2 \end{pmatrix} \right\}^{-1} \times \left\{ \sum_{i=1}^{n} \begin{pmatrix} \mathbf{Y}_i \\ \mathbf{Z}_i\mathbf{Y}_i \\ \mathbf{W}_i\mathbf{Y}_i \end{pmatrix} \right\}.$$

Solving this system of equations we find that

$$\widehat{\beta}_v(\lambda) = (\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{Y} - \frac{(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{W}\left(\mathbf{W}^t\mathbf{Y} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{Y}\right)}{\mathbf{W}^t\mathbf{W} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{W} + \lambda\sigma^2}, \tag{4.5}$$

$$\widehat{\beta}_x(\lambda) = \frac{\mathbf{W}^t\mathbf{Y} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{Y}}{\mathbf{W}^t\mathbf{W} - \mathbf{W}^t\mathbf{V}(\mathbf{V}^t\mathbf{V})^{-1}\mathbf{V}^t\mathbf{W} + \lambda\sigma^2}, \tag{4.6}$$

where $\beta_v = (\beta_1, \beta_z^t)^t$, $\mathbf{V}^t = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n)$ with $\mathbf{V}_i = (1, \mathbf{Z}_i^t)^t$. Note that all of the components of $\widehat{\Theta}(\lambda)$ are functions of λ of the form $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ for suitably defined, component-dependent $\Gamma = (\gamma_1, \gamma_2, \gamma_3)^t$.

It follows that if the models fit in the SIMEX extrapolation step have the form $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$, allowing different Γ for different components, then SIMEX results in the usual method-of-moments estimator of Θ.

4.6.2 Loglinear Mean Models

Suppose that $\mathbf{X}$ is a scalar and that $E(\mathbf{Y}|\mathbf{X}) = \exp(\beta_0 + \beta_x\mathbf{X})$, with variance function $\mathrm{Var}(\mathbf{Y} \mid \mathbf{X}) = \sigma^2 \exp\{\theta(\beta_0 + \beta_x\mathbf{X})\}$ for some constants σ^2 and θ. It follows from the appendix in Stefanski (1989) that if $(\mathbf{W}, \mathbf{X})$ has a bivariate normal distribution and generalized least squares is the method of estimation, then $\widehat{\beta}_0(\lambda)$ and $\widehat{\beta}_x(\lambda)$

consistently estimate

$$\beta_0(\lambda) = \beta_0 + (1+\lambda)\frac{\mu_x\sigma_u^2\beta_x + \beta_x^2\sigma_x^2\sigma_u^2/2}{\sigma_x^2 + (1+\lambda)\sigma_u^2}$$

and

$$\beta_x(\lambda) = \frac{\beta_x\sigma_x^2}{\sigma_x^2 + (1+\lambda)\sigma_u^2}$$

respectively, where $\mu_x = E(\mathbf{X})$, $\sigma_x^2 = \mathrm{Var}(\mathbf{X})$ and $\sigma_u^2 = \mathrm{Var}(\mathbf{W} \mid \mathbf{X})$.

The rational linear extrapolant is asymptotically exact for estimating both β_0 and β_x.

4.6.3 Quadratic Mean Models

Consider fitting a quadratic regression model using orthogonal polynomials and least square estimation. Components of the parameter vector $\Theta = (\beta_0, \beta_{x,1}, \beta_{x,2})^t$ are the coefficients in the linear model

$$\mathbf{Y}_i = \beta_0 + \beta_{x,1}(\mathbf{X}_i - \overline{\mathbf{X}}) + \beta_{x,2}(\mathbf{X}_i^2 - a - b\mathbf{X}_i) + \epsilon_i,$$

where $a = a\{(\mathbf{X}_i)_1^n\}$ and $b = b\{(\mathbf{X}_i)_1^n\}$ are the intercept and slope, respectively, of the least squares regression line of $\mathbf{X}_i^2$ on $\mathbf{X}_i$. The so-called naive estimator for this model is obtained by fitting the quadratic regression to $(\mathbf{Y}_i,\ \mathbf{W}_i)_1^n$ noting that $\mathbf{W}_i$ replaces $\mathbf{X}_i$, $i = 1, \ldots, n$, in the definitions of a and b.

Let $\mu_{x,j} = E(\mathbf{X}^j)$, $j = 1, \ldots, 4$. We assume for simplicity that $\mu_{x,1} = 0$ and $\mu_{x,2} = 1$. The exact functional form of $\widehat{\Theta}_b(\lambda)$ is known for this model and is used to show that asymptotically, $\widehat{\Theta}(\lambda)$ converges in probability to $\Theta(\lambda)$ given by

$$\begin{aligned}
\beta_0(\lambda) &= \beta_0, \\
\beta_{x,1}(\lambda) &= \frac{\beta_{x,1}\sigma_x^2}{\sigma_x^2 + \delta}, \\
\beta_{x,2}(\lambda) &= \frac{\mu_{x,3}\beta_{x,1}\delta + (1+\delta)\beta_{x,2}(\mu_{x,4} - 1) - \mu_{x,3}^2\beta_{x,2}}{(1+\delta)(\mu_{x,4} - 1 + 4\delta + 2\delta^2) - \mu_{x,3}^2},
\end{aligned}$$

where $\delta = (1+\lambda)\sigma_u^2$.

Note that both $\beta_0(\lambda)$ and $\beta_{x,1}(\lambda)$ are functions of λ of the form $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ whereas $\beta_{x,2}(\lambda)$ is not. For arbitrary choices of σ_u^2, $\mu_{x,3}$, $\mu_{x,4}$, $\beta_{x,1}$ and $\beta_{x,2}$, the shape of $\beta_{x,2}(\lambda)$ can vary dramatically for

$-1 \leq \lambda \leq 2$ thereby invalidating the extrapolation step employing an approximate extrapolant. However, in many practical cases the quadratic extrapolant corrects for most of the bias, especially for σ_u^2 sufficiently small. When $\mathbf{X}$ is normally distributed, $\beta_{x,2}(\lambda) = \beta_{x,2}/(1+\delta)^2$ which is monotone for all $\lambda \geq -1$ and reasonably well approximated by either a quadratic or $\mathcal{G}_{\mathrm{RL}}(\lambda, \Gamma)$ for a limited, but useful, range of values of σ_u^2.

4.6.4 Segmented Linear Regression Mean Models

A particularly difficult nonlinear model occurs when two unknown regression lines are joined at an unknown change point. We take the simplest case where one of the lines is known to have zero slope, and the change point is thus a threshold. The model for the mean is given by

$$E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1(\mathbf{X} - \beta_2)_+,$$

where

$$(x - \beta_2)_+ = \begin{cases} 0 & \text{if} \quad x < \beta_2; \\ x - \beta_2 & \text{if} \quad x \geq \beta_2. \end{cases}$$

Because β_2 is a threshold, it is called the threshold limiting value (TLV).

The effects of measurement error on estimating the TLV cannot be easily described mathematically. However, under the assumption that $\mathbf{X}$, $\mathbf{U}$, and the error in the regression of $\mathbf{Y}$ on $\mathbf{X}$ are independent and normally distributed, it is possible to compute the limiting values of the naive and SIMEX estimators, see Küchenhoff & Carroll (1995).

The limiting values of the SIMEX TLV estimators are given in Figure 4.4. While we have displayed only the quadratic extrapolant, with either the linear, quadratic or rational linear extrapolants, the SIMEX estimator provides estimates that are much closer to the actual value of β_2 in large samples than the naive estimate. The quadratic and rational linear extrapolants result in nearly consistent estimates of the change point.

4.7 Theory and Variance Estimation

The ease with which estimates can be obtained via SIMEX, even for very complicated and nonstandard models, is offset somewhat by

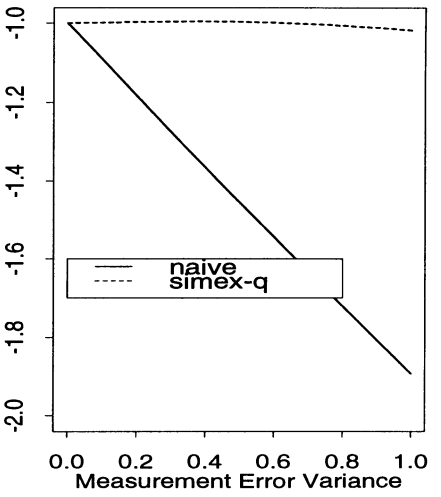

Figure 4.4. *The segmented regression model. For given amounts of measurement error, the actual limiting value of the naive estimator ("naive") and SIMEX estimator with quadratic extrapolant ("simex-q") for estimating the threshold limiting value (TLV)* $\beta_2 = -1$. *Deviations from the value* $\beta_2 = -1$ *on the vertical axis represent large-sample bias. Here* $\beta_0 = 0$, $\beta_1 = 2$ *and* $\sigma_x^2 = 1$.

the complexity of the resulting estimates, making the calculation of standard errors difficult or at least nonstandard. Except for the computational burden of nested resampling schemes, SIMEX is a natural candidate for the use of the bootstrap or a standard implementation of Tukey's jackknife to calculate standard errors.

We now describe two methods of estimating the covariance matrix of the asymptotic distribution of $\widehat{\Theta}_{\text{simex}}$ that avoid nested resampling. The first method uses the pseudo estimates, $\widehat{\Theta}_b(\lambda)$, generated during the SIMEX simulation step in a procedure akin to Tukey's jackknife variance estimate. Its applicability is limited to situations in which σ_u^2 is known or estimated well enough to justify such an assumption. The second method exploits the fact that $\widehat{\Theta}_{\text{simex}}$ is asymptotically equivalent to an M-estimator and makes

use of standard formulae from Appendix A. This method requires additional programming, but has the flexibility to accommodate situations in which σ_u^2 is estimated and the variation in $\widehat{\sigma}_u^2$ is not negligible.

4.7.1 Simulation Extrapolation Variance Estimation

Stefanski & Cook (1996) establish a close relationship between SIMEX inference and jackknife inference. In particular they identify a method of variance estimation applicable when σ_u^2 is known that closely parallels Tukey's jackknife variance estimation. We now describe the implementation of their method of estimating $\text{var}(\widehat{\Theta}_{\text{simex}})$.

It is convenient to introduce a function $\mathcal{T}$ to denote the estimator under study. For example, $\mathcal{T}\{(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i)_1^n\}$ is the estimator of Θ when $\mathbf{X}$ is observable, and $\mathcal{T}\{(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n\}$ is the naive estimator.

For theoretical purposes we redefine

$$\widehat{\Theta}(\lambda) = E\left\{\widehat{\Theta}_b(\lambda) \mid (\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n\right\}. \tag{4.7}$$

The expectation in (4.7) is with respect to the distribution of $(\mathbf{U}_{b,i})_{i=1}^n$ only, since we condition on the observed data. It can be obtained as the limit as $B \to \infty$ of the average $\{\widehat{\Theta}_1(\lambda) + \cdots + \widehat{\Theta}_B(\lambda)\}/B$. In effect, $\widehat{\Theta}(\lambda)$ is the estimator obtained when computing power is unlimited.

We now introduce a second function, $\mathcal{T}_{\text{var}}$ to denote an associated variance estimator, i.e.,

$$\mathcal{T}_{\text{var}}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\} = \widehat{\text{var}}(\widehat{\Theta}_{\text{true}}) = \widehat{\text{var}}[\mathcal{T}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\}]$$

where $\widehat{\Theta}_{\text{true}}$ denotes the "estimator" calculated from the "true" data $(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n$.

We allow $\mathcal{T}$ to be p-dimensional, in which case $\mathcal{T}_{\text{var}}$ is $(p \times p)$-matrix valued, and variance refers to the variance-covariance matrix. For example, $\mathcal{T}_{\text{var}}$ could be the inverse of the information matrix when $\widehat{\Theta}_{\text{true}}$ is a maximum likelihood estimator. Alternatively, $\mathcal{T}_{\text{var}}$ could be a sandwich estimator for either a maximum likelihood estimator or a general M-estimator (Appendix A).

We use τ^2 to denote the parameter $\text{var}(\widehat{\Theta}_{\text{true}})$, $\widehat{\tau}^2_{\text{true}}$ to denote the true variance estimator $\mathcal{T}_{\text{var}}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\}$, and $\widehat{\tau}^2_{\text{naive}}$ to

denote the naive variance estimator $\mathcal{T}_{\text{var}}\{(\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_i)_1^n\}$.

Stefanski & Cook (1996) show that

$$E\{\widehat{\Theta}_{\text{simex}} \mid (\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{X}_i)_1^n\} \approx \widehat{\Theta}_{\text{true}}, \tag{4.8}$$

where the approximation is due to both a large-sample approximation and to use of an approximate extrapolant function. We will make use of such approximations without further explanation; see Stefanski & Cook (1996) for additional explanation.

It follows from Equation (4.8) that

$$\text{var}(\widehat{\Theta}_{\text{simex}}) \approx \text{var}(\widehat{\Theta}_{\text{true}}) + \text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}}). \tag{4.9}$$

Equation (4.9) decomposes the variance of $\widehat{\Theta}_{\text{simex}}$ into a component due to sampling variability, $\text{var}(\widehat{\Theta}_{\text{true}}) = \tau^2$, and a component due to measurement error variability, $\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}})$.

SIMEX estimation can be used to estimate the first component τ^2. That is,

$$\widehat{\tau}_b^2(\lambda) = \mathcal{T}_{\text{var}}[\{\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_{b,i}(\lambda)\}_1^n]$$

is calculated for $b = 1, \ldots, B$, and upon averaging and letting $B \to \infty$, results in $\widehat{\tau}^2(\lambda)$. The components of $\widehat{\tau}^2(\lambda)$ are then plotted as functions of λ, extrapolant models are fit to the components of $\{\widehat{\tau}^2(\lambda_m), \lambda_m\}_1^M$ and the modeled values at $\lambda = -1$ are estimates of the corresponding components of τ^2.

The basic building blocks required to estimate the second component of the variance, $\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}})$, are the differences

$$\Delta_b(\lambda) = \widehat{\Theta}_b(\lambda) - \widehat{\Theta}(\lambda), \quad b = 1, \ldots, B. \tag{4.10}$$

Define

$$s_\Delta^2(\lambda) = (B-1)^{-1} \sum_{b=1}^{B} \Delta_b(\lambda)\Delta_b^t(\lambda), \tag{4.11}$$

i.e., the sample variance matrix of $\{\widehat{\Theta}_b(\lambda)\}_{b=1}^B$. Its significance stems from the fact that

$$\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}}) = -\lim_{\lambda \to -1} \text{var}\{\widehat{\Theta}_b(\lambda) - \widehat{\Theta}(\lambda)\} \tag{4.12}$$

see Stefanski & Cook (1996).

The variance matrix $s_\Delta^2(\lambda)$ is an unbiased estimator of the conditional variance $\text{var}\{\widehat{\Theta}_b(\lambda) - \widehat{\Theta}(\lambda) \mid (\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n\}$ for all $B > 1$ and

converges in probability to its conditional expectation as $B \to \infty$. Since $E\{\widehat{\Theta}_b(\lambda) - \widehat{\Theta}(\lambda) \mid (\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)_1^n\} = 0$ it follows that unconditionally $E\{s_\Delta^2(\lambda)\} = \text{var}\{\widehat{\Theta}_b(\lambda) - \widehat{\Theta}(\lambda)\}$.

Thus the component of variance we want to estimate is given by

$$\text{var}(\widehat{\Theta}_{\text{simex}} - \widehat{\Theta}_{\text{true}}) = -\lim_{\lambda \to -1} E\{s_\Delta^2(\lambda)\}.$$

This can be (approximately) estimated by fitting models to the components of $s_\Delta^2(\lambda)$ as functions of $\lambda > 0$ and extrapolating the component models back to $\lambda = -1$. We use $\widehat{s}_\Delta^2(-1)$ to denote the estimated variance matrix obtained by this procedure.

In light of (4.9), the definition of τ^2, and (4.12) the difference, $\widehat{\tau}^2_{\text{simex}} - \widehat{s}_\Delta^2(-1)$, is an estimator of $\text{var}\{\widehat{\Theta}_{\text{simex}}\}$. In practice, separate extrapolant functions are not fit to the components of both $\widehat{\tau}^2(\lambda)$ and $s_\Delta^2(\lambda)$, but rather the components of the difference, $\widehat{\tau}^2(\lambda) - s_\Delta^2(\lambda)$, are modeled and extrapolated to $\lambda = -1$.

In summary, for SIMEX estimation with known σ_u^2, the simulation step results in $\widehat{\Theta}(\lambda)$, $\widehat{\tau}^2(\lambda)$ and $s_\Delta^2(\lambda)$ for $\lambda \in \Lambda$. The model extrapolation of $\widehat{\Theta}(\lambda)$ to $\lambda = -1$, $\widehat{\Theta}_{\text{simex}}$, provides an estimator of Θ, and the model extrapolation of (the components of) the difference, $\widehat{\tau}^2(\lambda) - s_\Delta^2(\lambda)$ to $\lambda = -1$ provides an estimator of $\text{var}(\widehat{\Theta}_{\text{simex}})$. It should be emphasized that the entire procedure is approximate in the sense that it is generally valid only in large samples with small measurement error.

There is no guarantee that the estimated covariance matrix so obtained is positive definite. This is similar to the nonpositivity problems that arise in estimating components-of-variance. We have not encountered problems of this nature, although there is no guarantee that they will not occur. If it transpires that the estimated variance of a linear combination, say $\gamma^t\widehat{\Theta}$, is negative, a possible course of action is to plot, model and extrapolate directly the points $[\gamma^t\{\widehat{\tau}^2(\lambda_m) - s_\Delta^2(\lambda_m\}\gamma,\ \lambda_m]_1^M$.

4.7.2 Estimating Equation Approach to Variance Estimation

This section is based on the results in Carroll, Küchenhoff, Lombard & Stefanski (1996). Assuming iid sampling these authors show that $\widehat{\Theta}_{\text{simex}}$ is asymptotically normally distributed and propose an estimator of its asymptotic covariance matrix. We highlight the main points of the asymptotic analysis in order to motivate the

proposed variance estimator.

We describe the application of SIMEX in the setting of M-estimation, i.e., using unbiased estimating equations (Appendix A), assuming that in the absence of measurement errors, M-estimation produces consistent estimators.

The estimator obtained in the absence of measurement error is denoted $\widehat{\Theta}_{\text{true}}$ and solves the system of equations

$$0 = n^{-1} \sum_{i=1}^{n} \Psi \left(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \widehat{\Theta}_{\text{true}} \right). \tag{4.13}$$

This is just a version of (A.5), and is hence applicable to variance function and generalized linear models. In multiple linear regression, $\Psi(\cdot)$ represents the normal equations for a single observation, namely

$$\Psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta) = (\mathbf{Y} - \beta_0 - \beta_z^t \mathbf{Z} - \beta_x \mathbf{X}) \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{X} \end{pmatrix}.$$

In multiple logistic regression, with $H(\cdot)$ being the logistic distribution function,

$$\Psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta) = \left\{ \mathbf{Y} - H \left(\beta_0 + \beta_z^t \mathbf{Z} + \beta_x \mathbf{X} \right) \right\} \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{X} \end{pmatrix}.$$

The solution to (4.13) cannot be calculated, since it depends on the unobserved true predictors. The estimator obtained by ignoring measurement error is denoted by $\widehat{\Theta}_{\text{naive}}$ and solves the system of equations

$$0 = n^{-1} \sum_{i=1}^{n} \Psi \left(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}_{\text{naive}} \right).$$

For fixed b and λ, and large n a standard linearization (Appendix A) shows that

$$\begin{aligned} n^{1/2} \left\{ \widehat{\Theta}_b(\lambda) - \Theta(\lambda) \right\} \approx & -\mathcal{A}^{-1} \{ \sigma_u^2, \lambda, \Theta(\lambda) \} \\ & \times n^{-1/2} \sum_{i=1}^{n} \Psi \{ \mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_{b,i}(\lambda), \Theta(\lambda) \}, \end{aligned} \tag{4.14}$$

where

$$\mathcal{A}\{\sigma_u^2, \lambda, \Theta(\lambda)\} = E[\Psi_\Theta\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W}_{b,i}(\lambda), \Theta(\lambda)\}],$$

and

$$\Psi_\Theta\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W}_{b,i}(\lambda), \Theta\} = (\partial/\partial\Theta^t)\Psi\{\mathbf{Y},\ \mathbf{Z},\ \mathbf{W}_{b,i}(\lambda), \Theta\}.$$

Averaging (4.14) over b results in the asymptotic approximation

$$n^{1/2}\left\{\widehat{\Theta}(\lambda) - \Theta(\lambda)\right\} \approx -\mathcal{A}^{-1}(\cdot) \times n^{-1/2}\sum_{i=1}^{n}\chi_{B,i}\{\sigma_u^2,\ \lambda,\ \Theta(\lambda)\}, \tag{4.15}$$

where

$$\chi_{B,i}\{\sigma_u^2, \lambda, \Theta(\lambda)\} = B^{-1}\sum_{b=1}^{B}\Psi\{\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_{b,i}(\lambda), \Theta(\lambda)\},$$

and $\mathcal{A}^{-1}(\cdot) = \mathcal{A}^{-1}\{\sigma_u^2, \lambda, \Theta(\lambda)\}$. The summands $\chi_{B,i}(\cdot)$ in (4.15) are independent and identically distributed with mean zero.

Let $\Lambda = \{\lambda_1, \ldots, \lambda_M\}$ denote the grid of values used in the extrapolation step. Let $\widehat{\Theta}_*(\Lambda)$ denote the vector of estimators, $\left\{\widehat{\Theta}^t(\lambda_1), \ldots, \widehat{\Theta}^t(\lambda_M)\right\}^t$, which we also denote $\text{vec}\{\widehat{\Theta}(\lambda), \lambda \in \Lambda\}$. The corresponding vector of estimands is denoted by $\Theta_*(\Lambda)$. Define

$$\Psi_{B,i(1)}\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\} = \text{vec}[\chi_{B,i}\{\sigma_u^2, \lambda, \Theta(\lambda)\}, \lambda \in \Lambda]$$

and

$$\mathcal{A}_{11}\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\} = \text{diag}[\mathcal{A}\{\sigma_u^2,\ \lambda,\ \Theta(\lambda)\}, \lambda \in \Lambda].$$

Then, using (4.15), the joint limit distribution of $n^{1/2}\{\widehat{\Theta}_*(\Lambda) - \Theta_*(\Lambda)\}$ is seen to be multivariate normally distributed with mean zero and covariance Σ, where

$$\Sigma = \mathcal{A}_{11}^{-1}(\cdot)\mathcal{C}_{11}\left\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\right\}\left\{\mathcal{A}_{11}^{-1}(\cdot)\right\}^t \tag{4.16}$$

$$\mathcal{C}_{11}\left\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\right\} = \text{Cov}\left[\Psi_{B,1(1)}\left\{\sigma_u^2, \Lambda, \Theta_*(\Lambda)\right\}\right]. \tag{4.17}$$

Define

$$\mathcal{G}^*(\Lambda, \Gamma^*) = \text{vec}\left[\{\mathcal{G}(\lambda_m, \Gamma_j)\}_{m=1,\ldots,M,\ j=1,\ldots,p}\right]$$

where $\Gamma^* = (\Gamma_1^t, \ldots, \Gamma_p^t)^t$ and Γ_j is the parameter vector estimated in the extrapolation step for the jth component of $\widehat{\Theta}(\lambda)$, $j = 1, \ldots, p$.

Define $R(\Gamma^*) = \widehat{\Theta}_*(\Lambda) - \mathcal{G}^*(\Lambda, \Gamma^*)$. The extrapolation steps results in $\widehat{\Gamma}^*$, obtained by minimizing $R^t(\Gamma^*)R(\Gamma^*)$. The estimating equation for $\widehat{\Gamma}^*$ has the form $0 = s(\Gamma^*)R(\Gamma^*)$ where $s^t(\Gamma^*) = \{\partial/\partial(\Gamma^*)^t\}R(\Gamma^*)$. With $D(\Gamma^*) = s(\Gamma^*)s^t(\Gamma^*)$, standard asymptotic results show that

$$n^{-1/2}\left(\widehat{\Gamma}^* - \Gamma^*\right) \approx \mathrm{N}\left\{0,\ \Sigma(\Gamma^*)\right\}$$

where

$$\Sigma(\Gamma^*) = D^{-1}(\Gamma^*)s(\Gamma^*)\Sigma s^t(\Gamma^*)D^{-1}(\Gamma^*)$$

and Σ is given by (4.16). Now $\widehat{\Theta}_{\mathrm{simex}} = \mathcal{G}^*(-1, \widehat{\Gamma}^*)$ and thus by the Δ method, the $\sqrt{n}$-normalized SIMEX estimator is asymptotically normal with asymptotic variance,

$$\mathcal{G}^*_{\Gamma^*}(-1, \Gamma^*)\Sigma(\Gamma^*)\left\{\mathcal{G}^*_{\Gamma^*}(-1, \Gamma^*)\right\}^t$$

where $\mathcal{G}^*_{\Gamma^*}(\lambda, \Gamma^*) = \{\partial/\partial(\Gamma^*)^t\}\,\mathcal{G}^*(\lambda, \Gamma^*)$.

Note that the matrix $C_{11}(\cdot)$ in (4.17) is consistently estimated by $\widehat{C}_{11}(\cdot)$, the sample covariance matrix of $[\Psi_{B,i(1)}\{\sigma_u^2, \Lambda, \widehat{\Theta}_*(\Lambda)\}]_1^n$. Also, $A_{11}(\cdot)$ is consistently estimated by $\widehat{\mathcal{A}}_{11}(\cdot) = \mathrm{diag}\left\{\widehat{\mathcal{A}}_m(\cdot)\right\}$ for $m = 1, \ldots, M$, where

$$\widehat{\mathcal{A}}_m(\cdot) = (nB)^{-1}\sum_{i=1}^{n}\sum_{b=1}^{B}\Psi_\Theta\left\{\mathbf{Y}_i,\ \mathbf{Z}_i,\ \mathbf{W}_{b,i}(\lambda_m),\ \widehat{\Theta}(\lambda_m)\right\}.$$

The indicated variance estimator is

$$n^{-1}\mathcal{G}^*_{\Gamma^*}(-1, \widehat{\Gamma}^*)\widehat{\Sigma}(\widehat{\Gamma}^*)\left\{\mathcal{G}^*_{\Gamma^*}(-1, \widehat{\Gamma}^*)\right\}^t, \tag{4.18}$$

where

$$\widehat{\Sigma}(\widehat{\Gamma}^*) = \widehat{D}^{-1}(\widehat{\Gamma}^*)s(\widehat{\Gamma}^*)\widehat{\Sigma}s^t(\widehat{\Gamma}^*)\widehat{D}^{-1}(\widehat{\Gamma}^*);$$
$$\widehat{D}(\widehat{\Gamma}^*) = s(\widehat{\Gamma}^*)s^t(\widehat{\Gamma}^*);$$
$$\widehat{\Sigma} = \widehat{\mathcal{A}}_{11}^{-1}(\cdot)\widehat{C}_{11}^{-1}(\cdot)\left\{\widehat{\mathcal{A}}_{11}^{-1}(\cdot)\right\}^t.$$

When σ_u^2 is estimated, the estimating equation approach is modified by the inclusion of additional estimating equations employed in the estimation of $\widehat{\sigma}_u^2$. We illustrate the case in which each $\mathbf{W}_i$ is the mean of two replicate measurements, $\mathbf{W}_{ij}$, $j = 1, 2$ where

$$\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{i,j}, \qquad j = 1, 2, \quad i = 1, \ldots, n.$$

With replicates, $\mathbf{W}_i$ is replaced by $\mathbf{W}_i^* = \overline{\mathbf{W}}_{i,\cdot}$ and σ_u^2 by $\sigma_{u,*}^2 = \sigma_u^2/2$.

Let

$$\Psi_{(i)2}(\sigma_{u,*}^2, \mu) = \begin{Bmatrix} (\mathbf{D}_i - \mu)^2 - \sigma_{u,*}^2 \\ \mathbf{D}_i - \mu \end{Bmatrix},$$

where $\mathbf{D}_i = (\mathbf{W}_{i1} - \mathbf{W}_{i2})/2$. Then solving $\sum \Psi_{i(2)}(\sigma_{u,*}^2, \mu) = 0$, results in the estimators $\widehat{\mu} = \overline{\mathbf{D}}$ and $\widehat{\sigma}_{u,*}^2 = (n-1)s_d^2/n$ where s_d^2 is the sample variance of $(\mathbf{D}_i)_1^n$ and consistently estimates $\sigma_{u,*}^2$.

By combining $\Psi_{B,i(1)}$ and $\Psi_{i(2)}$ into a single estimating equation and applying standard theory, the covariance matrix of the joint distribution of $\widehat{\Theta}_*(\Lambda)$, $\widehat{\sigma}_{u,*}^2$ and $\widehat{\mu}$ is

$$n^{-1} \begin{Bmatrix} \mathcal{A}_{11}(\cdot) & \mathcal{A}_{12}(\cdot) \\ 0 & \mathcal{A}_{22}(\cdot) \end{Bmatrix}^{-1} \begin{Bmatrix} \mathcal{C}_{11}(\cdot) & \mathcal{C}_{12}(\cdot) \\ \mathcal{C}_{12}^t(\cdot) & \mathcal{C}_{22}(\cdot) \end{Bmatrix} \times \begin{Bmatrix} \mathcal{A}_{11}(\cdot) & \mathcal{A}_{12}(\cdot) \\ 0 & \mathcal{A}_{22}(\cdot) \end{Bmatrix}^{-t}, \tag{4.19}$$

where

$$\begin{Bmatrix} \mathcal{C}_{11}(\cdot) & \mathcal{C}_{12}(\cdot) \\ \mathcal{C}_{12}^t(\cdot) & \mathcal{C}_{22}(\cdot) \end{Bmatrix} = \mathcal{C}_*(\cdot) = \operatorname{cov}\left[\begin{matrix} \Psi_{B,1(1)}\left\{\sigma_{u,*}^2, \Lambda, \Theta_*(\Lambda)\right\} \\ \Psi_{1(2)}\left(\sigma_{u,*}^2, \mu\right) \end{matrix} \right];$$

$$\mathcal{A}_{12}\{\sigma_{u,*}^2, \Lambda, \Theta_*(\Lambda)\} = n^{-1} \sum_{i=1}^{n} E\left[\frac{\partial}{\partial(\sigma_{u,*}^2, \mu)} \Psi_{B,i(1)}\left\{\sigma_{u,*}^2, \Lambda, \Theta_*(\Lambda)\right\} \right];$$

and

$$\begin{aligned} \mathcal{A}_{22}\left(\sigma_{u,*}^2, \mu\right) &= n^{-1} \sum_{i=1}^{n} E\left\{ \frac{\partial}{\partial(\sigma_{u,*}^2, \mu)} \Psi_{i(2)}\left(\sigma_{u,*}^2, \mu\right) \right\} \\ &= -n^{-1} \sum_{i=1}^{n} E \begin{Bmatrix} 1 & 2(\mathbf{D}_i - \mu) \\ 0 & 1 \end{Bmatrix} = - \begin{Bmatrix} 1 & 0 \\ 0 & 1 \end{Bmatrix}. \end{aligned}$$

Estimating these quantities via the sandwich method is straight forward. For $\mathcal{A}_{12}(\cdot)$ remove the expectation symbol and replace $\left\{\sigma_{u,*}^2, \Theta_*(\Lambda), \mu\right\}$ by the estimates $\left\{\widehat{\sigma}_{u,*}^2, \widehat{\Theta}_*(\Lambda), \widehat{\mu}\right\}$. The covariance matrix $\mathcal{C}_*(\cdot)$ can be estimated by the sample covariance matrix of

the vectors

$$\begin{bmatrix} \Psi_{B,i(1)}\left\{\widehat{\sigma}^2_{u,*}, \Lambda, \widehat{\Theta}_*(\Lambda)\right\} \\ \Psi_{i(2)}\left(\widehat{\sigma}^2_{u,*}, \widehat{\mu}\right) \end{bmatrix}.$$

These estimates are substituted into (4.19) thereby obtaining an estimate of the joint covariance matrix of $\widehat{\Theta}_*(\Lambda)$, $\widehat{\sigma}^2_{u,*}$ and $\widehat{\mu}$. The submatrix corresponding to the components of $\widehat{\Theta}_*(\Lambda)$ is now employed in (4.18) in place of $\widehat{\Sigma}$.

CHAPTER 5

INSTRUMENTAL VARIABLES

5.1 Overview

In previous chapters we assumed that it was possible to estimate the measurement error variance, say with replicate measurements or validation data. However, it is not always possible to obtain replicates or validation data and thus direct estimation of the measurement error variance is sometimes impossible. In the absence of information about the measurement error variance, estimation of the regression model parameters is still possible provided the data contain an *instrumental variable* $\mathbf{T}$, in addition to the unbiased measurement $\mathbf{W} = \mathbf{X} + \mathbf{U}$.

There are three basic requirements that an instrumental variable (IV) must satisfy: (i) it must be correlated with $\mathbf{X}$; (ii) it must be independent of $\mathbf{W} - \mathbf{X}$; (iii) it must be a surrogate, i.e., independent of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$.

One possible source of an instrumental variable is a second measurement of $\mathbf{X}$ obtained by an independent method. This second measurement need not be unbiased for $\mathbf{X}$. Thus the assumption that a variable is an instrument is weaker than the assumption that it is a replicate measurement.

In the example of Chapter 4 it was explicitly assumed that transformed blood pressure measurements from successive exam periods were replicate measurements, even though a test of the replicate measurements assumption was found to be statistically (although not practically) significant. The same data can also be analyzed under the weaker assumption that the Exam #2 blood pressure measurements are instrumental variables. We do this in section 5.3 to illustrate the instrumental variable methods.

In this chapter we restrict attention to the important and common case in which there is a generalized linear model relating $\mathbf{Y}$

to $(\mathbf{Z}, \mathbf{X})$, i.e., the mean and variance functions depend on a linear function of the covariates and predictors, and there is a linear regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{T}, \mathbf{W})$.

Instrumental variable estimation in linear models is covered in depth Fuller (1987). There are a number of approaches to instrumental variable estimation in nonlinear models. In this chapter we describe an approach that is closely related to the regression calibration method of Chapter 3. It summarizes the work of Carroll & Stefanski (1994) and Stefanski & Buzas (1995). The methods in this chapter can also be viewed as extensions of the results for probit regression with multivariate normal predictors and covariates by Buzas & Stefanski (1996a). Other work on instrumental variable estimation not described in this chapter includes the general nonlinear model methods of Amemiya (1985, 1990a,b), and the methods of Buzas & Stefanski (1996b) for generalized linear models in canonical form.

We introduce the estimators in section 5.2, and apply them to the Framingham data in section 5.3. In sections 5.4 and 5.5 we derive the estimators and obtain their asymptotic distributions.

5.2 Approximate Instrumental Variable Estimation

In this chapter it is necessary to indicate numerous regression parameters and we adopt the notation employed by Stefanski and Buzas (1995). For example, $\beta_{Y|\underline{1}ZX}$ is the coefficient of $\mathbf{1}$, i.e., the intercept, in the generalized linear regression of $\mathbf{Y}$ on $\mathbf{1}$, $\mathbf{Z}$ and $\mathbf{X}$; $\beta^t_{Y|1\underline{Z}X}$ is the coefficient of $\mathbf{Z}$ in the regression of $\mathbf{Y}$ on $\mathbf{1}$, $\mathbf{Z}$ and $\mathbf{X}$. This notation allows representation of subsets of coefficient vectors, e.g., $\beta^t_{Y|1\underline{Z}X} = (\beta_{Y|\underline{1}ZX},\ \beta^t_{Y|1\underline{Z}X})$ and $\beta^t_{X|1\underline{ZT}} = (\beta_{X|\underline{1}ZT},\ \beta^t_{X|1\underline{Z}T},\ \beta^t_{X|1Z\underline{T}})$.

Our analysis is based upon generalized linear models, or more generally on mean/variance models. Examples of these models are linear, logistic and Poisson regression. As described more fully in sections A.4 and A.5, such models depend on a linear combination of the predictors plus possibly a parameter θ that describes the variability in the response. The sections listed above give details for model fitting when there is no measurement error. It might be useful upon first reading to simply think of this chapter as dealing with a class of important models, the details of fitting of which are

standard in many computer programs.

These models can be written in general form as

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = f(\beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}\mathbf{X}), \tag{5.1}$$

$$\text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = \sigma^2 g^2(\beta_{Y|\underline{1}ZX} + \beta^t_{Y|1\underline{Z}X}\mathbf{Z} + \beta^t_{Y|1Z\underline{X}}\mathbf{X}, \theta). \tag{5.2}$$

Those unfamiliar with generalized linear models might wish to focus on linear regression, where $f(v) = v$ and $g \equiv 1$. In logistic regression, f is the logistic distribution function and g^2 is the Bernoulli variance $f(1-f)$. The only notational change with other parts of the book is that the parameters β_0, β_z and β_x have been replaced by $\beta_{Y|\underline{1}ZX}$, $\beta_{Y|1\underline{Z}X}$ and $\beta_{Y|1Z\underline{X}}$, respectively.

The approximate models and estimation algorithms are best described in terms of the composite vectors

$$\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{Z}^t, \mathbf{X}^t)^t, \quad \tilde{\mathbf{W}} = (\mathbf{1}, \mathbf{Z}^t, \mathbf{W}^t)^t \text{ and } \tilde{\mathbf{T}} = (\mathbf{1}, \mathbf{Z}^t, \mathbf{T}^t)^t.$$

If we define $\beta_{Y|\underline{\tilde{X}}} = (\beta_{Y|\underline{1}ZX}, \beta^t_{Y|1\underline{Z}X}, \beta^t_{Y|1Z\underline{X}})^t$, the basic model (5.1)–(5.2) becomes

$$E(\mathbf{Y}|\tilde{\mathbf{X}}) = f(\beta^t_{Y|\underline{\tilde{X}}}\tilde{\mathbf{X}}), \tag{5.3}$$

$$\text{var}(\mathbf{Y}|\tilde{\mathbf{X}}) = \sigma^2 g^2(\beta^t_{Y|\underline{\tilde{X}}}\tilde{\mathbf{X}}, \theta). \tag{5.4}$$

The goal is to estimate $\beta_{Y|\underline{\tilde{X}}}$, θ and σ^2.

The assumptions that are necessary for our methods are stated more precisely in section 5.4, but we note here that in addition to the conditions stated in section 5.1, we will also assume that the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{T}, \mathbf{W})$ is approximately linear. This restricts the applicability of our methods somewhat, but is sufficiently general to encompass many potential applications.

5.2.1 First Regression Calibration Instrumental Variable Algorithm

In section 5.4.1 it is shown that approximately

$$E(\mathbf{Y}|\tilde{\mathbf{T}}) = f\{\beta^t_{Y|\underline{\tilde{X}}}E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}})\} = f(\beta^t_{Y|\underline{\tilde{X}}}\beta^t_{\tilde{X}|\underline{\tilde{T}}}\tilde{\mathbf{T}}),$$

and under the crucial assumption that $\beta_{\tilde{W}|\underline{\tilde{T}}} = \beta_{\tilde{X}|\underline{\tilde{T}}}$ it follows that (approximately)

$$\beta_{Y|\underline{\tilde{T}}} = \beta_{\tilde{W}|\underline{\tilde{T}}}\beta_{Y|\underline{\tilde{X}}}. \tag{5.5}$$

That is, the coefficient of $\tilde{\mathbf{T}}$ in the generalized linear regression of $\mathbf{Y}$ on $\tilde{\mathbf{T}}$ is the product of $\beta^t_{Y|\underline{\tilde{X}}}$ and $\beta^t_{\tilde{W}|\underline{\tilde{T}}}$. Starting with this basic approximation there are two ways to derive estimates of $\beta_{Y|\underline{\tilde{X}}}$.

The first and simplest starts with a multivariate regression of $\tilde{\mathbf{W}}$ on $\tilde{\mathbf{T}}$ to obtain $\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}$. Then the generalized linear regression of $\mathbf{Y}$ on the predicted values $\widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}}\tilde{\mathbf{T}}$ results in an estimator of $\beta_{Y|\underline{\tilde{X}}}$ which we denote $\widehat{\beta}^{IV1,RC}_{Y|\underline{\tilde{X}}}$.

This estimator is easily computed as it requires only linear regressions of the components of $\tilde{\mathbf{W}}$ on $\tilde{\mathbf{T}}$, and then quasilikelihood and variance function estimation of $\mathbf{Y}$ on the "predictors" $\widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}}\tilde{\mathbf{T}}$.

The second means of exploiting the basic regression calibration approximation works directly from the identity (5.5). For a fixed nonsingular matrix M_1, let $\widehat{\beta}^{-(M_1)}_{\tilde{W}|\underline{\tilde{T}}} = (\widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}} M_1 \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})^{-1} \widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}} M_1$. The second estimator is

$$\widehat{\beta}^{IV1,(M_1)}_{Y|\underline{\tilde{X}}} = \widehat{\beta}^{-(M_1)}_{\tilde{W}|\underline{\tilde{T}}} \widehat{\beta}_{Y|\underline{\tilde{T}}}, \tag{5.6}$$

where $\widehat{\beta}_{Y|\underline{\tilde{T}}}$ is the estimated regression coefficient when the generalized model is fit to the $(\mathbf{Y}, \tilde{\mathbf{T}})$ data. Note that (5.6) makes evident the requirement that $\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}$ be of full rank.

When $\mathbf{T}$ and $\mathbf{W}$ are the same dimension, this estimator does not depend on M_1 and is identical to the first estimator, but not otherwise. When there are more instruments than variables measured with error the choice of M_1 matters. In section 5.5.1 we derive an estimate $\widehat{M}_1$ that minimizes the asymptotic variance of $\widehat{\beta}^{IV1,(M_1)}_{Y|\underline{\tilde{X}}}$.

5.2.2 Second Regression Calibration Instrumental Variable Algorithm

The second algorithm exploits the fact that both $\mathbf{W}$ and $\mathbf{T}$ are surrogates. The derivation of the estimator is involved (section 5.4.2), but the estimator is not difficult to compute.

Let $\dim(\mathbf{Z})$ be the number of components of $\mathbf{Z}$. Define

$$\begin{aligned} \beta_{Y|\underline{\tilde{T}}\tilde{W}} &= \beta_{Y|\underline{1ZT}W}, \\ \beta_{Y|\tilde{T}\underline{\tilde{W}}} &= (0_{1\times d},\ \beta^t_{Y|1ZT\underline{W}})^t, \end{aligned}$$

where $d = 1 + \dim(\mathbf{Z})$. Then, for a given matrix M_2, the second

instrumental variables estimator is

$$\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}} = \widehat{\beta}^{-(M_2)}_{\tilde{W}|\underline{\tilde{T}}}(\widehat{\beta}_{Y|\underline{\tilde{T}}\tilde{W}} + \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}\widehat{\beta}_{Y|\tilde{T}\underline{\tilde{W}}}). \tag{5.7}$$

When $\mathbf{T}$ and $\mathbf{W}$ are the same dimension, $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}}$ does not depend on M_2. In section 5.5.1 we derive an estimate, $\widehat{M_2}$, that minimizes the asymptotic variance of $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}}$ for the case $\dim(\mathbf{T}) > \dim(\mathbf{W})$.

5.3 An Example

We now illustrate the methods presented in this chapter. We employ the same data from the Framingham heart study used in the example of section 4.5, wherein two systolic blood pressure measurements from each of two exams were employed. It was assumed that the two transformed variates

$$\mathbf{W}_1 = \log\{(\mathrm{SBP}_{3,1} + \mathrm{SBP}_{3,2})/2 - 50\}$$

and

$$\mathbf{W}_2 = \log\{(\mathrm{SBP}_{2,1} + \mathrm{SBP}_{2,2})/2 - 50\},$$

where $\mathrm{SBP}_{i,j}$ is the jth measurement of SBP from the ith exam, $j = 1, 2$, $i = 2, 3$, were replicate measurements of the long-term average transformed SBP.

Table 5.1 displays estimates of the same logistic regression model fit in section 4.5.2 with the difference that $\mathbf{W}_2$ was employed as an instrumental variable, not as a replicate measurement, i.e., in the notation of this section, $\mathbf{W} = \mathbf{W}_1$ and $\mathbf{T} = \mathbf{W}_2$, and no subsampling was done.

Because $\mathbf{T}$ has the same dimension as $\mathbf{W}$, the estimate $\beta^{IV1,(M_1)}_{Y|\underline{\tilde{X}}}$ does not depend on M_1 and is equivalent to $\beta^{IV1,RC}_{Y|\underline{\tilde{X}}}$. This common estimate is listed under IV1 in Table 5.1. Also $\beta^{IV2,(M_2)}_{Y|\underline{\tilde{X}}}$ does not depend on M_2 and is listed under IV2 in the table.

Table 5.2 displays estimates of the same logistic regression model with the difference that the instrumental variable $\mathbf{T}$ was taken to be the two-dimensional variate

$$\mathbf{T} = \{\log(\mathrm{SBP}_{2,1}),\ \log(\mathrm{SBP}_{2,2})\}.$$

The purpose of this analysis is to illustrate the differences between

	Age	Smoke	Chol	LSBP
Naive	.056	.573	.0078	1.524
Std. Err.	.010	.243	.0019	.364
IV1	.054	.577	.0076	2.002
Std. Err.	.011	.244	.0020	.517
IV2	.054	.579	.0077	1.935
Std. Err.	.011	.244	.0020	.513

Table 5.1. *Estimates and standard errors from the Framingham data instrumental variable logistic regression analysis. This analysis used the one-dimensional instrumental variable* $LSBP = log\{(SBP_{2,1} + SBP_{2,2})/2 - 50\}$. *"Smoke" is smoking status and "Chol" is cholesterol level. Standard errors calculated using the sandwich method.*

the estimators when $\dim(\mathbf{T}) > \dim(\mathbf{X})$, and to emphasize that $\mathbf{T}$ need only be correlated with $\mathbf{X}$, and not a second measurement, for the methods to be applicable.

5.4 Derivation of the Estimators

In this section, we derive the estimators presented in section 5.2. We start with the following assumptions:

$$E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}, \mathbf{W}) = \beta^t_{X|\underline{1}ZTW} + \beta^t_{X|1\underline{Z}TW}\mathbf{Z} + \beta^t_{X|1Z\underline{T}W}\mathbf{T} + \beta^t_{X|1ZT\underline{W}}\mathbf{W}; \quad (5.8)$$

$$E(\mathbf{X} - \mathbf{W} \mid \mathbf{Z}, \mathbf{X}, \mathbf{T}) = 0; \quad (5.9)$$

$$E(\mathbf{Y} \mid \mathbf{Z},\ \mathbf{T},\ \mathbf{W}) = E\{E(\mathbf{Y} \mid \mathbf{Z},\ \mathbf{X}) \mid \mathbf{Z},\ \mathbf{T},\ \mathbf{W}\}. \quad (5.10)$$

We have discussed each of these previously. Note that (5.8) and (5.9) imply that $E(\mathbf{X} \mid \mathbf{Z}, \mathbf{T}) = E(\mathbf{W} \mid \mathbf{Z}, \mathbf{T})$ and also that $\beta_{W|\underline{1}ZT} = \beta_{X|\underline{1}ZT}$, $\beta_{W|1\underline{Z}T} = \beta_{X|1\underline{Z}T}$ and $\beta_{W|1Z\underline{T}} = \beta_{X|1Z\underline{T}}$.

When validation data are available, i.e., complete data on $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{T}$ for some units in either the primary sample or an external sample, it is possible to check some or all of (5.8)–(5.10) depending

	Age	Smoke	Chol	LSBP
Naive	.056	.573	.0078	1.524
Std. Err.	.010	.243	.0019	.364
IV1,RC	.054	.577	.0076	1.877
Std. Err.	.011	.244	.0020	.481
IV1,(M_1)	.054	.577	.0076	1.884
Std Err.	.011	.244	.0020	.483
IV2,(M_2)	.054	.579	.0077	1.860
Std. Err.	.011	.244	.0020	.484

Table 5.2. *Estimates and standard errors from the Framingham data instrumental variable logistic regression analysis. This analysis used the two-dimensional instrumental variable* $\{log(SBP_{2,1}),\ log(SBP_{2,2})\}$. *"Smoke" is smoking status and "Chol" is cholesterol level. Standard errors calculated using the sandwich method.*

on the nature and extent of the validation data. Furthermore, with validation data it is sometimes possible to determine transformations so that (5.8)–(5.10) hold approximately.

5.4.1 First Regression Calibration Instrumental Variable Algorithm

The first algorithms are simple to describe once (5.5) is justified, which we do now. Making use of the fact that $\mathbf{T}$ is a surrogate, (5.10) and the standard regression calibration approximation results in the approximate model

$$E(\mathbf{Y}|\tilde{\mathbf{T}}) = f\{\beta^t_{Y|\underline{\tilde{X}}} E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}})\} = f(\beta^t_{Y|\underline{\tilde{X}}} \beta^t_{\tilde{X}|\underline{\tilde{T}}} \tilde{\mathbf{T}}), \quad (5.11)$$

$$\begin{aligned} \text{var}(\mathbf{Y}|\tilde{\mathbf{T}}) &= \sigma^2 g^2\{\beta^t_{Y|\underline{\tilde{X}}} E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}}), \theta\} \quad (5.12) \\ &= \sigma^2 g^2(\beta^t_{Y|\underline{\tilde{X}}} \beta^t_{\tilde{X}|\underline{\tilde{T}}} \tilde{\mathbf{T}}, \theta). \end{aligned}$$

It follows from (5.11)–(5.12) that the coefficient of $\tilde{\mathbf{T}}$ in the generalized linear regression of $\mathbf{Y}$ on $\tilde{\mathbf{T}}$ is $\beta^t_{Y|\underline{\tilde{T}}} = \beta^t_{Y|\underline{\tilde{X}}}\beta^t_{\tilde{X}|\underline{\tilde{T}}}$. By (5.9) $\beta_{\tilde{W}|\underline{\tilde{T}}} = \beta_{\tilde{X}|\underline{\tilde{T}}}$, and (5.5) follows.

5.4.2 Second Regression Calibration Instrumental Variable Algorithm

The derivation of the second algorithm is somewhat involved, but the estimator is relatively easy to compute. Remember that the strategy is to exploit the fact that both $\mathbf{W}$ and $\mathbf{T}$ are surrogates.

Making use of the fact that both $\mathbf{T}$ and $\mathbf{W}$ are surrogates, application of the standard regression calibration approximation produces the approximate model

$$E(\mathbf{Y}|\tilde{\mathbf{T}}, \tilde{\mathbf{W}}) = f\{\beta^t_{Y|\underline{\tilde{X}}} E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}}, \tilde{\mathbf{W}})\}, \tag{5.13}$$

$$\text{var}(\mathbf{Y}|\tilde{\mathbf{T}}, \tilde{\mathbf{W}}) = \sigma^2 g^2\{\beta^t_{Y|\underline{\tilde{X}}} E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}}, \tilde{\mathbf{W}}), \theta\}. \tag{5.14}$$

Under the linear regression assumption (5.8), there exist coefficient matrices $\beta^t_{\tilde{X}|\underline{\tilde{T}}\tilde{W}}$ and $\beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}}$ such that

$$E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}}, \tilde{\mathbf{W}}) = \beta^t_{\tilde{X}|\underline{\tilde{T}}\tilde{W}}\tilde{\mathbf{T}} + \beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}}\tilde{\mathbf{W}}. \tag{5.15}$$

Taking conditional expectations of both sides of (5.15) with respect to $\tilde{\mathbf{T}}$ and using the fact that $E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}}) = \beta^t_{\tilde{X}|\underline{\tilde{T}}}\tilde{\mathbf{T}}$ results in the identity

$$\beta^t_{\tilde{X}|\underline{\tilde{T}}}\tilde{\mathbf{T}} = \beta^t_{\tilde{X}|\underline{\tilde{T}}\tilde{W}}\tilde{\mathbf{T}} + \beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}}\beta^t_{\tilde{W}|\underline{\tilde{T}}}\tilde{\mathbf{T}}.$$

Equating coefficients of $\tilde{\mathbf{T}}$ and using the fact that $\beta_{\tilde{W}|\underline{\tilde{T}}} = \beta_{\tilde{X}|\underline{\tilde{T}}}$ we find that

$$\beta^t_{\tilde{X}|\underline{\tilde{T}}} = \beta^t_{\tilde{X}|\underline{\tilde{T}}\tilde{W}} + \beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}}\beta^t_{\tilde{X}|\underline{\tilde{T}}}. \tag{5.16}$$

Solving (5.16) for $\beta^t_{\tilde{X}|\underline{\tilde{T}}\tilde{W}}$ and then substitution into (5.15) shows that

$$E(\tilde{\mathbf{X}} \mid \tilde{\mathbf{T}}, \tilde{\mathbf{W}}) = (I - \beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}})\beta^t_{\tilde{X}|\underline{\tilde{T}}}\tilde{\mathbf{T}} + \beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}}\tilde{\mathbf{W}}. \tag{5.17}$$

By convention $\beta^t_{Y|\underline{\tilde{T}\tilde{W}}}$ is the regression coefficient of $(\tilde{\mathbf{T}}^t, \tilde{\mathbf{W}}^t)^t$ in the generalized linear regression of $\mathbf{Y}$ on $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{W}}$. The indicated model is over-parameterized and thus the components of $\beta^t_{Y|\underline{\tilde{T}\tilde{W}}}$ are not uniquely determined. Although other specifications are possible

we define the components of $\beta_{Y|\tilde{T}\tilde{W}}$ uniquely as

$$\begin{aligned}\beta_{Y|\underline{\tilde{T}}\tilde{W}} &= \beta_{Y|\underline{1ZT}W},\\ \beta_{Y|\tilde{T}\underline{\tilde{W}}} &= (0_{1\times d},\ \beta^t_{Y|1ZT\underline{W}})^t,\end{aligned}$$

where $d = 1 + \dim(\mathbf{Z})$. Let H_1 and H_2 be the matrices that define $\beta_{Y|\underline{\tilde{T}}\tilde{W}}$ and $\beta_{Y|\tilde{T}\underline{\tilde{W}}}$ in terms of $\beta_{Y|\underline{1ZTW}}$, so that $\beta_{Y|\underline{\tilde{T}}\tilde{W}} = H_1\beta_{Y|\underline{1ZTW}}$ and $\beta_{Y|\tilde{T}\underline{\tilde{W}}} = H_2\beta_{Y|\underline{1ZTW}}$. Also note that because $\tilde{\mathbf{T}} = (1, \mathbf{Z}^t, \mathbf{T}^t)^t$, our notation allows us to write $\beta^t_{Y|\underline{1ZTW}} = \beta^t_{Y|\underline{\tilde{T}W}}$.

Substitution of (5.17) into (5.13) and equating coefficients of $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{W}}$ results in the two equations,

$$\beta^t_{Y|\underline{\tilde{T}}\tilde{W}} = \beta^t_{Y|\underline{\tilde{X}}}(I - \beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}})\beta^t_{\tilde{X}|\underline{\tilde{T}}}, \tag{5.18}$$

$$\beta^t_{Y|\tilde{T}\underline{\tilde{W}}} = \beta^t_{Y|\underline{\tilde{X}}}\beta^t_{\tilde{X}|\tilde{T}\underline{\tilde{W}}}. \tag{5.19}$$

Post-multiplying (5.19) by $\beta^t_{\tilde{X}|\underline{\tilde{T}}}$ and adding the resulting equation to (5.18) results in the single equation,

$$\beta_{Y|\underline{\tilde{T}}\tilde{W}} + \beta_{\tilde{X}|\underline{\tilde{T}}}\beta_{Y|\tilde{T}\underline{\tilde{W}}} = \beta_{\tilde{X}|\underline{\tilde{T}}}\beta_{Y|\underline{\tilde{X}}},$$

which upon using the definitions of H_1 and H_2 and the identity $\beta_{\tilde{X}|\underline{\tilde{T}}} = \beta_{\tilde{W}|\underline{\tilde{T}}}$, is shown to be equivalent to

$$H_1\beta_{Y|\underline{\tilde{T}W}} + \beta_{\tilde{W}|\underline{\tilde{T}}}H_2\beta_{Y|\underline{\tilde{T}W}} = \beta_{\tilde{W}|\underline{\tilde{T}}}\beta_{Y|\underline{\tilde{X}}}.$$

Let $\widehat{\beta}_{Y|\underline{\tilde{T}W}}$ be the estimated regression parameter from the generalized linear regression of $\mathbf{Y}$ on $(\mathbf{1}, \mathbf{Z}, \mathbf{T}, \mathbf{W})$, and let $\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}$ be as before. Under the identifiability assumption that for a given matrix M_2, $(\widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}}M_2\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})$ is asymptotically nonsingular, it follows that the estimator (5.7), namely

$$\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}} = \widehat{\beta}^{-(M_2)}_{\tilde{W}|\underline{\tilde{T}}}(H_1\widehat{\beta}_{Y|\underline{\tilde{T}W}} + \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}H_2\widehat{\beta}_{Y|\underline{\tilde{T}W}}),$$

is approximately consistent for $\beta_{Y|\underline{\tilde{X}}}$.

When $\mathbf{T}$ and $\mathbf{W}$ are the same dimension, $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}}$ does not depend on M_2. In section 5.5.1 we derive an estimate $\widehat{M}_2$ that minimizes the asymptotic variance of $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}}$ for the case $\dim(\mathbf{T}) > \dim(\mathbf{W})$.

5.5 Asymptotic Distribution Approximations

We first derive the asymptotic distributions assuming that M_1 and M_2 are fixed and that M-estimation is used in the generalized linear and linear regression modeling steps. We then show how to estimate M_1 and M_2 for efficient asymptotic inference.

Let ψ denote the score function for the generalized linear model under consideration (5.3)–(5.4). This score function has as many as three components, the first corresponding to the unknown regression parameter, the second and third to the parameters in the variance function. All of the components are functions of the unknown parameters, the response variable and the vector of covariate/predictor variables. For example, with logistic regression there are no variance parameters and $\psi(y, x, \beta) = \{y - H(\beta^t x)\}\, x$ where $H(t) = 1/\{1 + \exp(-t)\}$.

Let

$$\psi_{1i} = \psi\left\{\mathbf{Y}_i, \tilde{\mathbf{T}}_i, \beta_{Y|\underline{\tilde{T}}}, \sigma_1^2, \theta_1\right\}$$

denote the ith score function employed in fitting the approximate model (5.11)–(5.12) to $(\mathbf{Y}_i, \tilde{\mathbf{T}}_i)_1^n$.

Let

$$\psi_{2i} = \psi\left\{\mathbf{Y}_i, (\tilde{\mathbf{T}}_i^t, \mathbf{W}_i^t)^t, \beta_{Y|\underline{\tilde{T}W}}, \sigma_2^2, \theta_2\right\}$$

denote the ith score function employed in fitting the approximate model (5.13)–(5.14) to $\left\{\mathbf{Y}_i, (\tilde{\mathbf{T}}_i^t, \mathbf{W}_i^t)^t\right\}_1^n$. Note that each fit of the generalized linear model produces estimates of the variance parameters as well as the regression coefficients. These are denoted with subscripts as above, e.g., σ_1^2, θ_1, etc.

Let ψ_{3i} denote the ith score function used to estimate $\text{vec}(\beta_{\tilde{W}|\underline{\tilde{T}}})$, e.g., for least squares estimation

$$\psi_{3i} = \left(\tilde{\mathbf{W}}_i - \beta_{\tilde{W}|\underline{\tilde{T}}}^t \tilde{\mathbf{T}}_i\right) \otimes \tilde{\mathbf{T}}_i,$$

and let

$$\psi_{4i} = \psi\left\{\mathbf{Y}_i, (\beta_{\tilde{W}|\underline{\tilde{T}}}^t \tilde{\mathbf{T}}_i), \beta_{Y|\underline{\tilde{X}}}^{IV1,RC}, \sigma_3^2, \theta_3\right\}.$$

Finally, define ψ_{5i} and ψ_{6i} as

$$\psi_{5i} = \left(\beta_{\tilde{W}|\underline{\tilde{T}}}^t M_1 \beta_{\tilde{W}|\underline{\tilde{T}}}\right) \beta_{Y|\underline{\tilde{X}}}^{IV1,(M_1)} - \beta_{\tilde{W}|\underline{\tilde{T}}}^t M_1 \beta_{Y|\underline{\tilde{T}}},$$

and

$$\psi_{6i} = \left(\beta_{\tilde{W}|\underline{\tilde{T}}}^t M_2 \beta_{\tilde{W}|\underline{\tilde{T}}}\right) \beta_{Y|\underline{\tilde{X}}}^{IV2,(M_2)}$$

$$-\beta^t_{\tilde{W}|\underline{\tilde{T}}} M_2 (H_1 \beta_{Y|\underline{\tilde{T}W}} + \beta_{\tilde{W}|\underline{\tilde{T}}} H_2 \beta_{Y|\underline{\tilde{T}W}}).$$

Note that neither ψ_{5i} nor ψ_{6i} depends on i.

Define the composite parameter

$$\Theta = \left\{ \beta^t_{Y|\underline{\tilde{T}}}, \sigma_1^2, \theta_1^t, \beta^t_{Y|\underline{\tilde{T}W}}, \sigma_2^2, \theta_2^t, \mathrm{vec}^t \left(\beta^t_{\tilde{W}|\underline{\tilde{T}}} \right), \right. \tag{5.20}$$

$$\left. \left(\beta^{IV1,RC}_{Y|\underline{\tilde{X}}} \right)^t, \sigma_3^2, \theta_3^t, \left(\beta^{IV1,(M_1)}_{Y|\underline{\tilde{X}}} \right)^t, \left(\beta^{IV2,(M_2)}_{Y|\underline{\tilde{X}}} \right)^t \right\}^t,$$

and the ith composite score function

$$\psi_i(\Theta) = \left(\psi^t_{1i}, \psi^t_{2i}, \psi^t_{3i}, \psi^t_{4i}, \psi^t_{5i}, \psi^t_{6i} \right)^t. \tag{5.21}$$

It follows that $\widehat{\Theta}$ solves

$$\sum_{i=1}^{n} \psi_i(\widehat{\Theta}) = 0_{\dim(\Theta)\times 1},$$

showing that $\widehat{\Theta}$ is an M-estimator. Thus under fairly general conditions $\widehat{\Theta}$ is approximately normally distributed in large samples and the theory of Chapter A applies.

An estimate of the asymptotic covariance matrix of $\widehat{\Theta}$ is given by the sandwich formula $\widehat{A}_n^{-1} \widehat{B}_n (\widehat{A}_n^{-1})^t$ where $\widehat{A}_n = \sum_{i=1}^n \psi_{i\Theta}(\widehat{\Theta})$ with $\psi_{i\Theta}(\Theta) = \partial \psi_i(\Theta)/\partial \Theta^t$, and $\widehat{B}_n = \sum_{i=1}^n \psi_i(\widehat{\Theta}) \psi_i^t(\widehat{\Theta})$. Note that because we are fitting approximate (or misspecified) models, information-based standard errors, i.e., standard errors obtained by replacing $\widehat{A}_n$ and $\widehat{B}_n$ by model-based estimates exploiting the information identity, are generally not appropriate.

Let $\widehat{\Omega} = \widehat{A}_n^{-1} \widehat{B}_n (\widehat{A}_n^{-1})^t$ and let $\widehat{\Omega}_{i,j}$, $i,j = 1, \ldots, 12$ denote the (i,j)th submatrix of $\widehat{\Omega}$ corresponding to the natural partitioning induced by the components of Θ in (5.20). It follows that $\widehat{\Omega}_{8,8}$, $\widehat{\Omega}_{11,11}$ and $\widehat{\Omega}_{12,12}$ are estimates of the variance matrices of the asymptotic distributions of $\widehat{\beta}^{IV1,RC}_{Y|\underline{\tilde{X}}}$, $\widehat{\beta}^{IV1,(M_1)}_{Y|\underline{\tilde{X}}}$ and $\widehat{\beta}^{IV2,(M_2)}_{Y|\underline{\tilde{X}}}$, respectively.

5.5.1 Two-Stage Estimation

When $\mathbf{T}$ and $\mathbf{W}$ have the same dimension the estimators (5.6) and (5.7) do not depend on M_1 and M_2. However, when there are

more instruments than predictors measured with error it is possible to identify and consistently estimate matrices M_1 and M_2 that minimize the asymptotic variance matrix of the corresponding estimators. We give the results first and then sketch their derivations.

For an asymptotically efficient estimator (5.6) replace M_1 with

$$\widehat{M}_{1,\text{opt}} = \left(\widehat{\Omega}_{1,1} - \widehat{\Omega}_{1,7}\widehat{C}^t - \widehat{C}\widehat{\Omega}_{7,1} + \widehat{C}\widehat{\Omega}_{7,7}\widehat{C}^t\right)^{-1}$$

where $\widehat{C} = I_{d_{\tilde{T}}} \otimes \left(\widehat{\beta}^{IV1,(I)}_{Y|\underline{X}}\right)^t$, $I_{d_{\tilde{T}}}$ is the identity matrix of dimension $d_{\tilde{T}} = \dim(\tilde{T})$, and $\widehat{\beta}^{IV1,(I)}_{Y|\underline{X}}$ is the estimator obtained by setting M_1 equal to $I_{d_{\tilde{T}}}$.

For an asymptotically efficient estimator (5.7) replace M_2 with

$$\widehat{M}_{2,\text{opt}} = \Big\{(H_1 + \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})\widehat{\Omega}_{4,4}(H_1 + \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})^t + (H_1 + \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})\widehat{\Omega}_{4,7}\widehat{D}^t + \widehat{D}\widehat{\Omega}_{7,4}(H_1 + \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})^t + \widehat{D}\widehat{\Omega}_{7,7}\widehat{D}^t\Big\}^{-1}$$

where $\widehat{D} = I_{d_{\tilde{T}}} \otimes \left(H_2\widehat{\beta}_{Y|\underline{\tilde{T}W}}\right)^t - I_{d_{\tilde{T}}} \otimes \left(\widehat{\beta}^{IV2,(I)}_{Y|\underline{\tilde{X}}}\right)^t$, and $\widehat{\beta}^{IV2,(I)}_{Y|\underline{\tilde{X}}}$ is the estimator obtained by setting M_2 equal to $I_{d_{\tilde{T}}}$.

We now describe the main steps in the demonstrations of the asymptotic efficiency of $\widehat{M}_{1,\text{opt}}$ and $\widehat{M}_{2,\text{opt}}$.

The argument for $\widehat{M}_{1,\text{opt}}$ and the estimator (5.6) is simpler and is given first. We start with a heuristic derivation of the efficient estimator.

Consider the basic identity in (5.5), $\beta_{Y|\underline{\tilde{T}}} = \beta_{\tilde{W}|\underline{\tilde{T}}}\beta_{Y|\underline{\tilde{X}}}$. Replacing $\beta_{Y|\underline{\tilde{T}}}$ with $\widehat{\beta}_{Y|\underline{\tilde{T}}} - (\widehat{\beta}_{Y|\underline{\tilde{T}}} - \beta_{Y|\underline{\tilde{T}}})$ and $\beta_{\tilde{W}|\underline{\tilde{T}}}$ with $\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}} - (\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}} - \beta_{\tilde{W}|\underline{\tilde{T}}})$ and rearranging terms shows that this equation is equivalent to

$$\widehat{\beta}_{Y|\underline{\tilde{T}}} = \widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}\beta_{Y|\underline{\tilde{X}}} + (\widehat{\beta}_{Y|\underline{\tilde{T}}} - \beta_{Y|\underline{\tilde{T}}}) - (\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}} - \beta_{\tilde{W}|\underline{\tilde{T}}})\beta_{Y|\underline{\tilde{X}}}.$$

This equation has the structure of a linear model with response vector $\widehat{\beta}_{Y|\underline{\tilde{T}}}$, design matrix $\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}$, regression parameter $\beta_{Y|\underline{\tilde{X}}}$, and equation error, $(\widehat{\beta}_{Y|\underline{\tilde{T}}} - \beta_{Y|\underline{\tilde{T}}}) - (\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}} - \beta_{\tilde{W}|\underline{\tilde{T}}})\beta_{Y|\underline{\tilde{X}}}$. Let Σ denote the covariance matrix of this equation error. The best linear

unbiased estimator of $\beta_{Y|\underline{\tilde{X}}}$ in this pseudolinear model is

$$(\widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}}\Sigma^{-1}\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}})^{-1}\widehat{\beta}^t_{\tilde{W}|\underline{\tilde{T}}}\Sigma^{-1}\widehat{\beta}_{Y|\underline{\tilde{T}}},$$

which is exactly (5.6) with $M_1 = \Sigma^{-1}$. Note that the estimator $\widehat{M}_{1,\text{opt}}$ is a consistent estimator of Σ^{-1}.

Showing that the heuristic derivation is correct and that there is no penalty for using an estimated covariance matrix is somewhat more involved, but entails nothing more than linearization via Taylor series approximations and Δ-method arguments.

Let $\widehat{M}_1$ be a consistent estimator of the matrix M_1. Expanding the estimating equation for $\widehat{\beta}^{IV1,(\widehat{M}_1)}_{Y|\underline{\tilde{X}}}$ around the true parameters results in the approximation

$$\sqrt{n}\left\{\widehat{\beta}^{IV1,(\widehat{M}_1)}_{Y|\underline{\tilde{X}}} - \beta_{Y|\underline{\tilde{X}}}\right\} \approx \beta^{-(M_1)}_{\tilde{W}|\underline{\tilde{T}}}\left(\epsilon_2 - C\epsilon_3\right),$$

where

$$\begin{aligned}
\epsilon_2 &= \sqrt{n}\left(\widehat{\beta}_{Y|\underline{\tilde{T}}} - \beta_{Y|\underline{\tilde{T}}}\right), \\
\epsilon_3 &= \sqrt{n}\left\{\text{vec}\left(\widehat{\beta}_{\tilde{W}|\underline{\tilde{T}}}\right) - \text{vec}\left(\beta_{\tilde{W}|\underline{\tilde{T}}}\right)\right\}, \\
C &= I_{d_{\tilde{T}}} \otimes \beta^t_{Y|\underline{\tilde{X}}}.
\end{aligned}$$

This Taylor series approximation is noteworthy for the fact that it is the same for M_1 known as it is for M_1 estimated. Consequently, there is no penalty asymptotically for estimating M_1.

Thus, with AVAR denoting asymptotic variance, we have that

$$\text{AVAR}\left\{\sqrt{n}\,\widehat{\beta}^{IV1,(\widehat{M}_1)}_{Y|\underline{\tilde{X}}}\right\} = \beta^{-(M_1)}_{\tilde{W}|\underline{\tilde{T}}}\left\{\text{AVAR}\left(\epsilon_2 - C\epsilon_3\right)\right\}\left(\beta^{-(M_1)}_{\tilde{W}|\underline{\tilde{T}}}\right)^t.$$

That this asymptotic variance is minimized when

$$M_1 = \left\{\text{AVAR}\left(\epsilon_2 - C\epsilon_3\right)\right\}^{-1},$$

is a consequence of the optimality of weighted-least squares linear regression.

Let $\widehat{M}_2$ be a consistent estimator of the matrix M_2. Expanding the estimating equation for $\widehat{\beta}^{IV2,(\widehat{M}_2)}_{Y|\underline{\tilde{X}}}$ around the true parameters results in the approximation

$$\sqrt{n}\left\{\widehat{\beta}^{IV2,(\widehat{M}_2)}_{Y|\underline{\tilde{X}}} - \beta_{Y|\underline{\tilde{X}}}\right\} \approx \beta^{-(M_2)}_{\tilde{W}|\underline{\tilde{T}}}\left\{\left(H_1 + \beta_{\tilde{W}|\underline{\tilde{T}}}H_2\right)\epsilon_1 + D\epsilon_3\right\},$$

where

$$\begin{aligned} \epsilon_1 &= \sqrt{n}\left(\widehat{\beta}_{Y|\underline{\tilde{T}W}} - \beta_{Y|\underline{\tilde{T}W}}\right), \\ D &= I_{d_{\tilde{T}}} \otimes \left(H_2 \beta_{Y|\underline{\tilde{T}W}}\right)^t - I_{d_{\tilde{T}}} \otimes \beta^t_{Y|\underline{\tilde{X}}}. \end{aligned}$$

As before, estimating M_2 does not affect the asymptotic distribution of the parameter estimates.

From the approximation we find that

$$\begin{aligned} &\text{AVAR}\left(\sqrt{n}\,\widehat{\beta}^{IV2,(\widehat{M}_2)}_{Y|\underline{\tilde{X}}}\right) = \\ &\qquad \beta^{-(M_2)}_{\tilde{W}|\underline{\tilde{T}}} \left[\text{AVAR}\left\{\left(H_1 + \beta_{\tilde{W}|\underline{\tilde{T}}} H_2\right)\epsilon_1 + D\epsilon_3\right\}\right]\left(\beta^{-(M_2)}_{\tilde{W}|\underline{\tilde{T}}}\right)^t, \end{aligned}$$

which is minimized when

$$M_2 = \left[\text{AVAR}\left\{\left(H_1 + \beta_{\tilde{W}|\underline{\tilde{T}}} H_2\right)\epsilon_1 + D\epsilon_3\right\}\right]^{-1}.$$

5.5.2 Computing Estimates and Standard Errors

The two-stage estimates are only slightly more difficult to compute than the first-stage estimates. Here we describe an algorithm that results in both estimates.

Note that for fixed matrices M_1 and M_2 all of the components of $\widehat{\Theta}$ in (5.20) are calculated either directly as linear regression or generalized linear regression estimates, or are simple transformations of such estimates. So for fixed M_1 and M_2 obtaining $\widehat{\Theta}$ is straightforward.

Asymptotic variance estimation is most easily accomplished by first programming the two functions

$$\begin{aligned} G_1(\Theta) &= \sum_{i=1}^{n} \psi_i(\Theta), \\ G_2(\Theta) &= \sum_{i=1}^{n} \psi_i(\Theta)\psi_i(\Theta)^t, \end{aligned}$$

where $\psi_i(\Theta)$ is the ith composite score function from (5.21). Although we do not actually solve $G_1(\Theta) = 0$ to find $\widehat{\Theta}$, it should be true that $G_1(\widehat{\Theta}) = 0$. This provides a check on the programming of G_1.

Numerical differentiation of G_1 at $\Theta = \widehat{\Theta}$ results in the matrix $\widehat{A}_n$. Alternatively, analytical derivatives of $\psi_i(\Theta)$ can be used, but these are complicated and tedious to program. Evaluation of G_2 at $\Theta = \widehat{\Theta}$ is the matrix $\widehat{B}_n$. The covariance matrix of $\widehat{\Theta}$ is then found as $\widehat{\Omega} = \widehat{A}_n^{-1}\widehat{B}_n(\widehat{A}_n^{-1})^t$.

The algorithm described above is first used with M_1 and M_2 set to the identity matrix of dimension $\dim(\tilde{\mathbf{T}})$ resulting in the first-stage estimates and estimated asymptotic covariance matrix. Next M_1 and M_2 are set to $\widehat{M}_{1,\text{opt}}$ and $\widehat{M}_{2,\text{opt}}$, respectively, as described in section 5.5.1. A second implementation of the algorithm results in the second-stage estimates and estimated asymptotic covariance matrix.

CHAPTER 6

FUNCTIONAL METHODS

6.1 Overview

Regression calibration (Chapter 3) and SIMEX (Chapter 4) are easily applied general methods. Although the resulting estimators are consistent in important special cases such as linear regression and loglinear mean models, they are only approximately consistent in general.

For certain generalized linear models and measurement error distributions there are easily applied methods that are fully and not just approximately consistent, without making assumptions about the distribution of $\mathbf{X}$. This is an example of functional modeling. We describe such methods in this chapter.

We focus on the case of additive normally distributed measurement error, so that $\mathbf{W} = \mathbf{X} + \mathbf{U}$ with $\mathbf{U}$ distributed as a normal random vector with mean zero and covariance matrix Σ_{uu}. Although the problem has this parametric error assumption, it also has a nonparametric component, in that no assumptions are made about the true predictors, $(\mathbf{X}_i)_1^n$, which can be random, as in a structural model, or fixed unknown constants, as in a functional model.

Suppose for the sake of discussion that the measurement error covariance matrix Σ_{uu} is known. In the functional model, the unobservable $\mathbf{X}$'s are fixed constants, and hence the unknown parameters include the $\mathbf{X}$'s. With additive normally distributed measurement error, one strategy is to maximize the joint density of the observed data with respect to all of the unknown parameters including $(\mathbf{X}_i)_1^n$. While this works for linear regression (Gleser, 1981), it fails for more complex models such as logistic regression. Indeed, the logistic regression functional maximum likelihood estimator is both inconsistent and difficult to compute (Stefanski & Carroll,

1985). An alternative approach is to change to the structural model and apply likelihood techniques (Chapter 7), although this is not always appropriate.

In this chapter, we consider two functional methods referred to here as the conditional-score and corrected-score methods.

We start with linear, logistic and gamma-loglinear modeling as important examples for which the techniques of this chapter apply. In section 6.2 we show how one can compute estimators in these examples without making assumptions about the $\mathbf{X}$'s. The methods are illustrated with a logistic regression example in section 6.3. The remainder of the chapter shows how to obtain estimators for other problems, and goes into detail about the methods of derivation.

Outside of the previously mentioned examples and Poisson loglinear models, the estimators have an appearance of being more involved than the regression calibration or SIMEX estimators. This is really only a matter of algebra (they are algebraically more complex, to be sure), but the conditional-score and corrected-score methods have a general theoretical basis. The conditional methods exploit special structures in important models such as linear, logistic, Poisson loglinear and gamma-inverse, and then use a traditional statistical device, conditioning on sufficient statistics, to obtain estimators. The corrected-score method effectively estimates the estimator one would use if there were no measurement error.

The conditional-score method is presented in section 6.4, and the corrected-score method is introduced in section 6.5, for a class of problems which lead to easy computation. Inference for the parameters when Σ_{uu} is estimated are described in section 6.6. In section 6.7 we describe a broad class of infinite series corrected-score estimators.

6.2 Linear, Logistic and Gamma-Loglinear Models

Three models of wide interest are the linear, logistic and loglinear (especially the gamma-loglinear) models, the latter for responses following the gamma distribution. The methods described below for these three models share the property that the resulting estimators do not depend on the distribution of $\mathbf{X}$.

First consider the multiple linear regression model with mean $\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}$, and write the unknown regression parameter as $\Theta = (\beta_0, \beta_x, \beta_z)$. When the measurement error is additive with

nondifferential measurement error variance Σ_{uu}, the usual method-of-moments regression estimator (2.7) can be derived as the solution to the equation

$$\sum_{i=1}^{n} \psi_*(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}) = 0, \tag{6.1}$$

where

$$\begin{aligned}&\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu}) \\ &\quad = \left(\mathbf{Y} - \beta_0 - \beta_x^t \mathbf{X} - \beta_z^t \mathbf{Z}\right) \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{W} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \Sigma_{uu}\beta_x \end{pmatrix}\end{aligned}$$

is the *corrected score* for linear regression. In linear regression, the corrected-score method results in the usual method-of-moments estimator. If Σ_{uu} is unknown, one substitutes an estimate of it (section 3.4) into (6.1) and solves for the regression parameters.

The key point to note here is that in solving (6.1), we need know nothing about the $\mathbf{X}$'s. This feature is common to all the methods in this chapter.

Equation (6.1) is an example of an estimating equation approach for estimating a set of unknown parameters. The reader can consult section A.3 for an overview of estimating equations, although this is unnecessary for the purpose of using the methods. Asymptotic standard errors for the estimators can be derived using either the bootstrap or the sandwich formula as described in Appendix A.

For linear regression, solving (6.1) instead of simply writing down the method-of-moments estimate may appear purely algebraic, but the approach can be derived (and is derived in subsequent sections) from general principles. The general principles allow us to handle other problems, for example gamma regression with loglinear mean.

When $\mathbf{Y}$ has a gamma distribution with loglinear mean $\exp(\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z})$, it has variance which is ϕ times the square of the mean. For this important example, the corrected-score estimator is obtained from the corrected score

$$\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu}) = \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{W} \end{pmatrix} \tag{6.2}$$

$$-\exp\{\Delta(\mathbf{Z},\mathbf{W},\Theta,\Sigma_{uu})\}\begin{pmatrix}\mathbf{Y}\\ \mathbf{ZY}\\ \mathbf{Y}(\mathbf{W}+.5\Sigma_{uu}\beta_x)\end{pmatrix},$$

where $\Delta(\mathbf{Z},\mathbf{W},\Theta,\Sigma_{uu}) = -\beta_0 - \beta_x^t\mathbf{W} - \beta_z^t\mathbf{Z} - .5\beta_x^t\Sigma_{uu}\beta_x$.

Logistic regression is best handled using the conditional-score method, although under certain conditions it is also amenable to the corrected-score method, see section 6.7. For example, consider the usual linear-logistic model, where $\mathbf{Y}$ is binary and has success probability following the logistic model $H(\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z})$. The conditional score is

$$\begin{aligned}&\psi_*(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta,\Sigma_{uu}) \qquad (6.3)\\ &= \left[\mathbf{Y} - H\left\{\beta_0 - \beta_x^t\Delta(\cdot) - .5\beta_x^t\Sigma_{uu}\beta_x - \beta_z^t\mathbf{Z}\right\}\right]\begin{bmatrix}1\\ \mathbf{Z}\\ \Delta(\cdot)\end{bmatrix},\end{aligned}$$

where

$$\Delta(\cdot) = \Delta(\mathbf{Y},\mathbf{W},\beta_x,\Sigma_{uu}) = \mathbf{W} + \mathbf{Y}\Sigma_{uu}\beta_x.$$

Equation (6.3) is substituted into (6.1), and the resulting equation is solved numerically (section 6.4.3).

6.3 Framingham Data

We fit a logistic regression model to the Framingham data used in the examples of sections 3.3 and 4.5. All of the replicate measurements were used, and thus our variance estimate is based on 1614 degrees of freedom and we proceeded under the assumption that the sampling variability in the estimate was negligible, i.e., the known measurement error case.

Previously, we have fit logistic regression models to these data, and here we use the conditional estimator based on (6.3). The estimates and standard errors are in Table 6.1, compare with Table 4.1. In this example, there is little difference among the regression calibration, SIMEX and conditional estimators.

6.4 Unbiased Score Functions via Conditioning

In this section, we describe the conditional estimators of Stefanski & Carroll (1987), which apply to an important class of generalized

	Age	Smoke	Chol	LSBP
Naive	.055	.59	.0078	1.70
Std. Err.	.010	.24	.0019	.39
Conditional	.053	.58	.0078	1.93
Std. Err.	.011	.25	.0020	.46

Table 6.1. *Estimates and sandwich standard errors from the Framingham data logistic regression analysis. Here "Smoke" is smoking status, "Chol" is cholesterol and "LSBP" is log(SBP−50).*

linear models. The logistic regression conditional score presented in section 6.2 is the most noteworthy example. Here we extend the methods to Poisson-loglinear, gamma-inverse and other models.

Canonical generalized linear models (McCullagh & Nelder, 1989) for $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ have density or mass function

$$f(y|z,x,\Theta) = \exp\left\{\frac{y\eta - \mathcal{D}(\eta)}{\phi} + c(y,\phi)\right\}, \tag{6.4}$$

where $\eta = \beta_0 + \beta_z^t z + \beta_x^t x$ is called the *natural parameter*, and $\Theta = (\beta_0, \beta_z^t, \beta_x^t, \phi)$ is the unknown parameter to be estimated. The mean and variance of $\mathbf{Y}$ are $\mathcal{D}^{(1)}(\eta)$ and $\phi\mathcal{D}^{(2)}(\eta)$, respectively (the first and second derivatives). This class of models includes:

- linear regression: mean $= \eta$, variance $= \phi$, $\mathcal{D}(\eta) = \eta^2/2$, $c(y,\phi) = -y^2/(2\phi) - \log(\sqrt{2\pi\phi}\,)$;
- logistic regression: mean $= H(\eta)$, variance $= H'(\eta)$, $\phi \equiv 1$, $\mathcal{D}(\eta) = -\log\{1 - H(\eta)\}$, $c(y,\phi) = 0$;
- Poisson loglinear regression: mean = variance $= \exp(\eta)$, $\phi \equiv 1$, $\mathcal{D}(\eta) = \exp(\eta)$, $c(y,\phi) = -\log(y!)$;
- Gamma inverse regression: mean $= -1/\eta$, variance $= -\phi/\eta$, $\mathcal{D}(\eta) = -\log(-\eta)$, $c(y,\phi) = \phi^{-1}\log(y/\phi) - \log\{y\Gamma(1/\phi)\}$.

If $\mathbf{X}$ were observed, then Θ is estimated by solving

$$\sum_{i=1}^{n} \left\{ \mathbf{Y}_i - \mathcal{D}^{(1)}(\eta_i) \right\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \mathbf{X}_i \end{pmatrix} = 0 \qquad (6.5)$$

$$\sum_{i=1}^{n} \left[\left(\frac{n-p}{n} \right) \phi - \frac{\left\{ \mathbf{Y}_i - \mathcal{D}^{(1)}(\eta_i) \right\}^2}{\mathcal{D}^{(2)}(\eta_i)} \right] = 0, \qquad (6.6)$$

where $\eta_i = \beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \mathbf{X}_i$.

For certain models equations (6.5)–(6.6) result in maximum likelihood estimators (when $n-p$ is replaced by n), although in general they result in quasilikelihood estimators, see Appendix A.

Assume now that the measurement error is additive and normally distributed, with error covariance matrix Σ_{uu}. If $\mathbf{X}$ is regarded as an unknown parameter and all other parameters are assumed known, then it transpires that $\Delta = \mathbf{W} + \mathbf{Y}\Sigma_{uu}\beta_x/\phi$ is a sufficient statistic for $\mathbf{X}$ (Stefanski & Carroll, 1987). Furthermore, the conditional distribution of $\mathbf{Y}$ given $(\mathbf{Z}, \Delta) = (z, \delta)$ is also a canonical generalized linear model in exactly the same form as (6.4), except that we make the following substitutions when $(\mathbf{Y}, \mathbf{Z}, \Delta) = (y, z, \delta)$, namely to replace x by δ, and set

$$\begin{aligned} &\eta_* = \beta_0 + \beta_z^t z + \beta_x^t \delta; \\ &c_*(y, \phi, \beta_x^t \Sigma_{uu} \beta_x) = c(y, \phi) - (1/2)(y/\phi)^2 \beta_x^t \Sigma_{uu} \beta_x; \\ &\mathcal{D}_*(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) \\ &\quad = \phi \log \left[\int \exp \left\{ y\eta_*/\phi + c_*(y, \phi, \beta_x^t \Sigma_{uu} \beta_x) \right\} \, d\mu(y) \right], \end{aligned}$$

where as before the notation means that the last term is a sum if $\mathbf{Y}$ is discrete and an integral otherwise. This means that the conditional density or mass function is

$$\begin{aligned} &f(y|z, \delta, \Theta, \Sigma_{uu}) = \\ &\exp \left\{ \frac{y\eta_* - \mathcal{D}_*(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x)}{\phi} + c_*(y, \phi, \beta_x^t \Sigma_{uu} \beta_x) \right\}, \end{aligned} \qquad (6.7)$$

where $\eta_* = \beta_0 + \beta_z^t z + \beta_x^t \delta$.

The obvious correspondence between (6.4) and (6.7) suggests that one simply substitutes $\mathcal{D}_*(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x)$ for $\mathcal{D}(\eta)$ into (6.5)–(6.6), and then solves the resulting equations replacing η_i by $\eta_{*,i} = \beta_0 + \beta_x^t \Delta_i + \beta_z^t \mathbf{Z}_i$, noting that Δ_i depends on β_x and ϕ.

Here are the details of how to implement this procedure. The conditional mean and variance of $\mathbf{Y}$ given $(\mathbf{Z}, \Delta)$ are determined by the derivatives of $\mathcal{D}_*$ with respect to η_*, i.e.,

$$\begin{aligned} m(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) &= \frac{\partial}{\partial \eta_*} \mathcal{B}_*; \\ \phi v(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) &= \phi \frac{\partial^2}{\partial \eta_*^2} \mathcal{B}_*. \end{aligned} \tag{6.8}$$

The estimates of $\Theta = (\beta_0, \beta_x, \beta_z, \phi)$ are obtained by solving

$$\begin{aligned} \sum_{i=1}^{n} \{\mathbf{Y}_i - m(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix} &= 0 \\ \sum_{i=1}^{n} \left[\left(\frac{n-p}{n} \right) \phi - \frac{\{\mathbf{Y}_i - m(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)\}^2}{v(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)} \right] &= 0 \end{aligned} \tag{6.9}$$

where $\eta_{*,i} = \beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \Delta_i$, with $\Delta_i = \mathbf{W}_i + \mathbf{Y}_i \Sigma_{uu} \beta_x / \phi$. Solving (6.9) is generally more difficult than solving (6.5)–(6.6).

Stefanski & Carroll (1987) discuss a number of ways of deriving unbiased estimating equations from (6.7) and (6.8). The approach described here is the simplest to implement.

6.4.1 Linear and Logistic Regression

The functions $m(\cdot)$ and $v(\cdot)$ are easily obtained in linear and logistic regression. In linear regression, when (6.4) is a normal density,

$$m(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \frac{\partial}{\partial \eta_*} \mathcal{B}_* = \frac{\eta_*}{1 + \phi^{-1} \beta_x^t \Sigma_{uu} \beta_x};$$

$$v(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \phi \frac{\partial}{\partial \eta_*} m(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \frac{\phi}{1 + \phi^{-1} \beta_x^t \Sigma_{uu} \beta_x}.$$

For logistic regression (where $\phi \equiv 1$), $\mathcal{D}_*$ is a function of only η_* and $\beta_x^t \Sigma_{uu} \beta_x$. The conditional mean and variance functions are

$$m(\eta_*, \beta_x^t \Sigma_{uu} \beta_x) = \frac{\partial}{\partial \eta_*} \mathcal{B}_* = H(\eta_* - \beta_x^t \Sigma_{uu} \beta_x / 2);$$

$$v(\eta_*, \beta_x^t \Sigma_{uu} \beta_x) = \frac{\partial}{\partial \eta_*} m(\eta_*, \beta_x^t \Sigma_{uu} \beta_x) = H^{(1)}(\eta_* - \beta_x^t \Sigma_{uu} \beta_x / 2),$$

where $H^{(1)} = H(1 - H)$ is the logistic density function.

6.4.2 Other Canonical Models

Linear and logistic regression are the only common canonical models for which $\mathcal{D}_*^{(1)}$ and $\mathcal{D}_*^{(2)}$ have closed-form expressions. In general either numerical integration or summation is required to determine the moments (6.8). For example, for Poisson regression (for which $\phi \equiv 1$),

$$\mathcal{D}_*(\eta_*, \phi, \beta_x^t \Sigma_{uu} \beta_x) = \log \left\{ \sum_{y=0}^{\infty} (y!)^{-1} \exp(y\eta_* - y^2 \beta_x^t \Sigma_{uu} \beta_x / 2) \right\},$$

and the mean and variance $m(\cdot)$ and $v(\cdot)$ are the first and second derivatives of $\mathcal{D}_*$ with respect to η_*. In fact, $m = s_1$ and $v = s_2 - s_1^2$, where

$$s_j = E(\mathbf{Y}^j \mid \mathbf{Z} = z, \Delta = \delta) = \frac{\sum_{y=0}^{\infty} y^j (y!)^{-1} \exp\{y(\eta_*) - y^2 \beta_x^t \Sigma_{uu} \beta_x / 2\}}{\sum_{y=0}^{\infty} (y!)^{-1} \exp\{y(\eta_*) - y^2 \beta^t \Sigma_{uu} \beta_x / 2\}}. \quad (6.10)$$

Computation of both the mean and variance functions for the Poisson model entails summing infinite series. The series can be summed analytically when $\beta_x^t \Sigma_{uu} \beta_x = 0$; however, for $\beta_x^t \Sigma_{uu} \beta_x > 0$ numerical summation is required.

6.4.3 Computation

Define

$$\psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}) = \{\mathbf{Y}_i - m(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)\} \begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix}$$

and

$$\psi_2(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}) = \left(\frac{n-p}{n}\right) \phi - \frac{\{\mathbf{Y}_i - m(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)\}^2}{v(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)},$$

where $\eta_{*,i} = \beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \Delta_i$, with $\Delta_i = \mathbf{W}_i + \mathbf{Y}_i \Sigma_{uu} \beta_x / \phi$.

For linear regression and other models with variance parameter ϕ, define $\psi_C(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}) = (\psi_1^t, \ \psi_2)^t$. For logistic regression and other models with $\phi \equiv 1$, define $\psi_C(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}) =$

ψ_1. We call ψ_C a conditional score, and estimators derived from it, conditional-score estimators, are denoted $\widehat{\Theta}_C$.

With these definitions $\widehat{\Theta}_C$ is obtained by solving

$$\sum_{i=1}^{n} \psi_C(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}_C, \Sigma_{uu}) \quad = \quad 0 \tag{6.11}$$

Define

$$\begin{aligned} \widehat{A}_{n,1}(\Theta, \Sigma_{uu}) &= n^{-1} \sum_{i=1}^{n} \left[\begin{array}{c} -\begin{pmatrix} 1 \\ \mathbf{Z}_i \\ \Delta_i \end{pmatrix} \frac{\partial}{\partial \Theta^t} m(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x) \\ \frac{\partial}{\partial \Theta^t} \psi_2(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu}) \end{array} \right] \\ \widehat{A}_{n,2}(\Theta, \Sigma_{uu}) &= n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta^t} \psi_C(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu}) \end{aligned} \tag{6.12}$$

where $\eta_{*,i} = \beta_0 + \beta_z^t \mathbf{Z}_i + \beta_x^t \Delta_i$, with $\Delta_i = \mathbf{W}_i + \mathbf{Y}_i \Sigma_{uu} \beta_x / \phi$. Note that the derivatives in the definitions of $\widehat{A}_{n,j}(\Theta, \Sigma_{uu})$, $j = 1, 2$ are full derivatives with respect to Θ, and not with $\eta_{*,i}$ or Δ_i held fixed. Also note that in the case $\phi \equiv 1$, the second term of $\widehat{A}_{n,1}$ is omitted.

Estimates are calculated iteratively. Starting with an initial estimate $\widehat{\Theta}_C^{(0)}$, either the so-called naive estimate or the SIMEX or regression calibration estimates, successive estimates are obtained from the iteration

$$\widehat{\Theta}_C^{(k+1)} = \widehat{\Theta}_C^{(k)} - \widehat{A}_{n,j}^{-1} \left(\widehat{\Theta}_C^{(k)}, \Sigma_{uu} \right) \sum_{i=1}^{n} \psi_C \left(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}_C^{(k)}, \Sigma_{uu} \right).$$

Use of $\widehat{A}_{n,1}^{-1}$ in the iteration corresponds to the method of scoring, whereas use of $\widehat{A}_{n,2}^{-1}$ results in a standard Newton-Raphson iteration.

6.4.4 Inference

As usual, the bootstrap (section A.6) can be used for inference. However, conditional-score estimators admit analytical formulae for standard errors, as we now show. Define

$$b\{\mathbf{Z}, \Delta(\Theta, \Sigma_{uu}), \Theta, \Sigma_{uu}\} \quad = \quad \operatorname{cov}\{\psi_C(\cdot) \,|\mathbf{Z}, \Delta(\Theta, \Sigma_{uu})\};$$

$$\widehat{B}_{n,1}(\Theta, \Sigma_{uu}) = n^{-1} \sum_{i=1}^{n} b\{\mathbf{Z}_i, \Delta_i(\Theta, \Sigma_{uu}), \Theta, \Sigma_{uu}\}$$

$$\widehat{B}_{n,2}(\Theta, \Sigma_{uu}) = n^{-1} \sum_{i=1}^{n} \psi_C^t(\cdot)\psi_C(\cdot). \quad (6.13)$$

Note that $b(\cdot)$ can in theory be computed from the conditional distribution in (6.7). Except for linear and logistic regression this usually entails numerical summation or integration.

The asymptotic covariance matrix of $\widehat{\Theta}$ for the case that Σ_{uu} is known is consistently estimated by

$$n^{-1} \widehat{A}_{n,j}^{-1}\left(\widehat{\Theta}, \Sigma_{uu}\right) \widehat{B}_{n,j}\left(\widehat{\Theta}, \Sigma_{uu}\right) \widehat{A}_{n,j}^{-t}\left(\widehat{\Theta}, \Sigma_{uu}\right), \quad j = 1, 2.$$

When $j = 1$ the covariance matrix estimator is a (conditional) inverse information matrix-type estimator, whereas $j = 2$ is a standard sandwich estimator.

6.5 Exact Corrected Estimating Equations

The method of section (6.4) is limited in application to generalized linear models in canonical form. For example, the methods do not apply to gamma-loglinear regression with mean $\exp(\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z})$, which is not a canonical generalized linear model. We now describe a second approach that is applicable to a general class of generalized linear regression models including the gamma loglinear regression model. This corrected-score method yields the normal and gamma-loglinear examples in section 6.2.

The method of *corrected score functions* has been studied by Nakamura (1990) and Stefanski (1989). Suppose that in the absence of measurement error, one would estimate Θ by solving

$$0 = \sum_{i=1}^{n} \psi(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \Theta).$$

Typically ψ is a likelihood score from the model for the data without error. Now suppose that it is possible to find a function of the data, say $\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)$, having the property that

$$E\{\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta) \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}\} = \psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta), \quad (6.14)$$

for all $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{X}$ and Θ. Upon taking expectations in (6.14) it follows that $\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)$ is a Fisher-consistent score function. In

general, ψ_* depends on Σ_{uu} and we will indicate this by writing $\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu})$ when necessary to emphasize this dependence.

Corrected score functions satisfying (6.14) do not always exist and finding them when they do is not always easy. This problem is studied in detail in Stefanski (1989), where corrected functions are derived for some common models, and generally applicable approximate corrected score functions are given.

6.5.1 Likelihoods With Exponentials and Powers

One useful class of models that admit corrected functions contains those models with log-likelihoods of the form

$$\log\{f(y|z,x,\Theta)\} = \sum_{k=0}^{2}\{c_k(y,z,\Theta)(\beta_x^t x)^k\} + c_3(y,z,\Theta)\exp(\beta_x^t x), \quad (6.15)$$

see the examples given below. Then, using normal distribution moment generating function identities, the required function is

$$\psi_*(y,z,w,\Theta,\Sigma_{uu}) = \frac{\partial}{\partial\Theta^t}\left[\sum_{k=0}^{2}\{c_k(y,z,\Theta)(\beta_x^t w)^k\} - c_2(y,z,\Theta)\beta_x^t\Sigma_{uu}\beta_x + c_3(y,z,\Theta)\exp(\beta_x^t w - .5\beta_x^t\Sigma_{uu}\beta_x)\right].$$

Regression models in this class include:

- Normal linear with mean $= \eta$, variance $= \phi$, $c_0 = -(y-\beta_0-\beta_z^t z)^2/(2\phi) - \log(\sqrt{\phi})$, $c_1 = (y-\beta_0-\beta_z^t z)/\phi$, $c_2 = -(2\phi)^{-1}$, $c_3 = 0$;
- Poisson with mean $= \exp(\eta)$, variance $= \exp(\eta)$, $c_0 = y(\beta_0 + \beta_z^t z) - \log(y!)$, $c_1 = y$, $c_2 = 0$, $c_3 = -\exp(\beta_0 + \beta_z^t z)$;
- Gamma with mean $= \exp(\eta)$, variance $= \phi\exp(2\eta)$, $c_0 = -\phi^{-1}(\beta_0+\beta_z^t z)+(\phi^{-1}-1)\log(y)+\phi^{-1}\log(\phi^{-1})-\log\{\Gamma(\phi^{-1})\}$, $c_1 = \phi^{-1}$, $c_2 = 0$, $c_3 = -\phi^{-1}y\exp(-\beta_0 - \beta_z^t z)$.

6.5.2 Asymptotic Distribution Approximation

Let $\psi_*(Y, \mathbf{Z}, \mathbf{W}, \Theta)$ denote a corrected score and suppose that $\widehat{\Theta}$ is a solution to the corrected-score estimating equations

$$\sum_{i=1}^{n} \psi_*(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}) = 0.$$

Then $\widehat{\Theta}$ is asymptotically normally distributed with mean Θ and covariance matrix $n^{-1}A^{-1}B(A^{-1})^t$ where A and B are consistently estimated by

$$\widehat{A} = n^{-1} \sum_{i=1}^{n} \psi_{*\Theta}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}) \tag{6.16}$$

$$\widehat{B} = n^{-1} \sum_{i=1}^{n} \psi_*(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}) \psi_*^t(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \widehat{\Theta}), \tag{6.17}$$

with $\psi_{*\Theta}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta) = (\partial/\partial\Theta^t)\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)$.

6.6 Estimated Σ_{uu}

When Σ_{uu} is unknown, additional data are required to consistently estimate it and the asymptotic variance-covariance matrix of the estimators is altered. The bootstrap handles this issue directly, but analytical standard errors can also be obtained.

Let $\psi_{\text{CS}}(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu})$ denote either a conditional score or a corrected score and $\widehat{\Theta}_{\text{CS}}$ the corresponding estimator. Define $\gamma = \text{vech}(\Sigma_{uu})$, where "vech" is the vector-half of a symmetric matrix, i.e., its distinct elements.

If an independent estimate of the error covariance matrix is available the following method may be used. Let $\widehat{\gamma}$ be an estimate of γ which is assumed to be independent of $\widehat{\Theta}_{\text{CS}}$, with asymptotic covariance matrix $C_n(\Sigma_{uu})$. If we define

$$D_n(\Theta, \Sigma_{uu}) = \sum_{i=1}^{n} \frac{\partial}{\partial \gamma^t} \psi_{\text{CS}}\left(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}\right),$$

then a consistent estimate of the covariance matrix of $\widehat{\Theta}_{\text{CS}}$ is

$$n^{-1} A_n^{-1}\left(\widehat{\Theta}_{\text{CS}}, \widehat{\Sigma}_{uu}\right) \left\{ B_n\left(\widehat{\Theta}_{\text{CS}}, \widehat{\Sigma}_{uu}\right) + D_n(\widehat{\Theta}_{\text{CS}}, \widehat{\Sigma}_{uu}) C_n(\widehat{\Sigma}_{uu}) D_n^t(\widehat{\Theta}_{\text{CS}}, \widehat{\Sigma}_{uu}) \right\} A_n^{-t}\left(\widehat{\Theta}_{\text{CS}}, \widehat{\Sigma}_{uu}\right),$$

where $\widehat{A}_n$ and $\widehat{B}_n$ are from either (6.16)–(6.17), or (6.12)–(6.13), depending on the score function employed and, in the case of the conditional score, on the type of covariance matrix estimator.

Finally, a problem of considerable importance occurs when for each of the $i = 1, \cdots, n$ observations in the data set, there are k_i independent replicated $\mathbf{W}$'s: $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$, $j = 1, ..., k_i$. The most common situation is that much of the data is unreplicated ($k_i = 1$), but some of the data have a single replicate ($k_i = 2$). Constructing estimated standard errors for this problem has not been done previously, and the justification for our results is given in the appendix. The necessary changes are as follows. In computing the estimates, in the previous definitions, replace Σ_{uu} by Σ_{uu}/k_i and $\mathbf{W}_i$ by $\overline{\mathbf{W}}_{i\cdot}$, the sample mean of the replicates. The estimate of Σ_{uu} is the usual components of variance estimator,

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right) \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t}{\sum_{i=1}^{n} (k_i - 1)}.$$

While the components of variance estimator has a known asymptotic distribution (based on the Wishart distribution), it is easier in practice to use the sandwich estimator of its variance,

$$C_n(\widehat{\Sigma}_{uu}) = \frac{\sum_{i=1}^{n} d_i d_i^t}{\left\{\sum_{i=1}^{n} (k_i - 1)\right\}^2},$$

where

$$d_i = \text{vech}\left\{\left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right) \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^t\right\} - (k_i - 1)\text{vech}\left(\widehat{\Sigma}_{uu}\right).$$

6.7 Infinite Series Corrected Estimating Equations

Many models of interest do not have the form (6.15). For example, the canonical gamma regression model is not of the form (6.15), nor is the logistic regression model. Thus there is no easy method of deriving a corrected score for these models. In fact a corrected score function satisfying (6.14) for logistic regression does not exist in general (Stefanski, 1989). However, under certain restrictions it is possible to obtain a corrected score function for logistic regression and for many other models as well.

In this section we briefly describe extensions of the corrected-score method to certain generalized linear models whose mean and variance functions depend on $\mathbf{X}$ only through $\exp(\beta_x^t \mathbf{X})$. This sec-

tion summarizes the results in Buzas & Stefanski (1995), using examples to illustrate the basic idea.

6.7.1 Rare-Event Logistic Regression

If $\mathbf{Y} = 1$ is a rare event, by which is meant that $\sup \Pr(\mathbf{Y}_i = 1 \mid \mathbf{Z}_i, \mathbf{X}_i) < 1/2$ where the supremum is taken over all members of the population, then a corrected-score function exists. There is a corresponding method for frequent events, $\sup \Pr(\mathbf{Y}_i = 1 \mid \mathbf{Z}_i, \mathbf{X}_i) > 1/2$, obtained by considering $\mathbf{Y}^* = 1 - \mathbf{Y}$, so that we consider only the rare-event case.

Let $\eta_x = \beta_0 + \beta_z^t \mathbf{Z} + \beta_x^t \mathbf{X}$, and similarly define $\eta_w = \beta_0 + \beta_z^t \mathbf{Z} + \beta_x^t \mathbf{W}$. Note that $\Theta = (\beta_0, \beta_z^t, \beta_x^t)^t$. In terms of the logistic regression model, the rare-event assumption implies that $H(\eta_x) < 1/2$, which in turn means that the parameter space can be restricted to the set of parameters such that $\eta_{x,i} < 0$ for all members in the population. This restriction makes it possible to obtain the corrected-score function. We take ψ to be the logistic regression likelihood score,

$$\psi_{\mathrm{ml}}(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta) = \{\mathbf{Y} - H(\eta_x)\} \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{X} \end{pmatrix}.$$

Define the two functions

$$\tilde{H}_-(x) = \sum_{k=1}^{\infty} (-1)^{k+1} e^{(kx - k^2\sigma^2/2)}, \quad \text{and} \quad \tilde{H}_-^{(1)}(x) = \frac{d}{dx}\tilde{H}_-(x).$$

Buzas & Stefanski (1995) show that if $W \sim N(\mu, \sigma^2)$ and $\mu < 0$, then $E\left\{\tilde{H}_-(W)\right\} = H(\mu)$ and

$$E\left\{W\tilde{H}_-(W) - \sigma^2 \tilde{H}_-^{(1)}(W)\right\} = \mu H(\mu).$$

Now let $\tilde{H}_-$ and $\tilde{H}_-^{(1)}$ be defined as above with σ^2 replaced by $\beta_x^t \Sigma_{uu} \beta_x$. Under the assumed measurement error model $\eta_w \sim N(\eta_x, \beta_x^t \Sigma_{uu} \beta_x)$, and the rare-event assumption implies that $\eta_x < 0$. It follows that $E\left\{\tilde{H}_-(\eta_w)\right\} = H(\eta_x)$ and

$$E\left\{\eta_w \tilde{H}_-(\eta_w) - \beta_x^t \Sigma_{uu} \beta_x \tilde{H}_-^{(1)}(\eta_w)\right\} = \eta_x H(\eta_x).$$

Using these identities it can be shown that

$$E\left\{\psi_{\mathrm{ml},*}(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta) \mid \mathbf{Y},\mathbf{Z},\mathbf{X}\right\} = \psi_{\mathrm{ml}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\Theta),$$

where $\psi_{\mathrm{ml},*}(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta) =$

$$\mathbf{Y}\begin{pmatrix}1\\ \mathbf{Z}\\ \mathbf{W}\end{pmatrix} - \left\{\begin{pmatrix}1\\ \mathbf{Z}\\ \mathbf{W}\end{pmatrix}\tilde{H}_{-}(\eta_w) - \begin{pmatrix}0\\ 0\\ \Sigma_{uu}\beta_x\end{pmatrix}\tilde{H}_{-}^{(1)}(\eta_w)\right\}. \quad (6.18)$$

That is, $\psi_{\mathrm{ml},*}$ is a corrected score for ψ_{ml}. A corrected-score estimator $\widehat{\Theta}_{\mathrm{ml},*}$ is obtained by solving

$$\sum_{i=1}^{n}\psi_{\mathrm{ml},*}(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{W}_i,\widehat{\Theta}_{\mathrm{ml},*}) = 0. \quad (6.19)$$

When used to estimate the parameters of the logistic regression model for the Framingham data, the corrected-score method yields results very similar to the conditional-score method.

6.7.2 Extensions to Mean and Variance Function Models

Buzas & Stefanski (1995) describe extensions of the method in the previous subsection to mean and mean/variance function models of the type described by Carroll & Ruppert (1988) and McCullagh & Nelder (1989). The extensions are mathematically involved. Here we describe only a simple special case for a mean function model and a method-of-moments type score function.

We assume that $E(\mathbf{Y} \mid \mathbf{Z},\mathbf{X}) = G(\eta_x)$ where as before $\eta_x = \beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t\mathbf{X}$, $\Theta = (\beta_0,\beta_z^t,\beta_x^t)^t$, G is a function of the form $G(\mu) = g(e^{\mu})$, and g has the absolutely convergent series expansion

$$g(x) = \sum_{k=0}^{\infty} a_k x^k, \qquad 0 \le |x| < r_g.$$

In this case define $\tilde{G}(x) = \sum_{k=0}^{\infty} a_k e^{(kx-k^2\sigma^2/2)}$ and let $\tilde{G}^{(1)}$ denote its derivative. Then, provided $e^{\mu} < r_g$, expectation and summation can be interchanged, and with $W \sim N(\mu,\sigma^2)$ as before,

$$E\left\{\tilde{G}(W)\right\} = \sum_{k=0}^{\infty} a_k E\left\{e^{(kW-k^2\sigma^2/2)}\right\} = \sum_{k=0}^{\infty} a_k e^{k\mu} = G(\mu).$$

Define

$$\psi_{\mathrm{mm}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\Theta) = \{\mathbf{Y} - G(\eta_x)\}\begin{pmatrix}1\\ \mathbf{Z}\\ \mathbf{X}\end{pmatrix}.$$

The score ψ_{mm} is a type of method-of-moments score for Θ. Define $\psi_{\mathrm{mm},*}(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta) =$

$$\mathbf{Y}\begin{pmatrix}1\\ \mathbf{Z}\\ \mathbf{W}\end{pmatrix} - \left\{\begin{pmatrix}1\\ \mathbf{Z}\\ \mathbf{W}\end{pmatrix}\tilde{G}(\eta_w) - \begin{pmatrix}0\\ 0\\ \Sigma_{uu}\beta_x\end{pmatrix}\tilde{G}^{(1)}(\eta_w)\right\}. \quad (6.20)$$

where as before $\beta_x^t\Sigma_{uu}\beta_x$ replaces σ^2 in $\tilde{G}$ and $\tilde{G}^{(1)}$.

It can be shown that

$$E\{\psi_{\mathrm{mm},*}(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta) \mid \mathbf{Y},\mathbf{Z},\mathbf{X}\} = \psi_{\mathrm{mm}}(\mathbf{Y},\mathbf{Z},\mathbf{X},\Theta)$$

for $\exp(\eta_x) < r_g$. Thus $\psi_{\mathrm{mm},*}$ is a corrected version of ψ_{mm} and can be used to obtain consistent estimators of Θ.

Derivations of the identities used above and further extensions of the method, including an application to extreme-value binary regression can be found in Buzas & Stefanski (1995).

6.8 Comparison of Methods

Poisson regression is amenable to both conditional- and corrected-score methods. For Poisson regression the corrected estimating equations are more convenient because they are explicit, whereas the conditional estimator involves numerical summation, see (6.10). For Poisson regression the conditional-score estimator is more efficient than the corrected-score estimator in some practical cases (Stefanski, 1989).

The corrected-score estimators have, in theory, a distributional robustness property not enjoyed by the conditional estimators. Ideally, a measurement error analysis should provide consistent estimators of the same parameters that would be consistently estimated in the absence of measurement error. The corrected-score estimator accomplishes this when the mean model is misspecified, whereas the conditional-score estimator in theory does not. Of course, this theoretical advantage of the corrected-score method depends critically on the assumed normality of the measurement error.

The conditional-score method and certain extensions thereof have a theoretical advantage in terms of efficiency. For the canonical generalized linear models of section 6.4, Stefanski & Carroll (1987) show that any unbiased estimating equation for $(\beta_0, \beta_z^t, \beta_x^t)^t$ must be conditionally unbiased given $(\mathbf{Z}, \Delta)$, and from this they deduce that the asymptotically efficient estimating equations for structural models are based on score functions of the form

$$\sum_{i=1}^{n} \{\mathbf{Y}_i - m(\eta_{*,i}, \phi, \beta_x^t \Sigma_{uu} \beta_x)\} \begin{Bmatrix} 1 \\ \mathbf{Z}_i \\ E(\mathbf{X}_i \mid \mathbf{Z}_i, \Delta_i) \end{Bmatrix} = 0. \quad (6.21)$$

This result shows that, in general, none of the methods we have proposed previously are asymptotically efficient in structural models, except when $E(\mathbf{X}|\mathbf{Z}, \Delta)$ is linear in $(\mathbf{Z}, \Delta)$. This is the case in linear regression with $(\mathbf{Z}, \mathbf{X})$ marginally normally distributed, and in logistic regression when $(\mathbf{Z}, \mathbf{X})$ given $\mathbf{Y}$ is normally distributed, i.e., the linear discriminant model.

The problem of constructing fully efficient conditional-score estimators based on simultaneous estimation of $E(\mathbf{X}_i \mid \mathbf{Z}_i, \Delta_i)$ has been studied (Lindsay, 1985; Bickel & Ritov, 1987; van der Vaart, 1988), although the methods are generally too specialized or too difficult to implement in practice routinely.

Both methods have further extensions not mentioned previously. The conditional-score method is easily extended to the case that the model for $\mathbf{W}$ given $\mathbf{X}$ is a canonical generalized linear model with natural parameter $\mathbf{X}$.

Buzas & Stefanski (1995) describe a simple extension of the methods in sections 6.5.1, 6.7.1, and 6.7.2 to additive non-normal error models. Suppose that $\mathbf{W} = \mathbf{X} + \mathbf{U}$, and that $m_u(t) = E\{\exp(t^t \mathbf{U})\}$, exists for some t and is known. For normal errors the corrected score is a function of terms of the form $\exp(j\beta_x^t \mathbf{W} - j^2 \beta_x^t \Sigma_{uu} \beta_x / 2)$. When $\mathbf{U}$ is normally distributed

$$\exp(j\beta_x^t \mathbf{W} - j^2 \beta_x^t \Sigma_{uu} \beta_x / 2) = \frac{\exp(j\beta_x^t \mathbf{W})}{m_u(j\beta_x)},$$

showing that for general error distributions it is sufficient to replace all terms of the form $\exp(j\beta_x^t \mathbf{W} - j^2 \beta_x^t \Sigma_{uu} \beta_x / 2)$ by terms of the form $\exp(j\beta_x^t \mathbf{W})/m_u(j\beta_x)$.

Extensions to nonadditive models are also possible in some cases.

Nakamura (1990) shows how to construct a corrected estimating equation for linear regression with multiplicative lognormal errors. He also suggests different methods of estimating standard errors.

6.9 Appendix

6.9.1 Technical Complements to Conditional Score Theory

We first justify (6.7). The joint density of $\mathbf{Y}$ and $\mathbf{W}$ is the product of (6.4) and the normal density, and hence is proportional to

$$\begin{aligned}
&\sim \exp\left\{\frac{y\eta - \mathcal{D}(\eta)}{\phi} + c(y,\phi) - (1/2)(w-x)^t\Sigma_{uu}^{-1}(w-x)\right\} \\
&\sim \exp\left\{y(\beta_0 + \beta_z^t z)/\phi + c(y,\phi) - \right. \\
&\qquad \left. (1/2)w^t\Sigma_{uu}^{-1}w + x^t\Sigma_{uu}^{-1}(w + y\Sigma_{uu}\beta_x/\phi\right\},
\end{aligned}$$

where by $\sim$ we mean terms that do not depend on y or w. Now set $\delta = w + y\Sigma_{uu}\beta_x/\phi$ and make a change of variables (which has Jacobian 1). The joint density of $(\mathbf{Y}, \Delta)$ given $(\mathbf{Z}, \mathbf{X})$ is thus seen to be proportional to

$$\begin{aligned}
&\sim \exp\left\{y(\beta_0 + \beta_x^t\delta + \beta_z^t z)/\phi + \right. \\
&\qquad \left. c(y,\phi) - (1/2)(y/\phi)^2\beta_x^t\Sigma_{uu}\beta_x\right\} \\
&= \exp\left\{y\eta_*/\phi + c_*(y,\phi,\beta_x^t\Sigma_{uu}\beta_x)\right\}, \qquad (6.22)
\end{aligned}$$

The conditional density of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X}, \Delta)$ is (6.22) divided by its integral with respect to y, which is necessarily in the form (6.7) as claimed, with

$$\begin{aligned}
&\mathcal{D}_*(\eta_*, \phi, \beta_x^t\Sigma_{uu}\beta_x) = \\
&\quad \phi\log\left[\int \exp\left\{y\eta_*/\phi + c_*(y,\phi,\beta_x^t\Sigma_{uu}\beta_x)\right\}\, d\mu(y)\right], \quad (6.23)
\end{aligned}$$

where as before the notation means that (6.23) is a sum if $\mathbf{Y}$ is discrete and an integral otherwise.

6.9.2 Technical Complements to Distribution Theory for Estimated Σ_{uu}

Next we justify the estimated standard errors for $\widehat{\Theta}$ when there is partial replication. Recall that with normally distributed observations, the sample mean and the sample covariance matrix are independent. Hence, $\widehat{\Sigma}_{uu}$ and $\widehat{\gamma} = \text{vech}(\widehat{\Sigma}_{uu})$ are independent of all

the terms $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \overline{\mathbf{U}}_{i\cdot})$, and also independent of $(\mathbf{Y}_i, \mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot})$. By a Taylor series expansion,

$$A_n(\cdot)\left(\widehat{\Theta} - \Theta\right) \approx$$
$$\sum_{i=1}^{n} \{\psi_C(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu})\} + D_n(\Theta, \Sigma_{uu})\left(\widehat{\gamma} - \gamma\right).$$

Because the two terms in the last sum are independent, the total covariance is the sum of the two covariances, namely $B_n(\cdot) = D_n(\cdot)C_n(\cdot)D_n^t(\cdot)$ as claimed.

CHAPTER 7

LIKELIHOOD AND QUASILIKELIHOOD

7.1 Introduction

This chapter describes the use of likelihood methods in nonlinear measurement error models. There have been only a few examples in the literature based on likelihood, see Carroll, et al. (1984) and Schafer (1988, 1993) for probit regression, Whittemore & Gong (1991) in a Poisson model, Crouch & Spiegelman (1990) and Wang, Carroll & Liang(1995) in logistic regression, and Küchenhoff & Carroll (1995) in a change point problem. The relatively small literature belies the importance of the topic and the potential for further applications.

Except where noted, we assume nondifferential measurement error (section 1.6). For a review of maximum likelihood methods in general see Appendix A.

Fully specified likelihood problems, including problems where $\mathbf{X}$ is not observable or is observable for only a subset of the data are discussed in sections 7.3, 7.4, and 7.5. The use of likelihood ideas in quasilikelihood and variance function models (QVF) (section A.4) is covered in section 7.8.

In section 7.2, we point out the relationships and differences between nonlinear measurement error models and missing data problems.

There are number of important differences between the likelihood methods in this chapter and the methods described in previous chapters.

- The previous methods are based on additive or multiplicative measurement error models, possibly after a transformation. Typically, few if any distributional assumptions are required.

Likelihood methods require stronger distributional assumptions, but they can be applied to more general problems, including those with discrete covariates subject to misclassification.

- The likelihood for a fully specified parametric model can be used to obtain likelihood ratio confidence intervals. In methods not based on likelihoods, inference is based on bootstrapping or on normal approximations. In highly nonlinear problems, likelihood-based confidence intervals are generally more reliable than those derived from normal approximations.
- Likelihood methods are often computationally more demanding, whereas the previous methods require little more than the use of standard statistical packages.
- Robustness to modeling assumptions is a concern for both approaches, but generally more difficult to understand with likelihood methods.
- Traditional folklore suggests that in many statistical models, especially for the most common generalized linear models, the simpler methods described previously perform just as well in practice as likelihood methods. Somewhat amazingly, there is little documentation as to whether the folklore is realistic. The only evidence for this folklore that we know of is given for logistic regression by Stefanski & Carroll (1990b), who contrast the maximum likelihood estimate and the conditional scores estimate of Chapter 6. They find that the conditional score estimates are usually fairly efficient relative to the maximum likelihood estimate unless the measurement error is "large" or the logistic coefficient is "large," where the definition of large is somewhat vague. One should be aware though that their calculations indicate that there are situations where *properly parameterized* maximum likelihood estimates are considerably more efficient than estimates derived from functional modeling considerations (see also section 7.7).

We organize our discussion of likelihood methods based on the type and extent of data that are available. Although a simplification, in practice it is useful to think of three cases. In the first $\mathbf{X}$ is not observable, but there are sufficient data, either internal or external, to characterize the distribution of $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Z})$. In the second case $\mathbf{X}$ is unobservable and it is known that the Berkson model holds. The third case we consider is that where $\mathbf{X}$ is

observable for a subset of the data.

The benefit of this trichotomy is that it depends on the data that are available, as well as any subject-matter information at hand. Like any simplification, there are exceptions, but the reader can understand the main ideas simply by focusing on these cases.

To perform a likelihood analysis, one must specify a parametric model for every component the data. Likelihood analysis starts with a model for the distribution of the response given the true predictors. The likelihood (density or mass) function of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ will be called $f_{Y|Z,X}(y|z, x, \mathcal{B})$ here, and interest lies in estimating $\mathcal{B}$.

The form of the likelihood function can generally be specified by reference to any standard statistics text. For example, if $\mathbf{Y}$ is normally distributed with mean $\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}$ and variance σ^2, then $\mathcal{B} = (\beta_0, \beta_x, \beta_z, \sigma^2)$ and

$$f_{Y|Z,X}(y|z, x, \mathcal{B}) = \sigma^{-1} \phi \left\{ (y - \beta_0 + \beta_x^t x + \beta_z^t z)/\sigma \right\},$$

where $\phi(v) = (2\pi)^{-1/2} \exp(-.5v^2)$ is the standard normal density function. If $\mathbf{Y}$ follows a logistic regression model with mean $H(\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z})$, then $\mathcal{B} = (\beta_0, \beta_x, \beta_z)$ and

$$\begin{aligned} f_{Y|Z,X}(y|z, x, \mathcal{B}) &= H^y \left(\beta_0 + \beta_x^t x + \beta_z^t z\right) \\ &\quad \times \left\{1 - H\left(\beta_0 + \beta_x^t x + \beta_z^t z\right)\right\}^{1-y}. \end{aligned}$$

7.1.1 Identifiable Models

In some problems, the parameters are identifiable without any extra information other than measures of $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$, i.e., without validation or replications. Brown (1992) discusses this issue, considering both likelihood and quasilikelihood techniques.

One should not be overly impressed by all claims of identifiability. Many problems of practical importance actually are identifiable, but only barely so, and estimation without additional data is not practical. For instance, in linear regression it is known that the regression parameters can be identified without validation or replication as long as $\mathbf{X}$ is *not* normally distributed (Fuller, 1987, pp. 72–73). However, this means that the parameter estimates will be very unstable if $\mathbf{X}$ is at all close to being normally distributed. In binary regression with a normally distributed calibration, it is known that the probit model is not identified (Carroll, et al., 1984)

but that the logistic model is (Küchenhoff, 1990). The difference between these two models is so slight (Figure 3.5) that there is really no useful information about the parameters without some additional validation or replication data. But there are exceptions, e.g., Rudemo, et al. (1989) describe a highly nonlinear model that is both identifiable and informative, see section 3.7.

7.2 Measurement Error Models and Missing Data

The usual interpretation of the classical missing data problem (Little & Rubin, 1987) is that the values of some of the variables of interest may not be observable for all study participants. For example, a variable may be observed for 80% of the study, but unobserved for the other 20%. The techniques for analyzing missing data are continually evolving, but it is fair to say that most of the recent advances (multiple imputation, data augmentation, etc.) have been based on likelihood (and Bayesian) methods.

The classical measurement error problem discussed to this point is one in which one set of variables, which we call $\mathbf{X}$, is *never* observable, i.e., always missing. As such, the classical measurement error model is an extreme form of a missing data problem, but with *supplemental information* about $\mathbf{X}$ in the form of a surrogate, which we call $\mathbf{W}$, and possibly a second measure, which we call $\mathbf{T}$. Part of the art in measurement error modeling concerns how the supplemental information is related to the unobservable covariate.

Because there is a formal connection between the two fields, and because missing data analysis has become increasingly parametric, it is important to consider likelihood analysis of measurement error models, and this is the subject of this chapter.

At this point the reader should be struck by a seeming contradiction. Missing data analyses are becoming increasingly likelihood-based, but none of the measurement error techniques described in earlier chapters are based upon likelihood analysis. The separate development of two formally connected fields is intriguing historically, but has its roots in the distinction between *functional* statistical modeling and *structural* statistical modeling.

As we will indicate in this chapter, likelihood methods require statistical models for the distribution of $\mathbf{X}$, sometimes conditional on the observed covariates. Because these models describe the structure of $\mathbf{X}$, then are called structural models. There has tradi-

tionally been considerable concern in the measurement error literature about the robustness of estimation and inferences based upon structural models for unobservable variates. Fuller (1987, page 263) discusses this issue briefly in the classical nonlinear regression problem, and basically concludes that the results of structural modeling "may depend heavily on the (assumed) form of the $\mathbf{X}$ distribution". In probit regression, Carroll, et al. (1984) report that if one assumes that $\mathbf{X}$ is normally distributed, and it really follows a chi-squared distribution with one degree of freedom, then the effect on the likelihood estimate is "markedly negative". Essentially all research workers in the measurement error field come to a common conclusion: likelihood methods can be of considerable value, but the possible nonrobustness of inference due to model misspecification is a vexing and difficult problem.

The issue of model robustness is hardly limited to measurement error modeling. Indeed, it pervades statistics, and has led to the rise of a variety of semiparametric and nonparametric techniques. From this general point of view, *functional modeling* may be thought of as a group of semiparametric techniques. Functional modeling uses parametric models for the response, but makes no assumptions about the distribution of the unobserved covariate. In previous chapters, we have reviewed these functional techniques.

There is simply no agreement in the statistical literature as to whether functional or structural modeling is more appropriate. Many researchers strongly believe that one should make as few model assumptions as possible, and in our context would thus favor functional modeling. The argument here is that any extra efficiency gained by structural modeling is more than offset by the need to perform careful and often time-consuming sensitivity analyses. Other researchers believe that appropriate statistical analysis requires one to do one's best to model every feature of the data, and thus favor structural modeling.

We take a somewhat more relaxed view of these issues. There are many problems, e.g., linear and logistic regression with additive measurement error, where functional techniques are easily computed and fairly efficient, and we have a strong bias in such circumstances towards functional modeling. In other problems, for example the segmented regression problem in section 7.7, structural modeling clearly has an important role to play, and should not be neglected.

This and the next chapter can be thought of as presenting the basic ideas for structural modeling. In Chapter 9, we describe functional (semiparametric) methods when $\mathbf{X}$ is observed in an internal validation study.

7.3 Likelihood Methods when X is Unobserved

Often, $\mathbf{X}$ is unobservable even for a subset of the data. For example, it is practically impossible to observe a person's yearly dietary intake, long-term blood pressure, etc. These are the types of problems which earlier chapters have studied, through the additive and multiplicative error models. In this section, we allow for general error models, and for the possibility that a second measure $\mathbf{T}$ is available.

A likelihood analysis starts with determination of the joint distribution of $\mathbf{Y}$, $\mathbf{W}$ and $\mathbf{T}$ given $\mathbf{Z}$, as these are the observed variates. We condition on $\mathbf{Z}$ throughout, because its distribution does not depend on the unknown parameters. We first consider a simple problem wherein $\mathbf{Y}$, $\mathbf{W}$ and $\mathbf{X}$ are discrete random variables, no second measure $\mathbf{T}$ is observed, and there are no other covariates $\mathbf{Z}$. From basic probability, we know that

$$\begin{aligned}\mathrm{pr}(\mathbf{Y}=y,\mathbf{W}=w) &= \sum_x \mathrm{pr}(\mathbf{Y}=y,\mathbf{W}=w,\mathbf{X}=x)\\ &= \sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{W}=w,\mathbf{X}=x)\mathrm{pr}(\mathbf{W}=w,\mathbf{X}=x). \quad (7.1)\end{aligned}$$

When $\mathbf{W}$ is a surrogate (nondifferential measurement error, see section 1.6), it provides no additional information about $\mathbf{Y}$ when $\mathbf{X}$ is known, so that (7.1) is

$$\begin{aligned}&\mathrm{pr}\,(\mathbf{Y}=y,\mathbf{W}=w)\\ &= \sum_x \mathrm{pr}(\mathbf{Y}=y|\mathbf{X}=x,\mathcal{B})\mathrm{pr}(\mathbf{W}=w,\mathbf{X}=x), \quad (7.2)\end{aligned}$$

where we have now indicated the unknown parameter $\mathcal{B}$ in the underlying model. Thus, in addition to the underlying model, we must specify a model for the joint distribution of $\mathbf{W}$ and $\mathbf{X}$. How we do this depends on the model relating $\mathbf{W}$ and $\mathbf{X}$.

7.3.1 Error Models

For additive and multiplicative error models, it is natural to specify the the joint distribution of $\mathbf{W}$ and $\mathbf{X}$ in terms of the conditional distribution of $\mathbf{W}$ given $\mathbf{X}$. Using the result from elementary probability that $\text{pr}(\mathbf{W} = w, \mathbf{X} = x) = \text{pr}(\mathbf{W} = w|\mathbf{X} = x)\text{pr}(\mathbf{X} = x)$, (7.2) becomes

$$\sum_x \text{pr}(\mathbf{Y} = y|\mathbf{X} = x, \mathcal{B})\text{pr}(\mathbf{W} = w|\mathbf{X} = x)\text{pr}(\mathbf{X} = x). \qquad (7.3)$$

Equation (7.3) has three components: (a) the underlying model of primary interest; (b) the error model for $\mathbf{W}$ given the true covariates; and (c) the distribution of the true covariates. Both (a) and (b) are expected; almost all the methods we have discussed so far require an underlying model and an error model. However, (c) is unexpected, in fact a bit disconcerting, because it requires a model for the distribution of the unobservable $\mathbf{X}$. It is (c) that causes almost all the practical problems of implementation and model selection with maximum likelihood methods.

When there are covariates $\mathbf{Z}$ measured without error, or when there are second measures $\mathbf{T}$, (7.3) changes in two ways. The second measure is appended to $\mathbf{W}$, and all probabilities are conditional on $\mathbf{Z}$. In general, in problems where $\mathbf{X}$ is not observed but there is a natural error model, then in addition to specifying the underlying model and the error model, we must hypothesize a distribution for $\mathbf{X}$ given $\mathbf{Z}$.

The error model has a density or mass function which we will denote by $f_{W,T|Z,X}(w, t|z, x, \tilde{\alpha}_1)$. The density or mass function of $\mathbf{X}$ given $\mathbf{Z}$ will be denoted by $f_{X|Z}(x|z, \tilde{\alpha}_2)$. These densities depend on the unknown parameter vectors $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$.

In many applications, the error model does not depend on z. For example, in the classical additive measurement error model (1.1) with normally distributed measurement error, σ_u^2 is the only component of $\tilde{\alpha}_1$, there is no second measure $\mathbf{T}$, and the error model density is $\sigma_u^{-1}\phi\{(w - x)/\sigma_u\}$, where $\phi(\cdot)$ is the standard normal density function. If $\mathbf{W}$ is binary, a natural error model is the logistic where, for example, $\tilde{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13})$ and $\text{pr}(\mathbf{W} = 1|\mathbf{X} = x, \mathbf{Z} = z) = H(\alpha_{11} + \alpha_{12}x + \alpha_{13}^t z)$. Multiplicative models occur when $\mathbf{W} = \mathbf{XU}$, where typically $\mathbf{U}$ has a lognormal or gamma distribution with $E(\mathbf{U}) = 1$.

As seen in previous chapters, often the surrogate is replicated k times. In the classical error model with independent replicates, $\mathbf{W}$ consists of the k replicates, and $f_{W|Z,X}$ is the k-variate normal density function with mean zero, common variance σ_u^2, and zero correlation. A generalization of this error model that allows for correlations among the replicates has been studied (Wang, Carroll & Liang, 1995).

Statisticians are trained in the area of error modeling and thus specifying a sensible error model is often relatively easy. When parameters of the error model are estimated from external data the issue of transportability (Chapter 1), by which we mean that error models apply across different populations, is important. In some application areas, error model structures are studied independently of their role in measurement error modeling, and one can use this research to identify candidate error models for the problem at hand.

Specifying a model for the distribution of the true covariate $\mathbf{X}$ given all the other covariates $\mathbf{Z}$ is more difficult. Difficulties arise due to: (a) the distribution is usually not transportable, so that different studies yield very different models; and (b) $\mathbf{X}$ is not observed.

Nevertheless there are some obvious candidates for modeling $\mathbf{X}$ given $\mathbf{Z}$. When $\mathbf{X}$ is univariate, generalized linear models (section A.5) are natural and useful. More complex models are easily generated. For example, in many applications the distribution of $\mathbf{X}$ or $\log(\mathbf{X})$ appears to come from two populations. This can be modeled by the mixture of normals density function, as follows. Let $\tilde{\alpha}_2 = (\alpha_{21}, \mu_{x,1}, \mu_{x,2}, \sigma_{x,1}, \sigma_{x,2}, p)$. Then

$$\begin{aligned} f_{X|Z}(x,z) &= \frac{1-p}{\sigma_{x,1}}\phi\left(\frac{x-\alpha_{21}^t z-\mu_{x,1}}{\sigma_{x,1}}\right) \\ &\quad + \frac{p}{\sigma_{x,2}}\phi\left(\frac{x-\alpha_{21}^t z-\mu_{x,2}}{\sigma_{x,2}}\right). \end{aligned}$$

The density has mean $(1-p)\mu_{x,1} + p\mu_{x,2} + \alpha_{21}^t \mathbf{Z}$. The major problem in working with mixtures of normals is computational, as the parameters can be extraordinarily difficult to estimate. Because of this, Davidian & Gallant (1993, page 478) suggest another way of generating mixture distributions, see their Figure 3(d).

When $\mathbf{X}$ is multivariate, models for the true covariates become

more complex. Davidian & Gallant's mixture model generalizes easily to the case that all components of **X** are continuous. For mixtures of discrete and continuous variables, the models of Zhao, Prentice & Self (1992) hold considerable promise. Otherwise, one can proceed on a case-by-case basis. For example, one can split **X** into discrete and continuous components. The distribution of the continuous component given the discrete components might be modeled by multivariate normal linear regression, while that of the discrete component given **Z** could be any multivariate discrete random variable. We would be remiss in not pointing out that multivariate discrete models can be difficult to specify.

Having hypothesized the various models, the likelihood that $(\mathbf{Y} = y, \mathbf{W} = w, T = t)$ given that $\mathbf{Z} = z$ is then

$$f_{Y,W,T|Z}(y,w,t|z,\mathcal{B},\tilde{\alpha}_1,\tilde{\alpha}_2) \tag{7.4}$$

$$= \int f_{Y|Z,X,W,T}(y|z,x,w,t,\mathcal{B}) f_{W,T|Z,X}(w,t|z,x,\tilde{\alpha}_1) \times f_{X|Z}(x|z,\tilde{\alpha}_2) d\mu(x)$$

$$= \int f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{W,T|Z,X}(w,t|z,x,\tilde{\alpha}_1) \times f_{X|Z}(x|z,\tilde{\alpha}_2) d\mu(x). \tag{7.5}$$

The notation $d\mu(x)$ indicates that the integrals are sums if **X** is discrete and integrals if **X** is continuous. The assumption of nondifferential measurement error (section 1.6), which is equivalent to assuming that **W** and **T** are surrogates for **X**, was used in going from (7.4) to (7.5), and will be used without mention elsewhere in this chapter. The likelihood for the problem is just the product over the sample of the terms (7.5) evaluated at the data.

Of interest in applications is the density or mass function of **Y** given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$, which is (7.5) divided by its integral or sum over y. This density is an important tool in the process of model criticism, because it allows us to compute such diagnostics as the conditional mean and variance of **Y** given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$, so that standard model verification techniques from regression analysis can be used.

7.3.2 Likelihood and External Second Measures

If the second measure comes from an external data set for which the response is not observed, one can often estimate $\tilde{\alpha}_1$ from the ex-

ternal data, substitute the estimate of in $\tilde{\alpha}_1$ in (7.5), and maximize this (pseudo) likelihood with respect to the remaining parameters. We have emphasized the need for caution in assuming transportability. This requires that we allow for a possibly different value of the parameter $\tilde{\alpha}_2^*$ that determines the distribution of the true predictor given the other covariates. The likelihood for such external data is

$$f_{W,T|Z}(w,t|z,\tilde{\alpha}_1,\tilde{\alpha}_2^*) = \int f_{W,T|Z,X}(w,t|z,x,\tilde{\alpha}_1) f_{X|Z}(x|z,\tilde{\alpha}_2^*) d\mu(x). \quad (7.6)$$

7.3.3 The Berkson Model

In the Berkson model, a univariate $\mathbf{X}$ is not observed but it is related to a univariate $\mathbf{W}$ by $\mathbf{X} = \mathbf{W} + \mathbf{U}$, perhaps after a transformation. There are no other covariates. Usually, $\mathbf{U}$ is taken to be independent of $\mathbf{W}$ and normally distributed with mean zero and variance σ_u^2, but more complex models are possible. For example, in the bioassay data of Chapter 3, the variance might be $\sigma_u^2 \mathbf{W}^{2\theta}$.

The additive model is not a requirement. In some cases, it might be more reasonable to assume that $\mathbf{X} = \mathbf{WU}$, where $\mathbf{U}$ has mean 1.0 and is either lognormal or gamma.

The Berkson additive model has an unusual feature, in that for linear regression the naive analysis ignoring measurement error gives correct inference about the regression parameters (Berkson, 1950). The reason for this is quite simple, namely that $E(\mathbf{X}|\mathbf{W}) = \mathbf{W}$. This means that in the Berkson context, the naive analysis is the same as the regression calibration analysis of Chapter 3. Thus, as in regression calibration, the additive Berkson model with homoscedastic errors leads to consistent estimates of nonintercept parameters in loglinear models, and often nearly consistent estimates in logistic regression. In the latter case, the exceptions occur with severe measurement error and a strong predictive effect, see Burr (1988).

The likelihood of the observed data is (7.2) because $\mathbf{W}$ is a surrogate. At this point, however, the analysis changes. When the Berkson model holds, it should not be forced into an additive error model, and so in place of (7.3) we write

$$\text{pr}(\mathbf{Y} = y, \mathbf{W} = w) \quad (7.7)$$

$$= \sum_x \text{pr}(\mathbf{Y} = y|\mathbf{X} = x, \mathcal{B})\text{pr}(\mathbf{X} = x|\mathbf{W} = w)\text{pr}(\mathbf{W} = w).$$

The third component of (7.7) is the distribution of **W**, and conveys no information about the critical parameter $\mathcal{B}$. Thus, we will divide both sides of (7.7) by $\text{pr}(\mathbf{W} = w)$ to get likelihoods conditional on **W**. In general problems, we must specify the conditional density or mass function of **X** given **W**, which we denote by $f_{X|W}(x|w, \tilde{\gamma})$. In the usual Berkson model, $\tilde{\gamma}$ is σ_u^2, and the density is $\sigma_u^{-1}\phi\{(x-w)/\sigma_u\}$. In a Berkson model where the variance is proportional to $\mathbf{W}^{2\theta}$, the density is $(w^\theta\sigma_u)^{-1}\phi\{(x-w)/(w^\theta\sigma_u)\}$. The likelihood function then becomes

$$f_{Y|Z,W}(y|z, w, \mathcal{B}, \tilde{\gamma}) = \int f_{Y|Z,X}(y|z, x, \mathcal{B}) f_{X|W}(x|z, \tilde{\gamma}) d\mu(x). \quad (7.8)$$

The likelihood for the problem is the product over the sample of the terms (7.8) evaluated at the data.

As a practical matter, there is rarely a direct "second measure" in the Berkson additive or multiplicative models. This means that the parameters in the Berkson model can be estimated only through the likelihood (7.8). In some cases, such as linear regression, not all of the parameters can be identified (estimated). For nonlinear models, identification usually is possible.

In classical generalized linear models, a likelihood analysis of a homoscedastic, additive Berkson model can be shown to be equivalent to a random coefficients analysis with random intercept for each study participant.

7.3.4 Error Model Choice

Modeling always has options. Even when **X** is unobserved, one can use either the error model likelihood (7.5), or the Berkson likelihood (7.8) and its extension (7.9) described below. With additive or multiplicative measurement error, the former seems to us the most natural. The reasons are twofold: (i) the error model can be checked by replicates (section 7.6) or external data; and (ii) the error model focuses the indeterminacy of the likelihood on the distribution of **X**.

There is, however, nothing illegal in simply specifying a mod-

el for $\mathbf{X}$ given $\mathbf{W}$, as in equation (7.8), or a model for $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$, as in equation (7.9). One can even specify reasonably flexible models for such distributions, as in section 7.3.1 or using the Davidian & Gallant models. There is effectively no literature on whether such an approach can yield sensible answers when additive/multiplicative error models hold.

7.4 Likelihood When X is Partly Observed

For problems in which $\mathbf{X}$ is observed for some study participants, i.e., internal validation, a modification of the Berkson model is required. Because $\mathbf{X}$ is sometimes observed, one has data to model its distribution given the observed covariates. Let $f_{X|Z,W}(x|z,w,\tilde{\gamma})$ be the appropriate model, in which case (7.8) becomes

$$\begin{aligned} & f_{Y|Z,W}(y|z,w,\mathcal{B},\tilde{\gamma}) \\ & = \int f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{X|Z,W}(x|z,w,\tilde{\gamma}) d\mu(x). \end{aligned} \tag{7.9}$$

We assume that in a sample of size n, we observe $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$. For a (usually small) subset of the data, $\mathbf{X}$ is also observed ($\Delta = 1$), while in all other cases $\mathbf{X}$ is not observed ($\Delta = 0$).

As we have pointed out in section 7.2, when $\mathbf{X}$ is partially observed we are in the context of a classical missing data problem, with supplementary information coming from $\mathbf{W}$. As is discussed by Little & Rubin (1987, Chapter 5), the mechanism for observing $\mathbf{X}$ is critical to the validity of likelihood inferences. As they discuss, $\mathbf{X}$ must be *missing at random*, i.e., whether or not $\mathbf{X}$ is observed depends only on the values of $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$, and not otherwise on the value of $\mathbf{X}$ itself. Somewhat more formally, we must assume that the probability that $\mathbf{X}$ is observed is $\pi(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ (there is also a technical matter called "parameter distinctness" which holds almost universally in our context, and will be ignored).

With the proviso that $\mathbf{X}$ is missing at random, the likelihood of the observed data is proportional to

$$\begin{aligned} \prod_{i=1}^{n} \Big[& \{ f_{Y|Z,X}(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) f_{X|Z,W}(\mathbf{X}_i|\mathbf{Z}_i, \mathbf{W}_i, \tilde{\gamma}) \}^{\Delta_i} \\ & \times f_{Y|Z,W}^{1-\Delta_i}(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{W}_i, \mathcal{B}, \tilde{\gamma}) \Big]. \end{aligned} \tag{7.10}$$

The actual likelihood is (7.10) multiplied by the likelihood of all the observable covariates, but since this latter likelihood contains no information about $\mathcal{B}$, it can be ignored.

Satten & Kupper (1993) describe a likelihood method for logistic regression when $\mathbf{X}$ is observed only when $\mathbf{Y} = 0$.

7.5 Numerical Computation of Likelihoods

The overall likelihood based on a sample of size n is the product over the sample of (7.4) when $\mathbf{X}$ is unobserved, the product over the sample of (7.8) in the Berkson model, or (7.10). Typically one maximizes the logarithm of the overall likelihood in the unknown parameters. There are two ways one can maximize the likelihood function. The most direct is to compute the likelihood function itself, and then use numerical optimization techniques to maximize the likelihood. Below we provide a few details about computing the likelihood function. The second general approach is to view the problem as a missing data problem (section 7.2), and then use missing data techniques, see for example Little & Rubin (1987), Tanner (1993) and Geyer & Thompson (1992).

Computing the likelihoods (7.5)–(7.9) analytically is easy if $\mathbf{X}$ is discrete, as the conditional expectations are simply sums of terms. For example, consider (7.9), and suppose that $\mathbf{X}$ has possible values $(x_1, ..., x_K)$ with probabilities $p(x_k|z, w, \tilde{\gamma})$. Then (7.9) is given by

$$\sum_{k=1}^{K} p(x_k|z, w, \tilde{\gamma}) f_{Y|Z,X}(y|z, x_k, \mathcal{B}).$$

Likelihoods in which $\mathbf{X}$ has some continuous components can be computed using a number of different approaches. In some problems the loglikelihood can be computed or very well approximated analytically, e.g., linear, probit and logistic regression with $(\mathbf{W}, \mathbf{X})$ normally distributed, see section 7.9.2. In most problems that we have encountered, $\mathbf{X}$ is a scalar or a 2×1 vector. In these cases, standard numerical methods such as Gaussian quadrature can be applied, although they are not always very good. When sufficient computing resources are available, the likelihood can be computed using Monte-Carlo techniques (section 7.9.1).

7.6 Framingham Data

The Framingham heart study was described in section 4.5. Here $\mathbf{X}$ is not observable, and the likelihoods of section 7.3.1 are appropriate. The sample size is $n = 1{,}615$. As before, $\mathbf{Z}$ includes age, smoking status, and serum cholesterol. Transformed systolic blood pressure (SBP) is log(SBP-50).

At Exam #2, the mean and standard deviation of transformed systolic blood pressure are 4.374 and .226, respectively, while the corresponding figures at Exam #3 are 4.355 and .229. The difference between measurements at Exam #2 and Exam #3 has mean 0.019 and standard deviation .159, indicating a statistically significantly difference in means due largely to the sample size ($n = 1,615$). However, the following analysis will allow for differences in the means. The standard deviations are sufficiently similar that we will assume that the two exams have the same variability.

We write $\mathbf{W}$ and $\mathbf{T}$ for the transformed SBP at Exams 3 and 2, respectively. Since Exam #2 is not a true replicate, we are treating it as a second measure, differing from Exam #3 only in the mean. Thus, $\mathbf{W} = \mathbf{X} + \mathbf{U}$ and $\mathbf{T} = \alpha_{11} + \mathbf{X} + \mathbf{V}$, where $\mathbf{U}$ and $\mathbf{V}$ have common measurement error variance σ_u^2, and α_{11} represents the (small) difference between the two exams.

There is justification for the assumption that transformed systolic blood pressure can be modeled reasonably by an additive model with normally distributed, homoscedastic measurement error. For example, if the additive normal error model holds, the differences in the systolic blood pressures at Exams #2 and #3 should be approximately normally distributed. In Figure 7.1, we provide the normal quantile–quantile plot of these differences in the original (left plot) and transformed (right plot) scales. The former plot indicates some skewness, suggesting the need for a transformation, while the latter plot is nearly linear (except for a small number of observations in the tails). In addition, the intra-individual standard deviation is plotted against the mean in Figure 7.2, with a lowess line. The lack of pattern is further confirmation that the transformation is a reasonable one.

Since the transformed systolic blood pressures are themselves approximately normally distributed, we will also assume that $\mathbf{X}$ given $\mathbf{Z}$ is normally distributed with mean $\alpha_{21}^t Z$ and variance σ_x^2.

Using the probit approximation to the logistic (section 3.9.2), it

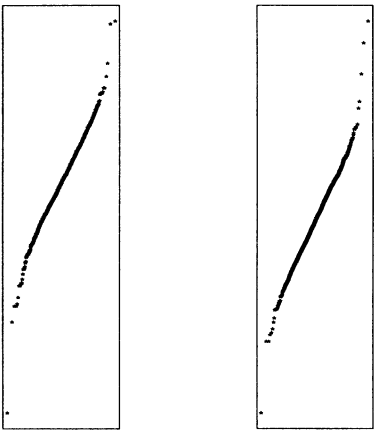

Figure 7.1. *Framingham data. Normal q-q plot of the difference between the second and third exams. Plot on left is in the original SBP scale, plot on right uses the transformation log(SBP-50).*

is possible to compute (7.5) analytically, see section 7.9.2 in the appendix. We used this analytical calculation, rather than numerical integration. When using all the data, the likelihood estimate for systolic blood pressure had a logistic coefficient of 2.013 with an (information) estimated standard error of 0.496, which is essentially the same as the regression calibration analysis, compare with Table 4.1.

We repeated the likelihood analysis but with the partially replicated data, where Exam #2 was used for only 30 randomly selected individuals The logistic coefficient for SBP is now 2.146, with (information) standard error 0.604. For comparison, regression calibration gives similar answers; coefficient estimate 2.074, sandwich standard error 0.533 and bootstrap standard error 0.566, see Table 4.2.

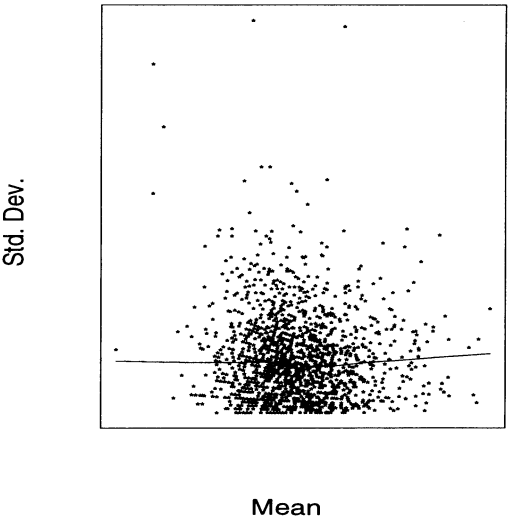

Figure 7.2. *Framingham data. Plot of intra-individual standard deviation versus mean, with lowess line. Lack of pattern indicates proper transformation to additivity. Variable is log(SBP-50).*

7.7 Bronchitis Example

In occupational medicine an important problem is the assessment of the health hazard of specific harmful substances in a working area. One approach to modeling assumes that there is a threshold concentration, called the threshold limiting value (TLV) under which there is no risk due the substance. Estimating the TLV is of particular interest in the industrial workplace. We consider here the specific problem of estimating the TLV in a dust-laden mechanical engineering plant in Munich.

The regressor variable $\mathbf{X}$ is the logarithm of 1.0 plus the average dust concentration in the working area over the period of time in question. In addition, the duration of exposure $\mathbf{Z}_1$ and smoking status $\mathbf{Z}_2$ are also measured. Following Ulm (1991), we based our analysis upon the segmented logistic model

$$\text{pr}(\mathbf{Y} = 1|\mathbf{X}, \mathbf{Z})$$

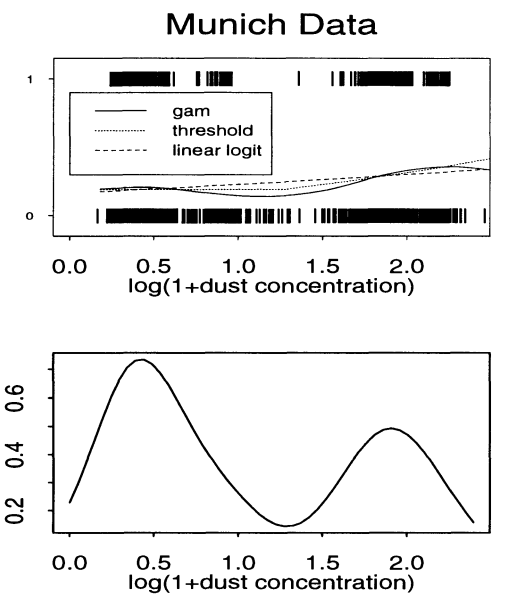

Figure 7.3. *The Munich plant. Top figure shows binary regression models of bronchitis incidence on log(1 + dust concentration) using (i) gam (generalized additive model), (ii) segmented logistic regression and ordinary logistic regression. The bottom figure is a kernel density estimate of the observed concentrations.*

$$= H\left\{\beta_0 + \beta_{x,1}(\mathbf{X} - \beta_{x,2})_+ + \beta_{z,1}\mathbf{Z}_1 + \beta_{z,2}\mathbf{Z}_2\right\}. \quad (7.11)$$

where $(a)_+ = a$ if $a > 0$ and $= 0$ if $a \leq 0$. The parameter of primary interest is $\beta_{x,2}$, the TLV.

It is impossible to measure $\mathbf{X}$ exactly, and instead sample dust concentrations were obtained several times between 1960 and 1977. The resulting measurements are $\mathbf{W}$. There were 1,246 observations: 23% of the workers reported chronic bronchitis, and 74% were smokers. Measured dust concentration had a mean of 1.07 and a standard deviation of 0.72. The durations were effectively independent of concentrations, with correlation 0.093, compare with Ulm's (1991) Figure 3. Smoking status is also effectively independent of dust concentration, with the smokers having mean concentration 1.068, and the nonsmokers having mean 1.083. Thus, in this example, for likelihood calculations we will treat the $\mathbf{Z}$'s as if they were

independent of $\mathbf{X}$.

A preliminary segmented regression analysis ignoring measurement error suggested an estimated TLV $\widehat{\beta}_{x,2} = 1.27$. We will call this the naive TLV. In Figure 7.3 we show the results of such an analysis when regressing bronchitis only on dust concentration. A generalized additive model fit using S-plus suggests a threshold in the neighborhood of the estimated value. Note also that an ordinary logistic regression is sufficiently different from the generalized additive model fit to suggest the need for a changepoint.

In Figure 7.3 we also plot a kernel density estimate of the observed dust concentrations, with a Gaussian kernel and bandwidth equal to 0.25. The dust concentrations appear strongly bimodal, with almost no observations in the vicinity of the naive TLV. With such a clear indication of two subpopulations, one would expect a naive TLV of between 1.0 and 1.5, the range that separates the two subpopulations. We fit a two-population mixture normal model to the data, i.e., one having density function

$$f_W(w) = \frac{p}{\sigma_1}\phi\left(\frac{w-\mu_1}{\sigma_1}\right) + \frac{1-p}{\sigma_2}\phi\left(\frac{w-\mu_2}{\sigma_2}\right). \qquad (7.12)$$

A similar model in which dust concentration is not assumed independent of smoking and duration is discussed in section 7.3.1. The maximum likelihood estimate of p is 0.607, the means are $(0.520, 1.927)$ and the variance are $(0.236^2, 0.215^2)$.

We computed the bias of the naive TLV estimator and that of the SIMEX estimators with linear and quadratic extrapolant functions, for a simulated data set designed to approximate the Munich data. For the distribution of $\mathbf{X}$, we used a mixture normal distribution with $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p) = (0.45, 1.90, 0.03, 0.03, 0.60)$, which has mean 1.03 and variance 0.524. We added small amounts of measurement error with variance σ_u^2 ranging from 0.0 to 0.04; at the extreme end of the scale, we have a situation that $\mathbf{X}$ comes from two subpopulations, both of which are estimated with large measurement error.

The biases are exhibited in Figure 7.4. The naive estimator and the SIMEX estimator with linear extrapolant are both considerably more biased than the SIMEX estimator with quadratic extrapolant. Note that the SIMEX estimator with rational linear extrapolant has very poor bias behavior. In this example, the regression calibration estimator is very badly biased (not shown).

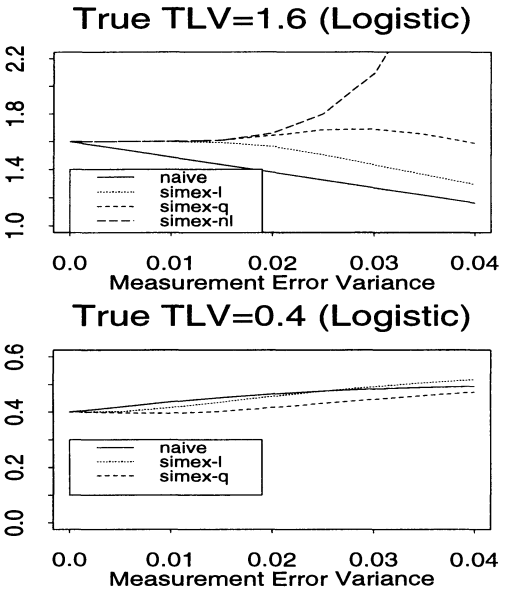

Figure 7.4. *Limit of estimated TLV in a mixture normal model for the naive estimator and various SIMEX estimators. See text for details.*

As far as we can ascertain, there are no data available to fit an error model relating $\mathbf{W}$ to $\mathbf{X}$. In the absence of such information, for illustration we used an additive error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$, and we assumed that $\sigma_u^2 = 0.035$, making $\widehat{\sigma}_x^2 = 0.489$. While the error variance σ_u^2 is rather small relative to the marginal variance of $\mathbf{X}$ (σ_x^2), it is fairly large relative to the variance of each of the components of the mixture.

The likelihood estimator was computed assuming that $\mathbf{X}$ has a mixture normal distribution, and making the probit approximation to the logistic. We used the quadratic extrapolant for SIMEX.

Our theoretical bias calculations suggest a substantial downward bias in the naive estimator, and so, as expected, the maximum likelihood estimator taking measurement error into account gives a substantial correction to the naive estimator. The maximum likelihood estimate is $\widehat{\beta}_{x,2} = 1.76$, with a bootstrap standard error of 0.21 and a profile likelihood 95% confidence interval from 0.50 to 2.00. Somewhat surprisingly, the SIMEX estimator is 1.40, with

bootstrap estimated standard error 0.34. It is possible that the lack of a substantial correction in the SIMEX estimated value is due to its variability.

In section 7.1, we raised the issue of the efficiency of functional versus structural modeling using maximum likelihood estimation. Küchenhoff & Carroll (1995) investigated this issue in the context of segmented linear regression. In simulations with $\mathbf{X}$ and $\mathbf{W}$ normally distributed, they found that the maximum likelihood estimated was typically far more efficient than the SIMEX estimate with linear, quadratic or rational linear extrapolants. By assuming that the $\mathbf{X}$'s follow a mixture of normals distribution, we have thus added considerable information to the problem, and likelihood takes advantage of this information. The smaller variance of the maximum likelihood estimator is essentially the result of modeling assumptions.

7.8 Quasilikelihood and Variance Function Models

In a quasilikelihood and variance function (QVF) model, recall that we model only the mean and variance functions of the response, and not its entire distribution. As before, we write the mean and variances as $E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = f(\mathbf{Z},\mathbf{X},\mathcal{B})$ and $\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) = \sigma^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)$.

We are concerned here particularly with the case when $\mathbf{X}$ is unobservable, and that only a surrogate can be observed. The surrogate of course is $\mathbf{W}$, and one should remember that the surrogate might have more than one component if there are replicates.

Quasilikelihood and variance function techniques require that we compute the mean and variance functions of the *observed* data (and not the unobservable data). As we have seen before, these are

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = E\left\{f(\cdot)|\mathbf{Z},\mathbf{W}\right\}. \tag{7.13}$$

$$\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = \sigma^2 E\left\{g^2(\cdot)|\mathbf{Z},\mathbf{W}\right\} + \mathrm{var}\left\{f(\cdot)|\mathbf{Z},\mathbf{W}\right\}. \tag{7.14}$$

Equations (7.13)–(7.14) define a variance function model. If we knew the functional forms of the mean and variance functions, then we could apply the fitting and model criticism techniques discussed in section A.4. Note how both (7.13) and (7.14) require an estimate of a model for the distribution of the unobserved covariate given the observed covariates and the surrogate.

Even if $\mathbf{X}$ is unobserved, one can estimate the QVF parameters. This is possible provided, as in section 7.3.1, one has a model for $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$, and a model for $\mathbf{X}$ given $\mathbf{Z}$. The (reduced) likelihood for a single observation based upon only the observed covariates is

$$\int f_{W|Z,X}(\mathbf{W}|\mathbf{Z}, x, \tilde{\alpha}_1) f_{X|Z}(x|\mathbf{Z}, \tilde{\alpha}_2) d\mu(x),$$

where again the integral is replaced by a sum if $\mathbf{X}$ is discrete. The $(\mathbf{W}, \mathbf{Z})$ data are used to estimate $(\tilde{\alpha}_1, \tilde{\alpha}_2)$ by multiplying this reduced likelihood over the observations, and maximizing. The density or mass function of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is then given by

$$f_{X|Z,W}(x|z,w) = \frac{f_{W|Z,X}(w|z,x,\tilde{\alpha}_1) f_{X|Z}(x|z,\tilde{\alpha}_2)}{\int f_{W|Z,X}(w|z,v,\tilde{\alpha}_1) f_{X|Z}(v|z,\tilde{\alpha}_2) d\mu(v)}.$$

From this, one can obtain (7.13)–(7.14) by integration either analytically or numerically. The sandwich method or the bootstrap can be used for inference, although of course one must take into account the estimation of $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$.

When there is internal validation, there are functional modeling (semiparametric) techniques for QVF estimation. The basic estimating functions for QVF estimation are given in section A.4, and they can be applied to the semiparametric methods of Chapter 9.

7.9 Appendix

7.9.1 Monte-Carlo Computation of Integrals

If one can easily generate observations from the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ (error model) or given $(\mathbf{Z}, \mathbf{W})$ (calibration model), an appealing and easily programmed Monte–Carlo approximation due to McFadden (1989) can be used to compute likelihoods. The error model likelihood (7.5) can be approximated as follows. Generate on a computer a sample $(\mathbf{X}_1^s, \cdots, \mathbf{X}_N^s)$ of size N from the density $f(x|z, \tilde{\alpha}_2)$ of $\mathbf{X}$ given $\mathbf{Z} = z$. Then for large enough N,

$$\begin{aligned} & f_{Y,W|Z}(y, w|\mathbf{z}, \mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2) \qquad (7.15) \\ & \quad \approx N^{-1} \sum_{i=1}^{N} f_{Y|Z,X}(y|z, \mathbf{X}_i^s, \mathcal{B}) f_{W|Z,X}(w|z, \mathbf{X}_i^s, \tilde{\alpha}_1). \end{aligned}$$

The dependence of (7.15) on $\tilde{\alpha}_2$ comes from the fact that the distribution of $\mathbf{X}$ given $\mathbf{Z}$ depends on $\tilde{\alpha}_2$.

We approximate (7.9) by generating a sample $(\mathbf{X}_1^s, \cdots, \mathbf{X}_N^s)$ of size N from the distribution $f(x|z, w, \tilde{\gamma})$ of $\mathbf{X}$ given $(\mathbf{Z} = z, \mathbf{W} = w)$. Then for large enough N,

$$f_{Y|Z,W}(y|z, w, \mathcal{B}, \tilde{\gamma}) \approx N^{-1} \sum_{i=1}^{N} f_{Y|Z,X}(y|z, \mathbf{X}_i^s, \mathcal{B}).$$

This "brute force" Monte–Carlo integration method is computing intensive. There are two reasons for this. First, one has to generate random observations for each value of $(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)$, which may be a formidable task if the sample size is large. Second, and somewhat less important, maximum likelihood is an iterative algorithm, and one must generate simulated $\mathbf{X}$'s at each iteration. Brown (1992) suggests that N must be fairly large compared to $n^{1/2}$ in order to eliminate the effects of Monte-Carlo variance. He also suggests a modification which will be less computing intensive.

7.9.2 Linear, Probit and Logistic Regression

In some cases, the required likelihoods can be computed exactly or very nearly so. Suppose that $\mathbf{W}$ and $\mathbf{T}$ are each normally distributed unbiased replicates of $\mathbf{X}$, being independent given $\mathbf{X}$, and each having covariance matrix Σ_{uu} $(= \tilde{\alpha}_1$ in our general notation). Suppose also that $\mathbf{X}$ itself is normally distributed with mean $\alpha_{21}^t \mathbf{Z}$ and covariance matrix Σ_{xx} $(= \alpha_{22}$ in our general notation). As elsewhere, all distributions are conditioned on $\mathbf{Z}$.

In normal linear regression where the response has mean $\beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}$ and variance σ^2, the joint distribution of $(\mathbf{Y}, \mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ is multivariate normal with means $\beta_0 + \beta_x^t \gamma_{\text{em},1}^t \mathbf{Z} + \beta_z^t \mathbf{Z}$, $\gamma_{\text{em},1}^t \mathbf{Z}$ and $\gamma_{\text{em},1}^t \mathbf{Z}$, and covariance matrix

$$\Sigma_{y,w,t} = \begin{bmatrix} \sigma^2 + \beta_x^t \Sigma_{xx} \beta_x & \beta_x^t \Sigma_{xx} & \beta_x^t \Sigma_{xx} \\ \Sigma_{xx} \beta_x & \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xx} \\ \Sigma_{xx} \beta_x & \Sigma_{xx} & \Sigma_{xx} + \Sigma_{uu} \end{bmatrix}.$$

For probit and logistic regression, we compute the joint density using the formulas $f_{Y,W|Z} = f_{Y|Z,W} f_{W|Z}$ and $f_{Y,W,T|Z} = f_{Y|Z,W,T} f_{W,T|Z}$. This requires a few preliminary calculations.

First consider $\mathbf{W}$ alone. Our model says that $\mathbf{W}$ given $\mathbf{Z}$ is nor-

mally distributed with mean $\alpha_{21}^t \mathbf{Z}$ and covariance matrix $\Sigma_{xx} + \Sigma_{uu}$. Define $\Lambda_w = \Sigma_{xx}(\Sigma_{xx} + \Sigma_{uu})^{-1}$, $m(\mathbf{Z}, \mathbf{W}) = (I - \Lambda_w)\alpha_{21}^t \mathbf{Z} + \Lambda_w \mathbf{W}$ and $\Sigma_{x|z,w} = (I - \Lambda_w)\Sigma_{xx}$. From linear regression theory, $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ is normally distributed with mean $m(\mathbf{Z}, \mathbf{W})$ and covariance matrix $\Sigma_{x|z,w}$.

Next consider $\mathbf{W}$ and $\mathbf{T}$ together. Our model says that given $\mathbf{Z}$ they are jointly normally distributed with common mean $\gamma_{\text{em},1}^t \mathbf{Z}$, common individual covariances $(\Sigma_{xx} + \Sigma_{uu})$ and cross-covariance matrix Σ_{xx}. If we define

$$\Lambda_{w,t} = (\Sigma_{xx}, \Sigma_{xx}) \begin{bmatrix} \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xx} \\ \Sigma_{xx} & \Sigma_{xx} + \Sigma_{uu} \end{bmatrix}^{-1} = (\Sigma_{xx}, \Sigma_{xx})\Gamma_{w,t}^{-1},$$

then $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$ is normally distributed with mean and covariance matrix given by

$$\begin{aligned} m(\mathbf{Z}, \mathbf{W}, \mathbf{T}) &= \gamma_{\text{em},1}^t \mathbf{Z} \\ &\quad + \Lambda_{w,t} \left\{ (\mathbf{W} - \gamma_{\text{em},1}^t \mathbf{Z})^t, (\mathbf{T} - \gamma_{\text{em},1}^t \mathbf{Z})^t \right\}^t; \\ \Sigma_{x|z,w,t} &= \Sigma_{xx} - \Lambda_{w,t}(\Sigma_{xx}, \Sigma_{xx})^t, \end{aligned}$$

respectively.

Now we return to probit and logistic regression. In probit regression, exact statements are possible. We have indicated that given either $(\mathbf{Z}, \mathbf{W})$ or $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$, $\mathbf{X}$ is normally distributed with mean $m(\cdot)$ and covariance matrix $\Sigma_{x|\cdot}$, where $\Sigma_{x|\cdot}$ is either $\Sigma_{x|z,w}$ or $\Sigma_{x|z,w,t}$, and similarly for $m(\cdot)$. From the calculations in the appendix to Chapter 3, it follows that

$$\text{pr}(\mathbf{Y} = 1 | \mathbf{Z}, \mathbf{W}, \mathbf{T}) = \Phi \left[\frac{\beta_0 + \beta_x^t m(\cdot) + \beta_z^t \mathbf{Z}}{(1 + \beta_x^t \Sigma_{x|\cdot} \beta_x)^{1/2}} \right].$$

For logistic regression (section 3.9.2), a good approximation is

$$\text{pr}(\mathbf{Y} = 1 | \mathbf{Z}, \mathbf{W}, \mathbf{T}) \approx H \left[\frac{\beta_0 + \beta_x^t m(\cdot) + \beta_z^t \mathbf{Z}}{(1 + \beta_x^t \Sigma_{x|\cdot} \beta_x / c^2)^{1/2}} \right], \qquad (7.16)$$

where $c = 15\pi/(3^{1/2}16)$; see also Monahan & Stefanski (1992).

Write $\Theta = (\mathcal{B}, \Sigma_{uu}, \alpha_{21}, \Sigma_{xx})$, and $r(\mathbf{W}) = r(\mathbf{W}, \alpha_{21}) = (\mathbf{W} - \alpha_{21}^t \mathbf{Z})$. Using (7.16), except for a constant in logistic regression the logarithm of the approximate likelihood for $(\mathbf{Y}, \mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ is

$$\begin{aligned} &\ell(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{T}, \Theta) = -(1/2)\log\{\det(\Gamma_{w,t})\} \qquad (7.17) \\ &+ \mathbf{Y}\log\{H(\cdot)\} + (1 - \mathbf{Y})\log\{1 - H(\cdot)\} \end{aligned}$$

$$-(1/2)\left\{r^t(\mathbf{W}), r^t(\mathbf{T})\right\}\Gamma_{w,t}^{-1}\left\{r^t(\mathbf{W}), r^t(\mathbf{T})\right\}^t.$$

A similar result applies if only $\mathbf{W}$ is measured, namely

$$\begin{aligned}\ell(\mathbf{Y},\mathbf{Z},\mathbf{W},\Theta) &= -(1/2)\log\left\{\det\left(\Sigma_{xx}+\Sigma_{uu}\right)\right\}\\ &+\mathbf{Y}\log\left\{H(\cdot)\right\}+(1-\mathbf{Y})\log\left\{1-H(\cdot)\right\}\\ &-(1/2)r^t(\mathbf{W},\gamma_{\mathrm{em},1})\left(\Sigma_{uu}+\Sigma_{xx}\right)^{-1}r(\mathbf{W},\gamma_{\mathrm{em},1}).\end{aligned}$$

CHAPTER 8

BAYESIAN METHODS

8.1 Overview

The Bayesian approach to statistics treats all parameters as random variables, with the randomness of a parameter representing uncertainty about its value. In this section, we give a quick introduction to the Bayesian paradigm. The reader is referred to Box & Tiao (1973) or Berger (1985) for a thorough introduction.

Bayesian analysis of parametric models requires specifying a likelihood (Chapter 7) that is then interpreted as the conditional density of the data given the parameters. It also requires a prior distribution for the parameters, representing knowledge about the parameters prior to data collection. The product of the prior and likelihood is the joint density of the data and the parameters. Often, one uses noninformative priors, meaning that the prior tells us extremely little about the parameters, relative to what is learned from the sample. However, if there is substantial prior knowledge about some parameters, then using informative priors for them leads to a more effective analysis.

Given the joint density of the data and parameters, one can integrate out the parameters to get the marginal density of the data. One can then divide the joint density by this marginal density to get the posterior density, i.e., the conditional density of the parameters given the data. The posterior summarizes all of the information about the values of the parameters and is the basis for all Bayesian inference. For example, the mean, median, or mode of the posterior density are all suitable point estimators. A region with probability $(1 - \alpha)$ under the posterior is called a "credible set," and is a Bayesian analog to a confidence region.

Computing the posterior distribution is often a non-trivial problem, because it usually requires high-dimensional numerical inte-

gration. This computational problem is the subject of much recent research, with many major advances. The method currently receiving the most attention in the literature is the Gibbs sampler (Hastings, 1970; Geman & Geman, 1984; Gelfand & Smith, 1990), good introductions to which are given by Smith & Gelfand (1992) and Casella & George (1992). Also, see Tanner (1993) for a book-length introduction to modern methods for computing posteriors, including the Gibbs sampler.

The Gibbs sampler generates a Markov chain whose stationary distribution is the posterior distribution. The key feature of the Gibbs sampler is that this chain can be simulated using only the joint density of the parameters and the data, e.g., the product of the likelihood and the prior, and not the unknown posterior density. If the chain is run long enough, then the observations in a sample from the chain are approximately identically distributed with common distribution equal to the posterior. Thus posterior moments, the posterior density, and other posterior quantities can be estimated from a sample from the chain.

Because of its current popularity we use the Gibbs sampler in the examples of this section. The examples are chosen to illustrate two general approaches to the Bayesian analysis of measurement error models. Data from a study of cervical cancer are used to illustrate an analysis based on "filling in" the missing $\mathbf{X}$'s. The Framingham data are used to illustrate a more standard Bayesian analysis based on approximate calculation of the likelihood of the observed data exploiting the regression calibration approximation (3.1). In both examples we use the Gibbs sampler with the intent of illustrating its application in the Bayesian analysis of measurement error models, and not specifically to promote or endorse its use to the exclusion of other computational methods.

The usual distinction between classical structural and functional models, namely whether unknown covariates ($\mathbf{X}_i$'s) are random variables or fixed parameters, is blurred in the Bayesian framework, where all parameters are random. Instead, the Bayesian distinction between functional and structural models is that under the latter the $\mathbf{X}_i$'s have a common parametric distribution. This agrees with our more modern contrast between functional modeling and structural modeling discussed in section 1.2.

As usual $\mathbf{Z}$ and $\mathbf{X}$ are the error-free covariate and the covariate measured with error, respectively. For most of this chapter we use

Gibbs notation to indicate density functions, so that for example $[\mathbf{W}|\mathbf{X}, \mathbf{Z}, \tilde{\alpha}_1]$ denotes the density of $\mathbf{W}$ given $(\mathbf{Z}, \mathbf{X})$ and the parameter $\tilde{\alpha}_1$, while $[\mathbf{X}|\mathbf{Z}, \tilde{\alpha}_2]$ denotes the density of $\mathbf{X}$ given $\mathbf{Z}$ and the parameter $\tilde{\alpha}_2$.

With this notation a Bayesian structural model has $[\mathbf{X}|\mathbf{Z}, \tilde{\alpha}_2]$ independent of i, that is, of the same form for all i and with a common parameter $\tilde{\alpha}_2$. We use such structural models in this chapter. Examples of this approach are given by Schmid & Rosner (1993), Richardson & Gilks (1993) and Stephens & Dellaportas (1992).

There are at least several ways to formulate a Bayesian functional model. One way would allow $[\mathbf{X}|\mathbf{Z}, \tilde{\alpha}_2]$ to depend on the observation number, i; a possible approach to this would be a "hierarchical" model, where the form of $[\mathbf{X}|\mathbf{Z}, \tilde{\alpha}_2]$ is independent of i and the observation-specific $\tilde{\alpha}_2$'s are identically distributed. Müller & Roeder (1995) use this idea for the case that $\mathbf{X}$ is partially observed. They assume that the $(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ are jointly normally distributed with mean μ_i and covariance matrix Σ_i, where $\theta_i = (\mu_i, \Sigma_i)$ is modeled by a Dirichlet process distribution which itself has unknown hyperparameters. Lindley and El Sayyad (1968) is the first Bayesian paper on functional models, covering the linear regression case. Because of their complexity, we do not consider Bayesian functional models here.

A second possibility intermediate between functional and structural approaches is to specify a flexible distributions, much as we suggested in sections 7.3.1 and 7.7. For instance, Mallick & Gelfand (1995) work with the likelihood (7.9), which is still a conditional likelihood even though $\mathbf{X}$ is unobserved. When $\mathbf{X}$ is scalar, they construct a model assuming that $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$ follows a generalized linear model with $\tilde{\alpha}_2 = (\alpha_{2,0}, \alpha_{2,1}, \alpha_{2,2})$, mean function $g^{-1}(\alpha_{2,0} + \alpha_{2,1}^t \mathbf{Z} + \alpha_{2,2}^t \mathbf{W})$ and a scale parameter σ^2, where $g(\cdot)$ is a monotone function. If $g(\cdot)$ is fully specified, this would be a standard structural modeling situation. Their compromise between the structural and functional models is to let $g(\cdot)$ be of a flexible form, namely a mixture of beta distribution functions with unknown parameters.

In this chapter, the $\mathbf{Z}_i$'s are treated as fixed constants, as before. This makes perfect sense, since Bayesians only treat unknowns as random variables. Thus, the likelihood is the conditional density of the $\mathbf{Y}_i$'s, $\mathbf{W}_i$'s, and any $\mathbf{X}_i$'s that are observed, given the parameters and the $\mathbf{Z}_i$'s. The posterior is the conditional density of

the parameters given all data, i.e., the $\mathbf{Z}_i$'s, $\mathbf{Y}_i$'s, $\mathbf{W}_i$'s, and any observed $\mathbf{X}_i$'s.

8.2 The Gibbs Sampler

8.2.1 Direct Sampling without Measurement Error

The Gibbs sampler is most easily understood when there is no measurement error. In this case the likelihood is

$$\prod_{i=1}^{n} f_{Y|Z,X}(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}).$$

Letting $(\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}})$ refer to the ensemble of complete data, we can write this likelihood in Gibbs sampling notation as $\left[\tilde{\mathbf{Y}}\middle|\tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \mathcal{B}\right]$. If $[\mathcal{B}]$ denotes a prior distribution for $\mathcal{B}$, then the density of $(\tilde{\mathbf{Y}}, \mathcal{B})$ given $(\tilde{\mathbf{Z}}, \tilde{\mathbf{X}})$ is

$$\left[\tilde{\mathbf{Y}}\middle|\tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \mathcal{B}\right][\mathcal{B}].$$

The posterior distribution of $\mathcal{B}$ is then

$$\left[\mathcal{B}\middle|\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}\right] = \frac{\left[\tilde{\mathbf{Y}}\middle|\tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \mathcal{B}\right][\mathcal{B}]}{\int \left[\tilde{\mathbf{Y}}\middle|\tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, v\right][v]dv}, \tag{8.1}$$

The practical problem is that the denominator of (8.1) may be very difficult to compute. Numerical integration typically fails to provide an adequate approximation even when there are as few as three or four components to $\mathcal{B}$.

The Gibbs sampler is one solution to the dilemma, although other methods are possible. The Gibbs sampler is an iterative, Monte-Carlo method consisting of two main steps:

- Form a sequence of computer-generated observations $\mathcal{B}_1, \mathcal{B}_2, ...$ from the posterior distribution of $[\mathcal{B}|\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}]$;
- Quantities such as the posterior mean are estimated by the sample mean of $\mathcal{B}_1, \mathcal{B}_2, ...$, while kernel density estimates are used to approximate the entire posterior density or the marginal posterior density of a single parameter or subset of parameters.

Here is how the iteration works. Start the iteration at any value of $\mathcal{B}$, and suppose that the current value in the iteration is $\mathcal{B} = (\beta_0, \beta_1, ..., \beta_M)$ In the Gibbs sampler, one generates an updated random variable β_0^* from its conditional posterior distribution given $(\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \beta_1, ..., \beta_M)$; see subsection 8.2.2. Then one generates an updated random variable β_1^* from its posterior distribution given $(\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \beta_0^*, \beta_2, ..., \beta_M)$. Continue until β_M^* has been generated. One then repeats this process a "large" number of times, see subsection 8.2.2. After discarding the first "few" observations in order to eliminate the influence of the starting value, one is left with a sequence of observations $(\mathcal{B}_1, \mathcal{B}_2, ...,)$ from the posterior distribution.

The mechanics of each step work as follows. The posterior distribution of β_j given $(\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}})$ and the other components of $\mathcal{B}$ is

$$\left[\beta_j \middle| \tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \beta_k \text{ for } k \neq j\right] = \frac{\left[\tilde{\mathbf{Y}} \middle| \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \beta_0, ..., \beta_M\right] [\beta_0, ..., \beta_M]}{\int \left[\tilde{\mathbf{Y}} \middle| \tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, g_j(u, \mathcal{B})\right] [g_j(u, \mathcal{B})] du}, \tag{8.2}$$

where $$g_j(u, \mathcal{B}) = (\beta_0, ..., \beta_{j-1}, u, \beta_{j+1}, ...\beta_M).$$

Generating pseudo random observations from (8.2) is the first step of the Gibbs sampler. Smith & Gelfand (1992) discuss the rejection method and the weighted bootstrap method. Ritter & Tanner (1992) and references therein discuss ways of drawing samples from (8.2), including the griddy Gibbs sampler, which effectively discretizes the components of $\mathcal{B}$ in a clever way; this can be useful since sampling from a multinomial distribution is trivial.

The mechanics of stopping the Gibbs sampler, and whether one should use one long sequence as described here or a number of shorter sequences, are currently a matter of controversy and is not discussed here; however, we note that Gelman & Rubin (1992) and Geyer (1992) give exactly opposite recommendations.

8.2.2 The Weighted Bootstrap

In our nondiscrete example we use the weighted bootstrap, and for the sake of completeness we now provide an explanation of this method. Suppose we want to sample from a density $f(\theta)$ that is

represented as the ratio $h(\theta)/\int h(v)dv$. Examination of (8.1) and (8.2) indicates that this is the relevant problem to consider. Let $g(\theta)$ be another distribution from which it is easy to generate data, and let $\theta_1, ..., \theta_N$ be a computer-generated sample from g. Now calculate $\omega_i = h(\theta_i)/g(\theta_i)$ and then $q_i = \omega_i / \sum_{j=1}^{N} \omega_j$. Draw θ^* from the discrete distribution that has probability q_i at θ_i. As N becomes large, the distribution of θ^* approaches f. This method is easy to program when, as is typical in practice, h is easy to compute. The closer g resembles the shape of h, the smaller the value of N that is needed. The weighted bootstrap is also called "sampling/importance resampling" and is sometimes given the same acronym, SIR, as the totally unrelated method of "sliced inversion regression" discussed in Chapter 10.

8.2.3 Forming Complete Data

Now suppose that $\mathbf{X}$ cannot be observed, and that instead the surrogate $\mathbf{W}$ has an independent replicate $\mathbf{T}$ at every observation. The likelihood for an individual observation is given by (7.5). If this likelihood is computable or very nearly so as in logistic regression (section 7.9.2), then the Gibbs sampler can be implemented as described above. The ensemble of data is $(\tilde{\mathbf{Y}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{W}}, \tilde{\mathbf{T}})$ and the parameters are $[\mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2]$, and with these substitutions the same idea as in section 8.2.1 applies.

However, as we know, computing (7.9) (either analytically or exactly) can sometimes be difficult, and in this case a missing-data technique may be helpful. The device is to treat the unobserved $\mathbf{X}$'s as unobserved random effects (parameters) with distribution $[\mathbf{X}|\mathbf{Z}, \tilde{\alpha}_2]$ where $\tilde{\alpha}_2$ has a prior distribution $[\tilde{\alpha}_2]$. Treating the $\mathbf{X}$'s just like any other unobserved parameter, the joint density of $(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{W}}, \tilde{\mathbf{T}}, \mathcal{B}, \tilde{\alpha}_2, \tilde{\alpha}_1)$ given $\tilde{\mathbf{Z}}$ becomes

$$[\tilde{\mathbf{Y}}|\tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \mathcal{B}][\tilde{\mathbf{W}}, \tilde{\mathbf{T}}|\tilde{\mathbf{Z}}, \tilde{\mathbf{X}}, \tilde{\alpha}_1][\tilde{\mathbf{X}}|\tilde{\mathbf{Z}}, \tilde{\alpha}_2][\mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2]. \tag{8.3}$$

Note that (8.3) is computable. One applies the Gibbs sampler to all the unknown parameters, namely $\mathcal{B}$, $\tilde{\alpha}_1$, $\tilde{\alpha}_2$ and the $\mathbf{X}$'s. The major burden lies in generating samples from the posterior distribution of the unobserved $\mathbf{X}$'s. If the sample size is large, then this might require much computer time.

8.3 Importance Sampling

Although high-dimensional integrals are difficult to compute accurately by quadrature and other deterministic numerical methods, they can often be computed quite accurately by Monte Carlo simulation. In fact, a Monte Carlo study of a statistical method on samples of size n amounts to estimating n-dimensional integrals. Importance sampling is a widely applicable method for numerical integration by simulation. It can, for example, be used to find posterior moments and to estimate posterior densities.

Let $f(\mathcal{B})$ be the product of the prior and the likelihood, e.g., in the case of no measurement error, the numerator of (8.1). Although $f(\mathcal{B})$ depends on the data, this dependence will be suppressed in the notation since the data are fixed. Let $h(\mathcal{B})$ be any function of the parameter, and suppose that we need to find the expected value of $h(\mathcal{B})$ with respect to the posterior distribution. This quantity is

$$\frac{\int h(v)f(v)\,dv}{\int f(v)\,dv}. \tag{8.4}$$

Importance sampling allows us to estimate both the numerator and the denominator of (8.4) with one simulation. Let g be a density somewhat similar to the posterior, but easy to sample from, e.g., a normal density. We discuss the choice of g below. Let $V_1, \ldots, V_N$ be an iid Monte Carlo sample from g. Define $w_i = f(V_i)/g(V_i)$, $i = 1, \ldots, N$. The w_i's are called the importance sampling weights. For models considered in this book, the prior and the likelihood and, hence, f are easy to evaluate; it is only $\int f(v)\,dv$ that is difficult to determine. Thus, computing the V_i's and the w_i's is not a problem. Then, $E(\mathcal{B}|\text{the data})$ is estimated by a weighted average of the $h(V_j)$'s using the importance sampling weights:

$$\frac{N^{-1}\sum_{j=1}^{N} h(V_j)w_j}{N^{-1}\sum_{j=1}^{N} w_j}. \tag{8.5}$$

The numerator and denominator of (8.5) estimate the corresponding quantities is (8.4). To see this, note that

$$E\{h(V_j)w_j\} = \int h(v)\frac{f(v)}{g(v)}g(v)dv = \int h(v)f(v)dv, \tag{8.6}$$

which shows the correspondence both between the numerators and (using $h(v) \equiv 1$) between the denominators.

Let $w_j^* = w_j / \sum_{k=1}^N w_k$. Then the vector of posterior means of $\mathcal{B}$ is estimated by

$$\widehat{E}(\mathcal{B}) = \sum_{j=1}^N w_j^* V_j, \tag{8.7}$$

and the posterior variance-covariance matrix is estimated by

$$\sum_{j=1}^N w_j^* \Big(V_j - \widehat{E}(\mathcal{B})\Big)\Big(V_j - \widehat{E}(\mathcal{B})\Big)^t. \tag{8.8}$$

One can estimate posterior densities by weighted kernel estimators, using the importance sampling weights. Let $V_{j,k}$ and $\mathcal{B}_k$ be the kth components of V_j and $\mathcal{B}$, respectively. Then, the posterior density of $\mathcal{B}_k$ evaluated at B_k is estimated by

$$\sum_{j=1}^N \frac{w_j^*}{b} K\Big\{\frac{V_{j,k} - B_k}{b}\Big\}, \tag{8.9}$$

where K is a kernel and b is a bandwidth. As in ordinary kernel density estimation (see Silverman (1986)), a "kernel" can be any function that integrates to 1, and typically K is chosen to be a symmetric probability density function.

Now we address the choice of g. Importance sampling gives unbiased estimates of posterior expectations, provided that the support of f is a subset of the support of g, so g should be positive on the entire parameter space. This positivity can be achieved by reparameterizing so that all components of $\mathcal{B}$ range from $-\infty$ to ∞, e.g., logging variances, and then letting g be Gaussian, or at least have a Gaussian component. The accuracy of the importance sampling increases as the variance of the importance sampling weights decreases. Therefore, we want g to be close to f, and it is especially important that g have tails as least as heavy as f. In our applications, we approximate f by a Gaussian density with mean equal to the MLE and variance-covariance matrix equal to $\widehat{I}_n^{-1}$, where $\widehat{I}_n$ is the observed Fisher information matrix defined in section A.2.2. This approximation of f is used *only* for guidance in choosing g. To ensure sufficiently heavy tails, we let g be the mixture of this density and the Gaussian density with the same mean but with variance-covariance matrix equal to $\sigma_*^2 \mathcal{I}^{-1}$ where $\sigma_* > 1$, with mixing probabilities of $(1 - \alpha)$ and α, respectively. In the Fram-

ingham study, we found after some experimenting that $\alpha = .1$ and $\sigma_* = 2$ roughly minimized the coefficient of variation of the w_j^*'s.

The use of importance sampling for Bayesian inference is discussed in more detail by Geweke (1989).

8.4 Cervical Cancer

The cervical cancer data are listed in Carroll, Gail, and Lubin (1993). The response $\mathbf{Y}$ is the indicator of invasive cervical cancer, $\mathbf{X}$ is exposure to herpes simplex virus, type 2 (HSV-2) measured by a refined western blot procedure, and $\mathbf{W}$ is exposure to HSV-2 measured by the western blot procedure. See Hildesheim et al. (1991) for biological background to this problem. There are 115 complete observations where $(\mathbf{Y}, \mathbf{X}, \mathbf{W})$ is observed and 1929 incomplete observations where only $(\mathbf{Y}, \mathbf{W})$ is observed. There are 39 cases $(\mathbf{Y} = 1)$ among the complete data and 693 cases among the incomplete data. Among the complete data, there is substantial misclassification, i.e., observations where $\mathbf{X} \neq \mathbf{W}$. Also, there is evidence of differential error.

We now describe a Bayesian analysis of the cervical cancer data using the Gibbs sampler with the strategy of filling in missing data.

In this example, $(\mathbf{W}, \mathbf{Y}, \mathbf{X})$ are all binary, there is no variable $\mathbf{Z}$, and the prospective model is

$$\Pr(\mathbf{Y} = 1|\mathbf{X}) = H(\beta_0^* + \beta_x \mathbf{X}). \tag{8.10}$$

This problem is particularly easy to parameterize retrospectively in terms of the distributions of $\mathbf{X}$ given $\mathbf{Y}$, and $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Y})$, and we show how to implement the Gibbs sampler here.

With differential measurement error, the six free parameters are $\alpha_{xd} = \Pr(\mathbf{W} = 1|\mathbf{X} = x, \mathbf{Y} = d)$ and $\gamma_d = \Pr(\mathbf{X} = 1|\mathbf{Y} = d)$, $x = 0, 1$ and $d = 0, 1$. We use beta priors with parameters (a_{xd}, b_{xd}) for the α's and (a_d^*, b_d^*) for the γ's, with the α's and γ's being mutually independent. If we impose the constraints, $\alpha_{x0} = \alpha_{x1}$ for $x = 0, 1$, then we have a four-parameter, nondifferential measurement error model. Following the usual odds-ratio formulation, the logistic slope is related to the γ's by

$$\beta_x = \log\left[\{\gamma_1/(1-\gamma_1)\} / \{\gamma_0/(1-\gamma_0)\}\right].$$

Thus, the posterior distribution of β_x can be found via transformation from the posterior distribution of the γ's.

If we could observe all the $\mathbf{X}$'s, the joint density of the parameters and all the data would be proportional to

$$\prod_{x=0}^{1}\prod_{d=0}^{1}\left[\alpha_{xd}^{a_{xd}-1}\left(1-\alpha_{xd}\right)^{b_{xd}-1}\right. \tag{8.11}$$
$$\left.\times\prod_{i=1}^{n}\left\{\alpha_{xd}^{\mathbf{W}_i}\left(1-\alpha_{xd}\right)^{1-\mathbf{W}_i}\right\}^{I(\mathbf{X}_i=x,\mathbf{Y}_i=d)}\right]$$
$$\times\prod_{d=0}^{1}\left[\gamma_d^{a_d^*-1}\left(1-\gamma_d\right)^{b_d^*-1}\prod_{i=1}^{n}\left\{\gamma_d^{\mathbf{X}_i}\left(1-\gamma_d\right)^{1-\mathbf{X}_i}\right\}^{I(\mathbf{Y}_i=d)}\right].$$

We can use (8.2) and (8.11) to note that the posterior distribution of γ_d is a beta distribution with parameters $\sum_{i=1}^{n}\mathbf{X}_iI(\mathbf{Y}_i=d)+a_d^*$ and $\sum_{i=1}^{n}(1-\mathbf{X}_i)I(\mathbf{Y}_i=d)+b_d^*$. The posterior distribution of α_{xd} is also a beta distribution but with parameters $\sum_{i=1}^{n}\mathbf{W}_iI(\mathbf{X}_i=x,\mathbf{Y}_i=d)+a_{xd}$ and $\sum_{i=1}^{n}(1-\mathbf{W}_i)I(\mathbf{X}_i=x,\mathbf{Y}_i=d)+b_{xd}$. The conditional distribution of a missing $\mathbf{X}_i$, given the $(\mathbf{W}_i,\mathbf{Y}_i)$ and the parameters, is Bernoulli with success probability $p_{1i}/(p_{0i}+p_{1i})$, where

$$p_{xi}=\gamma_{\mathbf{Y}_i}^{x}\left(1-\gamma_{\mathbf{Y}_i}\right)^{1-x}\alpha_{x\mathbf{Y}_i}^{\mathbf{W}_i}\left(1-\alpha_{x\mathbf{Y}_i}\right)^{1-\mathbf{W}_i}.$$

Thus, in order to implement the Gibbs sampler, we need to simulate observations from the Bernoulli and beta distributions, both of which are easy to do using standard programs, so the weighted bootstrap was not needed.

For nondifferential measurement error, the only difference in these calculations is that $\alpha_{x0}=\alpha_{x1}=\alpha_x$, which have a beta prior with parameters (a_x,b_x) and a beta posterior with parameters $\sum_{i=1}^{n}\mathbf{W}_iI(\mathbf{X}_i=x)+a_x$ and $\sum_{i=1}^{n}(1-\mathbf{W}_i)I(\mathbf{X}_i=x)+b_d$.

Using the retrospective formulation of section 14.1, maximum likelihood analysis yielded $\widehat{\beta}_1=.609$ (std. error $=.350$), and, under the nondifferential error model $\widehat{\beta}_1=.958$ (std. error $=.237$).

In Figure 8.1 we plot kernel density estimates of the posterior distribution of β_1 for both differential and nondifferential measurement errors. We used uniform priors throughout, so that $a_{xd}=b_{xd}=a_d^*=b_d^*=1$. We ran the Gibbs sampling with an initial burn-in period of 2,000 simulations, and then recorded every 50th simulation thereafter. The posterior modes were 0.623 and 0.927, respectively, these being very close to the maximum likelihood es-

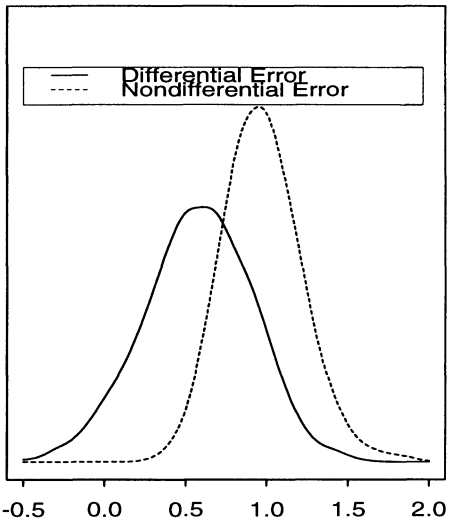

Figure 8.1. *Kernel posterior density estimates of β_1 for differential (solid) and nondifferential (dashed) measurement error in the cervical cancer data.*

timates. Note the large difference between the estimates for $d = 1$ and for $d = 0$, indicating the critical nature of assuming whether the error is differential or not.

In Figure 8.2 we plot kernel density estimates of the posterior of $\alpha_{0d} = \Pr(\mathbf{W} = 1|\mathbf{X} = x, \mathbf{Y} = d)$ for $d = 1$ and $d = 0$ with differential measurement error, the upper plot for $x = 0$ and the lower for $x = 1$. For each x, the posteriors for $d = 1$ and for $d = 0$ are clearly different, lending added strength to our earlier assertion that the assumption of nondifferential measurement error is problematic. We have found Figures 8.1–8.2 useful graphical diagnostics for detecting differential measurement error in the $2 \times 2 \times 2$ problem.

8.5 Framingham Data

The Framingham data also may be analyzed using the Gibbs sampler. We use the strategy here of approximately calculating the

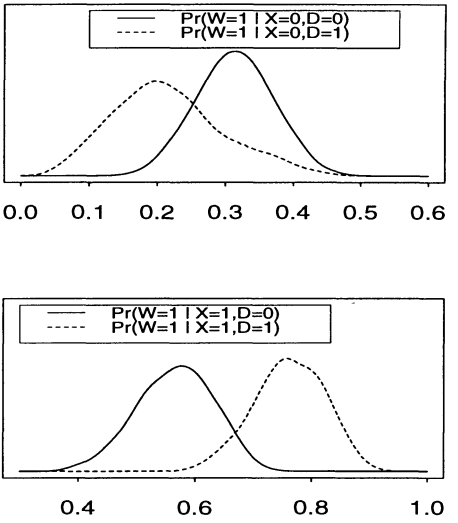

Figure 8.2. *Kernel posterior density estimates of* $Pr(\mathbf{W} = 1|\mathbf{X}, D = d)$ *for* $d = 1$ *(dashed) and* $d = 0$ *(solid) in the cervical cancer data with differential measurement error.*

likelihood, without filling in the missing $\mathbf{X}$'s. As an illustration, we consider only those males ages 45+ whose cholesterol values at Exam #3 ranged from 200 to 300, giving a data set of $n = 641$ observations. Recall that $\mathbf{Y}$ is the indicator of coronary heart disease. Initial frequentist analysis of this data set showed no evidence of age or cholesterol effects, so we work only with two covariates, smoking status ($\mathbf{Z}$) and $\mathbf{X} = \log(\text{SBP}-50)$, where SBP is long-term average systolic blood pressure. The main surrogate $\mathbf{W}$ is the measurement of $\log(\text{SBP}-50)$ at Exam #3, while the replicate $\mathbf{T}$ is $\log(\text{SBP}-50)$ measured at Exam #2. Given $(\mathbf{Z}, \mathbf{X})$, $\mathbf{W}$ and $\mathbf{T}$ are assumed independent and normally distributed with mean $\mathbf{X}$ and variance σ_u^2; $\sigma_u^2 = \tilde{\alpha}_1$ in the general notation of Chapter 7. The distribution of $\mathbf{X}$ given $\mathbf{Z}$ is assumed to be normal with mean $\alpha_{2,0} + \alpha_{2,1}\mathbf{Z}$ and variance $\sigma^2_{x|z}$ ($\tilde{\alpha}_2$ in the general notation). We also assume that $\sigma^2_{x|z}$ is constant, i.e., independent of $\mathbf{Z}$. Let

$\Theta = (\sigma_u^2, \alpha_{2,0}, \alpha_{2,1}, \sigma_{x|z}^2)$. Then the mean of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W}, \mathbf{T})$ is

$$m(\mathbf{Z}, \mathbf{W}, \mathbf{T}, \Theta) = \alpha_{2,0} + \alpha_{2,1}\mathbf{Z} + \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2/2}\left\{\frac{\mathbf{W}+\mathbf{T}}{2} - (\alpha_{2,0} + \alpha_{2,1}\mathbf{Z})\right\}. \tag{8.12}$$

Marginally, $(\mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ has a bivariate normal distribution with means $\alpha_{2,0} + \alpha_{2,1}\mathbf{Z}$, variances $\sigma_u^2 + \sigma_{x|z}^2$, and covariance $\sigma_{x|z}^2$. We use the regression calibration approximation (3.1) so that $\mathbf{Y}$ given $(\mathbf{W}, \mathbf{T})$ is treated as being logistic with mean

$$H\{\beta_0 + \beta_x m(\mathbf{Z}, \mathbf{W}, \mathbf{T}, \Theta) + \beta_z \mathbf{Z}\}.$$

Previous analysis suggested that the measurement error variance is less than 50% of the variance of the true long-term SBP given smoking status. It makes some sense to use this prior information, so we define $\lambda = \sigma_u^2/\sigma_{x|z}^2$ to be the ratio of these variances.

The unknown parameters are $(\beta_0, \beta_x, \beta_z, \alpha_{2,0}, \alpha_{2,1}, \alpha_{2,2} = \sigma_{x|z}^2, \lambda)$. The first five of these are given diffuse (noninformative) locally uniform priors, the next-to-last has a diffuse inverse Gamma prior, the density functions being proportional to $1/\sigma_{x|z}^2$, and λ has a uniform prior on the interval between zero and one half. Restricting the range here makes sense, and we would not credit an analysis that suggested that the measurement error variance is larger than the variance of true long-term SBP given smoking status.

Then, the joint density of the parameters and the observed data, conditional on $\tilde{\mathbf{Z}}$, is

$$\begin{aligned}
&[\tilde{\mathbf{Y}}|\tilde{\mathbf{Z}}, \tilde{\mathbf{W}}, \tilde{\mathbf{T}}, \mathcal{B}, \Theta][\tilde{\mathbf{W}}, \tilde{\mathbf{T}}|\tilde{\mathbf{Z}}, \Theta][\mathcal{B}, \Theta] \\
&= \frac{I(0 < \lambda < 1/2)}{\sigma_{x|z}^2} \prod_{i=1}^{n} \Big(f_2(\mathbf{W}_i, \mathbf{T}_i|\mathbf{Z}_i, \Theta) \\
&\quad \times H^{\mathbf{Y}_i}\{\beta_0 + \beta_x m(\mathbf{Z}_i, \mathbf{W}_i, \mathbf{T}_i, \Theta) + \beta_z \mathbf{Z}_i\} \\
&\quad \times [1 - H\{\beta_0 + \beta_x m(\mathbf{Z}_i, \mathbf{W}_i, \mathbf{T}_i, \Theta) + \beta_z \mathbf{Z}_i\}]^{1-\mathbf{Y}_i} \Big),
\end{aligned} \tag{8.13}$$

where $f_2(w, t|z, \Theta)$ is the bivariate normal density with common means $\alpha_{2,0} + \alpha_{2,1}\mathbf{Z}$, common variances $\sigma_u^2 + \sigma_{x|z}^2$, and covariance $\sigma_{x|z}^2$.

A frequentist analysis of these data yields the parameter estimates and bootstrap standard errors as given in Table 8.1.

In the Gibbs sampler, we used a weighted bootstrap (Smith & Gelfand, 1992) to generate observations from the univariate conditional posterior distributions given by (8.2). The variance parameters, σ_u^2 and $\sigma_{x|z}^2$, were log transformed to avoid positivity constraints. When generating observations by the weighted bootstrap, values of λ exceeding $1/2$ cannot occur, since they have prior probabilities of 0.

The weighted bootstrap was applied to each of the seven parameters in turn, which we call one cycle. The estimated posterior means and standard deviations from the Gibbs sampler are given in Table 8.1.

In Table 8.1, the frequentist estimates of β_0, β_x, and β_z and their bootstrap standard errors differ substantially from the posterior means and variances by the Gibbs sampler. We feel that this is due to inaccuracy of the Gibbs output; see subsection 8.5.1. Therefore, we recomputed the posterior means and variances by importance sampling, using $N = 10,000$, $\alpha = .1$, and $\sigma_* = 2$ in the notation of section 8.3. The results are in Table 8.1. In Figure 8.3, weighted kernel estimates using importance sampling of some posterior distributions are plotted. In this example, we found little difference between the frequentist and Bayes analyses using importance sampling, an exception being that the bootstrap standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_x$ are smaller than the posterior standard deviations of these parameters.

The maximum likelihood and importance sampling results in Table 8.1 are similar to the likelihood and regression calibration results given in section 7.6, and the differences are easily due to our use here of only 641 out of the 1,615 subjects analyzed in section 7.6.

8.5.1 Details of the Gibbs Sampler and Weighted Bootstrap

When implementing the Gibbs sampler on the Framingham data, we used three independent sequences through the seven parameters, each of 4000 cycles and each started at the MLE. We sampled every 10th cycle, for a total of 1200 observations of the seven-parameter posterior distribution. When sampling from (8.2) using the weighted bootstrap, g was the normal distribution with mean equal to the current value of β_j and standard deviation from the observed Fisher information matrix. Also, $N = 20$ θ's were gener-

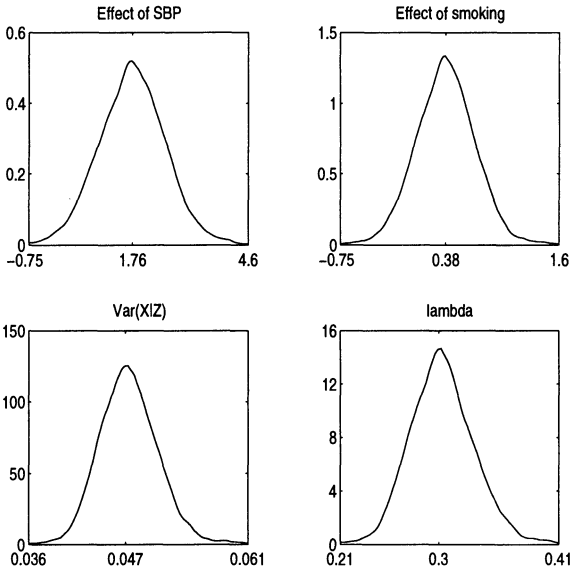

Figure 8.3. *Framingham data. The top left plot is the estimated posterior density of the effect of SBP, while the top right is for the effect of smoking. The bottom left is the posterior density of the variance of* $\mathbf{X}$ *given* $\mathbf{Z}$*, while the bottom right is the ratio of the measurement error variance to the variance of* $\mathbf{X}$ *given* $\mathbf{Z}$*. The bandwidth of each kernel density estimate was 0.4 times the sample standard deviation. The central tick mark on the x-axis is at the sample median.*

ated in the weighted bootstrap. The Gibbs sampler algorithm was implemented in MATLAB and the computations took about one day on a SPARC 20.

We used autocorrelations to check for dependence among the 1200 observations. For β_x, there was strong dependence, with autocorrelation coefficients of .96, .92, .89, .86, and .82 at lags of 1, 2, 3, 4, and 5, respectively. A lag of i in the sample corresponds, of course, to a lag of $10i$ in the original Markov chain. The parameter β_0 exhibited similar dependence. For β_z, $\alpha_{2,0}$, and $\alpha_{2,1}$, respectively, there were autocorrelations of .28, .25, and .25 at lag 1, but small autocorrelations at lags of 2 and higher. For $\sigma^2_{x|z}$ and λ, all autocorrelations were small, suggesting independence.

The strong dependence exhibited by β_x and β_0 means that their

Parameter	ML. est.	Boot. se	Posterior mean		Posterior std. dev.	
			IS	GS	IS	GS
β_0	−10.10	2.400	−10.10	−13.60	4.100	7.100
β_x	1.76	0.540	1.76	2.49	0.870	1.570
β_z	0.38	0.310	0.38	0.61	0.340	0.520
$\alpha_{2,\text{em}}$	4.42	0.019	4.43	4.42	0.021	0.019
$10 \times \alpha_{2,\text{em}}$	−0.19	0.210	−0.20	−0.20	0.240	0.222
$10 \times \sigma^2_{x\|z}$	0.47	0.033	0.47	0.47	0.036	0.031
$10 \times \sigma^2_u$	0.14	0.011	0.14	0.14	0.009	0.008
λ_{em}	0.30	0.031	0.30	0.30	0.032	0.028

Table 8.1. *Framingham data. The effects of SBP and smoking are given by β_x and β_z, respectively. The measurement error variance is σ^2_u. The mean of long-term SBP given smoking status is linear with intercept $\alpha_{2,0}$, slope $\alpha_{2,1}$ and variance $\sigma^2_{x|z}$. Also, $\lambda = \sigma^2_u/\sigma^2_{x|z}$. "ML" = maximum likelihood, "se" = standard error, "Boot." = bootstrap, "GS" = Gibbs sampling, and "IS" = importance sampling.*

posteriors are not accurately estimated by the Gibbs sampler with the amount of sampling we have used. For example, the sample means of the β_x's in the three sequences, 3.04, 2.25, and 2.96, vary among themselves far more than we would expect under independence within the sequences. Using the posterior standard deviation from Table 8.1, the standard error of each mean would be $1.09/\sqrt{400} = .055$, if the observations in the sample were independent.

Far more and far longer sequences might be contemplated, with sampling of only every 50th or perhaps every 100th cycle. However, that would require considerable computation, certainly on the order of a week on a SUN SPARC 20 with our MATLAB implementation. In comparison, the bootstrap took about two hours for very high accuracy (1000 *independent* replicates). The slowness of the Gibbs sampler is caused primarily by the weighted bootstrap, which requires 20 evaluations of the likelihood to obtain a single sample from (8.2). Thus, a single cycle takes 140 evaluations of the likelihood, so 1400 evaluations are needed to get a single observa-

tion when we sample every 10th cycle.

We note that the computing times reported above are specific to our implementation in MATLAB. Greater speed could likely be achieved with other languages, e.g., C or FORTRAN.

CHAPTER 9

SEMIPARAMETRIC METHODS

In Chapter 7 we described likelihood methods of inference for measurement error models. Especially in sections 7.1 and 7.2, we noted a formal relationship between measurement error models and missing data problems, and that when $\mathbf{X}$ is observed on a subset of the study participants, the measurement error problem is a missing data problem with supplementary information.

An important distinction was made in section 7.2 between functional modeling, which makes no assumptions about the distribution of $\mathbf{X}$, and structural modeling, in which this distribution is given a parametric form. While much of the missing data literature takes the form of structural modeling, there are important functional (semiparametric) techniques that have been developed recently. This chapter describes some of these functional techniques.

The focus of this chapter is on methods for problems with internal validation or replication data, in which no parametric assumptions are made about the calibration distribution. With the exception of the material in section 9.6, the techniques discussed in this section are relevant to the missing data problem, where $\mathbf{X}$ is observable in a subset of the study participants. Some of the methods have been developed only recently, especially those in section 9.5, and there is little in the way of studies documenting their performance in applications.

We focus on two-stage validation designs in which $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ are measured for all study participants at the first stage of the study. In the second stage, $\mathbf{X}$ is also observed for a subset of the study participants. These are the complete data, which we identify with the indicator variable Δ, i.e., $\Delta = 1$ if $\mathbf{X}$ is observed, otherwise

$\Delta = 0$. Admission into the second stage is assumed to depend only on the values of $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ observed at the first stage, and not on the value of $\mathbf{X}$ itself. In the terminology explained in Chapter 7, $\mathbf{X}$ is missing at random.

9.1 Using Only Complete Data

The simplest functional approach is to use standard methods, but analyze only the complete data, i.e., the data for which $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ are all observed. This is a functional analysis because no assumptions about the distribution of $\mathbf{X}$ are invoked. The main drawback is that an analysis of only the complete data disregards information in the incomplete observations.

A second problem is that naive use of the complete data can lead to biases in parameter estimates, under the circumstance that selection into the second stage depends on the response. This may seem contradictory, in that the data are missing at random, but a complete case analysis can be invalid. The reason for this is fairly technical, as indicated in (9.1) below. The reader should keep in mind, however, that the assumption that $\mathbf{X}$ is missing at random implies that we can ignore the pattern of missing data only for a full likelihood analysis (Chapter 7); using the complete data only is not a full likelihood analysis, and hence one needs to take into account the pattern of missing data.

There are two simple ways to correct this problem. Let $\mathbf{L} = (\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ denote the combined response and the observed covariates. Suppose that one selects a participant into the second stage with probability $\pi(\mathbf{L})$, where $\pi(\cdot)$ is a known function. One way to obtain consistent estimation is to perform a weighted, complete data analysis, with the weights inversely proportional to the selection probabilities, see Little & Rubin (1987, p. 55) and Zhao & Lipsitz (1992). This is the so-called Horvitz-Thompson approach from survey sampling. A second approach is to compute the actual density or mass function of the complete data and then maximize the complete data likelihood. The density or mass function of the complete data is

$$f_{\mathbf{Y}|\mathbf{Z},\mathbf{X},\mathbf{W},\Delta}(\mathbf{Y}|\mathbf{Z},\mathbf{X},\mathbf{W},\Delta = 1,\mathcal{B}) =$$

$$\frac{\pi(\mathbf{Y},\mathbf{Z},\mathbf{W})f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(\mathbf{Y}|\mathbf{Z},\mathbf{X},\mathcal{B})}{\int \pi(y,\mathbf{Z},\mathbf{W})f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|\mathbf{Z},\mathbf{X},\mathcal{B})d\mu(y).} \tag{9.1}$$

The complete data likelihood is the product over the *complete* data of the terms (9.1). This can be treated as any likelihood, and inference is standard, see Appendix A.

Equation (9.1) takes a very simple form in logistic regression, namely that of a logistic regression model but with an additional known "offset" term, $\log\{\pi(\mathbf{Y}=1,\mathbf{Z},\mathbf{W})/\pi(\mathbf{Y}=0,\mathbf{Z},\mathbf{W})\}$, added to the intercept. The likelihood can be maximized treating this offset as a new variable in the logistic regression whose parameter is constrained to equal 1.0.

In the usual context of measurement error models, selection into the second stage is under the control of the investigator. This need not always be the case, especially in classical missing data problems where the data are observational, and the selection mechanism can only be estimated. Because it is more technical, this material is discussed in the appendix, section 9.8.1. One of the more interesting features of this problem is that (asymptotically) it is better to estimate the selection probabilities even when they are known.

9.2 Special Two-Stage Designs for Binary Responses

While designing two-stage studies is reviewed briefly in Chapter 14, at this point is it useful to mention a particularly important class of two-stage designs. Recall that in a two-stage design, $(\mathbf{Y},\mathbf{Z},\mathbf{W})$ are observed for all study participants, and then $\mathbf{X}$ is observed for a subset of the study participants. When $\mathbf{Y}$ is binary (or more generally categorical), it can be particularly convenient to select an observation for admission into the second stage of the study (at which point $\mathbf{X}$ is observed) on the basis of the response and a categorical covariate, the latter usually being a function of $(\mathbf{W},\mathbf{Z})$. For purposes of this section, label this covariate as $\mathbf{Z}_*$, in which case selection into the second stage depends only upon $(\mathbf{Y},\mathbf{Z}_*)$. The convenience here is that the $(\mathbf{Y},\mathbf{Z}_*)$ data form a cross-classified table, from each cell of which one can select a predetermined or random number into the second stage of the study.

Such designs are particularly important in practice, although not a particularly important form of measurement error modeling per se. There is a large literature on analyzing such designs, both in

the context discussed here and also in case-control studies. Hsieh, Manski & McFadden (1985), Breslow & Cain (1988), Flanders & Greenland (1991), Scott & Wild (1991), Wild (1991), Zhao & Lipsitz (1992), and Breslow & Holubkov(1995) should be consulted for details of the analysis. Flanders & Greenland suggest weighting inversely with the selection probabilities. Zhao & Lipsitz, Scott & Wild, Wild, and Breslow & Holubkov discuss efficient estimation methods which are closely connected to the efficient methods outlined in section 9.5.

9.3 Pseudolikelihood

In functional modeling, we avoid parametric formulation of the distribution of $\mathbf{X}$. One way to do this is to use nonparametric techniques to estimate the distribution in question. We call such techniques *pseudolikelihood*, because they retain a likelihood and quasilikelihood flavor.

The key to these ideas is to remind oneself that in likelihood, and in quasilikelihood & variance function (QVF) models (Appendix A), the distribution of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$, and the moments of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$, can be written as regression functions. For example, from (7.9), the likelihood of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$ is just

$$f_{Y|Z,W}(y|z,w,\mathcal{B}) = E\left\{f_{Y|Z,X}(y|z,\mathbf{X},\mathcal{B})|\mathbf{Z}=z,\mathbf{W}=w\right\}. \quad (9.2)$$

The mean and variance functions in a QVF model are explicitly written as regressions as follows:

$$E(\mathbf{Y}|\mathbf{Z},\mathbf{W}) = E\left\{f(\mathbf{Z},\mathbf{X},\mathcal{B})|\mathbf{Z},\mathbf{W}\right\}. \quad (9.3)$$

$$\begin{aligned}\mathrm{var}(\mathbf{Y}|\mathbf{Z},\mathbf{W}) &= \sigma^2 E\left\{g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta)|\mathbf{Z},\mathbf{W}\right\} \\ &\quad + E\left\{f^2(\mathbf{Z},\mathbf{X},\mathcal{B})|\mathbf{Z},\mathbf{W}\right\} \\ &\quad - \left[E\left\{f(\mathbf{Z},\mathbf{X},\mathcal{B})|\mathbf{Z},\mathbf{W}\right\}\right]^2. \end{aligned} \quad (9.4)$$

The pseudolikelihood algorithm estimates the quantities in equations (9.2)–(9.4) nonparametrically, but otherwise employs the standard estimation scheme, i.e., maximizing likelihoods or solving QVF estimating equations. It applies as long as selection into the second stage of the study depends only on $(\mathbf{Z}, \mathbf{W})$ but not on the response (a special case of missing at random).

For example, suppose that we can estimate the loglikelihood, i.e., the logarithm of (9.2) as a function of $(y, z, w, \mathcal{B})$, by $\widehat{\ell}(y, z, w, \mathcal{B})$.

Then, following (7.10), the pseudo-maximum likelihood estimator of $\mathcal{B}$ maximizes

$$\sum_{i=1}^{n} \Big[\Delta_i \log \left\{ f_{Y|Z,X}(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) \right\} + (1 - \Delta_i) \widehat{\ell}(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \mathcal{B}) \Big].$$

In effect, we substitute an estimated likelihood for the density of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$, and then proceed as if this were the actual likelihood. Similarly, in QVF models, pseudolikelihood replaces the moment functions for $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{W})$ by their estimated values. Pseudolikelihood was introduced independently by Pepe & Fleming (1991) when $(\mathbf{Z}, \mathbf{W})$ are discrete, and by Carroll & Wand (1991) when $(\mathbf{Z}, \mathbf{W})$ are continuous.

This method requires estimation of functions like those in (9.2)–(9.4), i.e., functions like

$$\mathcal{H}(y, z, w, \mathcal{B}) = E \left\{ \mathcal{G}(y, z, \mathbf{X}, \mathcal{B}) | \mathbf{Z} = z, \mathbf{W} = w \right\}.$$

When $(\mathbf{Z}, \mathbf{W})$ are discrete, the obvious estimate is

$$\widehat{\mathcal{H}}(y, z, w, \mathcal{B}) = \frac{\sum_{i=1}^{n} \Delta_i I(\mathbf{Z}_i = z, \mathbf{W}_i = w) \mathcal{G}(y, z, \mathbf{X}_i, \mathcal{B})}{\sum_{i=1}^{n} \Delta_i I(\mathbf{Z}_i = z, \mathbf{W}_i = w)}.$$

When $(\mathbf{Z}, \mathbf{W})$ is not discrete, the function $\mathcal{H}$ can be estimated by nonparametric regression techniques, regressing the function $\mathcal{G}(y, z, \mathbf{X}, \mathcal{B})$ on $(\mathbf{Z}, \mathbf{W})$. There are several ways to do this regression, ranging from kernel methods to generalized additive models (Hastie & Tibshirani, 1990). The major practical difficulty in implementation is the well-known curse of dimensionality in high-dimensional nonparametric regression. Generalized additive models address this problem directly. Carroll, Knickerbocker & Wang (1995) perform a direct dimension reduction of $\mathbf{X}$ predicted by $(\mathbf{Z}, \mathbf{W})$ using sliced inverse regression (Li, 1991; Duan & Li, 1991).

Standard error formulae are given by Pepe & Fleming for the discrete case, and by Sepanski & Carroll (1993) for QVF models. The bootstrap is asymptotically justified in the discrete case. We conjecture that it will give acceptable results for the other methods, although this has not been investigated.

There is one somewhat paradoxical point about pseudolikelihood. In theory it is possible for the pseudolikelihood method to yield *less* efficient estimates when compared to using the complete

data only (Pepe, 1992; Robins, Hsieh & Newey, 1995). This problem can be avoided by weighting the validation and nonvalidation terms in the estimating equation, or by computing the pseudolikelihood and complete data estimates, and taking a weighted average of the two. These modifications are seldom necessary in applications.

9.4 Mean Score Method

When selection into a validation study ($\Delta = 1$) depends on the response, pseudolikelihood no longer applies. The distributions of $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W})$, and $\mathbf{X}$ given $(\mathbf{Z}, \mathbf{W}, \Delta)$ are not the same, and hence naive use of pseudolikelihood leads to inconsistent estimates.

Reilly & Pepe (1994) describe a modified pseudolikelihood approach for the case that $(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ are all discrete, called the *mean score* method. No results are yet available in the continuous case. Suppose for the moment that we have a parametric calibration model and define the complete-data likelihood,

$$\ell(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}, \mathcal{B}, \gamma_{\text{cm}}) = f_{Y|Z,X}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \mathcal{B}) f_{X|Z,W}(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \gamma_{\text{cm}}).$$

The EM algorithm (Little & Rubin, 1987, p. 130) for this parametric problem involves the iterative maximization of

$$\sum_{i=1}^{n} \left(\Delta_i \log \{\ell(\mathbf{L}_i, \mathbf{X}_i, , \mathcal{B}, \gamma_{\text{cm}})\} + (1 - \Delta_i) E \left[\log \{\ell(\mathbf{L}_i, \mathbf{X}, \mathcal{B}, \gamma_{\text{cm}})\} \,\middle|\, \mathbf{L}_i, \mathcal{B}_*, \gamma_* \right] \right),$$

where $(\mathcal{B}_*, \gamma_*)$ are the current values in the iteration. If we make no assumptions about the distribution of $\mathbf{X}$ given the other covariates, then in order to implement the EM algorithm we need the expectation of the loglikelihood given the observed incomplete data $\mathbf{L}_i = (\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i)$. This expectation can be estimated using pseudolikelihood techniques, leading to the iterative maximization of

$$\sum_{i=1}^{n} \left[\Delta_i \log \{ f_{Y|Z,X}(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) \}

$$+ (1 - \Delta_i) \frac{\sum_{j=1}^{n} \Delta_j I(\mathbf{L}_j = \mathbf{L}_i) \log \{ f_{Y|Z,X}(\mathbf{Y}_i | \mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) \}}{\sum_{j=1}^{n} \Delta_j I(\mathbf{L}_j = \mathbf{L}_i)} \Bigg].$$

The resulting estimate is asymptotically normally distributed, and its covariance may be estimated by $n^{-1}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$, where $\widehat{A}$ and $\widehat{B}$ are defined as follows. First let the derivative of the loglikelihood of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ be $\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathcal{B})$. Then define

$$n(\mathbf{L}) = \sum_{i=1}^{n} I(\mathbf{L}_i = \mathbf{L}); \quad n_v(\mathbf{L}) = \sum_{i=1}^{n} \Delta_i I(\mathbf{L}_i = \mathbf{L});$$

$$U(\mathbf{L}) = \sum_{i=1}^{n} \frac{\Delta_i}{n_v} I(\mathbf{L}_i = \mathbf{L}) \psi(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) \psi^t(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \mathcal{B});$$

$$\widehat{A} = n^{-1} \sum_{i=1}^{n} U(\mathbf{L}_i); \quad n_p(\mathbf{L}) = \sum_{i=1}^{n} (1 - \Delta_i) I(\mathbf{L}_i = \mathbf{L});$$

$$V(\mathbf{L}) = n_v^{-1} \sum_{i=1}^{n} \Delta_i I(\mathbf{L}_i = \mathbf{L}) \psi^t(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \mathcal{B});$$

$$\widehat{B} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{n^2(\mathbf{L}_i) U(\mathbf{L}_i)}{n_v(\mathbf{L}_i)} - \frac{n(\mathbf{L}_i) n_p(\mathbf{L}_i)}{n_v(\mathbf{L}_i)} V(\mathbf{L}_i) V^t(\mathbf{L}_i) \right\}.$$

9.5 General Unbiased Estimating Functions

Robins, et al. (1995) describe yet another method of functional estimation. Their method differs from pseudolikelihood and its variants as it is not based on nonparametric regression.

The material discussed in this section has potential importance in practice, but it is complex and requires a good understanding of unbiased estimating equations (Appendix A). The intent of the material is to improve upon complete data analyses without making any assumptions about the joint distribution of $(\mathbf{W}, \mathbf{X})$ given $\mathbf{Z}$. The methods are thus intermediate in efficiency between likelihood and complete data analyses.

Remember that $\Delta = 1$ means that $\mathbf{X}$ has been observed. Let $\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathcal{B})$ be any unbiased estimating function (section A.3) for $\mathcal{B}$ in an ordinary study with $\mathbf{X}$ observed everywhere, e.g., a likelihood score or quasilikelihood and variance function (QVF) estimating function. Let $\mathbf{L} = (\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ and let $\Omega(\mathbf{L}) = \Omega(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$

be any function of the same dimension as $\psi(\cdot)$. From Robins, et al. (1995), the following are unbiased estimating functions for $\mathcal{B}$ given $(\mathbf{Z}, \mathbf{W})$:

$$\Psi_1(\mathbf{L}, \mathbf{X}, \Delta, \mathcal{B}) = \Delta \left[\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathcal{B}) - \frac{E\{\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathcal{B})\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\}}{E\{\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\}} \right]; \tag{9.5}$$

$$\begin{aligned} \Psi_2(\mathbf{L}, \mathbf{X}, \Delta, \mathcal{B}) &= \Delta r(\mathbf{L}) - \Delta \frac{E\{\pi(\mathbf{L})r(\mathbf{L})|\mathbf{X}, \mathbf{Z}, \mathbf{W}\}}{E\{\pi(\mathbf{L})|\mathbf{X}, \mathbf{Z}, \mathbf{W}\}} \\ &\quad - \frac{\Delta - \pi(\mathbf{L})}{1 - \pi(\mathbf{L})} r(\mathbf{L}), \end{aligned} \tag{9.6}$$

where $r(\mathbf{L}) = \{1 - \pi(\mathbf{L})\}\,\Omega(\mathbf{L})/\pi(\mathbf{L})$. The simplest choice for Ω is just the naive score $\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathcal{B})$. That (9.5)–(9.6) are unbiased estimating functions is shown in the appendix, as are the other theoretical claims of this section.

It is not obvious that these estimating functions depend on all the data, so some explanation is required. The function Ψ_1 uses only the validation data. The first two terms in Ψ_2 also use only the validation data. However, since $\mathbf{L} = (\mathbf{Y}, \mathbf{Z}, \mathbf{W})$, the third term in Ψ_2 uses all the data.

It is also not obvious which terms depend on $\mathcal{B}$. In Ψ_2, only the second does, because for any function $g(\cdot)$,

$$\begin{aligned} &E\{g(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})|\mathbf{X}, \mathbf{Z}, \mathbf{W}\} \\ &\quad = \int g(y, \mathbf{Z}, \mathbf{X}, \mathbf{W}) f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|\mathbf{Z}, \mathbf{X}, \mathcal{B}) d\mu(y). \end{aligned}$$

For any given function $\Omega(\cdot)$, an unbiased estimating equation is

$$0 = \sum_{i=1}^{n} \sum_{j=1}^{2} \Psi_j(\mathbf{L}_i, \mathbf{X}_i, \Delta_i, \mathcal{B}). \tag{9.7}$$

Because (9.7) is an unbiased estimating equation, the asymptotic theory of section A.3 applies. Because $\Psi_1(\cdot)$ and $\Psi_2(\cdot)$ are uncorrelated, the covariance matrix of $\widehat{\mathcal{B}}$ can be estimated by

$$\frac{1}{n}\left\{\sum_{j=1}^{2} A_{nj}(\widehat{\mathcal{B}})\right\}^{-1}\left\{\sum_{j=1}^{2} B_{nj}(\widehat{\mathcal{B}})\right\}\left[\left\{\sum_{j=1}^{2} A_{nj}(\widehat{\mathcal{B}})\right\}^{-1}\right]^{t},$$

where

$$A_{nj}(\mathcal{B}) = n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial \mathcal{B}^t}\Psi_j(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \mathbf{X}_i, \Delta_i, \mathcal{B});$$

$$B_{nj}(\mathcal{B}) = \widehat{\text{cov}}\left\{\Psi_j(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \mathbf{X}_i, \Delta_i, \mathcal{B})\right\}_{i=1}^{n}. \tag{9.8}$$

In (9.8), $\widehat{\text{cov}}$ is the sample covariance matrix of the indicated terms.

The key question is how to choose $\Omega(\cdot)$. We address this issue briefly below.

9.5.1 Using Polynomials

Let $\xi(\mathbf{Y}, \mathbf{Z}, \mathbf{W}) = \xi(\mathbf{L})$ be a vector of size k whose elements include the arguments plus polynomial functions of them. In principle, this could be anything, but we use $\mathbf{L}$ itself for simplicity. Let $\Psi_{2*}(\cdot)$ be defined exactly as (9.6) but using $\xi(\cdot)$ instead of $\Omega(\cdot)$. Define

$$\mathcal{C} = E\left\{(\partial/\partial\mathcal{B})\Psi_{2*}^t(\mathbf{L})\right\}\left[E\left\{\Psi_{2*}(\mathbf{L})\Psi_{2*}^t(\mathbf{L})\right\}\right]^{-1}; \tag{9.9}$$

$$\mathcal{D} = E\left\{(\partial/\partial\mathcal{B}^t)\Psi_1(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathcal{B})\right\}; \tag{9.10}$$

$$\mathcal{F} = \mathcal{D} - \mathcal{C}E\left\{(\partial/\partial\mathcal{B})\Psi_{2*}^t(\mathbf{L})\right\}^t.$$

We show the following in the appendix. With $\Psi_{2*}(\cdot)$ fixed, suppose we use $\Psi_2 = A\Psi_{2*}$ for some matrix A. Then the asymptotically efficient choice of A is $A = -\mathcal{C}$. The resulting estimate of $\mathcal{B}$ would be asymptotically normally distributed with mean $\mathcal{B}$ and covariance matrix

$$\frac{1}{n}\mathcal{F}^{-1}\left[E\Psi_1(\mathbf{L})\Psi_1^t(\mathbf{L}) + \mathcal{C}E\left\{(\partial/\partial\mathcal{B})\Psi_{2*}^t(\mathbf{L})\right\}^t\right]\left(\mathcal{F}^{-1}\right)^t. \tag{9.11}$$

Note that if $\Psi_1(\cdot)$ is a likelihood score, then $\mathcal{D} = -E\Psi_1\Psi_1^t$ and the asymptotic covariance is just $-n^{-1}\mathcal{F}^{-1}$.

The following steps are used to implement this algorithm.

(i) Use the complete data only to obtain an estimate $\widehat{\mathcal{B}}$ of $\mathcal{B}$.

(ii) Estimate $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{F}$ by replacing $\mathcal{B}$ by its estimate and replacing the expectations in (9.9)–(9.10) by averages over the

data.

(iii) Define $\Psi_2(\cdot) = -\widehat{\mathcal{C}}\Psi_{2*}(\cdot)$.

(iv) Reestimate $\mathcal{B}$ by solving (9.7) with Ψ_2 as in step (iii).

A consistent estimate of the asymptotic covariance matrix of $\widehat{\mathcal{B}}$ is

$$\widehat{\mathcal{F}}^{-1}\left[nB_{n1}(\widehat{\mathcal{B}}) + \widehat{\mathcal{C}}\sum_{i=1}^{n}\left\{(\partial/\partial\mathcal{B})\Psi_{2*}^{t}(\mathbf{L}_i)\right\}^{t}\right]\widehat{\mathcal{F}}^{-1},$$

see (9.8). For a likelihood score, one can also use $n^{-1}\widehat{\mathcal{F}}^{-1}$.

9.5.2 Optimal Moment-Based Estimators

Robins, et al. (1995) show that there is a globally optimal choice of $\Omega(\mathbf{L})$, defined as follows. Refer to (9.5) and define $\Psi_1(\cdot) = \Delta\Psi_{1*}(\cdot)$. Then the optimal choice is the solution $\Omega(\mathbf{L})$ to the functional equation

$$\begin{aligned}\Omega(\mathbf{L},\mathcal{B}) &= E\left\{\Psi_{1*}(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathbf{W},\mathcal{B})|\mathbf{L}\right\} \\ &\quad -E\left(\frac{E\left[\{1-\pi(\mathbf{L})\}\,\Omega(\mathbf{L},\mathcal{B})|\mathbf{X},\mathbf{Z},\mathbf{W}\right]}{E\left\{\pi(\mathbf{L})|\mathbf{X},\mathbf{Z},\mathbf{W}\right\}}\middle|\mathbf{L}\right).\end{aligned} \tag{9.12}$$

The proof of this result is a nice example of semiparametric theory, but the technical details are beyond the scope of this book. The practical hurdle is to find the function $\Omega(\cdot)$ that solves (9.12). Robins, et al. (1995) show how to do this when $\mathbf{L}$ is discrete, but effectively it remains an open problem otherwise. More work on this topic is needed.

9.5.3 Mean Based Moment-Based Estimators

Robins, Rotnitzky & Zhao (1994) addressed estimation of regression parameters when only the mean function is specified, although the methods apply also to the quasilikelihood and variance function (QVF) models described in Appendix A when the variance function parameters are known. The estimating function (A.23) for such a QVF model is of the form

$$\psi(\mathbf{Y},\mathbf{Z},\mathbf{X}) = h(\mathbf{Z},\mathbf{X},\mathcal{B})\left\{\mathbf{Y} - f(\mathbf{Z},\mathbf{X},\mathcal{B})\right\}.$$

A set of unbiased estimating functions indexed by function ϕ which use all the data are

$$\frac{\Delta\psi(\mathbf{Y},\mathbf{Z},\mathbf{X}) + \{\Delta - \pi(\mathbf{Y},\mathbf{Z},\mathbf{W})\}\,\phi(\mathbf{Y},\mathbf{Z},\mathbf{W})}{\pi(\mathbf{Y},\mathbf{Z},\mathbf{W})}.$$

The optimal choice for ϕ is $E\{\psi(\cdot)|\mathbf{Y},\mathbf{Z},\mathbf{W}\}$. This is obviously not a known function, and instead has to be estimated from the observed data. Zhao, Lipsitz & Lew (1994) describe methods for estimating ϕ. One method, also mentioned by Robins, et al., is to fit a flexible regression model for the regression of $\psi(\cdot)$ on $(\mathbf{Y},\mathbf{Z},\mathbf{W})$. Another is to hypothesize a distribution for $\mathbf{X}$ given $(\mathbf{Y},\mathbf{Z},\mathbf{W})$. In both cases, the resulting estimates of $\mathcal{B}$ are consistent and asymptotically normal, and we may proceed as if $\phi(\cdot)$ were known. This is a promising approach which deserves further investigation.

9.6 Semiparametric Regression Calibration

To this point, the techniques of this section have assumed that $\mathbf{X}$ is observable in a subset of the study design, a situation that is not always possible. We have already described three general methods for handling this problem, regression calibration (Chapter 3), SIMEX (Chapter 4), and likelihood (Chapter 7).

For regression calibration, the basic idea is to replace $\mathbf{X}$ by an estimate of $m(\mathbf{Z},\mathbf{W}) = E(\mathbf{X}|\mathbf{Z},\mathbf{W})$ and then proceed as if the approximation were exact. Parametric linear and quadratic regression methods for estimating the calibration function were described in section 3.4.

Instead of estimating the calibration function using parametric models, one can use nonparametric regression (Sepanski, et al. 1994, Carroll, Knickerbocker & Wang, 1995). The techniques to be used depend on the available data.

When there is an unbiased instrument $\mathbf{T}$ for $\mathbf{X}$ measured either externally or internally in a subset of the data, one can regress $\mathbf{T}$ on $(\mathbf{Z},\mathbf{W})$ nonparametrically. Remembering, however, that nonparametric regression with multivariate predictors is difficult, we again suggest the use of dimension reduction (section 9.3).

For these algorithms, the previously cited authors construct an asymptotic distribution theory and estimated standard errors. Although the use of the bootstrap has not been investigated in this context, we conjecture that it will yield asymptotically correct in-

ference.

9.7 Comparison of the Methods

The area is one of much promise but little in the way of numerical results and programs. It is difficult at this moment to recommend one method over the others.

Reilly & Pepe (1994) compared the pseudolikelihood method of section 9.3 and the mean score method of section 9.4. When selection into the second stage of the study is independent of the response, they concluded that, asymptotically, the former was more efficient for a "spectrum of reasonable models".

The optimal semiparametric method of Robins, et al. outlined in this chapter is asymptotically more efficient than either pseudolikelihood or the mean score method. They performed a small simulation of simple logistic regression with intercept $\beta_0 = -1$ and slope $\beta_x = 1$ or 2. They set $\mathbf{X}$ to have a standard normal distribution, and the surrogate $\mathbf{W}$ as a binary variable taking on the value 1 with probability $\Phi(\mathbf{X})$, where Φ is the standard normal distribution function. They used a sample of size $n = 2,000$, with validation sample sizes of 100, 200 and 400. When $\beta_x = 1$, both methods had coverage probabilities near the nominal 95%, with the optimal semiparametric method being about 10% less variable. When $\beta_x = 2$, the pseudolikelihood coverage probabilities deteriorated somewhat from the nominal, and pseudolikelihood was about 1/3 more variable.

In a Texas A&M Ph.D. thesis, Knickerbocker (1993) reported on a logistic regression model with continuous covariates. He set $\mathbf{Z}$ to be a three-dimensional standard normal random variable, $\mathbf{W}$ to be standard normal, and $\mathbf{X} = \gamma^t(\mathbf{Z}^t, \mathbf{W})^t + \mathbf{U}$ (additive model) or $\log(\mathbf{X}) = \gamma^t(\mathbf{Z}^t, \mathbf{W})^t + \mathbf{U}$ (multiplicative model), where $\gamma = (.5, .5, .5, .5)^t$, and where $\mathbf{U}$ is normally distributed with mean zero and variance $\sigma_u^2 = 0.25$ and 1.0. He set $\beta_z = 0$ and (β_0, β_x) as in section 3.9.2, namely a 10% overall response rate and a relative risk of 3.0. The total sample size was $n = 150$, of which 50 were selected at random into the validation study; simulations were performed also when these sample sizes were 300 and 100, respectively.

Knickerbocker compared: (i) pseudolikelihood (section 9.3) with dimension reduction using sliced inverse regression; (ii) semiparametric regression calibration (section 9.6) with dimension reduc-

tion using sliced inverse regression; and (iii) the polynomial unbiased estimating function method (section 9.5.1).

The first two methods used kernel regression with a Gaussian kernel and bandwidth $\widehat{\sigma} n_v^{-1/3}$, where $\widehat{\sigma}$ is the sample standard deviation of the reduced-dimension variable outputted from sliced inverse regression, and n_v is the validation sample size. A striking feature of these simulations was the extraordinarily poor performance of the polynomial method, and so a modification was used, namely to take only one step in the method of scoring towards solving (9.7).

In this simulation, semiparametric regression calibration and the polynomial method were roughly comparable, although the former often had about 1/3 less variability. The pseudolikelihood estimate was clearly more variable than either of its competitors

When $\mathbf{L} = (\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ is discrete, all the methods are fairly straightforward to program. As we have indicated, by using dimension reduction, pseudolikelihood has been worked out for continuous covariates or response. Dimension reduction will presumably be a useful idea in extending the other methods to multiple continuous covariates.

Reilly & Pepe also discuss the use of their method in designing studies, see also Tosteson & Ware (1990).

9.8 Appendix

9.8.1 Use of Complete Data Only

In two-stage sampling, one uses only the complete data to estimate $\mathcal{B}$. We suppose that selection into the validation study ($\Delta = 1$) occurs with probability $\pi(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \alpha) = \pi(\mathbf{L}, \alpha)$, where α is an unknown parameter. As discussed in section 9.1, such selection occurs in missing data problems from observational studies. The complete data likelihood (9.1) must be modified to include the parameter α.

Let $\psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathcal{B})$ be any estimating equation that would be appropriate if $\mathbf{X}$ could have been observed for all the data, e.g., a likelihood score, or the unbiased estimating functions for quasi-likelihood & variance function models (section A.4).

An unbiased estimating function for $\mathcal{B}$ which uses only the vali-

dation data is the same as (9.5), namely

$$\begin{aligned}\Psi(\mathbf{L},\mathbf{X},\Delta,\mathcal{B},\alpha) &= \Delta\Bigg[\psi(\mathbf{Y},\mathbf{Z},\mathbf{X},\mathcal{B}) \\ &\quad - \frac{E\{\psi(\cdot,\mathcal{B})\pi(\mathbf{L},\alpha)|\mathbf{Z},\mathbf{X},\mathbf{W}\}}{E\{\pi(\mathbf{L},\alpha)|\mathbf{Z},\mathbf{X},\mathbf{W}\}}\Bigg].\end{aligned} \tag{9.13}$$

The maximum likelihood estimate for the parameter α has the estimating function

$$\begin{aligned}\ell(\mathbf{L},\Delta,\alpha) &= \{(\partial/\partial\alpha)\pi(\mathbf{L},\alpha)\}\{\Delta-\pi(\mathbf{L},\alpha)\} \\ &\quad \times[\pi(\mathbf{L},\alpha)\{1-\pi(\mathbf{L},\alpha)\}]^{-1}.\end{aligned} \tag{9.14}$$

Equation (9.14) is a special case of binary regression.

Let $\pi_\alpha(\mathbf{L},\alpha)$ be the derivative of $\pi(\mathbf{L},\alpha)$ with respect to α. Calculations outlined at the end of this subsection and based on the work in Robins, et al. (1995) show that $\widehat{\mathcal{B}}$ is asymptotically normally distributed with mean $\mathcal{B}$ and covariance matrix

$$\begin{aligned}\mathrm{cov}(\widehat{\mathcal{B}}) &\approx n^{-1}A^{-1}\left[E\left\{\Psi(\cdot)\Psi^t(\cdot)\right\}-CS^{-1}C^t\right]A^{-t}; \\ A &= E\left\{(\partial/\partial\mathcal{B}^t)\Psi(\mathbf{L},\mathbf{X},\Delta,\mathcal{B},\alpha)\right\}; \\ C &= E\left\{\Psi(\mathbf{L},\mathbf{X},\Delta,\mathcal{B},\alpha)\ell^t(\mathbf{L},\Delta,\alpha)\right\}; \\ S &= E\left[\frac{\pi_\alpha(\mathbf{L},\alpha)\pi_\alpha^t(\mathbf{L},\alpha)}{\pi(\mathbf{L},\alpha)\{1-\pi(\mathbf{L},\alpha)\}}\right].\end{aligned} \tag{9.15}$$

There are two ways to estimate this asymptotic covariance matrix. The usual approach is to estimate each of the terms in (9.15), as follows:

$$\begin{aligned}\widehat{A} &= \frac{1}{n}\sum_{i=1}^{n}(\partial/\partial\mathcal{B}^t)\Psi(\mathbf{L}_i,\mathbf{X}_i,\Delta_i,\widehat{\mathcal{B}},\widehat{\alpha}); \\ \widehat{C} &= \frac{1}{n}\sum_{i=1}^{n}\Psi(\mathbf{L}_i,\mathbf{X}_i,\Delta_i,\widehat{\mathcal{B}},\widehat{\alpha})\ell^t(\mathbf{L}_i,\Delta_i,\widehat{\alpha}); \\ \widehat{S} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_\alpha(\mathbf{L}_i,\widehat{\alpha})\pi_\alpha^t(\mathbf{L}_i,\widehat{\alpha})}{\pi(\mathbf{L}_i,\widehat{\alpha})\{1-\pi(\mathbf{L}_i,\widehat{\alpha})\}}; \\ \widehat{E}\left\{\Psi(\cdot)\Psi^t(\cdot)\right\} &= \frac{1}{n}\sum_{i=1}^{n}\Psi(\mathbf{L}_i,\mathbf{X}_i,\Delta_i,\widehat{\mathcal{B}},\widehat{\alpha})\Psi^t(\mathbf{L}_i,\mathbf{X}_i,\Delta_i,\widehat{\mathcal{B}},\widehat{\alpha}).\end{aligned}$$

Sometimes the resulting estimate is not positive semidefinite and an alternative estimator is required. Note that (9.13)–(9.14) form

a set of unbiased estimating functions for the parameters. Thus, the sandwich method (section A.3) of covariance estimation can be employed to obtain a consistent estimate of the joint covariance matrix of $(\widehat{\mathcal{B}}, \widehat{\alpha})$, from which the covariance matrix of $\widehat{\mathcal{B}}$ can be extracted.

The result (9.15) leads to an unusual fact. If α were known and not estimated, then the asymptotic covariance matrix of $\widehat{\mathcal{B}}$ would equal (9.15) *except* that one would set $C = 0$. The net effect then of estimating α is to make the asymptotic covariance matrix *smaller!* Typically, $C = 0$ only if selection into the second stage of the study does not depend on $\mathbf{Y}$.

9.8.2 Theory for Complete Data Only

Refer to (9.14) and the definitions after (9.15). By properties of the information matrix,

$$\begin{aligned} S = \operatorname{cov}\{\ell(\mathbf{L}, \Delta, \alpha)\} &= E\left[\{\ell(\mathbf{L}, \Delta, \alpha)\}\{\ell(\mathbf{L}, \Delta, \alpha)\}^t\right] \\ &= -E\left\{(\partial/\partial\alpha^t)\ell(\mathbf{L}, \Delta, \alpha)\right\}. \end{aligned}$$

From the theory of unbiased estimating function (section A.3), we have the asymptotic expansions

$$n^{1/2}(\widehat{\alpha} - \alpha) \approx n^{-1/2} \sum_{i=1}^{n} S^{-1}\ell(\mathbf{L}_i, \Delta_i, \alpha);$$

$$\begin{aligned} n^{1/2}\left(\widehat{\mathcal{B}} - \mathcal{B}\right) \approx -A^{-1}n^{-1/2} \sum_{i=1}^{n} \Bigg(&\Psi(\mathbf{L}_i, \mathbf{X}_i, \Delta_i, \mathcal{B}, \alpha) \\ &+ \left[E\left\{(\partial/\partial\alpha)\Psi(\mathbf{L}, \mathbf{X}, \Delta, \mathcal{B}, \alpha)\right\}\right] S^{-1}\ell(\mathbf{L}_i, \Delta_i, \alpha) \Bigg). \end{aligned}$$

We will show later that

$$E\left\{(\partial/\partial\alpha)\Psi(\cdot)\right\} = -C = -E\left\{\Psi(\cdot)\ell^t(\mathbf{L}, \Delta, \alpha)\right\}. \tag{9.16}$$

Making the substitution, we find that

$$\begin{aligned} &n^{1/2}\left(\widehat{\mathcal{B}} - \mathcal{B}\right) \\ &= -\frac{A^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \Psi(\mathbf{L}_i, \mathbf{X}_i, \Delta_i, \mathcal{B}, \alpha) - CS^{-1}\ell(\mathbf{L}_i, \Delta_i, \alpha) \right\}. \end{aligned} \tag{9.17}$$

The covariance matrix of the term on the right hand side of (9.17) is easily seen to equal (9.15), completing the argument.

It thus suffices to prove (9.16). First note that since the estimating equation is unbiased,

$$0 = E\Psi(\cdot) = \int \Psi(l, x, \delta, \mathcal{B}, \alpha) f_{\Delta|\mathbf{L}}(\delta|l, \alpha) f_{\mathbf{L}}(l|\mathcal{B}) d\mu(l, \delta).$$

Since this holds as a function of α, differentiate to find that

$$\begin{aligned} 0 &= E\left\{(\partial/\partial\alpha^t)\Psi(\cdot)\right\} \\ &\quad + \int \Psi(l, x, \delta, \mathcal{B}, \alpha) \left\{(\partial/\partial\alpha) f_{\Delta|\mathbf{L}}(\delta|l, \alpha)\right\} f_{\mathbf{L}}(l|\mathcal{B}) d\mu(l, \delta). \end{aligned}$$

Since $(\partial/\partial\alpha) f_{\Delta|\mathbf{L}}(\delta|l, \alpha) = \ell(l, \delta, \alpha) f_{\Delta|\mathbf{L}}(\delta|l, \alpha)$, this means that

$$0 = E\left\{(\partial/\partial\alpha^t)\Psi(\cdot)\right\} + E\left\{\Psi(\cdot)\ell(\mathbf{L}, \Delta, \alpha)\right\},$$

which verifies (9.16).

9.8.3 Theory of Moment-Estimating Functions

We first show that (9.5) is an unbiased estimating function. Similar calculations show that (9.6) is also unbiased. Dropping the arguments where it should cause no confusion, we have that

$$\begin{aligned} E\Psi_1(\mathbf{L}, \mathbf{X}, \Delta, \mathcal{B}) &= E\left[E\left\{\Psi_1(\mathbf{L}, \mathbf{X}, \Delta, \mathcal{B})|\mathbf{L}, \mathbf{X}\right\}\right] \\ &= E\left(\pi(\mathbf{L})\left[\psi(\cdot) - \frac{E\left\{\psi(\cdot)\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}}{E\left\{\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}}\right]\right). \end{aligned}$$

Writing this last term as $E\chi(\mathbf{L}, \mathbf{X})$, we note that

$$E\chi(\mathbf{L}, \mathbf{X}) = E\left[E\left\{\chi(\mathbf{L}, \mathbf{X})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}\right] = 0,$$

because the inner conditional expectation identically equals zero, i.e., if we write

$$\begin{aligned} R_1(\mathbf{Z}, \mathbf{X}, \mathbf{W}) &= \frac{E\left\{\psi(\cdot)\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}}{E\left\{\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}}; \\ R_2(\mathbf{Z}, \mathbf{X}, \mathbf{W}) &= \frac{E\left\{r(\cdot)\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}}{E\left\{\pi(\mathbf{L})|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right\}}, \end{aligned}$$

then

$$0 = E\left[\pi(\mathbf{L})\left\{\psi(\cdot) - R_1(\mathbf{Z}, \mathbf{X}, \mathbf{W})\right\}|\mathbf{Z}, \mathbf{X}, \mathbf{W}\right]. \tag{9.18}$$

We next show that Ψ_1 and Ψ_2 are uncorrelated. Because they are both unbiased estimating functions, their covariance is

$$\begin{aligned}
&E\left[\pi(\mathbf{L})\left\{\psi(\cdot)-R_1(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}\left\{r(\cdot)-R_2(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}^t\right] \\
&\quad -E\left[\pi(\mathbf{L})\left\{\psi(\cdot)-R_1(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}r^t(\mathbf{L})\right] \\
&= -E\left[\pi(\mathbf{L})\left\{\psi(\cdot)-R_1(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}\left\{R_2(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}^t\right] \\
&= -E\Bigg(E\left[\pi(\mathbf{L})\left\{\psi(\cdot)-R_1(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}|\mathbf{Z},\mathbf{X},\mathbf{W}\right] \\
&\qquad\qquad \times\left\{R_2(\mathbf{Z},\mathbf{X},\mathbf{W})\right\}^t\Bigg)=0,
\end{aligned}$$

by (9.18).

Because Ψ_1 and Ψ_2 are uncorrelated, for any Ψ_2 it follows from section A.3 that $n^{1/2}(\widehat{\mathcal{B}}-\mathcal{B})$ is asymptotically normally distributed with mean zero and covariance matrix

$$\begin{aligned}
&E\left\{\left(\frac{\partial}{\partial\mathcal{B}^t}\right)(\Psi_1+\Psi_2)\right\}^{-1}E\left(\Psi_1\Psi_1^t+\Psi_2\Psi_2^t\right) \\
&\qquad \times E\left\{\left(\frac{\partial}{\partial\mathcal{B}^t}\right)(\Psi_1+\Psi_2)\right\}^{-t}.
\end{aligned} \tag{9.19}$$

Equation (9.11) now follows from (9.19) and simple algebra.

CHAPTER 10

UNKNOWN LINK FUNCTIONS

10.1 Overview

Generalized linear models for a response $\mathbf{Y}$ as a function of a predictor $(\mathbf{Z}, \mathbf{X})$ are a special case of the general model

$$\mathbf{Y} = \mathcal{F}(\beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}, \epsilon), \tag{10.1}$$

where $\mathcal{F}(\cdot)$ is a "link" function and ϵ is a random variable independent of $(\mathbf{Z}, \mathbf{X})$. Here, ϵ can have any convenient distribution, say uniform $(0, 1)$, by suitable definition of $\mathcal{F}$.

In the current chapter, we will explore the question: *are there circumstances under which it is it possible to estimate and make inferences about (β_x, β_z) in the presence of measurement error even if the link function $\mathcal{F}(\cdot)$ is completely unknown*? Perhaps surprisely, the basic answer is "yes." In other words, if we assume that the response depends only on a linear combination of the basic predictors $(\mathbf{Z}, \mathbf{X})$, then we need not assume a fully specified model for the relationship between $\mathbf{Y}$ and $(\mathbf{Z}, \mathbf{X})$.

When there is no measurement error, there are a variety of methods for solving this problem. For binary outcomes, the maximum score estimator (Manski, 1985; Manski & Thompson, 1986) is a well-established technique. For other problems, such techniques as projection pursuit (Friedman & Stuetzle , 1981; Hall, 1989), average derivative estimation (Härdle & Stoker, 1989) and sliced inverse regression (Li, 1991; Duan & Li, 1991) have been proposed.

The primary reason to seek link-free solutions in measurement error models is the issue of *model robustness*. How do we know that a hypothesized model is correct, especially since many different link functions $\mathcal{F}(\cdot)$ for $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ lead to models for $\mathbf{Y}$ given

$(\mathbf{Z}, \mathbf{W})$ which fit the observed data? How do we know that our estimates and inferences about (β_x, β_z) are insensitive to whether we correctly specify the underlying link function?

One way to answer such model robustness questions is to postulate different link functions, and then see if the resulting inferences about (β_x, β_z) change very much. While this technique is the most used in practice, there is a need for easily implementable methods which estimate (β_x, β_z) with a minimum of assumptions.

The work proceeds under two basic assumptions. First, we will assume the classical form of the measurement error model, namely

$$\mathbf{W} = \gamma_{0,\text{em}} + \gamma_{1,\text{em}}\mathbf{X} + \mathbf{U}, \quad E(\mathbf{U}|\mathbf{Z}, \mathbf{X}, \epsilon) = 0. \tag{10.2}$$

Here, $\gamma_{1,\text{em}}$ is a square matrix and we require that it be invertible.

The second assumption is the one that replaces knowledge of the link function. We assume that for every (b_1, b_2), there are scalar constants (c_1, c_2) such that

$$E(b_1^t\mathbf{X} + b_2^t\mathbf{Z}|\beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}) = c_1 + c_2(\beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}). \tag{10.3}$$

Assumption (10.3) holds if $(\mathbf{Z}, \mathbf{X})$ has a multivariate normal distribution, but as pointed out by Li (1991), it holds under much more general circumstances than the normal. However, it is important to note that (10.3) does not always hold, e.g., when $(\mathbf{Z}, \mathbf{X})$ has discrete components.

10.1.1 Constants of Proportionality

Because the link function $\mathcal{F}$ in (10.1) is unspecified, the best one can hope for is to estimate (β_x, β_z) up to an *unknown* global constant of proportionality. For example, if we had written (10.1) as $\mathcal{F}\{c(\beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}), \epsilon\}$ for an arbitrary constant of proportionality c, since $\mathcal{F}$ is completely unknown this new specification is no different from (10.1).

Even more vexing, one cannot even directly estimate the sign of this constant! However, in many problems the sign can be ascertained from outside considerations, say knowledge that $\mathbf{Y}$ is an increasing function of one of the covariates. (One needs to add the constraint that $\mathcal{F}$ is an increasing function of its first argument in order to make the sign well defined.)

10.2 Estimation Methods

For $i = 1, ..., n$, suppose that there are k_i independent replicates of the surrogate with sample mean $\overline{\mathbf{W}}_{i\cdot}$, which has covariance matrix $\Sigma_{ww,i} = \gamma_{1,\mathrm{em}}\Sigma_{xx}\gamma_{1,\mathrm{em}}^t + \Sigma_{uu}/k_i$.

10.2.1 Some Basic Facts

The theory, due to Li (1991) and Carroll & Li (1992), is sketched in the appendix. Here we state the main results, and then show how they lead to easily implemented methods. There are two main *theoretical* results:

1 The slope (β_x, β_z) is estimated (consistently) up to a constant of proportionality by ordinary linear regression of $\mathbf{Y}_i$ on $\mathbf{Q}_i$ for $i = 1, ..., n$, where $\mathbf{Q}_i = \mathbf{L}_i(\overline{\mathbf{W}}_{i\cdot}^t, \mathbf{Z}_i^t)^t$ and

$$\mathbf{L}_i = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \gamma_{1,\mathrm{em}}^t & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{ww,i} & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{pmatrix}^{-1};$$

2 Up to a constant of proportionality, $(\beta_x^t, \beta_z^t)^t$ is the eigenvector corresponding to the sole nonzero eigenvalue of $\mathrm{cov}\{E(\xi_i|\mathbf{Y}_i)\}$, where

$$\begin{aligned} \Sigma_{qq,i} &= \mathrm{cov}(\mathbf{Q}_i) = \mathbf{L}_i \begin{pmatrix} \Sigma_{ww,i} & \Sigma_{xz} \\ \Sigma_{xz}^t & \Sigma_{zz} \end{pmatrix} \mathbf{L}_i^t; \\ \xi_i &= \Sigma_{qq,i}^{-1}\{\mathbf{Q}_i - E(\mathbf{Q}_i)\}. \end{aligned}$$

10.2.2 Least Squares and Sliced Inverse Regression

Of course, $\mathbf{L}_i$ is unknown, and in practice we have to estimate it. This requires estimating $(\gamma_{1,\mathrm{em}}, \Sigma_{xx}, \Sigma_{uu}, \Sigma_{xz}, \Sigma_{zz})$, a topic we take up below. For the moment, suppose that such estimators have been constructed, with $\widehat{\mathbf{L}}_i$ the estimate of $\mathbf{L}_i$.

Result 1 says that ordinary least squares regression of $\mathbf{Y}_i$ on $\widehat{\mathbf{Q}}_i = \widehat{\mathbf{L}}_i(\overline{\mathbf{W}}_{i\cdot}^t, \mathbf{Z}_i^t)^t$ consistently estimates (β_x, β_z) up to a constant of proportionality: note the simplicity of the method! Even more is true. If $(\mathbf{Z}, \mathbf{W})$ is unbiased for $(\mathbf{Z}, \mathbf{X})$ $(\gamma_{1,\mathrm{em}} = I)$, the estimate is the usual method of moments correction for attenuation in linear regression. Put differently, under the design condition (10.3), the usual correction for attenuation estimates (β_x, β_z) up to a global constant of proportionality for all generalized linear models!

For the second result, we use the ideas of sliced inverse regression (Li, 1991; Carroll & Li, 1992). Let

$$\begin{aligned} \widehat{\xi}_i &= \widehat{\Sigma}_{qq,i}^{-1}\left(\mathbf{Q}_i - \overline{\mathbf{Q}}_\cdot\right); \text{ where} \\ \widehat{\Sigma}_{qq,i} &= \widehat{\mathbf{L}}_i \begin{pmatrix} \widehat{\Sigma}_{ww,i} & \widehat{\Sigma}_{xz} \\ \widehat{\Sigma}_{xz}^t & \widehat{\Sigma}_{zz} \end{pmatrix} \widehat{\mathbf{L}}_i^t. \end{aligned}$$

Divide the range of $\mathbf{Y}$ into H intervals (slices in the usual jargon), say $I_1, ..., I_H$. Let $\overline{\xi}$ be the sample mean of the terms $\widehat{\xi}_i$. Then use the following algorithm to estimate $\text{cov}\{E(\xi|\mathbf{Y})\}$:

- Let $\widehat{p}_h$ be the observed proportion of $\mathbf{Y}_i$'s falling into the hth slice I_h.
- Within each slice compute the mean $\overline{\xi}_h = (n\widehat{p}_h)^{-1} \sum_{\mathbf{Y}_i \in I_h} \widehat{\xi}_i$, $h = 1, ..., H$.
- Form the covariance matrix $\widehat{\Sigma}_\xi = \sum_{h=1}^H \widehat{p}_h(\overline{\xi}_h - \overline{\xi})(\overline{\xi}_h - \overline{\xi})^t$.
- Compute the eigenvector associated with the largest eigenvalue of $\widehat{\Sigma}_\xi$.

10.2.3 Details of Implementation

In order to implement the methods, one needs estimates ($\widehat{\gamma}_{1,\text{em}}$, $\widehat{\Sigma}_{ww,i}, \widehat{\Sigma}_{zz}, \widehat{\Sigma}_{xx}, \Sigma_{xz}$). The sample covariance matrix of the $\mathbf{Z}$'s serves as the estimate $\widehat{\Sigma}_{zz}$, while the sample covariance matrix between the $\overline{\mathbf{W}}_{i\cdot}$'s and the $\mathbf{Z}$'s serves as the estimate $\widehat{\Sigma}_{xz}$.

Estimation of Σ_{uu} has already been described in section 3.4. If it cannot be assumed that $(\mathbf{Z}, \mathbf{W})$ is unbiased, then estimation of $\gamma_{1,\text{em}}$ requires additional data, usually an external validation data set. Freedman, et al. (1991) describe a method which does not require validation, but does require replicated unbiased measures of $(\mathbf{Z}, \mathbf{X})$.

Finally, we turn to constructing the estimate $\widehat{\Sigma}_{xx}$ and $\widehat{\Sigma}_{ww,i}$. First, Σ_{uu} can be estimated by the methods of section 3.4. If we denote the estimate of Σ_{xx} from that section as Σ_{xx*}, then a consistent estimate of Σ_{xx} in our context is

$$\widehat{\Sigma}_{xx} = \gamma_{1,\text{em}}^{-1} \Sigma_{xx*} \gamma_{1,\text{em}}^{-t}.$$

Finally, $\widehat{\Sigma}_{ww,i} = \widehat{\Sigma}_{xx} + \widehat{\Sigma}_{uu}/k_i$.

10.3 Framingham Heart Study

This is a continuation of the example in section 4.5. Since for a binary regression ordinary least squares regression of $\mathbf{Y}$ on $\mathbf{Q}$ and sliced inverse regression are the same, we used the former.

Figure 10.1 shows estimated densities of transformed saturated fat for the cases of CHD and the non-cases (the "controls"). One can see in Figure 10.1 the small but clear effect that those suffering from CHD have larger systolic blood pressures than those without CHD.

We repeated the analysis of section 4.5.1 using all the data and with the two exams treated as replicates of one another. The method of this chapter only estimates the regression parameter up to an *unknown* constant of proportionality. Thus, we suggest the following strategy. First, standardize all numerical random variables to have sample mean zero and sample variance 1.0. After producing the estimates, normalize them by dividing each by the square root of the sum of the squares of the estimated regression coefficients. When we did this to the method of this chapter and to the regression calibration estimator, we obtained similar parameter estimates. For example, regression calibration's parameter estimate for systolic blood pressure was 0.84 with a bootstrap standard error of 0.09, while the sliced inverse regression method has a parameter estimate of 0.89 with a bootstrap standard error of 0.07.

10.4 Appendix

10.4.1 Basic Theory

For ease of notation, in discussing the theory it will not be useful to make a distinction between covariates measured exactly and covariates measured with error. Hence, we will combine $\mathbf{Z}$ and $\mathbf{X}$ into $\mathbf{X}$, and also combine $\mathbf{Z}$ and $\mathbf{W}$ into $\mathbf{W}$.

The classical error model (10.2) assumes that $\mathbf{W}$ is a surrogate, and that in particular $\mathbf{U}$ is independent of the response given the underlying true predictors.

The theory is based on the fact that there is a scalar function $c(\cdot)$ such that

$$E(\mathbf{W}|\mathbf{Y}) = E(\mathbf{W}) + c(\mathbf{Y})\gamma_{1,\mathrm{em}}\Sigma_{xx}\beta_x. \tag{10.4}$$

We first prove (10.4). Assume without loss of generality that $E(\mathbf{X}) =$

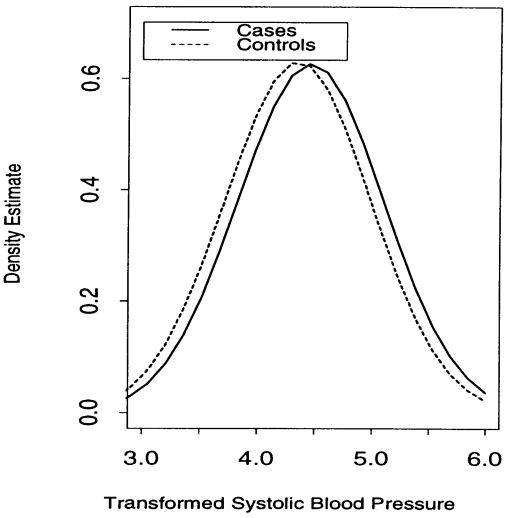

Figure 10.1. *Density estimates for transformed systolic blood pressure in the Framingham Heart Study. "Cases" are those with CHD, while "Controls" do not have CHD.*

0 and $\text{cov}(\mathbf{X}) = I$, the identity matrix. Because of the form of model (10.2), it suffices to to show that $E(\mathbf{X}|\mathbf{Y}) = c(\mathbf{Y})\Sigma_{xx}\beta_x$. However, Li (1991) shows that for any vector b such that $b^t\beta_x = 0$, $b^tE(\mathbf{X}|\mathbf{Y} = y) = 0$ (with probability 1). The result thus follows from the fact that $b^tE(\mathbf{X}|\mathbf{Y} = y) = 0$ for all b such that $b^t\beta_x = 0$ implies that $E(\mathbf{X}|\mathbf{Y} = y) = c(y)\beta_x$ for some scalar function $c(y)$.

We next prove the first theoretical result in section 10.2. The ordinary least squares slope is

$$\left\{ n^{-1} \sum_{i=1}^{n} \left(\mathbf{Q}_i - \overline{\mathbf{Q}}_\cdot\right) \left(\mathbf{Q}_i - \overline{\mathbf{Q}}_\cdot\right)^t \right\}^{-1} n^{-1} \sum_{i=1}^{n} \left(\mathbf{Q}_i - \overline{\mathbf{Q}}_\cdot\right) \mathbf{Y}_i.$$

The term in brackets is asymptotically the same as

$$n^{-1} \sum_{i=1}^{n} \left\{\mathbf{Q}_i - E(\mathbf{Q}_i)\right\} \left\{\mathbf{Q}_i - E(\mathbf{Q}_i)\right\}^t = n^{-1} \sum_{i=1}^{n} \mathbf{L}_i \Sigma_{ww,i} \mathbf{L}_i^t$$

$$= \Sigma_{xx}\gamma_{1,\mathrm{em}}^t n^{-1} \sum_{i=1}^{n} \Sigma_{ww,i}^{-1}\gamma_{1,\mathrm{em}}\Sigma_{xx}.$$

Using (10.4), the second term is asymptotically the same as

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\{\mathbf{Q}_i - E(\mathbf{Q}_i)\}\,\mathbf{Y}_i \\
&= n^{-1}\sum_{i=1}^{n}\mathbf{L}_i E\left[\mathbf{Y}_i\left\{E\left(\overline{\mathbf{W}}_{i\cdot} - \mu_w\right)|\mathbf{Y}_i\right\}\right] \\
&= n^{-1}\sum_{i=1}^{n}\mathbf{L}_i\gamma_{1,\mathrm{em}}\Sigma_{xx}\beta_x E\left\{\mathbf{Y}_i c(\mathbf{Y}_i)\right\} \\
&= n^{-1}\sum_{i=1}^{n} c_*\Sigma_{xx}\gamma_{1,\mathrm{em}}^t n^{-1}\sum_{i=1}^{n}\Sigma_{ww,i}^{-1}\gamma_{1,\mathrm{em}}\Sigma_{xx}\beta_x,
\end{aligned}$$

where $c_* = E\left\{\mathbf{Y}_i c(\mathbf{Y}_i)\right\}$. Thus, the OLS slope estimate converges to $c_*\beta_x$ as claimed.

Now consider the second result. We have that $E(\xi_i) = 0$ and, using (10.4),

$$E(\xi_i|\mathbf{Y}_i) = \Sigma_{qq,i}^{-1}\mathbf{L}_i\gamma_{1,\mathrm{em}}\Sigma_{xx}\beta_x c(\mathbf{Y}_i) = \beta_x c(\mathbf{Y}_i),$$

say. Thus, the terms $E(\xi_i|\mathbf{Y}_i)$ form a set of (marginally) independent and identically distributed random variables, and

$$\mathrm{cov}\left\{E(\xi_i|\mathbf{Y}_i)\right\} = c_{**}\beta_x\beta_x^t,$$

where $c_{**} = E\left\{c^2(\mathbf{Y}_i)\right\}$. The eigenvector of this matrix associated with the nonzero eigenvalue is proportional to β_x, as claimed.

CHAPTER 11

HYPOTHESIS TESTING

11.1 Overview

In this chapter, we discuss hypothesis tests concerning regression parameters. To keep the exposition simple, we will focus on linear regression. However, the results of sections 11.2.1, 11.2.3 and 11.4 hold in general, and the results of sections 11.2.2 and 11.3 hold at an approximate level for all generalized linear models, including logistic regression, under the regression calibration approximation. More generally, the same can be said of any problem for which the mean and variance of the response depends only upon a linear combination of the predictors. We assume nondifferential measurement error, $\mathbf{W} = \mathbf{X} + \mathbf{U}$, throughout the chapter.

Assuming that one or more of the estimation methods described in the previous chapters is applicable, the simplest approach to hypothesis testing forms the required test statistic from the parameter estimates and their estimated standard errors. Such tests are justified whenever the estimators themselves are justified. However, this approach to testing is only possible when the indicated methods of estimation are possible, and thus require either knowledge of the measurement error variance, or the presence of validation data, or replicate measurements, or instrumental variables, etc.

There are certain situations in which naive hypothesis tests are justified and thus can be performed without additional data or information of any kind. Here "naive" means that we ignore measurement error and substitute $\mathbf{W}$ for $\mathbf{X}$ in a test that is valid when $\mathbf{X}$ is observed. This chapter studies naive tests, describing when they are and are not acceptable, and indicates how supplementary data, when available, can be used to improve the efficiency of naive tests.

We use the criterion of asymptotic validity to distinguish be-

tween acceptable and nonacceptable tests. We say a test is asymptotically valid if its Type I error rate approaches its nominal level as the sample size increases. Asymptotic validity, which we shorten to validity, of a test is a minimal requirement for acceptability.

The main results on the validity of naive tests under nondifferential measurement error are as follows. The naive test of no effects due to $\mathbf{X}$ is valid, as is the naive test for no effects due to $(\mathbf{Z}^t, \mathbf{X}^t)^t$, i.e., that none of the covariates affect $\mathbf{Y}$. The naive test of no effects due to $\mathbf{Z}$ is not valid in general, but is valid under some restrictive assumptions, and the same is true for the naive test of no effects due to a specified subvector of $\mathbf{X}$, e.g., the first component of $\mathbf{X}$. These results are obtained using the regression calibration approximation, which takes the regression model for $\mathbf{Y}$ given $\mathbf{Z}$ and $\mathbf{X}$ and replaces $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$.

When $\mathbf{Y}$ follows a generalized linear model (section A.5) in $\mathbf{Z}$ and $\mathbf{X}$, then we show that the efficient score test of no effects due to $\mathbf{X}$ is easily obtained: one takes the efficient score test when $\mathbf{X}$ is observed and replaces $\mathbf{X}$ by a parametric estimate of $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$. Put another way, a null hypothesis test based on regression calibration is (asymptotically) efficient.

11.2 The Regression Calibration Approximation

In linear regression, the mean of the response given the true covariates is $\beta_0 + \beta_z^t \mathbf{Z} + \beta_x^t \mathbf{X}$. Under the additional assumption that the possibly multivariate regression of $\mathbf{X}$ on $\mathbf{Z}$ and $\mathbf{W}$ is linear, i.e.,

$$E(\mathbf{X} \mid \mathbf{Z}, \mathbf{W}) = \alpha_0 + \alpha_z^t \mathbf{Z} + \alpha_w^t \mathbf{W},$$

we have that the observed data also have a linear mean, namely

$$E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_x^t \alpha_0 + (\beta_z^t + \beta_x^t \alpha_z^t)\mathbf{Z} + \beta_x^t \alpha_w^t \mathbf{W}. \quad (11.1)$$

Equation (11.1) is the starting point for our discussion of testing. One of the assumptions of our measurement error model will be that α_w^t is an invertible matrix.

A naive analysis of the data fits a linear model as well. We will write this model as

$$E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) = \gamma_0 + \gamma_z^t \mathbf{Z} + \gamma_w^t \mathbf{W}. \quad (11.2)$$

It is the correspondence between the naive model (11.2) and the

actual model (11.1) which is of interest here.

11.2.1 Testing $H_0 : \beta_x = 0$

Here we show that the naive test of no effect due to any of the predictors measured with error is asymptotically valid. The result holds in general, and not just for linear regression.

A comparison of (11.1) and (11.2) shows that $\beta_x = 0$ implies that $\alpha_w \beta_x = 0$ which in turn implies that $\gamma_w = 0$. The converse is also true, namely that $\gamma_w = 0$ implies that $\beta_x = 0$ because α_w is invertible.

Because $\gamma_w = 0$ if $\beta_x = 0$, it follows that the naive test, i.e., the test of $H_0 : \gamma_w = 0$, is a valid test of $H_0 : \beta_x = 0$.

Although $\gamma_w = 0$ only if $\beta_x = 0$, this reverse implication, though perhaps interesting, is not necessary for the validity of the naive test.

11.2.2 Testing $H_0 : \beta_z = 0$

Here we show that in linear regression, the naive tests for effects due to $\mathbf{Z}$ is typically invalid, except under special circumstances.

Further comparison of (11.1) and (11.2) shows that $\beta_z = 0$ implies that $\gamma_z = 0$, only if $\alpha_z \beta_x = 0$. It follows that the naive test of $H_0 : \beta_z = 0$ is valid if $\mathbf{X}$ is unrelated to $\mathbf{Y}$ in the model (11.6), i.e., $\beta_x = 0$, or if $\mathbf{Z}$ is unrelated to $\mathbf{X}$, i.e., $\alpha_z = 0$.

In generalized linear models, the naive test is valid when $\mathbf{Z}$ and $\mathbf{X}$ are independent, at least approximately at the level of the regression calibration approximation. Gail, Wieand & Piantadosi (1984) and Gail, Tan & Piantadosi (1988) show that when the regression calibration approximation fails for logistic regression, then the naive test is no longer even approximately valid.

The general conclusion is that the test of $H_0 : \beta_z = 0$ is invalid, although there are certain situations in which it is valid.

11.2.3 Testing $H_0 : (\beta_x^t, \beta_z^t)^t = 0$

A final comparison of (11.1) and (11.2) shows that $(\beta_x^t, \beta_z^t)^t = 0$ if and only if $(\gamma_z^t, \gamma_x^t)^t = 0$, so the naive test that none of the covariates affect $\mathbf{Y}$ is valid in general.

11.3 Hypotheses about Subvectors of β_x and β_z

There are situations in which interest focuses on testing for effects due to some subset of the predictors measured with error, or due to some subset of the error-free covariates. That is if $\mathbf{X} = (\mathbf{X}_1^t, \mathbf{X}_2^t)^t$, $\beta_x = (\beta_{x,1}^t, \beta_{x,2}^t)^t$, and $\mathbf{Z} = (\mathbf{Z}_1^t, \mathbf{Z}_2^t)^t$, $\beta_z = (\beta_{z,1}^t, \beta_{z,2}^t)^t$, then we may be interested in testing $H_0 : \beta_{x,1} = 0$ or $H_0 : \beta_{z,1} = 0$.

We have already seen that for testing $H_0 : \beta_z = 0$, the naive test is not valid in general, and it follows from similar reasoning that the same is true of naive tests of $H_0 : \beta_{z,1} = 0$. Therefore we will restrict attention to naive tests of $H_0 : \beta_{x,1} = 0$.

Suppose now that $\beta_x^t \mathbf{X} = \beta_{x,1}^t \mathbf{X}_1 + \beta_{x,2}^t \mathbf{X}_2$ and that

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2) &= \alpha_{1,0} + \alpha_{1,z}^t \mathbf{Z} + \\ &\quad \alpha_{1,w_1}^t \mathbf{W}_1 + \alpha_{1,w_2}^t \mathbf{W}_2; \\ E(\mathbf{X}_2 \mid \mathbf{Z}, \mathbf{W}_1, \mathbf{W}_2) &= \alpha_{2,0} + \alpha_{2,z}^t \mathbf{Z} + \\ &\quad \alpha_{2,w_1}^t \mathbf{W}_1 + \alpha_{2,w_2}^t \mathbf{W}_2, \end{aligned} \tag{11.3}$$

where $\mathbf{W} = (\mathbf{W}_1^t, \mathbf{W}_2^t)^t$ is partitioned as is $\mathbf{X}$.

With these changes (11.1) becomes

$$\begin{aligned} &E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) = \beta_0 + \beta_{x,1}^t \alpha_{1,0} + \beta_{x,2}^t \alpha_{2,0} \\ &+(\beta_z^t + \beta_{x,1}^t \alpha_{1,z}^t + \beta_{x,2}^t \alpha_{2,z}^t)\mathbf{Z} + (\beta_{x,1}^t \alpha_{1,w_1}^t + \beta_{x,2}^t \alpha_{2,w_1}^t)\mathbf{W}_1 \\ &\quad +(\beta_{x,1}^t \alpha_{1,w_2}^t + \beta_{x,2}^t \alpha_{2,w_2}^t)\mathbf{W}_2, \end{aligned} \tag{11.4}$$

and in a naive analysis of the data the mean model

$$E(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}) = \gamma_0 + \gamma_z^t \mathbf{Z} + \gamma_{w_1}^t \mathbf{W}_1 + \gamma_{w_1}^t \mathbf{W}_2 \tag{11.5}$$

is fit to the observed data.

Comparing (11.4) and (11.5) shows that $\beta_{x,1} = 0$ implies that $\gamma_{w_1} = 0$ only if $\alpha_{2,w_1}\beta_{x,2} = 0$. It follows that the naive test of $H_0 : \beta_{x,1} = 0$ is valid only if $\alpha_{2,w_1}\beta_{x,2} = 0$. If $\mathbf{X}_2$ is related to $\mathbf{Y}$, then $\beta_{x,2}$ will be nonzero. If $\mathbf{X}_2$ is related to $\mathbf{W}_1$ in (11.3) then α_{2,w_1} will be nonzero. This will be the case if some components of $\mathbf{X}_1$ are correlated with some components of $\mathbf{X}_2$.

For example, consider the NHANES study briefly introduced in Chapter 1 and discussed in more detail in Chapter 3. Let $\mathbf{X}$ be the vector of true total caloric intake (TC = $\mathbf{X}_1$) and saturated fat (SF = $\mathbf{X}_2$), and and let $\mathbf{Z}$ denote nondietary variables. The naive test for a SF effect simply substitutes observed TC and SF intake for true TC and SF intake, and it will be a valid test if there is no risk

of breast cancer due to TC ($\beta_{x,1} = 0$) or if the regression of true SF intake on observed SF, observed TC and nondietary variables has no component due to TC ($\alpha_{2,w_1} = 0$).

In general the conclusion is that the test of $H_0 : \beta_{x,1} = 0$ is invalid, although there are certain situations in which it is valid.

11.4 Efficient Score Tests of $H_0 : \beta_x = 0$

In this section, we assume that $\mathbf{Y}$ given $\mathbf{Z}$ and $\mathbf{X}$ follows a generalized linear model (section A.5). In particular, the mean and variance functions for these models are in the form

$$\begin{aligned} E(\mathbf{Y}|\mathbf{Z},\mathbf{X}) &= f(\mathbf{Z},\mathbf{X},\mathcal{B}) = f(\beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t\mathbf{X}); \quad (11.6) \\ \text{var}(\mathbf{Y}|\mathbf{Z},\mathbf{X}) &= \sigma^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta) \\ &= \sigma^2 g^2(\beta_0 + \beta_z^t\mathbf{Z} + \beta_x^t,\theta). \quad (11.7) \end{aligned}$$

We show that the naive score test of $H_0 : \beta_x = 0$, while asymptotically valid in general, is not generally an efficient score test. However, we do find a test that is asymptotically equivalent to the efficient score test and show that under certain conditions this test is equal to the naive score test.

Recall that the naive test simply substitutes $\mathbf{W}$ for $\mathbf{X}$. We show that if a parametric model for $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$ is appropriate, say $E(\mathbf{X}|\mathbf{Z},\mathbf{W}) = m(\mathbf{Z},\mathbf{W},\alpha)$, and if $\widehat{\alpha}$ is a $n^{1/2}$-consistent estimator of α, then the test that substitutes $m(\mathbf{Z},\mathbf{W},\widehat{\alpha})$ for $\mathbf{X}$ is asymptotically an efficient score test. It must be emphasized that this result about substituting $m(\mathbf{Z},\mathbf{W},\widehat{\alpha})$ for $\mathbf{X}$ requires the assumption of a generalized linear model.

The validity of naive null tests for predictors measured with error, and the efficiency for generalized linear models of tests which replace $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$, was shown by Tosteson & Tsiatis (1988). For the special case of models with canonical link functions, the efficiency of tests that replace $\mathbf{X}$ by $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$, follows from the form of the efficient score for generalized linear measurement error models given in Stefanski & Carroll (1987).

It follows from these results that the only time that the naive test of $H_0 : \beta_x = 0$ in generalized linear models is equivalent to the efficient score test occurs when $E(\mathbf{X}|\mathbf{Z},\mathbf{W})$ is independent of $\mathbf{Z}$ and linear in $\mathbf{W}$. Moreover, Tosteson and Tsiatis (1988) show that the asymptotic relative efficiency (ARE) of the naive test to the effi-

cient score test is always less than 1, unless the two tests are equivalent. They also show that for the special case where $\mathbf{X}$ is univariate and $\mathbf{Z}$ is not present, that this ARE is $\{\text{corr}\,(E(\mathbf{X}|\mathbf{W}), \mathbf{W})\}^2$. Thus, the naive test can be arbitrarily inefficient if $E(\mathbf{X}|\mathbf{W})$ is sufficiently nonlinear in $\mathbf{W}$.

The mathematical arguments supporting these statements are given in the following subsection. This subsection is fairly technical and can be omitted on first reading.

11.4.1 Generalized Score Tests

To define a generalized score test of $H_0 : \beta_x = 0$, let $H_i(\alpha)$ be any random vector depending on $(\mathbf{Z}_i, \mathbf{X}_i, \mathbf{W}_i)$ and the parameter α and having the same dimension as $\mathbf{X}_i$. Possible choices of $H_i(\alpha)$ will be discussed later. Define

$$\mathcal{L}(\beta_0, \beta_z, \alpha, \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_i(\alpha) d_i(\beta_0, \beta_z, \theta) \left\{ \mathbf{Y}_i - f(\beta_0 + \beta_z^t \mathbf{Z}_i) \right\}, \quad (11.8)$$

where d_i used here and c_i used below are defined by

$$\begin{aligned} d_i(\beta_0, \beta_z, \theta) &= f^{(1)}(\beta_0 + \beta_z^t \mathbf{Z}_i)/g^2(\beta_0 + \beta_z^t \mathbf{Z}_i, \theta) \\ c_i(\beta_0, \beta_z, \theta) &= d_i(\beta_0, \beta_z, \theta) f^{(1)}(\beta_0 + \beta_z^t \mathbf{Z}_i). \end{aligned}$$

Our test statistics will be $\mathcal{L}$ with the parameters β_0, β_z, α, and θ replaced by estimators. Also define

$$\begin{aligned} C_1(\beta_0, \beta_z, \alpha, \theta) &= n^{-1} \sum_{i=1}^{n} H_i(\alpha) H_i^t(\alpha) c_i(\beta_0, \beta_z, \theta); \\ C_2(\beta_0, \beta_z, \alpha, \theta) &= n^{-1} \sum_{i=1}^{n} H_i(\alpha)(1,\ \mathbf{Z}_i^t)^t c_i(\beta_0, \beta_z, \theta); \\ C_3(\beta_0, \beta_z, \alpha, \theta) &= n^{-1} \sum_{i=1}^{n} (1,\ \mathbf{Z}_i^t)^t (1,\ \mathbf{Z}_i^t) c_i(\beta_0, \beta_z, \theta); \\ D(\beta_0, \beta_z, \alpha, \theta) &= C_1 - C_2 C_3^{-1} C_2^t, \end{aligned}$$

where in the last equation the dependence of C_1, C_2 and C_3 on $(\beta_0, \beta_z, \alpha, \theta)$ has been suppressed for brevity.

Let $\widehat{\theta}$ be any $n^{1/2}$-consistent estimate of the variance parameter θ; see section A.4 or Carroll & Ruppert (1988, Chapter 3) for some

methods of estimating θ. If α is unknown, e.g., when

$$H_i(\alpha) = E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) = m(\mathbf{Z}, \mathbf{W}, \alpha),$$

then we assume a $n^{1/2}$-consistent estimator of α. Methods of estimating α are discussed in Chapter 3. The quasilikelihood and variance function (QVF) estimates of (β_0, β_z), $(\widehat{\beta}_0, \widehat{\beta}_z)$, satisfy

$$0 = \sum_{i=1}^{n} (1,\ \mathbf{Z}_i^t)^t d_i(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\theta}) \left\{ \mathbf{Y}_i - f(\widehat{\beta}_0 + \widehat{\beta}_z^t \mathbf{Z}_i) \right\}.$$

With $\dim(\mathbf{Z})$ denoting the dimension of $\mathbf{Z}$, define

$$\widehat{\sigma}^2 = \{n - 1 - \dim(\mathbf{Z})\}^{-1} \sum_{i=1}^{n} \frac{\left\{ \mathbf{Y}_i - f(\widehat{\beta}_0 + \widehat{\beta}_z^t \mathbf{Z}_i) \right\}^2}{g^2(\widehat{\beta}_0 + \widehat{\beta}_z^t \mathbf{Z}_i, \widehat{\theta})}.$$

We consider the test statistics of the form

$$\widehat{\sigma}^{-2} \mathcal{L}^t(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\alpha}, \widehat{\theta}) D^{-1}(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\alpha}, \widehat{\theta}) \mathcal{L}^t(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\alpha}, \widehat{\theta}). \tag{11.9}$$

When $\mathbf{X}$ is observable, then setting $H_i(\alpha) = \mathbf{X}_i$ in (11.8) results in (11.9) being the usual score test statistic of $H_0 : \beta_x = 0$. The naive score test statistic is obtained by setting $H_i(\alpha) = \mathbf{W}_i$ in (11.8). We show is this section that when $E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) = m(\mathbf{Z}, \mathbf{W}, \alpha)$, then setting $H_i(\alpha) = m(\mathbf{Z}_i, \mathbf{W}_i, \alpha)$ in (11.8) results in a test statistic that is asymptotically equivalent to the efficient score test statistic.

We now show that under the hypothesis $H_0 : \beta_x = 0$, the test statistic in (11.9) is asymptotically chi-square with degrees of freedom equal to the common dimension of $H_i(\alpha)$, $\mathbf{X}_i$ and β_x. It follows from Carroll & Ruppert (1988, Chapter 7) that to order $o_p(1)$,

$$\sqrt{n} \begin{pmatrix} \widehat{\beta}_0 - \beta_0 \\ \widehat{\beta}_z - \beta_z \end{pmatrix} \approx \frac{C_3^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ \mathbf{Z}_i \end{pmatrix} d_i \left\{ \mathbf{Y}_i - f(\beta_0 + \beta_z^t \mathbf{Z}_i) \right\}.$$

where the dependence of C_3 and d_i on the parameters has been suppressed. Since $E(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{W}_i) = E(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i) = f(\beta_0 + \beta_z^t \mathbf{Z}_i)$ under the null hypothesis, it is straightforward to show that to order $o_p(1)$,

$$\mathcal{L}^t(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\alpha}, \widehat{\theta}) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} d_i$$

$$\times \left\{ \mathbf{Y}_i - f(\beta_0 + \beta_z^t \mathbf{Z}_i) \right\} \left\{ H_i(\alpha) - C_2 C_3^{-1} \begin{pmatrix} 1 \\ \mathbf{Z}_i \end{pmatrix} \right\}, \quad (11.10)$$

and $\mathcal{L}^t(\widehat{\beta}_0, \widehat{\beta}_z, \widehat{\alpha}, \widehat{\theta})$ is hence asymptotically multivariate normal with mean zero and covariance matrix $\sigma^2 D(\beta_0, \beta_z, \alpha, \theta)$. In (11.10) $d_i = d_i(\beta_0, \beta_z, \theta)$. It follows that (11.9) has the indicated chi-square distribution.

It remains to show that for generalized linear models, substituting $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ for $H_i(\alpha)$ in (11.9) results in a test that is asymptotically equivalent to the efficient score test. The argument is adapted from Tosteson & Tsiatis (1988).

The density or mass function of generalized linear models is given by (A.27). Write $\xi = g(\eta)$ with $\eta = \beta_0 + \beta_x^t x + \beta_z^t z$. Using the assumption of nondifferential measurement error (conditional independence so that $\mathbf{Y}$ and $\mathbf{W}$ are independent given $\mathbf{X}$ and $\mathbf{Z}$), the density or mass function of the observed data is

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{Z},\mathbf{W}}(y|z,w) &= \int f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) d\mu(x) \\ &= \int \exp\left[\frac{y g(\eta) - C\{g(\eta)\}}{\phi} + c(y,\phi) \right] f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) d\mu(x). \end{aligned}$$

Write $h(y,z) = \exp\left(\left[y g(\beta_0 + \beta_z^t z) - C\{g(\beta_0 + \beta_z^t z)\} \right] / \phi \right)$. Since $c(y,\phi)$ does not depend on β_x, the likelihood score used in construction of the efficient score statistic is

$$\begin{aligned} &\frac{\partial}{\partial \beta_x} \log\left\{ f_{\mathbf{Y}|\mathbf{Z},\mathbf{W}}(y|z,w) \right\} \Big|_{\beta_x = 0} \\ &= \frac{1}{h(y,z)} \frac{\partial}{\partial \beta_x} \int f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) d\mu(x) \Big|_{\beta_x = 0} \\ &= \frac{1}{h(y,z)} \left[\int f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) \right. \\ &\qquad \left. \times \frac{\partial}{\partial \beta_x} \log\left\{ f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) \right\} d\mu(x) \right]_{\beta_x = 0} \\ &= \int f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) \frac{\partial}{\partial \beta_x} \log\left\{ f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x) \right\} \Big|_{\beta_x = 0} d\mu(x) \\ &= g^{(1)}(\beta_0 + \beta_z^t z) \left[y - C^{(1)}\left\{ g(\beta_0 + \beta_z^t z) \right\} \right] \end{aligned}$$

$$\times \int (x/\phi) f_{\mathbf{X}|\mathbf{Z},\mathbf{W}}(x|z,w) d\mu(x)$$

$$= \frac{1}{\phi}\left[y - C^{(1)}\left\{g^2(\beta_0 + \beta_z^t z)\right\}\right]$$

$$g^{(1)}(\beta_0 + \beta_z^t z) E(\mathbf{X}|\mathbf{Z} = z, \mathbf{W} = w). \tag{11.11}$$

If $\mathbf{X}$ were observable, the only difference in these calculations would be that $\mathbf{X}$ would replace $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ in (11.11). Hence, the efficient score tests for the observed data is obtained by substituting $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$ for $\mathbf{X}$.

For the case studied above there is a parametric model, $E(\mathbf{X}|\mathbf{Z}, \mathbf{W}) = m(\mathbf{Z}, \mathbf{W}, \alpha)$. As mentioned before, $n^{1/2}$-consistent estimation of α is possible by the methods in Chapter 3. It is also possible to construct asymptotically efficient or nearly efficient score tests based on nonparametric estimates of $E(\mathbf{X}|\mathbf{Z}, \mathbf{W})$. Stefanski & Carroll (1990a, 1991) construct semiparametric tests that achieve full or nearly full efficiency when $\mathbf{W}$ is unbiased for $\mathbf{X}$ and its measurement error variance is known or independently estimated. Sepanski (1992) uses nonparametric regression techniques to construct efficient tests when there exists an independent validation data set or an independent data set containing an unbiased instrumental variable.

CHAPTER 12

DENSITY ESTIMATION AND NONPARAMETRIC REGRESSION

In this chapter we give an overview of two nonparametric estimation problems that are of interest in their own right, and also arise as secondary problems in regression calibration and hypothesis testing. The first problem is the estimation of the density of a random variable $\mathbf{X}$, while the second is the nonparametric estimation of a regression, both when $\mathbf{X}$ is measured with error.

12.1 Deconvolution

The fundamental problem is that of estimating the density of $\mathbf{X}$ when $\mathbf{W} = \mathbf{X} + \mathbf{U}$ is observed and the density of $\mathbf{U}$ is known. Closely related is the problem of estimating the regression function, $m(\mathbf{w}) = E(\mathbf{X} \mid \mathbf{W} = w)$, when only $\mathbf{W} = \mathbf{X} + \mathbf{U}$ is observed and the density of $\mathbf{U}$ is known. The latter estimation problem is encountered in both regression calibration (Chapter 3) and hypothesis testing (Chapter 11).

Suppose that $\mathbf{X}$ is a continuous, scalar random variable, and that there are no covariates $\mathbf{Z}$ measured without error. When $\mathbf{X}$ is unobservable, likelihood methods (Chapter 7) require a model for the density of $\mathbf{X}$. Regression calibration (Chapter 3) consists of the usual analysis but with $\mathbf{X}$ replaced by

$$\begin{aligned} m(\mathbf{W}) &= E(\mathbf{X}|\mathbf{W}) = \frac{1}{f_w(\mathbf{W})} \int x f_x(x) f_{w|x}(\mathbf{W}|x) dx \\ &= \frac{1}{f_w(\mathbf{W})} \int x f_x(x) f_u(\mathbf{W} - x) dx. \end{aligned} \tag{12.1}$$

In Chapter 11, it was shown that when testing for the effect of the covariate measured with error, replacing $\mathbf{X}$ by an estimate of its regression $m(\mathbf{W})$ on $\mathbf{W}$ yields the hypothesis test with the highest local power (asymptotically).

Estimating the density function, f_x, of $\mathbf{X}$ is thus critical. The density function f_w is the convolution of f_x and f_u,

$$f_w(w) = \int f_x(x) f_u(w - x) dx,$$

and we refer to the problem of estimating f_x in the absence of parametric assumptions as *deconvolution.*

When both f_w and f_u are known, f_x is recovered by Fourier inversion. Letting ϕ_a denote the characteristic function of the random variable $\mathbf{A}$, e.g., $\phi_w(t) = \int e^{itw} f_w(w) dw$, we have that $\phi_x(t) = \phi_w(t)/\phi_u(t)$. Then by Fourier inversion,

$$f_x(x) = \frac{1}{2\pi} \int e^{-itx} \phi_x(t) dt = \frac{1}{2\pi} \int e^{-itx} \frac{\phi_w(t)}{\phi_u(t)} dt.$$

Even if, as we will now suppose, the density function f_u of $\mathbf{U}$ is known, the problem is complicated by the fact that the density of $\mathbf{W}$ is unknown and must be estimated. For the deconvolution problem under these assumptions, estimators with known rates of convergence were first obtained by Stefanski & Carroll (1986, 1990c) and Liu & Taylor (1989). Their research has since spawned a considerable literature, see for example Carroll & Hall (1988), Liu & Taylor (1990), Zhang (1990), Fan (1991a,b,c; 1992a), Fan, et al. (1991), Masry & Rice (1992), Fan & Truong (1993), Fan & Masry (1993) and Stefanski (1989,1990). An interesting econometric application using a modification of these methods is discussed by Horowitz & Markatou (1993).

We now describe the solution. Statisticians have studied kernel density estimates of f_w of the form

$$\widehat{f}_w(w) = \frac{1}{nh} \sum_{j=1}^{n} K\left\{(\mathbf{W}_j - w)/h\right\},$$

where $K(\cdot)$ is a density function and h is the bandwidth, both chosen by the user. The function $\widehat{f}_w$ is itself a density function, with characteristic function $\widehat{\phi}_w$. It has long been known that for estimation of $f_w(w)$ the choice of kernel is relatively unimportant, and commonly ease of use dictates the choice of $K(\cdot)$, e.g., the

standard normal density or a density with bounded support.

It transpires that for commonly used kernels, the estimated density $\widehat{f}_x(x)$ cannot be deconvolved, in that the integral encountered in Fourier inversion is not defined. Stefanski & Carroll (1987, 1990c) showed that for certain smooth kernels, Fourier inversion of $\widehat{f}_x(x)$ is possible, see also Stefanski (1989). With an appropriately smooth kernel, the estimator,

$$\widehat{f}_x(x) = \frac{1}{2\pi} \int e^{-itx} \frac{\widehat{\phi}_w(t)}{\phi_u(t)} dt,$$

exists, and for suitable choice of bandwidth is consistent for $f_x(x)$. The *deconvoluting kernel density estimator*, $\widehat{f}_x(x)$, integrates to one but is not always positive. It has the alternative representation

$$\widehat{f}_x(x) = \frac{1}{nh} \sum_{j=1}^{n} K_* \left(\frac{\mathbf{W}_j - x}{h}, h \right),$$

where

$$K_*(t, h) = \frac{1}{2\pi} \int e^{ity} \frac{\phi_K(y)}{\phi_u(y/h)} dy$$

is called the deconvoluting kernel.

The deconvoluting kernel density estimator has pointwise mean squared error

$$\begin{aligned} \text{MSE} &= E \left\{ \widehat{f}_x(x) - f_x(x) \right\}^2 \\ &\sim ch^4 + (2\pi hn)^{-1} \int \left\{ \frac{\phi_K(t)}{|\phi_u(t/h)|} \right\}^2 dt; \end{aligned}$$

where $$c = (1/4) \int x^2 K(x) dx \int \left\{ f_x''(x) \right\}^2 dx.$$

The best bandwidth, in the sense of minimizing MSE asymptotically, and the best MSE, depends on the error density through its characteristic function ϕ_u. It is well known that in the absence of measurement error ($\mathbf{U} \equiv 0$), when f_x has two continuous derivatives the best MSE converges to 0 at the rate $n^{-4/5}$. However, for nondegenerate $\mathbf{U}$ convergence rates are much slower in general. The best rate of convergence depends on the tail behavior of $|\phi_u(t)|$, with lighter tails resulting in slower rates of convergence. The tail behavior of $|\phi_u(t)|$ is in turn related to the smoothness of $f_u(u)$ at $u = 0$, with smoother densities having characteristic

functions with lighter tails.

For example, if $\mathbf{U}$ is normally distributed, then

$$|\phi_u(t)| = \exp(-\sigma_u^2 t^2/2)$$

is extremely light tailed, and the mean squared error converges to 0 at a rate no faster than the exceedingly slow rate of $\{\log(n)\}^{-2}$. The implication is that with normally distributed errors, it is not possible to estimate the actual value of $f_x(x)$ well.

If $\mathbf{U}$ has a more peaked density function than the normal, then $|\phi_u(t)|$ does not diminish to 0 as rapidly, and the deconvoluting kernel density estimator has better asymptotic performance. For example, consider the Laplace distribution with density function $f_u(u) = (1/\sigma_u\sqrt{2})\exp(-\sqrt{2}|u|/\sigma_u)$. In this case $\phi_u(t) = 2/(2 + \sigma_u^2 t^2)$, and the optimal mean squared error converges to zero at the rate $n^{-4/9}$, tolerably close to the rate in the absence of measurement error, i.e., $n^{-4/5}$.

The fact that smoothness of the error density determines how well f_x can be estimated is a disconcerting nonrobustness result.

We note that the slow rate of convergence of $\widehat{f}_x(x)$ is intrinsic to the deconvolution problem, and not specific to the deconvoluting kernel density estimator, which is known to achieve the best rate of convergence in general (Carroll & Hall, 1988; Stefanski & Carroll, 1990c).

However, rates of convergence are not always fully informative with regard to the adequacy of $\widehat{f}_x(x)$ for estimating the basic *shape* of $f_x(x)$. As shown in the examples below, despite the slow pointwise rate, the estimator itself can provide useful information about shape.

In applications, calculation of $\widehat{f}_x(x)$ requires specification or estimation of a bandwidth h. Stefanski & Carroll (1990c) describe a bandwidth estimator when the improper sinc kernel, $K(t) = (\pi t)^{-1}\sin(t)$, is used. Stefanski (1990) shows that for a large class of kernels and a large class of error densities that includes the normal densities, the mean squared error is minimized asymptotically by a known sequence of bandwidths — the optimal bandwidth is $h = h_G = \sigma_u \{\log(n)\}^{-1/2}$ for normal (Gaussian) error. For Laplace measurement error and the kernel with characteristic function $\phi_K(t) = (1-t^2)^3$ when $|t| \le 1$ and zero otherwise, Fan, et al. (1991) suggest taking $h_L = (1/2)\sigma_u n^{-1/9}$. The examples below used h_G and h_L according to the assumed form of the error density.

12.1.1 Parametric Deconvolution via Moments

Nonparametric deconvolution is not the only way to estimate the density of $\mathbf{X}$ in an additive model. Instead, one can estimate the first four moments of the distribution of $\mathbf{X}$ making minimal distributional assumptions about $\mathbf{U}$, and then fit a parametric distribution for $\mathbf{X}$ via method of moments.

To be specific, suppose that in a sample of size n, one observes replicate observations $\mathbf{W}_{i,j} = \mathbf{X}_i + \mathbf{U}_{i,j}$ $(i = 1, ..., n$ and $j = 1, 2)$, where it is assumed only that the distribution of the errors is symmetrically distributed about zero, something which can often be achieved by transformation.

Let $\widehat{\mu}_w = \overline{\mathbf{W}}_{\cdot\cdot}$ (the mean), and for $k = 2, 3, 4$ define $\widehat{\kappa}_{w,k}$ to be the sample mean of the terms $\left(\overline{\mathbf{W}}_{i,\cdot} - \widehat{\mu}_w\right)^k$. For $k = 2, 4$ define $\widehat{\kappa}_{u,k}$ to be the sample mean of the terms $\{(\mathbf{U}_{i,1} - \mathbf{U}_{i,2})/2\}^k$. The term $\widehat{\kappa}_{w,k}$ is an estimate of the kth central moment of the $\overline{\mathbf{W}}_{i,\cdot}$'s, while under symmetry $\widehat{\kappa}_{u,k}$ is an estimate of the kth moment of $(\mathbf{U}_{i,1} - \mathbf{U}_{i,2})/2$, which because of symmetry is the same as the kth moment of $(\mathbf{U}_{i1} + \mathbf{U}_{i2})/2 = \overline{W}_{i\cdot} - X_i$.

By equating moments we find the following consistent estimates of the moments of the distribution of $\mathbf{X}$,

$$\begin{aligned}
E(\mathbf{X}) = \mu_x &\approx \widehat{\mu}_w; \\
E(\mathbf{X} - \mu_x)^2 &\approx \widehat{\kappa}_{w,2} - \widehat{\kappa}_{u,2}; \\
E(\mathbf{X} - \mu_x)^3 &\approx \widehat{\kappa}_{w,3}; \\
E(\mathbf{X} - \mu_x)^4 &\approx \widehat{\kappa}_{w,4} - \widehat{\kappa}_{u,4} - 6(\widehat{\kappa}_{w,2} - \widehat{\kappa}_{u,2})\widehat{\kappa}_{u,2}.
\end{aligned}$$

12.1.2 Estimating Distribution Functions

The pessimistic nature of the results for density estimation with normally distributed error extends to estimating quantiles of the distribution of $\mathbf{X}$, e.g., $\mathrm{pr}(\mathbf{X} \leq x)$. Here the *optimal* achievable rate of convergence is of the order $\{\log(n)\}^{-3}$, hardly much of an improvement! This casts doubt on the feasibility of estimating quantiles of the distribution of $\mathbf{X}$ without making parametric assumptions.

There are at least two alternatives to a full-blown likelihood analysis. The moment matching method described previously starts from a model for the density function of $\mathbf{X}$, but makes no

assumptions about the density of **U**. Its output is an estimated density function which yields estimated quantiles.

Alternatively, with no model for the density of **X** but a good model for the error density of **U**, the SIMEX method can be applied. Previous applications of SIMEX have been to estimated parameters and nonparametric regression estimates, but here the basic input is an empirical distribution function (possibly presmoothed).

12.1.3 Optimal Score Tests

While estimating a density function nonparametrically is difficult in the presence of measurement error, estimating smooth functionals of the unknown density, e.g., $m(w) = E(\mathbf{X}|\mathbf{W} = w)$, is often not as difficult.

For estimating $m(w)$, we can simply replace f_x and f_w in (12.1) by their estimators. Stefanski & Carroll (1991) showed that this substitution works, in the sense that the resulting estimate of $m(w)$ when substituted into the score test typically achieves the same local power as if $m(w)$ were a known function.

The reason for this is that $m(w)$ is much easier to estimate than f_x, because of the extra integration in (12.1). In fact, with normally distributed measurement errors, the rate of convergence for estimating $m(w)$ is of order $n^{-4/7}$, while for Laplace error the rate is the usual nonparametric one, i.e., $n^{-4/5}$ (Stefanski and Carroll, 1991).

12.1.4 Framingham Data

We applied deconvoluting kernel density estimation techniques to the Framingham data, for both SBP and transformed SBP, log(SBP-50). We used SBP at Exam #2 only to estimate the measurement error variance, but deconvolved SBP measured at Exam #3 (**W**). In the original scale, observed SBP had mean 130.01, variance 395.65 and the estimated measurement error variance was 83.69. This leads to an estimate of the variance for long-term SBP (**X**) of 311.96, with the ratio of the measurement error variance to that of the underlying variability of long-term SBP estimated as 0.27. In the transformed scale, the corresponding numbers are 4.35, 0.053, 0.013, 0.040 and 0.32, respectively.

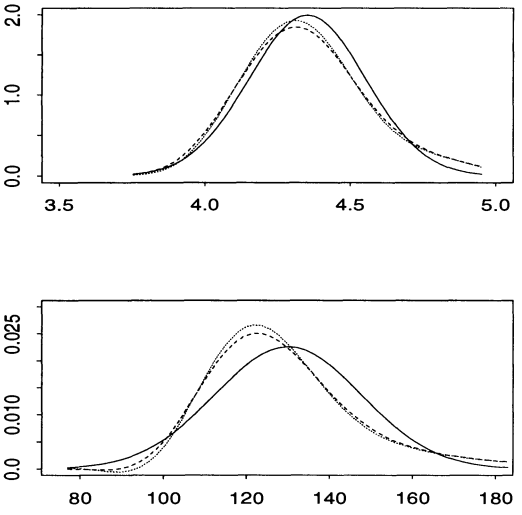

Figure 12.1. *Density estimates of transformed systolic blood pressure (top) and systolic blood pressure (bottom) for Framingham data. Solid line is best-fitting normal, short dashed line is deconvolution with normal errors, and long dashed line is deconvolution with Laplace errors.*

Figure 12.1 shows the two deconvoluting kernel density estimators, one assuming normally distributed errors and the other assuming Laplace errors. Also plotted is the normal distribution with means and variances for $\mathbf{X}$ as estimated above. The two deconvoluting density estimators are similar for the transformed (top plot) and untransformed (bottom plot) data. In the untransformed data, the deconvoluting density estimators differ noticeably from the best-fitting normal density, perhaps because the untransformed data have some skewness. In this case, the deconvoluting kernel density estimators correctly suggest that the data should be transformed.

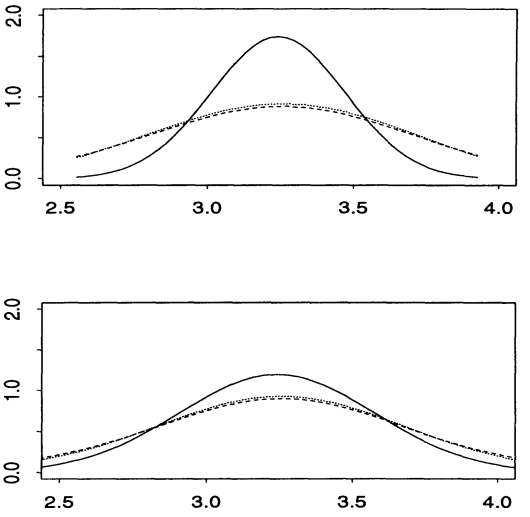

Figure 12.2. *Density estimates of transformed saturated fat for NHANES data. Top uses the estimated measurement error variance, bottom plot sets this variance equal to the variance of* $\mathbf{X}$. *Solid line is best-fitting normal, short dashed line is deconvolution with normal errors and long dashed line is deconvolution with Laplace errors.*

12.1.5 NHANES Data

The NHANES data (Chapter 3) exhibit considerably more measurement error, and consequently deconvolution is much harder. For these data we have earlier derived the variances $\widehat{\sigma}_w^2 = .223$, $\widehat{\sigma}_u^2 = .171$ and $\widehat{\sigma}_w^2 = .052$. We used the same methods as for the Framingham data. The top plot of Figure 12.2 gives the best-fitting normal density, along with the deconvolution density estimates for normal and Laplace errors. The reader will note that the deconvolution densities suggest that the underlying density for $\mathbf{X}$ is *much* heavier tailed than a normal density. This can be confirmed by an analysis of moments, as follows. The sample skewness of $\mathbf{W}$ is nearly zero (-0.05) and is ignored. The sample kurtosis is 3.32, where a kurtosis of 3 applies for the normal distribution. If the kurtosis of $\mathbf{X}$ is denoted by κ_x, then in the additive error model with normally

distributed errors the kurtosis for $\mathbf{W}$ is

$$\kappa_w = \left\{\sigma_x^4(\kappa_x + 3) + 6\sigma_x^2\sigma_u^2 + 3\sigma_u^4\right\}/\sigma_w^4.$$

Substituting sample estimates of $(\kappa_w, \sigma_x^2, \sigma_u^2, \sigma_w^2)$ and solving for κ_x, the kurtosis for $\mathbf{X}$ is estimated to be approximately 8.8, indicating very heavy tails consistent with Figure 12.2.

Also in Figure 12.2, we plot deconvoluting kernel density estimators under the assumption that the measurement error variance is the same as the variance of $\mathbf{X}$. Here $\kappa_x \approx 4.28$, a diminished kurtosis reflected in Figure 12.2.

12.2 Nonparametric Regression

Nonparametric regression has become a rapidly developing field as researchers have realized that parametric regression is not suitable for adequately fitting curves to all data sets that arise in practice. There have been several recent monographs on the topic (Müller, 1988; Härdle,1990; Hastie & Tibshirani, 1990), where it is shown that nonparametric regression techniques have much to offer in applications.

Nonparametric regression entails estimating the mean of $\mathbf{Y}$ as a function of $\mathbf{X}$,

$$E(\mathbf{Y}|\mathbf{X} = x_0) = f(x_0), \tag{12.2}$$

without the imposition of f belonging to a parametric family of functions.

We focus on the local-polynomial, least squares, kernel-regression estimator of f. When $\mathbf{X}$ is observable, the local, order-p polynomial estimator is $\widehat{\beta}_0(x)$, the solution for β_0 to the weighted least squares problem minimizing,

$$\sum_{i=1}^n \left\{Y_i - \beta_0 - \beta_1(\mathbf{X}_i - x) - \ldots - \beta_p(\mathbf{X}_i - x)^p\right\}^2 K_h(\mathbf{X}_i - x). \tag{12.3}$$

Here h is called the *bandwidth*, K is a kernel function such that $\int K(u)\,du = 1$, and $K_h(u) = h^{-1}K(u/h)$. The function $K(\cdot)$ and the bandwidth h are under the control of the investigator, and in practice it is the latter that is the more important.

Problem (12.3) is a straightforward weighted least squares problem, and hence is easily solved numerically. The local least squares

estimator of $f(x)$ is then

$$\widehat{f}(x,h) = \widehat{\beta}_0(x), \tag{12.4}$$

while for $j < p$, the estimator of the jth derivative of $f(x)$ is $j!\widehat{\beta}_j(x)$. Estimator (12.4) has had long use in time series analysis, and is a special case of the robust, local regression estimators in Cleveland (1979). Cleveland & Devlin (1988) discuss practical implementation and present several interesting case studies where local regression data analysis is considerably more insightful than classic linear regression analysis. Ruppert & Wand (1994) describe the asymptotic theory of these estimators.

As in parametric problems, ignoring measurement error causes inconsistent estimation of $f(x)$. The regression calibration and SIMEX methods of Chapters 3 and 4 provide simple means for constructing approximately consistent estimators of $f(x)$ in the case that $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ has mean zero. Hastie & Stuetzle (1989) describe an alternative method for an orthogonal regression problem wherein it is assumed that the conditional variances of $\mathbf{Y}$ and $\mathbf{W}$ given $\mathbf{X}$ are equal; we have already commented (section 2.3.2) on the general applicability of such an assumption.

In this section, we describe algorithms for nonparametric regression taking measurement error into account. Asymptotic theory is beyond the scope of this book and will be described elsewhere in research papers.

12.2.1 SIMEX

Use of SIMEX in nonparametric regression follows the same ideas as in parametric problems. We require an additive error model $\mathbf{W} = \mathbf{X} + \mathbf{U}$ where $\mathbf{U}$ is independent of $\mathbf{X}$ with variance σ_u^2. Sometimes, a transformation of the original surrogate is required to achieve additivity and homoscedasticity. The SIMEX algorithm for nonparametric regression is as follows.

(a) Fix values for $\lambda \in \Lambda = (0 < \lambda_1 < ... < \lambda_M)$.
(b) For $b = 1, ..., B$, let ϵ_{ib} be the non-iid pseudo errors.
(c) Define $W_{ib}(\lambda) = \mathbf{W}_i + \sigma_u \lambda^{1/2} \epsilon_{ib}$.
(d) For $b = 1, ..., B$ and $\lambda \in \Lambda$, compute the nonparametric regression estimate (12.4) by regressing $\mathbf{Y}_i$ on $\mathbf{W}_{ib}(\lambda)$. Call the resulting estimate $\widehat{f}(x, b, \lambda)$.

(e) Let $\widehat{f}(x, \lambda)$ be the sample mean of the terms $\widehat{f}(x, b, \lambda)$.
(f) For each x, extrapolate the values $\widehat{f}(x, \lambda)$ as a function of λ back to $\lambda = -1$, resulting in the SIMEX estimator $\widehat{f}(x)$.

For speed of computation in (d), we used a fixed bandwidth h corresponding to naive regression with $\lambda = 0$, although further research will likely suggest better methods.

12.2.2 Regression Calibration

The regression calibration approximation states that $E(\mathbf{Y}|\mathbf{W}) \approx f\{m(\mathbf{W})\}$, where $m(\mathbf{W}) = E(\mathbf{X}|\mathbf{W})$. Thus, the algorithm has only two steps.

(a) Estimate $m(w)$ by some estimate $\widehat{m}(w)$, see below.
(b) Estimate $f(\cdot)$ by a local linear regression of Y on $\widehat{m}(\mathbf{W})$.

Typically, if h_* is the bandwidth used in naive local linear regression ignoring measurement error, the bandwidth for regression calibration can be taken as h_* times the ratio of the sample standard deviation of $\widehat{m}(\mathbf{W})$ to the sample standard deviation of $\mathbf{W}$ itself. We used this simple algorithm in our calculations.

Because of systematic biases in quadratic and exponential models (section 3.10), use of the expanded approximations of Chapter 3 can be valuable. With additive homoscedastic measurement errors, in the normal case (3.14) becomes

$$E(\mathbf{Y}|\mathbf{W}) \approx f(\cdot) + (1/2)(\sigma_x^2\sigma_u^2/\sigma_w^2)f_{xx}(\cdot).$$

This suggests the corrected estimator

$$\widehat{f}_c = \widehat{f} - (1/2)(\widehat{\sigma}_x^2\widehat{\sigma}_u^2/\widehat{\sigma}_w^2)\widehat{f}_{xx}, \tag{12.5}$$

where $\widehat{f}_{xx} = 2\widehat{\beta}_2$ in the local, cubic regression of $\mathbf{Y}$ on $\widehat{m}(\mathbf{W})$. This estimator does correct for bias, but it adds variability, because it is more difficult to estimate the second derivative of a function then to estimate the function itself.

The expanded regression calibration algorithm has the drawback that it requires two different bandwidths, one to estimate f and the other to estimate f_{xx}. The latter can be particularly difficult in practice. The extra variability and the problems of selecting two bandwidths gives SIMEX an advantage here.

If the error model is that of classical additive measurement error, perhaps after transformation, then the simplest estimates of $m(w)$ are the linear and quadratic regressions described in section 3.4.

These are *global* calibration methods, in the sense that $E(\mathbf{X}|\mathbf{W})$ is estimated parametrically. The obvious potential drawback of this approach is that one assumes a linear or quadratic model adequately describes $m(w)$. In defense of the approach we believe that in many if not most additive error models this is the case, at least for most of the range of $\mathbf{W}$. The quadratic regression was used in the example.

12.2.3 QVF and Likelihood Models

Local linear nonparametric regression is easily extended to likelihood and quasilikelihood and variance function (QVF) models. The reason is that local linear regression can be looked at in two ways that permit immediate generalization. First, as seen in (12.3), local linear regression estimation of $f(x_0)$ at a value x_0 is equivalent to a weighted maximum likelihood estimate of the intercept in the model assuming that $\mathbf{Y}$ is normally distributed with mean $\beta_0+\beta_1(\mathbf{X}-x_0)$, constant variance and with the weights $K_h(\mathbf{X}-x_0)$. Thus, in other generalized linear models (logistic, Poisson, gamma, etc.), the suggestion is to perform a weighted likelihood analysis with a mean of the form $h\{\beta_0 + \beta_1(\mathbf{X} - x_0)\}$ for some function $h(\cdot)$.

Extending local linear nonparametric regression to QVF models is also routine. As seen in (12.4), local linear regression is a weighted QVF estimate based on a model with mean $\beta_0 + \beta_1(\mathbf{X} - x_0)$ and constant variance, and with extra weighting $K_h(\mathbf{X} - x_0)$. The suggestion in general problems is to do the QVF analysis with argument $\beta_0 + \beta_1(\mathbf{X} - x_0)$ and extra weighting $K_h(\mathbf{X} - x_0)$.

12.2.4 Framingham Data

We applied measurement error corrections for nonparametric regression to the Framingham data to estimate coronary heart disease (CHD) incidence from systolic blood pressure. We used local linear logistic regression as described above, with the kernel $K(t) = (3/4)(1 - t^2)$ for $|t| < 1$. Because regression calibration is often remarkably accurate in logistic regression, we use it here without using the expanded model.

In Chapter 7 we indicated that the classical error model holds reasonably well if the transformation $\log(\text{SBP} - 50)$ is used. In

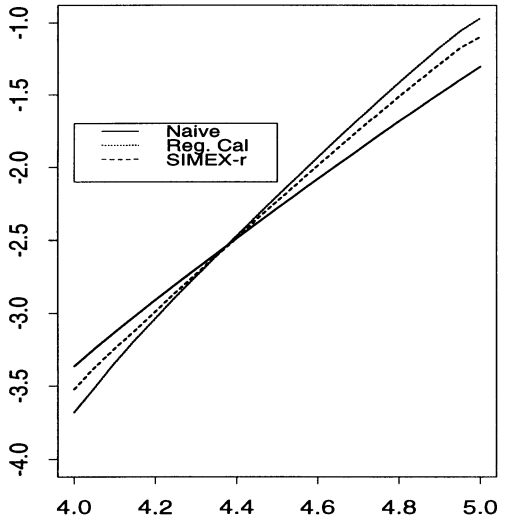

Figure 12.3. *Framingham data. In the transformed log(SBP-50) scale, this is a plot of the logits of the nonparametric regressions of CHD incidence against SBP. The solid line is the naive plot ignoring measurement errors. The dashed lines are the best linear fit regression calibration and SIMEX estimators.*

order to illustrate the behavior of the various estimators, we used transformed SBP at Exam #3 as the surrogate $\mathbf{W}$. We used transformed SBP at Exam #2 only to estimate the measurement error variance σ_u^2, obtaining the estimate $\widehat{\sigma}_u^2 = 0.01259$.

The bandwidth we used for naive local linear regression was $h = 1.2$, chosen to represent about 60% of the range of $\mathbf{W}$ in the data. This may represent oversmoothing, but the choice of bandwidth selection even for the naive local linear regression remains an area of active research development, and beyond the scope of this book. For the SIMEX with rational linear extrapolant and regression calibration estimators, we used bandwidths as described in the definition of the respective techniques.

The results are displayed in the logit scale, $\log\{p/(1-p)\}$, see Figure 12.3 where we plotted the regression estimates on the interval $[4, 5]$, representing SBP ranging from 100 to 200.

At least in this particular example, the SIMEX and regression calibration methods gave about the same answers, and provided a moderate correction to the naive estimator, in keeping with the moderate amount of measurement error in these data.

12.2.5 Other Methods

A globally consistent deconvoluting kernel regression function estimate can be obtained by replacing the kernel in (12.3) with a deconvoluting kernel (Fan & Truong, 1993), resulting in what we refer to as a deconvoluting kernel, local regression estimator.

However, the bandwidth selection problem associated with this approach is by no means trivial, and the rates of convergence for the resulting estimators are the same as for the density estimation problem. In our experience, the deconvoluting kernel, local regression estimators are typically inferior to the regression calibration and SIMEX methods.

A promising alternative approach is to apply regression calibration or SIMEX to generalized additive models (Hastie & Tibshirani, 1990) or roughness penalty estimators (Green & Silverman, 1994).

CHAPTER 13

RESPONSE VARIABLE ERROR

In preceding chapters we have focused exclusively on problems associated with measurement error in predictor variables. In this chapter we consider problems that arise when a true response is measured with error. For example, in a study of factors affecting dietary intake of fat, e.g., sex, race, age, socioeconomic status, etc., true long-term dietary intake is impossible to determine and instead it is necessary to use error-prone measures of long-term dietary intake. Wittes, et al. (1989) describe another example in which damage to the heart muscle caused by a myocardial infarction can be assessed accurately, but the procedure is expensive and invasive, and instead it is common practice to use peak cardiac enzyme level in the bloodstream as a proxy for the true response.

The exclusive attention paid to predictor measurement error in preceding chapters is explained by the fact that predictor measurement error is seldom ignorable, by which is meant that the usual method of analysis is statistically valid, whereas response measurement error is often ignorable. Here "ignorable" means that the model holding for the true response holds also for the proxy response with parameters unchanged, except that a measurement error variance component is added to the response variance. For example, in linear regression models with simple types of response measurement error, the response measurement error is confounded with equation error and the effect is simply to increase the variability of parameter estimates. Thus, response error is ignorable in these cases. However, in more complicated regression models, certain types of response error are not ignorable and it is important to explicitly account for the response error in the regression analysis. This chapter distinguishes between ignorable and nonignorable

cases and describes methods of analysis for the latter.

Although the details differ between methods for predictor error and response error, many of the basic ideas are similar. The main methods for the analysis of response error models are quasilikelihood and variance functions (QVF) and likelihood techniques. For QVF models (section A.4.1), the objective is still to model and estimate the mean and variance functions of the observed data. Likelihood methods for response error models are similar to those of Chapter 7, and there are close analogs to the pseudolikelihood and modified pseudolikelihood methods of Chapter 9.

Throughout this chapter, the response proxy is denoted by $\mathbf{S}$, the true response by $\mathbf{Y}$ and the predictors by $(\mathbf{Z}, \mathbf{X})$. We consider only the case of measurement error in the response, and not the more complex problem where both the response and some of the predictors are measured with error (although the complexity is largely notational). 'We start with a discussion of QVF models with simple measurement error models. Then parametric and semiparametric likelihood methods are discussed. The chapter concludes with an example application of some of the methods.

13.1 Additive/Multiplicative Error and QVF Models

In this section we consider the model (3.8)–(3.9), which is also called a quasilikelihood and variance function model (QVF model), see (A.21)–(A.22) in section A.4.1. We discuss the analysis of the observed data when the response is subject to independent additive or multiplicative measurement error. The basic conclusion is that if $\mathbf{S}$ is unbiased for $\mathbf{Y}$, then for either error model a standard QVF analysis is appropriate after modification of the variance function model. If $\mathbf{S}$ is not unbiased for $\mathbf{Y}$, then a validation study is required to understand the nature of the bias and to correct for it.

13.1.1 Unbiased Measures of True Response

The simplest case to handle is independent additive measurement error in the response. In this case a QVF analysis with the same mean function and a slightly modified variance function is appropriate, and there is no need to obtain replication or validation data.

Suppose that $\mathbf{S} = \mathbf{Y} + \mathbf{V}$, where $\mathbf{V}$ is independent of $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$

with mean zero and variance σ_v^2. Then the mean and variance of $\mathbf{S}$ is given by

$$\begin{aligned} E(\mathbf{S}|\mathbf{Z},\mathbf{X}) &= f(\mathbf{Z},\mathbf{X},\mathcal{B}); \\ \text{var}(\mathbf{S}|\mathbf{Z},\mathbf{X}) &= \sigma_v^2 + \sigma^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta). \end{aligned} \tag{13.1}$$

The addition of σ_v^2 to the variance function indicates that measurement error in the response increases the variability of the observed data, and consequently decreases the precision of parameter estimates.

The case of homoscedastic regression variance, $g \equiv 1$, provides an example in which response measurement error is ignorable. For then the variance function in (13.1) is again constant, $\sigma_v^2 + \sigma^2$, and ordinary nonlinear least squares is an appropriate method of estimation. The only effect of the measurement error is that the residual mean square is estimating $\sigma_*^2 = \sigma_v^2 + \sigma^2$ and not σ^2. Thus unless the separate variance components, σ_v^2 and σ^2, are of independent interest, the response error can be ignored and replication or validation data are not needed.

For heteroscedastic variance functions, $g \neq 1$, identifiability of the parameters in (13.1) depends on the form of g. For example, if g^2 is a constant plus some function of the mean, i.e., $g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta) = \tau^2 + h\{f(\mathbf{Z},\mathbf{X},\mathcal{B}),\theta\}$, then the variance function (13.1) is $\sigma^2(\tau_*^2 + h\{f(\mathbf{Z},\mathbf{X},\mathcal{B}),\theta\})$, where $\tau_*^2 = \sigma_v^2/\sigma^2 + \tau^2$. In this case neither τ^2 nor σ_v^2 are identifiable without replication or validation data, but all of the other parameters in (13.1) are. This is another example in which response measurement error is ignorable provided the variance components τ^2 and σ_v^2 are not of independent interest.

If $g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta) = \{f(\mathbf{Z},\mathbf{X},\mathcal{B})\}^{2\theta}$, i.e., the power-of-the-mean variance model, then the appropriate variance function given by (13.1) is $\sigma_v^2 + \sigma^2\{f(\mathbf{Z},\mathbf{X},\mathcal{B})\}^{2\theta}$, and all of the parameters in (13.1) are identifiable given sufficient data. It is worth noting that the additive variance component lends a certain robustness to the power-of-the-mean variance model. Without this component, use of the power-of-the-mean model can be dangerous, since the estimated variance of some observations may be near zero. In this case, these few observations are given near infinite weight and the other observations are essentially ignored when $\mathcal{B}$ is estimated.

A cautionary remark is in order here. Although for certain variance function models all of the parameters in (13.1) are formally identified, it should be remembered that identification does not

guarantee precise estimation. Since variance function parameter estimates are generally less precise than regression parameter estimates, the ability to isolate variance components, in this case σ_v^2, is limited by small to moderate sample sizes, as well as by the correctness of the assumed model for g.

Now consider the multiplicative measurement error model with $\mathbf{S} = \mathbf{YV}$, where $\mathbf{V}$ has mean 1 and variance σ_v^2, so that $\mathbf{S}$ is still an unbiased measure of the true response. In this case, the data follow the QVF model

$$\begin{aligned} E(\mathbf{S}|\mathbf{Z},\mathbf{X}) &= f(\mathbf{Z},\mathbf{X},\mathcal{B}); \\ \text{var}(\mathbf{S}|\mathbf{Z},\mathbf{X}) &= \sigma_v^2 f^2(\mathbf{Z},\mathbf{X},\mathcal{B}) + (1+\sigma_v^2)\sigma^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta) \\ &= \sigma_v^2 f^2(\mathbf{Z},\mathbf{X},\mathcal{B}) + \sigma_*^2 g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta), \qquad (13.2) \end{aligned}$$

where $\sigma_*^2 = (1+\sigma_v^2)\sigma^2$. The parameters $(\mathcal{B},\theta,\sigma_v^2,\sigma_*^2)$, and hence σ^2, in this QVF model are also formally identifiable in general without replication or validation data.

However, note that for the power-of-the-mean variance function model, $g^2(\mathbf{Z},\mathbf{X},\mathcal{B},\theta) = \{f(\mathbf{Z},\mathbf{X},\mathcal{B})\}^{2\theta}$, and the variance model (13.2) is $\sigma_v^2 f^2(\mathbf{Z},\mathbf{X},\mathcal{B}) + \sigma_*^2 \{f(\mathbf{Z},\mathbf{X},\mathcal{B})\}^{2\theta}$. Thus, when $\theta = 1$ it is only possible to estimate the sum $\sigma_v^2+\sigma_*^2$. Furthermore it is evident that for θ near 1, estimation of both σ_v^2 and σ_*^2 will be difficult. Thus for this model if it is expected that θ is near 1, then it may be necessary to use an approximate variance function model, in this case the power-of-the-mean model, recognizing its limitations.

We have seen that even when the error model for $\mathbf{S}$ given $\mathbf{Y}$ is fully specified, there will often be indeterminacy in the parameters of the variance function, and it may be necessary to settle for getting the variance function only approximately correct. In addition, in some instances, it may only be reliably assumed that $\mathbf{S}$ is unbiased for the true response, without specification of the error structure, i.e., multiplicative, additive, or other.

In such cases where the variance function is only approximate, it is still possible to estimate $\mathcal{B}$. In these situations, QVF estimation is still appropriate, but care must be taken with standard error estimation. The QVF-sandwich method of variance estimation provides asymptotically correct inferences, see section A.4.

13.1.2 Recommendations

There are two strategies that one can follow for additive or multiplicative response error. The first is basically what has been suggested here, namely to model the variance function as best one can, such as we have done in (13.1) and (13.2). One would then use the standard error and inference techniques as described in section A.4.2, working as if the variance function had been essentially correctly specified.

A second approach to modeling and estimation when the true variance function is unknown or only approximately known is to use the naive variance model, i.e., the variance function model that ignores measurement error, or postulate an additive or multiplicative error to develop a working or preliminary variance function model and then collapse the model with respect to parameters that are not identifiable or are nearly nonidentifiable. Then proceed with QVF estimation, with QVF-sandwich-based standard error estimation (section A.4.2).

13.1.3 Biased Responses

If $\mathbf{S}$ is not unbiased for $\mathbf{Y}$, then regression of it on the observed predictors leads to biased estimates of the main regression parameters. For example, suppose $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ follows a normal linear model with mean $\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z}$ and variance σ^2, while $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ follows a normal linear model with mean $\gamma_0 + \gamma_1\mathbf{Y}$ and variance σ_v^2. Here $\mathbf{S}$ is biased, and the observed data follow a normal linear model with mean $\gamma_0 + \beta_0\gamma_1 + \gamma_1\beta_x^t\mathbf{X} + \gamma_1\beta_z^t\mathbf{Z}$ and variance $\sigma_v^2 + \gamma_1^2\sigma^2$. Thus instead of estimating β_x, naive regression ignoring measurement error in the response estimates $\gamma_1\beta_x$.

13.1.4 Calibration

In a series of papers, Buonaccorsi (1991, 1993) and Buonaccorsi & Tosteson (1993) discuss the use of adjustments for a biased surrogate. We describe a modified version of their approach in the simplest possible case, namely that $\mathbf{S}$ is a linearly biased surrogate with mean $\gamma_0 + \gamma_1\mathbf{Y}$. If $\gamma = (\gamma_0, \gamma_1)$ were known, then the recommended procedure is to replace $\mathbf{S}$ by its adjusted value $\mathbf{Q}(\gamma) = (\mathbf{S} - \gamma_0)/\gamma_1$, and then proceed as in section 13.1.1.

When γ is unknown, it has to be estimated and appropriate methods of analysis depend on the data available for estimation of γ. For example, suppose that validation data are available on a simple random subsample of the primary data. The validation subsample data can be used to obtain estimates of $\mathcal{B}$ and γ, denoted $\widehat{\mathcal{B}}_1$ and $\widehat{\gamma}$. A second estimate of $\mathcal{B}$, $\widehat{\mathcal{B}}_2$, can be obtained via a QVF analysis of $\mathbf{Q}(\widehat{\gamma})$ on $(\mathbf{Z}, \mathbf{X})$ using all of the data.

The two estimates of $\mathcal{B}$ can then be combined to obtain a final and more efficient estimate. Suppose that the two estimators have a joint asymptotic covariance matrix Σ estimated by $\widehat{\Sigma}$. The best weighted combination of the two estimates is

$$(J^t\Sigma^{-1}J)^{-1}J^t\Sigma^{-1}(\widehat{\mathcal{B}}_1^t, \widehat{\mathcal{B}}_2^t)^t,$$

where $J = (I, I)$ and I is the identity matrix with the same number of rows as there are elements in $\mathcal{B}$. This best weighted combination is estimated by replacing Σ with $\widehat{\Sigma}$, resulting in

$$\widehat{\mathcal{B}} = (J^t\widehat{\Sigma}^{-1}J)^{-1}J^t\widehat{\Sigma}^{-1}(\widehat{\mathcal{B}}_1^t, \widehat{\mathcal{B}}_2^t)^t.$$

An estimate of the asymptotic covariance matrix of $\widehat{\mathcal{B}}$ is given by $(J^t\widehat{\Sigma}^{-1}J)^{-1}$.

The estimate of Σ required for this procedure can be obtained by application of delta-method techniques, but the resulting estimate is somewhat complicated. The bootstrap is ideally suited to the task of estimating Σ and is recommended on the basis of simplicity.

The method described above can be extended to the case where validation data on $\mathbf{Y}$ are impossible to obtain, but it is possible to obtain independent replicate unbiased measurements of $\mathbf{Y}$, $(\mathbf{S}_{1*}, \mathbf{S}_{2,*})$ on a simple random subsample of the primary data. Note that these unbiased replicates are in addition to the biased surrogate $\mathbf{S}$ measured on the complete sample.

In this case, $\widehat{\mathcal{B}}_1$ is obtained by a QVF analysis of $(\mathbf{S}_{1,*} + \mathbf{S}_{2,*})/2$ on $(\mathbf{Z}, \mathbf{X})$, while γ can be estimated using appropriate linear measurement error model techniques described in Chapter 2, because the replication data follow the model,

$$\begin{aligned} \mathbf{S} &= \gamma_0 + \gamma_1\mathbf{Y} + \mathbf{V}; \\ \mathbf{S}_{j,*} &= \mathbf{Y} + \mathbf{U}_{j,*} \text{ for } j = 1, 2, \end{aligned}$$

where $\mathbf{U}_{1,*}$ and $\mathbf{U}_{2,*}$ are independent with mean zero. This is a linear regression measurement error model with response $\mathbf{S}$ and "true covariate" $\mathbf{Y}$ and with replicate measurements $\mathbf{S}_{1,*}$ and $\mathbf{S}_{2,*}$

of $\mathbf{Y}$. The methods reviewed in Chapter 2 are used to estimate γ. Then $\mathbf{Q}(\widehat{\gamma})$ is constructed and $\widehat{\mathcal{B}}_2$ employed as described previously.

13.2 Likelihood Methods

13.2.1 General Likelihood Theory and Surrogates

Let $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma)$ denote the density or mass function for $\mathbf{S}$ given $(\mathbf{Y},\mathbf{Z},\mathbf{X})$. We will call $\mathbf{S}$ a *surrogate response* if its distribution depends only on the true response, i.e., $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma) = f_{\mathbf{S}|\mathbf{Y}}(s|y,\gamma)$. In both the additive and multiplicative error models, $\mathbf{S}$ is a surrogate. This definition of a surrogate response is the natural counterpart to a surrogate predictor, because it implies that all the information in the relationship between $\mathbf{S}$ and the predictors is explained by the underlying response.

In general, i.e., for a possibly nonsurrogate response, the likelihood function for the observed response is

$$f_{\mathbf{S}|\mathbf{Z},\mathbf{X}}(s|z,x,\mathcal{B},\gamma) = \int f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x,\mathcal{B}) f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma) d\mu(y). \quad (13.3)$$

If $\mathbf{S}$ is a surrogate, then $f_{\mathbf{S}|\mathbf{Y}}(s|y,\gamma)$ replaces $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma)$ in (13.3) showing that if there is no relationship between the true response and the predictors, then neither is there one between the observed response and the predictors. The reason for this is that under the stated conditions, neither term inside the integral depends on the predictors, the first because $\mathbf{Y}$ is not related to $(\mathbf{Z},\mathbf{X})$, and the second because $\mathbf{S}$ is a surrogate. However, if $\mathbf{S}$ is *not* a surrogate, then there may be no relationship between the true response and the covariates, but the observed response may be related to the predictors.

It follows that if interest lies in determining whether the predictors contain any information about the response, one can use naive hypothesis tests and ignore response error only if $\mathbf{S}$ is a surrogate. The resulting tests have asymptotically correct level, but decreased power relative to tests derived from true response data. This property of a surrogate is important in clinical trials, see Prentice (1989).

Note that one implication of (13.3) is that a likelihood analysis with mismeasured responses requires a model for the distribution

of response error.

Just as in the predictor-error problem, it is sometimes, but not always, the case that the parameters $(\mathcal{B}, \gamma)$ are identifiable, i.e., can be estimated from data on $(\mathbf{S}, \mathbf{Z}, \mathbf{X})$ alone. An example of the latter is the linear regression example of section 13.1, where the observed responses have mean $\gamma_0 + \beta_0\gamma_1 + \gamma_1\beta_x^t\mathbf{X} + \gamma_1\beta_z^t\mathbf{Z}$ and variance $\sigma_v^2 + \gamma_1^2\sigma^2$. As described above, because $\mathbf{S}$ is assumed to be a surrogate in this example, if there is no effect of predictors in the underlying true-data model ($\beta_x = \beta_z = 0$), then there is no effect of predictors in the observed-data model.

This example also shows that because $\mathbf{S}$ is a biased surrogate, only the product $\gamma_1\beta_x$, and not β_x, can be estimated from the observed data. Thus estimation of β_x requires either knowledge of γ_1 or sufficient data to estimate this parameter.

We now suppose that there is a validation subsample obtained by measuring $\mathbf{Y}$ on units in the primary sample selected with probability $\pi(\mathbf{S}, \mathbf{Z}, \mathbf{X})$. The presence (absence) of validation data on a primary-sample unit is indicated by $\Delta = 1$ (0). Then, based on a primary sample of size n, the likelihood of the observed data for a general proxy $\mathbf{S}$ is

$$\prod_{i=1}^{n} \Big[\{f(\mathbf{S}_i|\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \gamma) f(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})\}^{\Delta_i} \times \{f(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}, \gamma)\}^{1-\Delta_i} \Big], \tag{13.4}$$

where we have dropped the subscripts on the density functions for brevity.

The model for the distribution of $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ is a critical component of (13.4). If $\mathbf{S}$ is discrete, then one approach is to model this conditional distribution by a polytomous logistic model. For example, suppose the levels of $\mathbf{S}$ are $(0, 1, ..., \mathcal{S})$. A standard logistic model is

$$\text{pr}(\mathbf{S} \geq s|\mathbf{Y}, \mathbf{Z}, \mathbf{X}) = H(\gamma_{0s} + \gamma_1\mathbf{Y} + \gamma_2^t\mathbf{X} + \gamma_3^t\mathbf{Z}), \quad s = 1, ..., \mathcal{S}.$$

When $\mathbf{S}$ is not discrete, a simple strategy is to categorize it into $\mathcal{S}$ levels, and then use the logistic model above.

As described above, likelihood analysis is in principle straightforward. However, there are two drawbacks to this approach. First is the obvious one of requiring a model for the distribution of $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ and the attendant robustness issues. Second

is the numerical integration or summation required to compute $f(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}, \gamma)$ from (13.3).

13.2.2 Use of Complete Data Only

Section 9.1 describes methods for using only the complete (or validation) data when predictors are subject to error. As stated there, use of only validation data means that one need not worry about robustness issues arising from the modeling of the distribution of $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$, but at the cost of reduced efficiency for estimating the regression parameters. With response error, the analog to (9.1) is the likelihood of the validation data, given by

$$f(\mathbf{Y}, \mathbf{S}|\mathbf{Z}, \mathbf{X}, \Delta = 1) = \frac{\pi(\mathbf{S}, \mathbf{Z}, \mathbf{X}) f(\mathbf{S}|\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \gamma) f(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \mathcal{B})}{\int \pi(s, \mathbf{Z}, \mathbf{X}) f(s|y, \mathbf{Z}, \mathbf{X}, \gamma) f(y|\mathbf{Z}, \mathbf{X}, \mathcal{B}) d\mu(s) d\mu(y)}. \quad (13.5)$$

If selection into the second stage of the study depends only on the predictors and not on $\mathbf{S}$, then the joint likelihood has the following properties: (i) the denominator of (13.5) equals $\pi(\mathbf{Z}, \mathbf{X})$ which cancels the same term in the numerator; (ii) the likelihood factors into a product of terms involving γ only and terms involving $\mathcal{B}$ only; and (iii) valid estimates of $\mathcal{B}$ can be obtained from the complete $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ data only.

In particular, if selection into the validation study is by simple random sampling, i.e., $\pi(\mathbf{S}, \mathbf{Z}, \mathbf{X})$ is a constant, use of only the completed data is valid.

In general, (13.5) cannot be simplified, and in particular, regression of $\mathbf{Y}$ on $(\mathbf{Z}, \mathbf{X})$ is not valid if selection into the second stage depends on $\mathbf{S}$. Tosteson & Ware (1990) note an important exception, namely when: (a) $\mathbf{S}$ is a surrogate; (b) $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ follows a logistic regression model; and (c) selection into the second stage depends only on $\mathbf{S}$. It can be shown that when (a)–(c) hold, regression of $\mathbf{Y}$ on $(\mathbf{Z}, \mathbf{X})$ in the validation data alone is valid.

Especially for discrete responses, it is sometimes useful to consider the likelihood conditioned also on the value of $\mathbf{S}$,

$$f(\mathbf{Y}|\mathbf{S}, \mathbf{Z}, \mathbf{X}, \Delta = 1) = \frac{f(\mathbf{S}|\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \gamma) f(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \mathcal{B})}{\int f(\mathbf{S}|\mathbf{Y} = y, \mathbf{Z}, \mathbf{X}, \gamma) f(\mathbf{Y} = y|\mathbf{Z}, \mathbf{X}, \mathcal{B}) d\mu(y)}. \quad (13.6)$$

In the appendix it is shown that this does not depend on the selec-

tion probabilities. An important special case is logistic regression, where (13.6) takes the interesting form of a logistic regression with "offsets,"

$$\text{pr}(\mathbf{Y} = 1|\mathbf{S}, \mathbf{Z}, \mathbf{X}, \Delta = 1) = H\left\{q(\mathbf{S}, \mathbf{Z}, \mathbf{X}) + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}\right\},$$

where

$$q(\mathbf{S}, \mathbf{Z}, \mathbf{X}) = \beta_0 + \log\left\{f(\mathbf{S}|\mathbf{Y} = 1, \mathbf{Z}, \mathbf{X})/f(\mathbf{S}|\mathbf{Y} = 0, \mathbf{Z}, \mathbf{X})\right\}.$$

For the case that $\mathbf{S}$ is a discrete surrogate taking on the values $\mathbf{S} = 0, 1, ..., \mathcal{S}$, Tosteson & Ware (1990) suggest estimating $q(s, z, x) = q(s)$ by logistic regression with dummy variables for each of the values of $\mathbf{S}$. Of course, if $\mathbf{S}$ is discrete, as described previously an alternative approach is to construct a model for $\mathbf{S}$ given $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ and maximize the likelihood (13.4).

13.2.3 Other Methods

In some problems, it can occur that there are two data sets, a primary one in which $(\mathbf{S}, \mathbf{Z}, \mathbf{X})$ are observed ($\Delta = 0$), and an *independent* data set in which $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ are observed ($\Delta = 1$). This may occur when $\mathbf{Y}$ is a sensitive endpoint such as income, and $\mathbf{S}$ is reported income. Because of confidentiality concerns, it might be impossible to measure $\mathbf{Y}$ and $\mathbf{S}$ together. In such problems, the likelihood is

$$\prod_{i=1}^{n} \left\{f(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})\right\}^{\Delta_i} \left\{f(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}, \gamma)\right\}^{1-\Delta_i}.$$

13.3 Semiparametric Methods

13.3.1 Pseudolikelihood—Simple Random Subsampling

Suppose that selection into the second stage validation study is by simple random sampling. The similarity between (13.3) and (13.4) with the likelihood functions (7.10) and (7.9) for error–prone predictors led Pepe (1992) to construct a pseudolikelihood similar in spirit to that of Carroll & Wand (1991) and Pepe & Fleming (1991). The basic idea is to use the validation data to form a nonparametric estimator $\widehat{f}_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}$ of $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}$. One then substitutes this estimator into (13.3) to obtain an estimator $\widehat{f}_{\mathbf{S}|\mathbf{Z},\mathbf{X}}(s|z, x, \mathcal{B})$ and

then maximizes

$$\prod_{i=1}^{n} \{f(\mathbf{Y}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})\}^{\Delta_i} \left\{\widehat{f}(\mathbf{S}_i|\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})\right\}^{1-\Delta_i}$$

This approach requires an estimator of $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}$. If all the random variables are discrete, the nonparametric estimator of the probability that $\mathbf{S} = s$ given $(\mathbf{Y}, \mathbf{Z}, \mathbf{X}) = (y, z, x)$ is the fraction in the validation study which have $\mathbf{S} = s$ among those with $(\mathbf{Y}, \mathbf{Z}, \mathbf{X}) = (y, z, x)$, although as previously stated we prefer flexible parametric models in this case. Problems which have continuous components of $(\mathbf{S}, \mathbf{Y}, \mathbf{Z}, \mathbf{X})$ are more complicated. For example, suppose that $\mathbf{S}$ is continuous, but the other random variables are discrete. Then the density function of $\mathbf{S}$ in *each* of the cells formed by the various combinations of $(\mathbf{Y}, \mathbf{Z}, \mathbf{X})$ must be estimated. Even in the simplest case that there is no $\mathbf{Z}$ and $(\mathbf{Y}, \mathbf{X})$ are binary, this means estimating four density functions using validation data only. While the asymptotic theory of such a procedure has been investigated (Pepe, 1992), we know of no numerical evidence indicating that the density estimation methods will work adequately in finite samples, nor is there any guidance on the practical problems of bandwidth selection and dimension reduction when two or more components of $(\mathbf{S}, \mathbf{Y}, \mathbf{Z}, \mathbf{X})$ are continuous.

In practice, if $\mathbf{S}$ is not already naturally categorical, then an alternative strategy is to perform such categorization, fit a flexible logistic model to the distribution of $\mathbf{S}$ given the other variables, and maximize the resulting likelihood (13.4).

13.3.2 Modified Pseudolikelihood—Other Types of Subsampling

Just as in section 9.4, pseudolikelihood can be modified when selection into the second stage of the study is not by simple random sampling. As in section 9.4, the estimating equations for the EM–algorithm maximizing (13.4) are

$$\begin{aligned} 0 \quad = \quad & \sum_{i=1}^{n} \Delta_i \{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) + \Psi_2(\mathbf{S}_i, \mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \gamma)\} \\ & + \sum_{i=1}^{n} (1 - \Delta_i) E\{\Psi_1(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}) \\ & + \Psi_2(\mathbf{S}_i, \mathbf{Y}_i, \mathbf{Z}_i, \mathbf{X}_i, \gamma)|\mathbf{S}_i, \mathbf{Z}_i, \mathbf{X}_i\}, \end{aligned}$$

where

$$\begin{aligned}\Psi_1 &= ((\partial/\partial\mathcal{B})\log(f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}})^t,\ 0^t)^t,\\ \Psi_2 &= (0^t,(\partial/\partial\gamma)\log(f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}})^t)^t.\end{aligned}$$

The idea is to use the validation data to estimate

$$E\{\Psi_1(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{X}_i,\mathcal{B})|\mathbf{S}_i,\mathbf{Z}_i,\mathbf{X}_i\}$$

and then solve

$$\begin{aligned}0 = \sum_{i=1}^{n}[&\Delta_i\Psi_1(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{X}_i,\mathcal{B})+\\ &(1-\Delta_i)\widehat{E}\{\Psi_1(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{X}_i,\mathcal{B})|\mathbf{S}_i,\mathbf{Z}_i,\mathbf{X}_i\}\Big].\end{aligned}$$

For example, suppose that $(\mathbf{S},\mathbf{Z},\mathbf{X})$ are all discrete. Now define I_{ij} to equal one when $(\mathbf{S}_j,\mathbf{Z}_j,\mathbf{X}_j) = (\mathbf{S}_i,\mathbf{Z}_i,\mathbf{X}_i)$ and zero otherwise. Then

$$\widehat{E}\{\Psi_1(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{X}_i,\mathcal{B})|\mathbf{S}_i,\mathbf{Z}_i,\mathbf{X}_i\} = \frac{\sum_{j=1}^{n}\Delta_j\Psi_1(\mathbf{Y}_j,\mathbf{Z}_j,\mathbf{X}_j,\mathcal{B})I_{ij}}{\sum_{j=1}^{n}\Delta_j I_{ij}}.$$

In other cases, nonparametric regression can be used. In the discrete case, Pepe, et al. (1994) derive an estimate of the asymptotic covariance matrix of $\widehat{\mathcal{B}}$ as $A^{-1}(A+B)A^{-t}$, where

$$\begin{aligned}A &= -\sum_{i=1}^{n}\Delta_i(\partial/\partial\mathcal{B}^{\mathcal{T}})\Psi_1(\mathbf{Y}_i,\mathbf{Z}_i,\mathbf{X}_i,\widehat{\mathcal{B}})\\ &\quad -\sum_{i=1}^{n}(1-\Delta_i)\frac{\sum_{j=1}^{n}\Delta_j(\partial/\partial\mathcal{B}^{\mathcal{T}})\Psi_1(\mathbf{Y}_j,\mathbf{Z}_j,\mathbf{X}_j,\widehat{\mathcal{B}})I_{ij}}{\sum_{j=1}^{n}\Delta_j I_{ij}};\\ B &= \sum_{s,z,x}\frac{n(s,z,x)n_2(s,z,x)}{n_1(s,z,x)}r(s,z,x,\widehat{\mathcal{B}}),\end{aligned}$$

$n_1(s,z,x)$, $n_2(s,z,x)$, and $n(s,z,x)$ are the number of validation, nonvalidation and total cases with $(\mathbf{S},\mathbf{Z},\mathbf{X}) = (s,z,x)$, and where $r(s,z,x,\widehat{\mathcal{B}})$ is the sample covariance matrix of $\Psi_1(\mathbf{Y},\mathbf{Z},\mathbf{X},\widehat{\mathcal{B}})$ computed from observations with $(\Delta,\mathbf{S},\mathbf{Z},\mathbf{X}) = (1,s,z,x)$.

13.4 Example

In this section, we present an example where selection into the validation study depends on the proxy, **S**. We compare the valid

Validation Data **Z**	**S**	**Y**	Count
0	0	0	19
0	0	1	5
0	1	0	7
0	1	1	14
1	0	0	28
1	0	1	27
1	1	0	8
1	1	1	24
Nonvalidation Data			
0	0	–	47

Table 13.1. *GVHD data set. Here* $\mathbf{Y} = 1$ *if the patient develops chronic GVHD and* $= 0$ *otherwise, while* $\mathbf{S} = 1$ *if the patient develops acute GVHD. The predictor* $\mathbf{Z} = 1$ *if the patient is aged 20 or greater, and zero otherwise.*

modified pseudolikelihood estimate with the naive use of the complete data. The later is not valid and appears to be seriously biased in this case.

Pepe (1992) and Pepe, et al. (1994) describe a study of 179 aplastic anemia patients given bone marrow transplants. The objective of the analysis is to relate patient age to incidence of chronic graft versus host disease (GVHD). Patients who develop acute GVHD, which manifests itself early in the post-transplant period are at high risk of developing chronic GVHD. Thus, in this example $\mathbf{Y}$ is chronic GVHD, $\mathbf{S}$ is acute GVHD, and $\mathbf{Z} = 0, 1$ depending on whether a patient is less than 20 years of age or not. The data are given in Table 13.1. A logistic regression model for $\mathbf{Y}$ given $\mathbf{Z}$ is assumed.

The selection process as described by Pepe, et al. (1994) is to select only $1/3$ of low risk patients (less than 20 years old and no acute GVHD) into the validation study, while following all other patients. Thus, $\pi(\mathbf{S}, \mathbf{Z}) = 1/3$ if $\mathbf{S} = 0$ and $\mathbf{Z} = 0$, otherwise $\pi(\mathbf{S}, \mathbf{Z}) = 1$. Note that here selection into the validation study

	CDO	MP
$\widehat{\beta}_x$	0.66	1.13
Std. Err.	0.37	0.39
p-value	0.078	0.004

Table 13.2. *Analysis of GVHD data set. CDO, complete data only; MP, modified pseudolikelihood.*

depends on both $\mathbf{S}$ and $\mathbf{Z}$, so that an ordinary logistic regression analysis on the completed data ($\Delta = 1$) will be invalid.

We performed the following analyses: (i) use of complete data only, which is not valid in this problem because of the nature of the selection process, but is included for comparison, and (ii) modified pseudolikelihood.

The results of various analyses are listed in Table 13.2. We see that the complete-data analysis is badly biased relative to the valid, modified pseudolikelihood analysis, with markedly different significant levels.

CHAPTER 14

OTHER TOPICS

This chapter gives an overview of some topics which have not been covered, namely case-control studies, mixture methods for functional models, differential measurement error, design of two-stage studies, misclassification when all variables are discrete, and survival analysis.

14.1 Logistic Case-Control Studies

A *prospective* study is the usual kind, where subjects are randomly selected from the population. Selection may or may not depend on the covariates, but is independent of the response, and often selection occurs before the response is even observable, e.g., before a disease develops. A *retrospective study* is one in which sampling is conditioned on the response; it is useful to think that the response is first observed and only later are the predictors observed. These are called *case-control studies* in epidemiology and *choice-based samples* in econometrics; we will use the former terminology and concentrate on logistic regression models.

A distinguishing feature of case-control studies is that the measurement error may be differential; see section 1.6 for a definition. With the exception of the linear regression model in which the errors were correlated (section 2.3), this book has concentrated on nondifferential measurement error. Differential measurement error is discussed in section 14.2.

14.1.1 The Case that $\mathbf{X}$ *is Observed*

In a classical case-control study, $\mathbf{Y} = 1$ is called a "case", and $\mathbf{Y} = 0$ is a "control". Having observed case or control status, one observes $(\mathbf{Z}, \mathbf{X})$ in a random sample of controls and a random sam-

ple of cases.

Throughout, we will assume that if the data could be observed prospectively, then it would follow a logistic model:

$$\Pr(\mathbf{Y} = 1|\mathbf{Z}, \mathbf{X}) = H\left\{\beta_0^* + R(\mathbf{X}, \mathbf{Z}, \beta_x, \beta_z)\right\}. \tag{14.1}$$

The linear logistic model is the special case $R(\mathbf{X}, \mathbf{Z}, \beta_x, \beta_z) = \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z}$. Weinberg & Wacholder (1993) introduce the general model and discuss its importance.

Starting from (14.1), Prentice & Pyke (1979) and Weinberg & Wacholder (1993) show that when analyzing a classical case-control study one can ignore the case-control sampling scheme entirely. The one exception is that the intercept β_0^* cannot be estimated, because it depends on the underlying rate $\text{pr}(\mathbf{Y} = 1)$ in the source population, which is often unavailable. Furthermore, these authors show that if one *ignores the case-control sampling scheme and runs an ordinary logistic regression*, then the estimates $(\widehat{\beta}_x, \widehat{\beta}_z)$ that result are consistent and the standard errors are asymptotically correct.

14.1.2 Measurement Error

The effect of measurement error in logistic case-control studies is to bias (asymptotically) the estimates of the slopes (β_x, β_z). The question is how to correct for this bias.

In one sense, the problem of correcting for the bias is easily solved. Carroll, Wang & Wang (1995) show that for many problems, one can ignore the case-control study design and proceed as if one were analyzing a random sample from a population. This result can be stated (fairly loosely) as follows:

"Theorem" *In most problems, a prospective analysis which ignores the case-control study design leads to consistent estimates of* (β_x, β_z). *When it does, the standard errors derived from the prospective analysis are usually asymptotically correct, and they are at worst conservative (too large). Thus, in general, no new software is required to analyze case-control studies in the presence of measurement error or missing data.*

With nondifferential measurement error, this result applies to the conditional score methods of section 6.4, the likelihood method of section 7.4, the SIMEX method, and even a slight modification

of regression calibration, see Carroll, Gail & Lubin (1993).

14.1.3 Normal Discriminant Model

Michalek & Tripathi (1980), Armstrong, et al. (1989) and Buonaccorsi (1990b) consider the normal discriminant model. The latter's treatment is comprehensive, based on the model that given $\mathbf{Y} = y$, $(\mathbf{Z}, \mathbf{X}, \mathbf{W})$ has a multivariate normal distribution with mean μ_y and constant covariance matrix. Prospectively, such data lead to a logistic regression model, although prospective logistic models may hold even when the normal discriminant model fails. Buonaccorsi shows how differential and nondifferential measurement error models can be obtained as special cases of his discriminant model. He also shows how to compute maximum likelihood estimates for the parameters using all the data, and not just the complete data, and he describes an asymptotic theory.

14.2 Differential Measurement Error

Differential measurement error (section 1.6) means that $\mathbf{W}$ is no longer a surrogate. Differential measurement error poses special difficulties, both in technical details and in problem formulation. We know of no methods for differential measurement error in nonlinear models which do not require that $\mathbf{X}$ be observable in some subset of the study data. In linear models, differential measurement error can be overcome by method of moments techniques (section 2.3).

To appreciate the technical issues, remember that even in linear regression, differential measurement error means that regression calibration yields inconsistent estimates of the regression parameters.

14.2.1 Likelihood Formulation

The likelihood for differential measurement error differs slightly from (7.5), because the error density depends on the values of the response. Using the same notation as section 7.3, namely that we have measures $\mathbf{W}$ and $\mathbf{T}$, we will write the error model to have a density or mass function which we will denote by $f_{W,T|Y,Z,X}(w, t| y, z, x, \tilde{\alpha}_1)$. The density or mass function of $\mathbf{X}$ given $\mathbf{Z}$ will be denoted by $f_{X|Z}(x|z, \tilde{\alpha}_2)$. These densities depend on the unknown

parameter vectors $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$. The joint density of $(\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ is

$$f_{Y,X,W,T|Z}(y,x,w,t|z,\mathcal{B},\tilde{\alpha}_1,\tilde{\alpha}_2)$$
$$= f_{Y|Z,X}(y|z,x,\mathcal{B}) f_{W,T|Y,Z,X}(w,t|y,z,x,\tilde{\alpha}_1) f_{X|Z}(x|z,\tilde{\alpha}_2).$$

The density or mass function of $(\mathbf{Y}, \mathbf{W}, \mathbf{T})$ given $\mathbf{Z}$ is thus

$$f_{Y,W,T|Z}(y,w,t|z,\mathcal{B},\tilde{\alpha}_1,\tilde{\alpha}_2)$$
$$= \int f_{Y,X,W,T|Z}(y,x,w,t|z,\mathcal{B},\tilde{\alpha}_1,\tilde{\alpha}_2) d\mu(x), \quad (14.2)$$

where, as before, the notation $d\mu(x)$ indicates that the integrals are sums if $\mathbf{X}$ is discrete and integrals if $\mathbf{X}$ is continuous. In a two-stage study, where $\mathbf{X}$ is observed with probability depending only on $(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{T})$, the overall likelihood is proportional to

$$\prod_{i=1}^{n} \{f_{Y,X,W,T|Z}(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{W}_i, \mathbf{T}_i|\mathbf{Z}_i, \mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2)\}^{\Delta_i}$$
$$\times \prod_{i=1}^{n} \{f_{Y,W,T|Z}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{T}_i|\mathbf{Z}_i, \mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2)\}^{1-\Delta_i}.$$

14.2.2 Functional Methods in Two-Stage Studies

In a two-stage study, one observes $\mathbf{X}$ in a randomly chosen subsample; see section 9.5. Several methods have been proposed for applying functional methods to two-stage studies, and we discuss them in this section. Although classical functional models treat the $\mathbf{X}_i$'s as fixed, in section 1.2 we define a functional model as one where the $\mathbf{X}_i$'s are iid from a distribution F_X that is *not* parametrically modeled. However, methods designed for this situation are also appropriate when the $\mathbf{X}_i$'s are fixed parameters.

Carroll, Wang & Wang (1995) modify the general unbiased estimating equation methods discussed in section 9.5, based on specifying a parametric form for the error distribution of $(\mathbf{W}, \mathbf{T})$ given $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$. Robins, Rotnitzky & Zhao (1994) also modify the estimating equation methods, but they do so in such a way that the error distribution need not be parameterized; loss of efficiency may occur because of this. Implementation of these methods requires one to specify a function which is called $\Omega(\cdot)$ in section 9.5.

For illustration, we consider the second method. Define $\mathbf{L} =$

$(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{T})$, and suppose the data are obtained in a two-stage study in which $\mathbf{X}$ is observed with probability $\pi(\mathbf{L})$. Let $\psi(Y, \mathbf{Z}, \mathbf{X}, \mathcal{B})$ be an estimating function for the parameter $\mathcal{B}$. Then the functional estimating equation taking into account the two stage study is

$$\frac{\Delta\psi(\cdot) + \{\Delta - \pi(\mathbf{L})\}\, E\{\psi(\cdot)|\mathbf{L}\}}{\pi(\mathbf{L})}. \tag{14.3}$$

Implementation of this estimating function requires that one estimates functionals of $\mathbf{X}$ given $\mathbf{L}$. Robins, et al. (1994) and Zhao, Lipsitz & Lew (1994) propose different nonparametric ways to do this. Note that to even talk about such functionals, we need the assumption that $\mathbf{X}$ is random, not a fixed parameter.

14.2.3 Comparison of Functional and Likelihood Approaches

The essential differences between the likelihood and functional approaches in this missing data context can be seen by studying (14.2) and (14.3). The likelihood approach requires one to specify parametric models for the distribution of the errors and the distribution of $\mathbf{X}$ given $\mathbf{Z}$. However, once this is done, nothing need be known about the missing data mechanism. In contrast, the functional modeling approaches need not specify the indicated distributions, but they do require a model for the missing data mechanism. Depending on the context, one (distributions) or the other (missing data mechanism) may be more convenient to model.

14.3 Mixture Methods as Functional Modeling

14.3.1 Overview

When there are no covariates measured without error, the nonlinear measurement error problem can be viewed as a special case of what are called mixture problems; see Kiefer & Wolfowitz (1956), Laird (1978), Lindsay (1983), and Titterington, Smith & Makov (1985). Applications of nonparametric mixture methods to nonlinear measurement error models have only recently been described by Thomas, Gauderman & Kerber (1993) and Roeder, Carroll & Lindsay (1996); see also Thomas, Stram & Dwyer (1993).

The basic idea behind these methods is simple, but so far has only been worked out in detail when there are no covariates which

are measured exactly (although see Roeder, et al. for a suggestion when there such covariates). What is done is to approximate the distribution of the unknown $\mathbf{X}$ by a discrete distribution with $m \leq n$ points of positive probability (these are called the *support* points). Both the location of the support points and the probabilities attached to them are estimated. One possibility is to use the EM-algorithm (Titterington, et al., 1985), but when estimating the locations of support points and the values of the distribution the EM algorithm can be very slow. In this instance, gradient methods are often useful, see Lesperance & Kalbfleisch (1992) for a recent example. Lesperance (1989) discusses inference. The following material is fairly technical and can be skipped at first reading.

14.3.2 Nonparametric Mixture Likelihoods

First consider the case that $\mathbf{X}$ is not observed, but that one has a model for the error distribution. There is no specific restriction that the error be additive, multiplicative, etc., but a model is necessary. We have already covered a variety of functional and structural modeling techniques for this problem, including regression calibration (Chapter 3), SIMEX (Chapter 4), conditional and corrected scores (Chapter 6), likelihood and quasilikelihood (Chapter 7) and Bayesian methods (Chapter 8). The following methods appear to have some promise for situations where functional modeling is desired but the error model is not of a simple form.

In parametric models, the density or mass function for $(\mathbf{Y}, \mathbf{W})$ is given by (7.5), which in the special case considered here can be written as

$$f_{Y,W}(y, w, \mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2) = \int f_{Y|X}(y|x, \mathcal{B}) f_{W|X}(w|x, \tilde{\alpha}_1) f_X(x|\tilde{\alpha}_2) d\mu(x). \quad (14.4)$$

This is parametric structural modeling because the distribution of $\mathbf{X}$ has been parameterized. We have already discussed flexible parametric modeling of the distribution of $\mathbf{X}$ (section 7.3).

If we take a functional modeling approach and do not specify a parametric model for the distribution of $\mathbf{X}$, we can write the likelihood in general form as

$$f_{Y,W}(y, w, \mathcal{B}, \tilde{\alpha}_1, F_X)$$

$$= \int f_{Y|X}(y|v, \mathcal{B}) f_{W|X}(w|v, \tilde{\alpha}_1) dF_X(v), \qquad (14.5)$$

where $F_X(v)$ is the distribution function of $\mathbf{X}$. Equation (14.5) is called a *mixture* model because the the density or mass function of $(\mathbf{Y}, \mathbf{X}, \mathbf{W})$ is mixed across the unknown distribution of $\mathbf{X}$.

When $\mathbf{X}$ is observed on a subset of the study participants, the density of $(\mathbf{Y}, \mathbf{X}, \mathbf{W})$ can be written in mixture form as

$$f_{Y,X,W}(y, x, w, \mathcal{B}, \tilde{\alpha}_1, F_X)$$
$$= \int f_{Y|X}(y|v, \mathcal{B}) I(x = v) f_{W|X}(w|v, \tilde{\alpha}_1) dF_X(v), \qquad (14.6)$$

where $I(x = v)$ is the indicator function.

It thus follows that in a sample of size n, with $\Delta_i = 1$ meaning that $\mathbf{X}_i$ has been observed, the likelihood function in the unknowns $(\mathcal{B}, \tilde{\alpha}_1, F_X)$ is

$$\mathcal{L}(\mathcal{B}, \tilde{\alpha}_1, F_X) = \prod_{i=1}^{n} \{f_{Y,X,W}(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{W}_i, \mathcal{B}, \tilde{\alpha}_1, F_X)\}^{\Delta_i}$$
$$\times \prod_{i=1}^{n} \{f_{Y,W}(\mathbf{Y}_i, \mathbf{W}_i, \mathcal{B}, \tilde{\alpha}_1, F_X)\}^{1-\Delta_i}. \qquad (14.7)$$

The other functional modeling techniques we have discussed have tried to avoid consideration of F_X entirely. The mixing literature takes a different approach and tries to estimate this distribution nonparametrically. The basic result is that with n independent observations, the maximum likelihood estimate of F_X is discrete and has at most n support points, i.e., points at which $\widehat{F}_X$ has jumps. The distribution function F_X can be estimated by the EM-algorithm (Titterington, et al., 1985), or by gradient methods, see Lesperance & Kalbfleisch (1992) for a recent example.

Lesperance (1989) and Lesperance & Kalbfleisch (1992) discuss inference. They suggest that if $\mathcal{B} = (\beta_1, \beta_2)$ and one wants to make inference about a scalar β_1, then one should invert the semiparametric generalized likelihood ratio test statistic

$$\Lambda(\beta_1) = 2 \log \left\{ \frac{\sup_{b_1, \beta_2, \tilde{\alpha}_1, F_X} \mathcal{L}(b_1, \beta_2, \tilde{\alpha}_1, F_X)}{\sup_{\beta_2, \tilde{\alpha}_1, F_X} \mathcal{L}(\beta_1, \beta_2, \tilde{\alpha}_1, F_X)} \right\}.$$

A $(1-a)100\%$ profile confidence interval are all β_1's such that $\Lambda(\beta_1)$ is less than the $(1-a)$th percentile of the chi-squared distribution

with one degree of freedom. This is a semiparametric analog of the parametric profile likelihood interval discussed in section A.2.4.

Roeder, et al. (1996) provide a small simulation study for the multiplicative error model in logistic regression.

14.3.3 A Cholesterol Example

In this example we analyze a data set concerning the risk of coronary heart disease (CHD) as a function of blood cholesterol level. This data was extracted from the Lipids Research Clinics study which was previously discussed by Satten & Kupper (1993). We use a portion of these data involving men aged 60-70 who do not smoke (256 records: four outliers were removed). A subject is recorded as having CHD ($Y = 1$) 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 measurements of LDL levels is time-consuming and require 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 which is of primary interest; rather, it is the unknown quantity of the other components of TC (triglycerides and high density lipoproteins) that lead to the "measurement error". Henceforth CHD, LDL/100 and TC/100 play the roles of Y, X and W, respectively.

We will treat this study as if it were a case-control study. We already know (section 14.1) that this has no effect on the estimate of the slope β_x.

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. Using X as the predictor, the prospective logistic regression estimate for β_x 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.

We fit a 5-parameter lognormal measurement error model where

$$\log(\mathbf{W}) = \alpha_0 + \alpha_1 \log(\mathbf{X}) + \mathbf{U},$$

where $\mathbf{U}$ is $N(0, \sigma_u^2)$, and a 7-parameter differential measurement

error model where

$$\log(\mathbf{W}) = \alpha_0 + \alpha_1 \mathbf{Y} + \alpha_2 \log(\mathbf{X}) + \mathbf{U}.$$

Here $\mathbf{U}|\mathbf{Y}$ is $N(0, \sigma^2_{u,y})$, with $\sigma^2_{u,y}$ depending on the binary $\mathbf{Y}$ and therefore representing two parameters. In both models, β_0 and β_x are the extra two parameters. A 5-parameter model provided a good fit to the data with the exception of a slight increase in the variance of $\mathbf{W}|\mathbf{X}$ for small values of $\mathbf{X}$. The 7-parameter 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 a two-stage validation study design, 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 reduced observations. Using the complete data only, $\widehat{\beta}_x = .943$ with a standard error of 0.62. Te profile likelihood for the 5-parameter model when both complete and reduced data are used yields an estimate $\widehat{\beta}_x = .765$. Using one-fourth the length of the profile confidence interval, the standard error is is approximately 0.5, clearly smaller than when using only the complete data.

14.3.4 Covariates Measured Without Error

The mixture methods do not apply immediately when there are covariates measured without error. Carroll (1993) and Roeder, et al. (1996) both face this problem. With $\mathbf{X}$ partially observed, they suggest techniques based upon a dimension reduction scheme. This is a problem of clear long-term interest.

14.4 Design of Two-Stage Validation and Replication Studies

This book, and almost all the literature, focuses on the analysis of data in the presence of errors of measurement, There is, however, an emerging literature on the design of studies whose goal is the efficient estimation of parameters.

A good reference to this literature is in the review paper of Spiegelman (1994). Here we give only a brief overview of the main ideas, leaving the details to the cited literature.

Combining the ideas of Greenland (1988b) and Spiegelman & Gray (1991), the goal is to find the most cost-efficient study design which, for a fixed level-α two-sided test of the hypothesis that $\beta_x = \beta_{x,1}$, has power at least π at a prespecified alternative $\beta_{x,2}$, and vice-versa for $\beta_{x,2}$ and $\beta_{x,1}$.

For instance, consider regression models where the covariate $\mathbf{X}$ is measured subject to additive measurement error. All the methods we have discussed, including regression calibration, SIMEX or conditional scores, require an estimate of the measurement error variance σ_u^2 in order to make inference about the slope β_x for $\mathbf{X}$. Typically, this will be done via an internal replication substudy: using the replicated observations the components of variance estimate (3.2) of σ_u^2 is computed.

Any replication study will consist of n_1 observations at which only $(\mathbf{Y}, \mathbf{W})$ is observed, and n_2 observations at which a replicate for $\mathbf{W}$ will also be observed. If the unit cost of obtaining the first $(\mathbf{Y}, \mathbf{W})$ is C_1, and the unit cost of obtaining the replicate is C_2, then the total cost of the study is $C_1 n_1 + C_2 n_2$. One sees here that, for a fixed cost, the greater the number of replicates n_2, the smaller the available number of responses n_1.

Spiegelman & Gray (1991) note that if $V_1(n_1, n_2)$ and $V_2(n_1, n_2)$ are the large sample variance of $\widehat{\beta}_x$ when the actual values are $\beta_{x,1}$ and $\beta_{x,2}$, respectively, then the problem reduces to minimizing the cost $C_1 n_1 + C_2 n_2$ subject to the constraints

$$1 - \Phi\left\{\frac{z_{1-\alpha/2} V_1^{1/2}(n_1, n_2) - \beta_{x,2} + \beta_{x,1}}{V_2^{1/2}(n_1, n_2)}\right\} \geq \pi;$$

$$\Phi\left\{\frac{-z_{1-\alpha/2} V_2^{1/2}(n_1, n_2) - \beta_{x,1} + \beta_{x,2}}{V_1^{1/2}(n_1, n_2)}\right\} \geq \pi,$$

and $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. As Spiegelman (1994) notes, "Within this framework, it is simply a matter of supplying the necessary design specifications ... and of substituting the appropriate formula for V_1 and V_2 to obtain the optimal values of n_1 and n_2".

Spiegelman & Gray (1991) have worked out the details in logistic regression when $(\mathbf{W}, \mathbf{X})$ are jointly normally distributed. Buonaccorsi (1990b) did the same for the normal discriminant model. Both also have extensions to the case that $\mathbf{X}$ can be measured in a validation substudy.

When $\mathbf{X}$ is a binary variable subject to misclassification, Palmgren (1987) and Greenland (1988a) compute the necessary variances in different cases.

Two-stage validation studies, in which $\mathbf{X}$ is observed on a subset of the study design, have been investigated intensively. Breslow & Cain (1988) and Cain & Breslow (1988) discuss particular two-stage designs where selection into the second stage depends on results of the first stage. Zhao & Lipsitz (1992) provide a unification of possible first- and second-stage study designs. Tosteson & Ware (1990) investigate different second-stage designs in detail.

14.5 Misclassification

Situations in which discrete variables are measured with error are called *misclassification.* When all the variables are discrete, the data form a misclassified contingency table.

In principle, misclassified contingency tables can be handled by the method of maximum likelihood, which we have reviewed for misclassified covariates in Chapter 7 and for misclassified responses in Chapter 13. There is nothing new conceptually, and the main issues involve computation. Various papers along these lines include those by Espeland & Odoroff (1985), Ekholm, Green & Palmgren (1986), Espeland & Hui (1987), Palmgren (1987), Ekholm & Palmgren (1987), Chen (1989), Ekholm (1991, 1992), Baker (1991, 1992, 1994a, 1994b) and Baker, Wax & Patterson (1993). Computing the maximum likelihood estimates is discussed by Ekholm, et al. (1986), Clayton (1991), Baker (1992) and Baker, et al. (1993), among others.

When $\mathbf{X}$ is misclassified into $\mathbf{W}$, but $\mathbf{X}$ can be observed on a partial subset of the data, then the likelihood function is (7.10) when $\mathbf{W}$ is a surrogate; see section 14.2 for differential measurement error. When $\mathbf{X}$ cannot be observed but the misclassification probabilities are known, the appropriate likelihood function is the product over the observed data of the terms (7.4) (there is no second measure $\mathbf{T}$ in this context).

When the response $\mathbf{Y}$ is misclassified but also partially observed, the likelihood function is (13.4).

14.6 Survival Analysis

14.6.1 General Considerations

One of the earliest applications of the regression calibration method was discussed by Prentice (1982) in the context of survival analysis. Further results in survival analysis were obtained by Pepe, Self, & Prentice (1989) and Clayton (1991), Nakamura (1993) and Hughes (1993). While the details differ in substantive ways, the ideas are the same as put forward in the rest of this mongraph, and here we provide only a very brief overview of the proportional hazards model, in the case of covariates which do not depend on time.

Suppose that the instantaneous risk that the time $\mathbf{T}$ of an event equals t conditional on no events prior to time t and conditional on the true covariate $\mathbf{X}$ is denoted by

$$\psi(t, \mathbf{X}) = \psi_0(t)\exp\left(\beta_x^t \mathbf{X}\right), \tag{14.8}$$

where $\psi_0(t)$ is the baseline hazard function. When the baseline hazard is not specified, (14.8) is commonly called the Cox proportional hazards model (Cox, 1972). When $\mathbf{X}$ is observable, it is well-known that estimation of β_x is possible without specifying the form of the baseline hazard function.

If $\mathbf{X}$ is unobservable and instead we observe a surrogate $\mathbf{W}$, the induced hazard function is

$$\psi^*(t, \mathbf{W}, \beta_x) = \psi_0(t) E\left\{\exp\left(\beta_x^t \mathbf{X}\right) | T \geq t, \mathbf{W}\right\}. \tag{14.9}$$

As shown by Prentice (1982) and by Pepe, et al. (1989), the difficulty is that the expectation in (14.9) for the observed data depends upon the unknown baseline hazard function ψ_0. Thus, the hazard function does not factor into a product of an arbitrary baseline hazard times a term which depends only on observed data and an unknown parameter, and the technology for proportional hazards regression cannot be applied without modification.

14.6.2 Rare Events

As indicated above, the hazard function for the observed data does not factor nicely. The easiest route around this problem occurs when the event is rare, so that $T \geq t$ occurs with high probability for all t under consideration. As we now show, under certain circumstances this leads to the regression calibration algorithm.

Following section 3.2, if we write the distribution of $\mathbf{X}$ given $\mathbf{W}$ to depend on the parameter γ_{cm}, then for all practical purposes the rare event assumption means that the hazard of the observed data is approximated by

$$\psi^*(t, \mathbf{W}, \beta_x, \gamma_{\text{cm}}) = \psi_0(t) E\left\{\exp\left(\beta_x^t \mathbf{X}\right) | \mathbf{W}\right\}. \qquad (14.10)$$

The hazard function (14.10) requires a regression calibration formulation! If one specifies a model for the distribution of $\mathbf{X}$ given $\mathbf{W}$, then (14.10) is in the form of a proportional hazards model (14.8), but with $\beta_x^t \mathbf{X}$ replaced by

$$\log\left(E\left\{\exp\left(\beta_x^t \mathbf{X}\right) | \mathbf{W}\right\}\right).$$

Such models are easily fit if the regression calibration parameters γ_{cm} are known.

An important special case leads directly to the regression calibration model, namely when $\mathbf{X}$ given $\mathbf{W}$ is normally distributed with mean $m(\mathbf{W}, \gamma_{\text{cm}})$ and with constant covariance matrix Σ_{cm}. To see this, note that the hazard function is, from (14.10),

$$\psi^*(t, \mathbf{W}, \beta_x, \gamma_{\text{cm}}) = \psi_0^*(t) \exp\left\{\beta_x^t m(\mathbf{W}, \gamma_{\text{cm}})\right\},$$

where $\psi_0^*(t) = \psi_0(t) \exp(.5\, \beta_x^t \Sigma_{\text{cm}} \beta_x)$, which is still arbitrary since ψ_0 is arbitrary.

As another application of the rare event assumption, Prentice (1982) considers an example in which he assumes a heteroscedastic Berkson model, namely that $\mathbf{X}$ given $\mathbf{W}$ is normally distributed with mean $\mathbf{W}$ and variance $\sigma_{\text{cm}}^2 \mathbf{W}^2$. This leads to the relative risk model

$$\psi_0(t) \exp\left\{\beta_x \mathbf{W} + .5\, \sigma_{\text{cm}}^2 \beta_x^2 \mathbf{W}^2\right\},$$

which is just a quadratic proportional hazards model and thus easily fit using standard software if one combines $.5\sigma_{\text{cm}}^2 \beta_x^2$ into a separate unknown parameter. One does not even need replication to estimate σ_{cm}^2!

14.6.3 Risk Set Calibration

Clayton (1991) proposed a modification of regression calibration which does not require events to be rare. At each time t_i, $i = 1, ..., k$, for which an event occurs, define the risk set $\mathcal{R}_i \subseteq \{1, ..., n\}$ as the case numbers of those members of the study cohort for whom

an event has not occurred and who were still under study just prior to t_i. If the $\mathbf{X}$'s were observable, and if $\mathbf{X}_i$ is the covariate associated with the ith event, in the absence of ties the usual proportional hazards regression would maximize

$$\prod_{i=1}^{k} \frac{\exp(\beta_x^t \mathbf{X}_i)}{\sum_{j \in \mathcal{R}_i} \exp(\beta_x^t \mathbf{X}_j)}.$$

Clayton basically suggests using regression calibration within each risk set. He assumes that the true values $\mathbf{X}$ within the ith risk set are normally distributed with mean μ_i and variance σ_x^2, and that within this risk set $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is normally distributed with mean zero and variance σ_u^2. Neither σ_x^2 nor σ_u^2 depend upon the risk set in his formulation.

Given an estimate $\widehat{\sigma}_u^2$, one can construct an estimate of $\widehat{\sigma}_x^2$ just as in the equations following (3.3).

Clayton modifies regression calibration by using it within each risk set. Within each risk set, he applies the formula (3.4) for the best unbiased estimate of the $\mathbf{X}$'s. Specifically, in the absence of replication, for any member of the ith risk set, the estimate of the true covariate $\mathbf{X}$ is

$$\widehat{\mathbf{X}} = \widehat{\mu}_i + \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2} (\mathbf{W} - \widehat{\mu}_i),$$

where $\widehat{\mu}_i$ is the sample mean of the $\mathbf{W}$'s in the ith risk set.

As with regression calibration in general, the advantage of Clayton's method is that no new software need be developed, other than calculating the means within risk sets. Formula (3.4) shows how to generalize this method to multivariate covariates and covariates measured without error.

APPENDIX A

FITTING METHODS AND MODELS

A.1 Overview

This chapter collects some of the basic technical tools that are required for understanding the theory employed in this monograph. Section A.4 explains the general class of models upon which the monograph focuses.

Section A.2 reviews likelihood methods which will be familiar to most readers. Section A.3 is a brief introduction to the method of estimating equations, a widely applicable tool that is the basis of all estimators in this book. Section A.5 defines generalized linear models. The bootstrap is explained in section A.6, but one need only note while reading the text that the bootstrap is a computer-intensive method for performing inference.

A.2 Likelihood Methods

A.2.1 Notation

Denote the unknown parameter by Θ. The vector of observations, including response, covariates, surrogates, etc. is denoted by $(\tilde{\mathbf{Y}}_i, \mathbf{Z}_i)$ for $i = 1, ..., n$, where, as before, $\mathbf{Z}_i$ is the vector of covariates that is observable without error and $\tilde{\mathbf{Y}}_i$ collects all the other random variables into one vector. The data set $(\tilde{\mathbf{Y}}_i, \mathbf{Z}_i)$, $i = 1, ..., n$, is the aggregation of all data sets, primary and external, including replication and validation data. Thus, the composition of $\tilde{\mathbf{Y}}_i$ will depend on i, e.g., whether the ith case is a validation case, a replication case, etc. *We emphasize that $\tilde{\mathbf{Y}}_i$ is different from the response $\mathbf{Y}_i$ used throughout the book, and hence the use of tildes.* The $\tilde{\mathbf{Y}}_i$ are assumed independent with the density of $\tilde{\mathbf{Y}}_i$ depend-

ing both on $\mathbf{Z}_i$ and on the type of data set the ith case came from and denoted by $f_i(y|\Theta)$. We assume that f_i has two continuous derivatives with respect to Θ. The loglikelihood is

$$L(\Theta) = \sum_{i=1}^{n} \log f_i(\tilde{\mathbf{Y}}_i|\Theta).$$

A.2.2 Maximum likelihood Estimation

In practice, maximum likelihood is probably the most widely used method of estimation. It is reasonably easy to implement, efficient, and the basis of readily available inferential methods, such as standard errors by Fisher information and likelihood ratio tests. Also, many other common estimators are closely related to maximum likelihood estimators, e.g., the least squares estimator which is the maximum likelihood estimator under certain circumstances and quasilikelihood estimators. In this section, we quickly review some of these topics.

The maximum likelihood estimator (MLE) maximizes $L(\Theta)$. Under some regularity conditions, for example in Serfling (1980), the MLE has a simple asymptotic distribution. The "likelihood score" or "score function" is $s_i(y|\Theta) = (\partial/\partial\Theta)\log f_i(y|\Theta)$. The Fisher information matrix, or expected information, is

$$I_n(\Theta) = -\sum_{i=1}^{n} E\{(\partial/\partial\Theta^t)s_i(\tilde{\mathbf{Y}}_i|\Theta)\} \quad \text{(A.1)}$$

$$= \sum_{i=1}^{n} E\{s_i(\tilde{\mathbf{Y}}_i|\Theta)s_i^t(\tilde{\mathbf{Y}}_i,|\Theta)\}. \quad \text{(A.2)}$$

In large samples, the MLE is approximately normally distributed with mean Θ and covariance matrix $I_n^{-1}(\Theta)$, whose entries converge to 0 as $n \to \infty$. There are several methods of estimating $I_n(\Theta)$. The most obvious is $I_n(\widehat{\Theta})$. Efron & Hinkley (1978) present arguments in favor of using instead the observed Fisher information matrix, defined as

$$\widehat{I}_n = -\sum_{i=1}^{n} \partial/\partial\Theta^t s_i(\tilde{\mathbf{Y}}_i|\widehat{\Theta}), \quad \text{(A.3)}$$

which is an empirical version of (A.1). The empirical version of

(A.2) is

$$\widehat{B}_n = \sum_{i=1}^{n} s_i(\tilde{\mathbf{Y}}_i|\Theta)s_i^t(\tilde{\mathbf{Y}}_i|\Theta),$$

which is not used directly to estimate I_n, but is part of the so-called "sandwich formula," $\widehat{I}_n^{-1}\widehat{B}_n^{-1}\widehat{I}_n^{-1}$, used to estimate $I_n^{-1}(\Theta)$. As discussed in section A.3, the sandwich formula has certain "robustness" properties, but can be subject to high sampling variability.

A.2.3 Likelihood Ratio Tests

Suppose that $\dim\{\Theta\} = p$, that φ is a known function of Θ such that $\dim\{\varphi(\Theta)\} = p_1 < p$, and that we wish to test $H_0 : \varphi(\Theta) = 0$ against the general alternative that $\varphi(\Theta) \neq 0$. We suppose that $\text{rank}\{(\partial/\partial\Theta^t)\,\varphi(\Theta)\} = p_1$ so that the constraints imposed by the null hypothesis are linearly independent; otherwise p_1 is not well defined, i.e., we can add redundant constraints and increase p_1 without changing H_0, and the following result is invalid.

Let $\widehat{\Theta}_0$ maximize $L(\Theta)$ subject to $\varphi(\Theta) = 0$, and define $LR = 2\{L(\widehat{\Theta}) - L(\widehat{\Theta}_0)\}$, the log likelihood ratio. Under H_0, LR converges in distribution to the chi-squared distribution with p_1 degrees of freedom. Thus, an asymptotically valid test rejects the null hypothesis if LR exceeds $\chi^2_{p_1}(\alpha)$, the $(1-\alpha)$ quantile of the chi-squared distribution with p_1 degrees of freedom.

A.2.4 Profile Likelihood and Likelihood Ratio Confidence Intervals

Profile likelihood is used to draw inferences about a single component of the parameter vector. Suppose that $\Theta = (\theta_1, \Theta_2)$ where θ_1 is univariate. Let c be a hypothesized value of θ_1. To test $H_0 : \theta_1 = c$ using the theory of section A.2.3, we use $\varphi(\Theta) = \theta_1 - c$ and find $\widehat{\Theta}_2(c)$ so that $(c, \widehat{\Theta}_2(c))$ maximizes L subject to H_0. $L_{\max}(\theta_1) = L(\theta_1, \widehat{\Theta}_2(\theta_1))$ is called the profile likelihood function for θ_1—it does not involve Θ_2 since the log likelihood has been maximized over Θ_2. Then, $LR = L(\widehat{\Theta}) - L_{\max}(c)$ where, as before, $\widehat{\Theta}$ is the MLE. One rejects the null hypothesis if LR exceeds $\chi^2_1(\alpha)$.

Inference for θ_1 is typically based on the profile likelihood. In

particular, the likelihood ratio confidence region for θ_1 is

$$\{\theta_1 : L_{\max}(\theta_1) > L(\widehat{\Theta}) - \frac{\chi_1^2(\alpha)}{2}\}.$$

This region is also the set of all c such that we accept $H_0 : \theta_1 = c$. The confidence region is typically an interval, but there can be exceptions. An alternative large-sample interval is

$$\widehat{\theta}_1 \pm \Phi^{-1}(1 - \frac{\alpha}{2})\text{se}(\widehat{\theta}_1), \tag{A.4}$$

where $\text{se}(\widehat{\theta}_1)$ is the standard error of $\widehat{\theta}_1$, say from the Fisher information matrix or from bootstrapping as in section A.6. For nonlinear models, the accuracy of (A.4) is questionable, i.e., the true coverage probability is likely to be somewhat different than $(1-\alpha)$, and the likelihood ratio interval is preferred.

A.2.5 Efficient Score Tests

The efficient score test or simply the "score test," is due to Rao (1947). Under the null hypothesis, the efficient score test is asymptotically equivalent to the likelihood ratio test, e.g., the difference between the two test statistics converges to 0 in probability. The advantage of the efficient score test is that the MLE needs to be computed only under the null hypothesis, not under the alternative as for the likelihood ratio test. This can be very convenient when testing the null hypothesis of no effects for covariates measured with error, since these covariates, and hence measurement error, can be ignored when fitting under H_0.

To define the score test, start by partitioning Θ as $(\Theta_1^t, \Theta_2^t)^t$ where $\dim(\Theta_1) = p_1$, $1 \le p_1 \le p$. We will test the null hypothesis that $H_0 : \Theta_1 = 0$. Many hypotheses can be put into this form, possibly after reparametrization. Let $S(\Theta) = \sum_{i=1}^n s_i(\tilde{\mathbf{Y}}_i|\Theta)$ and partition S into S_1 and S_2 with dimensions p_1 and $(p - p_1)$, respectively. Let $\widehat{\Theta}_0 = (0^t, \widehat{\Theta}_{0,2}^t)^t$ be the MLE of Θ under H_0. Notice that $S_2(\widehat{\Theta}_0) = 0$ since $\widehat{\Theta}_{0,2}$ maximizes the likelihood over Θ_2 when $\Theta_1 = 0$. The basic idea behind the efficient score test is that under H_0 we expect $S_1(\widehat{\Theta}_0)$ to be close to 0, since the expectation of $S(\Theta)$ is 0 and $\widehat{\Theta}_0$ is consistent for Θ.

Let I_n^{11} be the upper left corner of $(I_n)^{-1}$ evaluated at $\widehat{\Theta}_0$. The efficient score test statistic measures the departure of $S_1(\widehat{\Theta}_0)$ from

0 and is defined as

$$R_n = S_1(\widehat{\Theta}_0)^t I_n^{11} S_1(\widehat{\Theta}_0) = S(\widehat{\Theta}_0) I_n^{-1} S(\widehat{\Theta}_0).$$

The equality holds because $S_2(\widehat{\Theta}_0) = 0$.

Under H_0, R_n asymptotically has a chi-squared distribution with p_1 degrees of freedom, so we reject H_0 is R_n exceeds $(1-\alpha)$ chi-squared quantile, $\chi^2_{p_1}(\alpha)$. See Cox and Hinkley (1974, section 9.3) for a proof of the asymptotic distribution.

A.3 Unbiased Estimating Equations

All of the estimators described in this book, including the MLE, can be characterized as solutions to *unbiased estimating equations.* Understanding the relationship between estimators and estimating equations is useful because it permits easy and routine calculation of estimated standard errors. The theory of estimating equations arose from two distinct lines of research, in Godambe's (1960) study of efficiency and Huber's (1964, 1967) work on robust statistics. Huber's (1967) seminal paper used estimating equations to understand the behavior of the MLE under model misspecification, but his work also applies to estimators that are not the MLE under any model. Over time, estimating equations became established as a highly effective, unified approach for studying wide classes of estimators; see, e.g., Carroll and Ruppert (1988) who use estimating equation theory to analyze a variety of transformation and weighting methods in regression.

This section reviews the basic ideas of estimating equations: see Huber (1967), Ruppert (1985), Carroll & Ruppert (1988), McLeish & Small (1988), Desmond (1989) or Godambe (1991) for more extensive discussion.

A.3.1 Introduction and Basic Large Sample Theory

As in section A.2, the unknown parameter is Θ, and the vector of observations, including response, covariates, surrogates, etc. is denoted by $(\tilde{\mathbf{Y}}_i, \mathbf{Z}_i)$ for $i = 1, ..., n$. For each i, let Ψ_i be a function of $(\tilde{\mathbf{Y}}_i, \Theta)$ taking values in p-dimensional space ($p = \dim(\Theta)$). Typically, Ψ_i depends on i through Z_i and the type of data set the ith case belongs to, e.g., whether that case is validation data, etc.

An estimating equation for Θ has the form

$$0 = n^{-1} \sum_{i=1}^{n} \Psi_i \left(\tilde{\mathbf{Y}}_i, \Theta \right). \tag{A.5}$$

The solution, $\widehat{\Theta}$, to (A.5) as Θ ranges across the set of possible parameter values is called an *M–estimator* of Θ, a term due to Huber (1964). In practice, one obtains an estimator by some principle, e.g., maximum likelihood, least squares, generalized least squares, etc. Then, one shows that the estimator satisfies an equation of form (A.5) and Ψ_i is identified. The point is that one doesn't choose the Ψ_i's directly, but rather that they are defined through the choice of an estimator.

In (A.5), the function Ψ_i is called an *estimating function* and depends on i through $\mathbf{Z}_i$. The estimating function (and hence the estimating equation) is said to be *conditionally unbiased* if it has mean zero when evaluated at the true parameter, i.e.,

$$0 = E \left\{ \Psi_i \left(\tilde{\mathbf{Y}}_i, \Theta \right) \right\}, \text{ for } i = 1, ..., n. \tag{A.6}$$

As elsewhere in this book, expectations and covariances are always conditional upon $\{\mathbf{Z}_i\}_1^n$.

If the estimating equations are unbiased, then under certain regularity conditions $\widehat{\Theta}$ is a consistent estimator of Θ. See Huber (1967) for the regularity conditions and proof in the iid case. The basic idea is that for each value of Θ the right hand side of (A.5) converges to its expectation by the law of large numbers, and the true Θ is a zero of the expectation of (A.5). One of the regularity conditions is that the true Θ is the only zero, so that $\widehat{\Theta}$ will converge to Θ under some additional conditions.

Moreover, if $\widehat{\Theta}$ is consistent then by a Taylor series approximation

$$0 \approx n^{-1} \sum_{i=1}^{n} \Psi_i(\tilde{\mathbf{Y}}_i, \Theta) + \left\{ n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \Theta^t} \Psi_i(\tilde{\mathbf{Y}}_i, \Theta) \right\} (\widehat{\Theta} - \Theta),$$

where Θ now is the true parameter value. Applying the law of large numbers to the term in curly brackets, we have

$$\widehat{\Theta} - \Theta \approx -A_n(\Theta)^{-1} n^{-1} \sum_{i=1}^{n} \Psi_i(\tilde{\mathbf{Y}}_i, \Theta), \tag{A.7}$$

where $A_n(\Theta)$ is given by (A.9) below. If follows that $\widehat{\Theta}$ is asymptotically normally distributed with mean Θ and covariance matrix $n^{-1}A_n^{-1}(\Theta)B_n(\Theta)A_n^{-t}(\Theta)$, where $A_n^{-t}(\Theta) = \left\{A_n^{-1}(\Theta)\right\}^t$ and

$$B_n(\Theta) = n^{-1}\sum_{i=1}^{n} \operatorname{cov}\left\{\Psi_i\left(\tilde{\mathbf{Y}}_i, \Theta\right)\right\}; \tag{A.8}$$

$$A_n(\Theta) = n^{-1}\sum_{i=1}^{n} E\left\{\frac{\partial}{\partial\Theta^t}\Psi_i\left(\tilde{\mathbf{Y}}_i, \Theta\right)\right\}. \tag{A.9}$$

See Huber (1967) for a proof. There are two ways to estimate this covariance matrix. The first uses *empirical expectation* and is often called the *sandwich estimator* or a *robust covariance estimator* (a term we do not like—see below); in the former terminology, B_n is sandwiched between the inverse of A_n. The sandwich estimator uses

$$\widehat{A}_n = n^{-1}\sum_{i=1}^{n} \frac{\partial}{\partial\Theta^t}\Psi_i\left(\tilde{\mathbf{Y}}_i, \widehat{\Theta}\right); \tag{A.10}$$

$$\widehat{B}_n = n^{-1}\sum_{i=1}^{n} \Psi_i\left(\tilde{\mathbf{Y}}_i, \widehat{\Theta}\right)\Psi_i^t\left(\tilde{\mathbf{Y}}_i, \widehat{\Theta}\right). \tag{A.11}$$

The second method, called the *model-based expectation method*, uses an underlying model to evaluate (A.8)–(A.9) exactly, and then substitutes the estimated value $\widehat{\Theta}$ for Θ, i.e., uses $A_n^{-1}\ B_n A_n^{-t}$.

If Ψ_i is the likelihood score, i.e., $\Psi_i = s_i$, where s_i is defined in section A.2.2, then $\widehat{\Theta}$ is the MLE. In this case, both $B_n(\Theta)$ and $A_n(\Theta)$ equal the Fisher information matrix, $I_n(\Theta)$. However, $\widehat{A}_n$ and $\widehat{B}_n$ are generally different, so the sandwich method differs from using the observed Fisher information.

As a general rule, the sandwich method provides a consistent estimate of the covariance matrix of $\widehat{\Theta}$, without the need to make any distribution assumptions. In this sense it is *robust*. However, in comparison with the model-based expectation method, when a distributional model is reasonable the sandwich estimator is typically inefficient, which can unnecessarily inflate the length of confidence intervals. This inefficiency is why we don't like to call the sandwich method "robust." Robustness usually means insensitivity to assumptions at the price of a *small* loss of efficiency, whereas the sandwich formula can lose a great deal of efficiency.

A.3.2 Sandwich Formula Example: Linear Regression Without Measurement Error

As an example, consider ordinary multiple regression without measurement errors so that $\mathbf{Y}_i = \beta_0 + \beta_z^t \mathbf{Z}_i + \epsilon_i$, where the ϵ's are independent, mean-zero random variables. Let $\mathbf{Z}_i^* = (1, \mathbf{Z}_i^t)^t$ and $\Theta = (\beta_0, \beta_z^t)^t$. Then the ordinary least squares estimator is an M-estimator with $\Psi_i(\mathbf{Y}_i, \Theta) = (\mathbf{Y}_i - \beta_0 - \beta_z^t \mathbf{Z}_i)\mathbf{Z}_i^*$. Also,

$$\begin{aligned} \frac{\partial}{\partial \Theta^t} \Psi_i(\mathbf{Y}_i, \Theta) &= -\mathbf{Z}_i^* (\mathbf{Z}_i^*)^t, \\ A_n &= -n^{-1} \sum_{i=1}^{n} \mathbf{Z}_i^* (\mathbf{Z}_i^*)^t, \end{aligned} \tag{A.12}$$

and if one assumes that the variance of ϵ_i is a constant σ^2 for all i, then

$$B_n = -\sigma^2 A_n. \tag{A.13}$$

Notice that A_n and B_n do not depend on Θ so they are known exactly except for the factor σ^2 in B_n. The model-based expectation method gives covariance matrix $-\sigma^2 A_n^{-1}$, the well-known variance of the least squares estimator. Generally, σ^2 is estimated by the residual mean square.

The sandwich formula uses $\widehat{A}_n = A_n$ and

$$\widehat{B}_n = n^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \widehat{\beta}_0 - \widehat{\beta}_z^t \mathbf{Z}_i)^2 \mathbf{Z}_i^* (\mathbf{Z}_i^*)^t. \tag{A.14}$$

We have not made distributional assumptions about ϵ_i, but we have assumed homoscedasticity, i.e., that $\text{var}(\epsilon_i) \equiv \sigma^2$. To illustrate the "robustness" of the sandwich formula, consider the heteroscedatic model where the variance of ϵ_i is σ_i^2 depending on $\mathbf{Z}_i$. Then B_n is no longer given by (A.13) but rather by

$$B_n = n^{-1} \sum_{i=1}^{n} \sigma_i^2 \mathbf{Z}_i^* (\mathbf{Z}_i^*)^t,$$

which is consistently estimated by (A.14). Thus, the sandwich formula is heteroscedasticity consistent. In contrast, the model-based estimator of B_n, which is $\widehat{\sigma}^2 A_n$ with A_n given by (A.12), is inconsistent for B_n. This makes model-based estimation of the covariance matrix of $\widehat{\Theta}$ inconsistent.

The inefficiency of the sandwich estimator can also be seen in this example. Suppose that there is a high leverage point, that is an observation with an outlying value of $\mathbf{Z}_i$. Then as seen in (A.14), the value of $\widehat{B}_n$ is highly dependent upon the squared residual of this observation. This makes $\widehat{B}_n$ highly variable, and indicates the additional problem that $\widehat{B}_n$ very sensitive to outliers.

A.3.3 Sandwich Method and Likelihood-type Inference

Likelihood ratio-type extensions of sandwich standard errors are also available, but not well known, see Huber (1967), Schrader & Hettmansperger (1980), Kent (1982), Ronchetti (1982) and Li & McCullagh (1994). This theory is essentially an extension of the theory of estimating equations, where the estimating equation is assumed to correspond to a criterion function, i.e., solving the estimating equation minimizes the criterion function.

In the general theory we consider inferences about a parameter vector Θ, and we assume that the estimate $\widehat{\Theta}$ maximizes an estimating criterion, $\ell(\Theta)$, which is effectively the working log likelihood, although it need not be the logarithm of an actual density function. Following Li & McCullagh (1994), we refer to $\exp(\ell) = \exp(\sum \ell_i)$ as the quasilikelihood function. (Here, ℓ_i is the log quasilikelihood for the ith case and ℓ is the log quasilikelihood for the entire data set.) Define the score function, a type of estimating function, as

$$\mathcal{U}_i(\Theta) = \frac{\partial}{\partial \Theta} \ell_i(\Theta | \tilde{\mathbf{Y}}_i),$$

the score covariance,

$$\mathcal{J}_n = \sum_{i=1}^{n} \mathrm{E}\{\mathcal{U}_i(\Theta)\, \mathcal{U}_i(\Theta)^t\}, \tag{A.15}$$

and the negative expected hessian,

$$\mathcal{H}_n = -\sum_{i=1}^{n} \mathrm{E}\left\{ \frac{\partial}{\partial \Theta^t} \mathcal{U}_i(\Theta) \right\}. \tag{A.16}$$

If ℓ were the true log likelihood, then we would have $\mathcal{H}_n = \mathcal{J}_n$, but this equality usually fails for quasilikelihood. As in the theory of estimating equations, the parameter Θ is determined by the equation $E\{\mathcal{U}_i(\Theta)\} = 0$ for all i (conditionally unbiased), or

possibly through the weaker constraint that $\sum_{i=1}^{n} E\{\mathcal{U}_i(\Theta)\} = 0$ (unbiased).

We partition $\Theta = (\gamma^t, \eta^t)^t$, where γ is the p–dimensional parameter vector of interest, and η is the vector of nuisance parameters. Partition $\mathcal{H}$, omitting the subscript n for ease of notation, similarly as

$$\mathcal{H} = \begin{pmatrix} \mathcal{H}_{\gamma\gamma} & \mathcal{H}_{\gamma\eta} \\ \mathcal{H}_{\eta\gamma} & \mathcal{H}_{\eta\eta} \end{pmatrix},$$

and define $\mathcal{H}_{\gamma\gamma\cdot\eta} = \mathcal{H}_{\gamma\gamma} - \mathcal{H}_{\gamma\eta}\mathcal{H}_{\eta\eta}^{-1}\mathcal{H}_{\eta\gamma}$.

Let $\widehat{\Theta}_0 = (\gamma_0^t, \widehat{\eta}_0^t)^t$ denote the maximum quasilikelihood estimate subject to $\gamma = \gamma_0$. We need the large sample distribution of the log quasilikelihood ratio,

$$\mathcal{L}(\gamma_0) = 2\{\ell(\widehat{\Theta}) - \ell(\widehat{\Theta}_0)\}.$$

The following result is well–known under various regularity conditions. For the basic idea of the proof see Kent (1982).

Theorem: *If $\gamma = \gamma_0$, then, as the number of independent observations increases, $\mathcal{L}(\gamma_0)$ converges in distribution to $\sum_{k=1}^{p} \lambda_k W_k$, where $W_1, \ldots, W_p$ are independently distributed as χ_1^2, and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\mathcal{H}_{\gamma\gamma\cdot\eta}(\mathcal{H}^{-1}\mathcal{J}\mathcal{H}^{-1})_{\gamma\gamma}$.*

To use this result in practice, either to perform a quasilikelihood ratio test of $H_0 : \gamma = \gamma_0$, or to compute a quasilikelihood confidence set for γ_0, we need to estimate the matrices $\mathcal{H}$ and $\mathcal{J}$. If all data are independent, an obvious approach is to replace the theoretical expectations in (A.15) and (A.16) by the analogous empirical averages.

We also need to compute quantiles of the distribution of $\sum_k \widehat{\lambda}_k W_k$. Observe that if $p = 1$ the appropriate distribution is simply a scaled χ_1^2 distribution. If $p > 1$, then algorithms given by Marazzi (1980) and Griffiths & Hill (1985) may be used. A quick and simple way to do the computation is to simulate from the distribution of $\sum_k \widehat{\lambda}_k W_k$, since chi-squared random variables are easy to generate.

A.3.4 Unbiased, But Conditionally Biased, Estimating Equations

It is possible to relax (A.6) to

$$0 = \sum_{i=1}^{n} E\left\{\Psi_i\left(\tilde{\mathbf{Y}}_i, \widehat{\Theta}\right)\right\},$$

and then the estimating function and estimating equation are *not* conditionally unbiased, but are still said to be *unbiased.* The theory of conditionally unbiased estimating equations carries over almost without change to estimating equations that are merely unbiased. The only difficulty is that if $E\{\Psi_i(\tilde{\mathbf{Y}}_i, \Theta)\} \neq 0$ then $E\{\Psi_i(\tilde{\mathbf{Y}}_i, \Theta)$ $\Psi_i(\tilde{\mathbf{Y}}_i, \Theta)^t\}$ does not equal $\text{cov}\{\Psi_i(\tilde{\mathbf{Y}}_i, \Theta)\}$ and (A.11) does not estimate B_n. Therefore, the sandwich formula does not lead to consistent standard errors unless modified appropriately, i.e., by computing the sample covariance matrix of the terms $\Psi_i(\tilde{\mathbf{Y}}_i, \widehat{\Theta})$.

A.3.5 Biased Estimating Equations

The estimation methods described in Chapters 3, 4, and 5 and later used in Chapters 14 and 9 are approximately consistent, in the sense that they consistently estimate a value that closely approximates the true parameter. These estimators are formed by estimating equations such as (A.5), but the estimating functions are *not* unbiased for the true parameter Θ. Usually there exists Θ_* which is close to Θ and which solves

$$0 = \sum_{i=1}^{n} E\left\{\Psi_i\left(\tilde{\mathbf{Y}}_i, \Theta_*\right)\right\}. \tag{A.17}$$

In such cases, $\widehat{\Theta}$ is *still* asymptotically normally distributed but with mean Θ_* instead of mean Θ. In fact, the theory of section A.3.4 is applicable since the equations are unbiased for Θ_*. If

$$0 = E\left\{\Psi_i\left(\tilde{\mathbf{Y}}_i, \Theta_*\right)\right\}, \text{ for } i = 1, ..., n,$$

then the the estimating functions are conditionally unbiased for Θ_* and the sandwich method yields asymptotically correct standard error estimators.

A.3.6 Stacking Estimating Equations: Using Prior Estimates of Some Parameters

To estimate the regression parameter, $\mathcal{B}$, in a measurement error model, one often uses the estimates of the measurement error parameters, α, obtained from another data set. How does uncertainty about the measurement error parameters affect the accuracy of the estimated regression parameter? In this subsection, we develop the

theory to answer this question. The fact that such complicated estimating schemes can be easily analyzed by the theory of estimating equations further illustrates the power of this theory.

We work generally in that α and $\mathcal{B}$ can be any parameter vectors in a statistical model, and we assume that both $\widehat{\alpha}$ and $\widehat{\mathcal{B}}$ are M-estimators. Suppose that that $\widehat{\alpha}$ solves the estimating equation

$$0 = \sum_{i=1}^{n} \phi_i(\tilde{\mathbf{Y}}_i, \alpha), \tag{A.18}$$

and $\widehat{\mathcal{B}}$ solves

$$0 = \sum_{i=1}^{n} \Psi_i(\tilde{\mathbf{Y}}_i, \mathcal{B}, \widehat{\alpha}), \tag{A.19}$$

with $\widehat{\alpha}$ in (A.19) fixed at the solution to (A.18). The estimating functions in (A.18) and (A.19) are assumed to be conditionally unbiased. Since $(\widehat{\alpha}, \widehat{\mathcal{B}})$ solves (A.18) and (A.19) simultaneously, the asymptotic distribution of $(\widehat{\alpha}, \widehat{\mathcal{B}})$ can be found by stacking (A.18) and (A.19) into a single estimating equation

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} \phi_i(\tilde{\mathbf{Y}}_i, \alpha) \\ \Psi_i(\tilde{\mathbf{Y}}_i, \mathcal{B}, \alpha) \end{pmatrix}. \tag{A.20}$$

One then applies the usual theory to (A.20). Partition $A_n = A_n(\Theta)$, $B_n = B_n(\Theta)$, and $A_n^{-1} B_n A_n^{-t}$ according to the dimensions of α and $\mathcal{B}$. Then the asymptotic variance of $\widehat{\mathcal{B}}$ is n^{-1} times the lower right submatrix of $A_n^{-1} B_n A_n^{-t}$. After some algebra, one gets

$$\begin{aligned} \operatorname{var}(\widehat{\mathcal{B}}) \approx n^{-1} A_{n,22}^{-1} \Big\{ & B_{n,22} - A_{n,21} A_{n,11}^{-1} B_{n,12} \\ & - B_{n,12}^t A_{n,11}^{-t} A_{n,21}^t + A_{n,21} A_{n,11}^{-1} B_{n,11} A_{n,11}^{-t} A_{n,21}^t \Big\} A_{n,22}^{-t} \end{aligned}$$

where

$$\begin{aligned} A_{n,11} &= \sum_{i=1}^{n} E\Big\{ \frac{\partial}{\partial \alpha^t} \phi_i(\tilde{\mathbf{Y}}_i, \alpha) \Big\}, \\ A_{n,21} &= \sum_{i=1}^{n} E\Big\{ \frac{\partial}{\partial \alpha^t} \Psi_i(\tilde{\mathbf{Y}}_i, \mathcal{B}, \alpha) \Big\}, \\ A_{n,22} &= \sum_{i=1}^{n} E\Big\{ \frac{\partial}{\partial \mathcal{B}^t} \Psi_i(\tilde{\mathbf{Y}}_i, \mathcal{B}, \alpha) \Big\}, \end{aligned}$$

$$B_{n,11} = \sum_{i=1}^{n} \phi_i(\tilde{\mathbf{Y}}_i, \alpha)\phi_i^t(\tilde{\mathbf{Y}}_i, \alpha),$$

$$B_{n,12} = \sum_{i=1}^{n} \phi_i(\tilde{\mathbf{Y}}_i, \alpha)\Psi_i^t(\tilde{\mathbf{Y}}_i, \alpha, \mathcal{B}), \quad \text{and}$$

$$B_{n,22} = \sum_{i=1}^{n} \Psi_i(\tilde{\mathbf{Y}}_i, \alpha, \mathcal{B})\Psi_i^t(\tilde{\mathbf{Y}}_i, \alpha, \mathcal{B}).$$

As usual, the components of A_n and B_n can be estimated by model-based expectations or by the sandwich method.

A.4 Quasilikelihood and Variance Function (QVF) Models

A.4.1 General Ideas

In the case of no measurement error, Carroll & Ruppert (1988) describe estimation based upon the mean and variance functions of the observed data, i.e., the conditional mean and variance of $\mathbf{Y}$ as functions of $(\mathbf{Z}, \mathbf{X})$. We will call these QVF methods, for Quasilikelihood and Variance Functions. The models include the important class of generalized linear models (McCullagh & Nelder, 1989 and section A.5 of this monograph), and in particular linear, logistic, Poisson, and gamma regression. QVF estimation is an important special case of estimating equations.

The typical regression model is a specification of the relationship between the mean of a response $\mathbf{Y}$ and the predictors $(\mathbf{Z}, \mathbf{X})$:

$$E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = f(\mathbf{Z}, \mathbf{X}, \mathcal{B}), \tag{A.21}$$

where $f(\cdot)$ is the *mean function* and $\mathcal{B}$ is the *regression parameter*. Generally, specification of the model is incomplete without an accompanying model for the variances,

$$\text{var}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = \sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta), \tag{A.22}$$

where $g(\cdot)$ is called the *variance function* and θ is called the *variance function parameter*. We find it convenient in (A.22) to separate the variance parameters into the scale factor σ^2 and θ, which determines the possible heteroscedasticity.

The combination of (A.21) and (A.22) includes many important special cases, among them:

- Homoscedastic linear and nonlinear regression, with $g(z, x, \mathcal{B}, \theta) \equiv 1$. For linear regression, $f(z, x, \mathcal{B}) = \beta_0 + \beta_x^t x + \beta_z^t z$.
- Generalized linear models, including Poisson and gamma regression, with

$$g(z, x, \mathcal{B}, \theta) = f^\theta(z, x, \mathcal{B})$$

  for some parameter θ. For example, $\theta = 1/2$ for Poisson regression, while $\theta = 1$ for gamma and lognormal models.
- Logistic regression, where $f(z, x, \mathcal{B}) = H(\beta_0 + \beta_x^t x + \beta_z^t z)$, $H(v) = 1/\{1+\exp(-v)\}$, and since $\mathbf{Y}$ is Bernoulli distributed, $g^2 = f(1 - f)$, $\sigma^2 = 1$ and there is no parameter θ.

Model (A.21)–(A.22) includes examples from fields including epidemiology, econometrics, fisheries research, quality control, pharmacokinetics, assay development, etc. See Carroll & Ruppert (1988, Chapters 2–4) for more details.

A.4.2 Estimation and Inference for QVF Models

Specification of only the mean and variance models (A.21)–(A.22) allows one to construct estimates of the parameters $(\mathcal{B}, \theta)$. No further detailed distributional assumptions are necessary. Given θ, $\mathcal{B}$ can be estimated by generalized (weighted) least squares (GLS) a term often now referred to as quasilikelihood estimation. The *conditionally unbiased estimating function* for estimating $\mathcal{B}$ by GLS is

$$\frac{\mathbf{Y} - f(\mathbf{Z}, \mathbf{X}, \mathcal{B})}{\sigma^2 g^2(\mathbf{Z}, \mathbf{X}, \mathcal{B}, \theta)} f_{\mathcal{B}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}), \tag{A.23}$$

where

$$f_{\mathcal{B}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = \frac{\partial}{\partial \mathcal{B}} f(\mathbf{Z}, \mathbf{X}, \mathcal{B})$$

is the vector of partial derivatives of the mean function. The *conditionally unbiased estimating equation* for $\mathcal{B}$ is the sum of (A.23) over the observed data.

To understand why (A.23) is the GLS estimating function, note that the nonlinear least squares (LS) estimator, which minimizes

$$\sum_{i=1}^{n} \{\mathbf{Y} - f(\mathbf{Z}, \mathbf{X}, \mathcal{B})\}^2,$$

solves

$$\sum_{i=1}^{n}\{\mathbf{Y} - f(\mathbf{Z}, \mathbf{X}, \mathcal{B})\} f_{\mathcal{B}}(\mathbf{Z}, \mathbf{X}, \mathcal{B}) = 0. \tag{A.24}$$

The LS estimator is inefficient and can be improved by weighting the summands in (A.24) by reciprocal variances; the result is (A.23).

There are many methods for estimating θ. These may be based on true replicates if they exist, or on functions of squared residuals. These methods are reviewed in Chapters 3 and 6 of Carroll & Ruppert (1988), see also Davidian & Carroll (1987) and Rudemo, et al. (1989). Let $(\cdot)$ stand for the argument $(\mathbf{Z}, \mathbf{X}, \mathcal{B})$. If we define

$$\mathbf{R}(\mathbf{Y}, \cdot, \theta, \sigma) = \{\mathbf{Y} - f(\cdot)\} / \{\sigma g(\cdot, \theta)\}, \tag{A.25}$$

then one such (approximately) conditionally unbiased score function for θ (and σ) given $\mathcal{B}$ is

$$\left\{\mathbf{R}^2(\mathbf{Y}, \cdot, \theta, \sigma) - \frac{n - \dim(\mathcal{B})}{n}\right\} \frac{\partial}{\partial (\sigma, \theta)^t} \log\{\sigma g(\cdot, \theta)\}, \tag{A.26}$$

where $\dim(\mathcal{B})$ is the number of components of the vector $\mathcal{B}$. The (approximately) *conditionally unbiased estimating equation* for θ and σ is the sum of (A.26) over the observed data. The resulting M-estimator is closely related to the REML estimator used in variance components modeling; see Searle, Casella, & McCulloch (1992).

As described by Carroll & Ruppert (1988), (A.23)–(A.26) are weighted least squares estimating equations, and nonlinear regression algorithms can be used to estimate the parameters.

There are two specific types of covariance estimates, depending on whether or not one believes that the variance model has been approximately correctly specified. We concentrate here on inference for the regression parameter $\mathcal{B}$, referring the reader to Chapter 3 of Carroll & Ruppert (1988) for variance parameter inference. Based on a sample of size n, $\widehat{\mathcal{B}}$ is generally asymptotically normally distributed with mean $\mathcal{B}$ and covariance matrix $n^{-1}A_n^{-1}B_nA_n^{-1}$, where if $(\cdot)$ stands for $(\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})$,

$$A_n = n^{-1}\sum_{i=1}^{n}\{f_{\mathcal{B}}(\cdot)\}\{f_{\mathcal{B}}(\cdot)\}^t\{\sigma^2 g^2(\cdot, \theta)\}^{-1};$$

$$B_n = n^{-1} \sum_{i=1}^{n} \{f_{\mathcal{B}}(\cdot)\} \{f_{\mathcal{B}}(\cdot)\}^t \frac{E\{\mathbf{Y}_i - f(\cdot)\}^2}{\sigma^4 g^4(\cdot, \theta)}.$$

The matrix B_n in this expression is the same as (A.8) in the general theory of unbiased estimating equations. The matrix A_n is the same as (A.9), but it is simplified somewhat by using the fact that $E(\mathbf{Y}|\mathbf{Z}, \mathbf{X}) = f(\mathbf{Z}, \mathbf{X}, \mathcal{B})$.

If the variance model is correct, then $E\{\mathbf{Y}_i - f(\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B})\}^2 = \sigma^2 g^2(\mathbf{Z}_i, \mathbf{X}_i, \mathcal{B}, \theta)$, $A_n = B_n$ and an asymptotically correct covariance matrix is $n^{-1}\widehat{A}_n^{-1}$, where $(\cdot)$ stands for $(\mathbf{Z}_i, \mathbf{X}_i, \widehat{\mathcal{B}})$ and

$$\widehat{A}_n = n^{-1} \sum_{i=1}^{n} \{f_{\mathcal{B}}(\cdot)\} \{f_{\mathcal{B}}(\cdot)\}^t \left\{\widehat{\sigma}^2 g^2(\cdot, \widehat{\theta})\right\}^{-1}.$$

If one has severe doubts about the variance model, one can use the sandwich method to estimate $E\{\mathbf{Y}_i - f(\cdot)\}^2$, leading to the covariance matrix estimate $\widehat{A}_n^{-1}\widehat{B}_n\widehat{A}_n^{-1}$, where

$$\widehat{B}_n = n^{-1} \sum_{i=1}^{n} \{f_{\mathcal{B}}(\cdot)\} \{f_{\mathcal{B}}(\cdot)\}^t \frac{\{\mathbf{Y}_i - f(\cdot)\}^2}{\widehat{\sigma}^4 g^4(\cdot, \widehat{\theta})}.$$

In some situations, the method of section A.3.3 can be used in place of the sandwich method.

With a flexible variance model which seems to fit the data fairly well, we prefer the covariance matrix estimate $n^{-1}\widehat{A}_n^{-1}$, because it can be much less variable than the sandwich estimator. Drum & McCullagh (1993) basically come to the same conclusion, stating that "unless there is good reason to believe that the assumed variance function is substantially incorrect, the model-based estimator seems to be preferable in applied work." Moreover, if the assumed variance function is clearly inadequate, most statisticians would find a better variance model and then use $n^{-1}\widehat{A}_n^{-1}$ with the better fitting model.

In addition to formal fitting methods, simple graphical displays exist to evaluate the models (A.21)–(A.22). Ordinary and weighted residual plots with smoothing can be used to understand departures from the assumed mean function, while absolute residual plots can be used to detect deviations from the assumed variance function. These graphical techniques are discussed in Chapter 2, section 7 of Carroll & Ruppert (1988).

A.5 Generalized Linear Models

Exponential families have density or mass function

$$f(y|\xi) = \exp\left\{\frac{y\xi - \mathcal{C}(\xi)}{\phi} + c(y,\phi)\right\}. \tag{A.27}$$

With superscripted (j) referring to the jth derivative, the mean and variance of $\mathbf{Y}$ are $\mu = \mathcal{C}^{(1)}(\xi)$ and $\phi\mathcal{C}^{(2)}(\xi)$, respectively. See, for example, McCullagh & Nelder (1989).

If ξ is a function of a linear combination of predictors, say $\xi = \Xi(\eta)$ where $\eta = (\beta_0 + \beta_x^t\mathbf{X} + \beta_z^t\mathbf{Z})$, then we have a generalized linear model. Generalized linear models include many of the common regression models, e.g., normal, logistic, Poisson and gamma. Consideration of specific models is discussed in detail in Chapter 6. Generalized linear models are mean and variance models in the observed data, and can be fit using QVF methods.

If we define $L = (\mathcal{C}^{(1)} \circ \Xi)^{-1}$, then $L(\mu) = \eta$; L is called the *link* function since it links the mean of the response and the linear predictor, η. If Ξ is the identity function, when we say that the model is canonical; this implies that $L = (\mathcal{C}^{(1)})^{-1}$, which is called the canonical link function. The link function L, or equivalently Ξ, should be chosen so that the model fits the data as well as possible. However, if the canonical link function fits reasonably well, then it is typically used, because doing so simplifies the analysis.

A.6 Bootstrap Methods

A.6.1 Introduction

The bootstrap is a widely used tool for analyzing the sampling variability of complex statistical methods. The basic idea is quite simple. One creates simulated data sets, called bootstrap data sets, whose distribution is equal to an estimate of the probability distribution of the actual data. Any statistical method that is applied to the actual data can also be applied to the bootstrap data sets. Thus, the empirical distribution of an estimator or test statistic across the bootstrap data sets can be used to estimate the actual sampling distribution of that statistic.

For example, suppose that $\widehat{\Theta}$ is obtained by applying some estimator to the actual data, and $\widehat{\Theta}^{(m)}$ is obtained by applying the same estimator to the mth bootstrap data set, $m = 1, ..., M$, where

M is the number of bootstrap data sets that we generate, and let $\bar{\Theta}$ be the average of $\hat{\Theta}^{(1)}, \ldots, \hat{\Theta}^{(m)}$. Then, the covariance matrix of $\hat{\Theta}$ can be estimated by

$$\widehat{\text{var}}(\hat{\Theta}) = (M-1)^{-1} \sum_{m=1}^{M} \left(\hat{\Theta}^{(m)} - \bar{\Theta}\right)\left(\hat{\Theta}^{(m)} - \bar{\Theta}\right)^t. \quad \text{(A.28)}$$

Despite this underlying simplicity, implementation of the bootstrap can be a complex, albeit fascinating, subject. There are many ways to estimate the probability distribution of the data, and it is not always obvious which way is most appropriate. Bootstrap standard errors are easily found from (A.28), and these can be plugged into (A.4) to get "normal theory" confidence intervals. However, these simple confidence intervals are not particularly accurate, and several improved bootstrap intervals have been developed. Comparing bootstrap standard errors and confidence intervals with traditional methods and comparing the various bootstrap intervals with each other requires the powerful methodology of Edgeworth expansions. Efron & Tibshirani (1993) give an excellent, comprehensive account of bootstrapping theory and applications. For more mathematical theory, including Edgeworth expansions, see Hall (1992). Here we give enough background so that the reader can understand how the bootstrap is applied to obtain standard errors in the examples.

A.6.2 Nonlinear Regression Without Measurement Error

To illustrate the basic principles of bootstrapping, we start with nonlinear regression without measurement error. Suppose that $\mathbf{Y}_i = f(\mathbf{Z}_i, \mathcal{B}) + \epsilon_i$ where the $\mathbf{Z}_i$ are, as usual, covariates measured without error, and the ϵ_i's are independent with the density of ϵ_i possibly depending on $\mathbf{Z}_i$. There are at least three distinct methods for creating the bootstrap data sets. Efron & Tibshirani (1993) call the first two methods *resampling pairs* and *resampling residuals.* The third method is a form of the *parametric bootstrap.*

Resampling pairs

Resampling pairs means forming a bootstrap data set by sampling at random *with replacement* from $\{(\mathbf{Y}_i, \mathbf{Z}_i)\}_i^n$. The advantage of this method is that it requires minimal assumptions. If ϵ_i has a dis-

tribution depending on $\mathbf{Z}_i$ in the real data, then this dependence is captured by the resampling since the $(\mathbf{Y}_i, \mathbf{Z}_i)$ pairs are never broken during the resampling. Therefore, standard errors and confidence intervals produced by this type of bootstrapping will be asymptotically valid in the presence of heteroscedasticity or other forms on nonhomogeneity. Besides this type of robustness, another advantage of resampling pairs is that it is easy to extend to more complex situations, such as measurement error models.

The disadvantage of resampling pairs is that the bootstrap data sets will have different sets of $\mathbf{Z}_i$'s than the original data. For example, if there is a high leverage point in the original data, it may appear several times or not at all in a given bootstrap data set. Therefore, this form of the bootstrapping estimates unconditional sampling distributions, not sampling distributions conditional on the $\mathbf{Z}_i$'s. Some statisticians will object to this, asking "even if the $\mathbf{Z}_i$'s are random, why should I care that I might have gotten different $\mathbf{Z}_i$'s than I did? I know the values of the $\mathbf{Z}_i$'s that I got, and I want to condition upon them." We feel that this objection is valid. However, as Efron & Tibshirani (1993) point out, often conditional and unconditional standard errors are nearly equal. In addition, unconditional variances are conservative in the sense of being larger than conditional variances.

Resampling residuals

The purpose behind resampling residuals is to condition upon the $\mathbf{Z}_i$'s. The ith residual is $e_i = \mathbf{Y}_i - f(\mathbf{Z}_i, \widehat{\mathcal{B}})$ where $\widehat{\mathcal{B}}$ is, say, the nonlinear least squares estimate. To create the mth bootstrap data set we first center the residuals by subtracting their sample mean, $\bar{e}$, and then draw $\{e_i^{(m)}\}_{i=1}^n$ randomly, with replacement, from $\{(e_i - \bar{e})\}_i^n$. Then we let $Y_i^{(m)} = f(\mathbf{Z}_i, \widehat{\mathcal{B}}) + e_i^{(m)}$. The mth bootstrap data set is $\{(Y_i^{(m)}, \mathbf{Z}_i)\}_{i=1}^n$. Notice that the bootstrap data sets have the same set of $\mathbf{Z}_i$'s as the original data, so that bootstrap sampling distributions are conditional on the $\mathbf{Z}_i$'s. By design, the distribution of the ith "error" in a bootstrap data set is independent of $\mathbf{Z}_i$. Therefore, resampling residuals is only appropriate when the ϵ_i's in the actual data are identically distributed, and is particularly sensitive to the homoscedasticity assumption.

The parametric bootstrap

The parametric bootstrap can be used when we assume a parametric model for the ϵ_i's. Let f be a known mean-zero density, say the standard normal density, ϕ. Assume that the density of ϵ_i is in the scale family $f(\cdot/\sigma)/\sigma$, $\sigma > 0$, and let $\widehat{\sigma}$ be a consistent estimator of σ, say the residual root-mean square if f is equal to ϕ. Then, as when resampling residuals, the bootstrap data sets are $\{(\mathbf{Y}_i^{(m)}, \mathbf{Z}_i)\}_i^n$, where $\mathbf{Y}_i = f(\mathbf{Z}_i, \widehat{\mathcal{B}}) + e_i^{(m)}$, but now the $\epsilon_i^{(m)}$s are, conditional on the observed data, iid from $f(\cdot/\widehat{\sigma})/\widehat{\sigma}$. Like resampling residuals, the parametric bootstrap estimates sampling distributions that are conditional on the $\mathbf{Z}_i$'s and requires that the ϵ_i's be independent of the $\mathbf{Z}_i$'s. In addition, like other parametric statistical methods, the parametric bootstrap is more efficient when the parametric assumptions are met, but possibly biased otherwise.

A.6.3 Bootstrapping Heteroscedastic Regression Models

Consider the QVF model

$$\mathbf{Y}_i = f(\mathbf{Z}_i, \mathcal{B}) + \sigma g(\mathbf{Z}_i, \mathcal{B}, \theta)\epsilon_i,$$

where the ϵ_i's are iid. The assumption of iid errors holds when $\mathbf{Y}_i$ given $\mathbf{Z}_i$ is normal, but this assumption precludes logistic, Poisson, and gamma regression, for example. This model can be fit by the methods of section A.4.2. To estimate the sampling distribution of the QVF estimators, bootstrap data sets can be formed by resampling from the set of pairs $\{(\mathbf{Y}_i, \mathbf{Z}_i)\}_i^n$ as discussed for nonlinear regression models in section A.6.2.

Resampling residual requires some reasonably obvious changes from section A.6.2. First, define the ith residual to be

$$e_i = \frac{\mathbf{Y}_i - f(\mathbf{Z}_i, \widehat{\mathcal{B}})}{\widehat{\sigma} g(\mathbf{Z}_i, \widehat{\mathcal{B}}, \widehat{\theta})} - \bar{e},$$

where $\bar{e}$ is defined so that the e_i's sum to 0. To form mth bootstrap data set, let $\{e_i^{(m)}\}_{i=1}^n$ be sampled with replacement from the residuals and then let

$$\mathbf{Y}_i^{(m)} = f(\mathbf{Z}_i, \widehat{\mathcal{B}}) + \widehat{\sigma} g(\mathbf{Z}_i, \widehat{\mathcal{B}}, \widehat{\theta}) e_i^{(m)}.$$

Note that $e_i^{(m)}$ is not the residual from the ith of the original observations, but rather is equally likely to be any of the n residuals

from the original observations. See Carroll and Ruppert (1991) for further discussion of bootstrapping heteroscedastic regression models, with application to prediction and tolerance intervals for the response.

A.6.4 Bootstrapping Logistic Regression Models

Consider the logistic regression model without measurement error,

$$\text{pr}(\mathbf{Y}_i = 1|\mathbf{Z}_i) = H(\beta_0 + \beta_z^T \mathbf{Z}_i),$$

where as elsewhere in this book, $H(v) = \{1+\exp(-v)\}^{-1}$. The general purpose technique of resampling pairs works here, of course. Resampling residuals is not applicable, since the residuals will have skewness depending on $\mathbf{Z}_i$ so are not homogeneous even after weighting as in section A.6.3. The parametric bootstrap, however, is easy to implement. To form the mth data set, fix the $\mathbf{Z}_i$'s equal to their values in the real data and let $\mathbf{Y}_i^{(m)}$ be Bernoulli with

$$\text{pr}(\mathbf{Y}_i^{(m)} = 1|\mathbf{Z}_i) = H(\widehat{\beta}_0 + \widehat{\beta}_z^t \mathbf{Z}_i).$$

A.6.5 Bootstrapping Measurement Error Models

In a measurement error problem, a typical data vector consists of $\mathbf{Z}_i$ and a subset of the following data: the response $\mathbf{Y}_i$, the true covariates $\mathbf{X}_i$, $\{\mathbf{W}_{i,j} : j = 1, ..., k_i\}$ which are replicate surrogates for $\mathbf{X}_i$, and a second surrogate $\mathbf{T}_i$. We divide the total collection of data into homogeneous data sets which have the same variables measured on each observation and are from a common source, e.g., primary data, internal replication data, external replication data, and internal validation data.

The method of "resampling pairs" ignores the various data subsets, and can often be successful (Efron, 1994). Taking into account the data subsets is better called "resampling vectors," and consists of resampling, with replacement, independently from each of the homogeneous data sets. This ensures that each bootstrap data set has the same amount of validation data, data with two replicates of $\mathbf{W}$, data with three replications, etc. as the actual data set. Although in principle we wish to condition on the $\mathbf{Z}_i$'s and resampling vectors does not do this, resampling vectors is a useful expedient and allows us to bootstrap any collection of data sets with minimal

assumptions. In the examples in this monograph, we have reported the "resampling pairs" bootstrap analyses, but because of the large sample sizes the reported results do not differ substantially from the "resampling vectors" bootstrap.

Resampling residuals is applicable to validation data when there are two regression models, one for $\mathbf{Y}_i$ given $(\mathbf{Z}_i, \mathbf{X}_i)$ and another for $\mathbf{W}_i$ given $(\mathbf{Z}_i, \mathbf{X}_i)$. One fits both models and resamples residuals from the first to create the bootstrap $\mathbf{Y}_i^{(m)}$'s and from the second to create the $\mathbf{W}_i^{(m)}$'s. This method generates sampling distributions that are conditional on the observed $(\mathbf{Z}_i, \mathbf{X}_i)$'s.

The parametric bootstrap can be used when the response, given the observed covariates, has a distribution in a known parametric family. For example, suppose one has a logistic regression model with internal validation data. One can fix the $(\mathbf{Z}_i, \mathbf{X}_i, \mathbf{W}_i)$ vectors of the validation data and create bootstrap responses as in section A.6.4 using $(\mathbf{Z}_i, \mathbf{X}_i)$ in place of $\mathbf{Z}_i$. Because $\mathbf{W}_i$ is a surrogate it is not used to create the bootstrap responses of validation data. For the nonvalidation data, one fixes the $(\mathbf{Z}_i, \mathbf{W}_i)$ vectors. Using regression calibration as described in Chapter 3, one fits an approximate logistic model for $\mathbf{Y}_i$ given $(\mathbf{Z}_i, \mathbf{W}_i)$ and again creates bootstrap responses distributed according to the fitted model. The bootstrap sampling distributions generated in this way are conditional on all observed covariates.

A.6.6 Bootstrap Confidence Intervals

As in section A.2.4, let $\Theta^t = (\theta_1, \Theta_2^t)$ where θ_1 is univariate, and suppose that we want a confidence interval for θ_1. The simplest bootstrap confidence interval is "normal based." The bootstrap covariance matrix in (A.28) is used for a standard error

$$\text{se}(\widehat{\theta}_1) = \sqrt{\widehat{\text{var}}(\widehat{\Theta})_{11}}.$$

This standard error is then plugged into (A.4) giving

$$\widehat{\theta}_1 \pm \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\widehat{\text{var}}(\widehat{\Theta}_{11})}. \tag{A.29}$$

The so-called "percentile" methods replace the normal approximation in (A.29) by percentiles of the empirical distribution of $\{(\widehat{\theta}_1^{(m)} - \widehat{\theta}_1)\}_1^M$. The best of these percentile methods are the so-

called BC_a and ABC intervals and they are generally more accurate than (A.29) in the sense of having a true coverage probability closer to the nominal $(1-\alpha)$; see Efron and Tibshirani (1993) for a full description of these intervals.

Hall (1992) has stressed the advantages of bootstrapping an asymptotically pivotal quantity, that is, a quantity whose asymptotic distribution is independent of unknown parameters. The percentile-t methods used the "studentized" quantity

$$t = \frac{\widehat{\theta}_1 - \theta_1}{\mathrm{se}(\hat{\theta}_1)}, \tag{A.30}$$

which is an asymptotic pivot with an large-sample standard normal distribution for all values of θ. Let $\mathrm{se}^{(m)}(\widehat{\theta}_1)$ be the standard error of $\widehat{\theta}_1$ computed from the mth bootstrap data set and let

$$t^{(m)} = \frac{\widehat{\theta}_1^{(m)} - \widehat{\theta}_1}{\mathrm{se}^{(m)}(\widehat{\theta}_1)}.$$

Typically, $\mathrm{se}^{(m)}(\widehat{\theta}_1)$ will come from an expression for the asymptotic variance matrix of $\widehat{\Theta}$ (e.g., the inverse of the observed Fisher information matrix given by (A.3)) rather than bootstrapping, since the latter would require two levels of bootstrapping, an outer level for $\{t^{(m)}\}_1^M$ and for each m an inner level for calculating the denominator of $t^{(m)}$. This would be very computationally expensive, especially for the nonlinear estimators in this monograph. Let $t_{1-\alpha}$ be the $(1-\alpha)$ quantile of $\{|t^{(m)}|\}_1^M$. Then the symmetric percentile-t confidence interval is

$$\widehat{\theta}_1 \pm \mathrm{se}(\widehat{\theta}_1)\, t_{1-\alpha}. \tag{A.31}$$

Note that $\mathrm{se}(\widehat{\theta}_1)$ is calculated from the original data in the same way that $\mathrm{se}^{(m)}(\widehat{\theta}_1)$ is calculated from the mth bootstrap data set.

References

Amemiya, Y. (1985). Instrumental variable estimator for the nonlinear errors in variables model. *Journal of Econometrics*, 28, 273-289.

Amemiya, Y. (1990a). Instrumental variable estimation of the nonlinear measurement error model. In *Statistical Analysis of Measurement Error Models and Application*, P. J. Brown and W. A. Fuller, editors. American Mathematics Society, Providence.

Amemiya, Y. (1990b). Two stage instrumental variable estimators for the nonlinear errors in variables model. *Journal of Econometrics*, 44, 311-332.

Amemiya, Y. & Fuller, W. A. (1988). Estimation for the nonlinear functional relationship. *Annals of Statistics*, 16, 147-160.

Armstrong, B. (1985). Measurement error in generalized linear models. *Communications in Statistics, Series B*, 14, 529-544.

Armstrong, B. G., Whittemore, A. S., & 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.

Baker, S. G. (1991). Evaluating a new test using a reference test with estimated sensitivity and specificity. *Communications in Statistics*, 20, 2739–2752.

Baker, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *Journal of Computational and Graphical Statistics*, 1, 63–76.

Baker, S. G. (1994a). Composite linear models for incomplete multinomial data, *Statistics in Medicine*, 13 609–622.

Baker, S. G. (1994b). Evaluating multiple diagnostic tests with partial verification, *Biometrics*, in press.

Baker, S. G., Wax, Y. & Patterson, B. H. (1993). Regression analysis of grouped survival data: informative censoring and double sampling. *Biometrics*, 49, 379–389.

Beaton, G. H., Milner, J. & Little, J. A. (1979). Sources of variation in 24-hour dietary recall data: implications for nutrition study design and interpretation. *American Journal of Clinical Nutrition*, 32, 2546-

2559.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis, second edition*, Springer-Verlag, New York.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45, 164-180.

Bickel, P. J. & Ritov, Y. (1987). Efficient estimation in the errors in variables model. *Annals of Statistics*, 15, 513-540.

Boggs, P. T., Spiegelman, C. H., Donaldson, J. R. and Schnabel, R. B. (1988). A computational examination of orthogonal distance regression. *Journal of Econometrics*, 38, 169-201.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Breslow, N. E. & Cain, K. C. (1988). Logistic regression for two–stage case–control data. *Biometrika*, 75, 11–20.

Breslow, N. E. & Holubkov, R. (1995). Two-sate case-control studies: relative efficiency of weighted likelihood and pseudolikelihood estimation methods. Preprint.

Britt, H. I. & Luecke, R. H. (1973). The estimation of parameters in nonlinear implicit models. *Technometrics*, 15, 233-247.

Brown, B. W. and Mariano, R. S. (1993). Stochastic simulations for inference in nonlinear errors-in-variables models. *Handbook of Statistics*, Vol. 11, 611–627. North Holland, New York.

Buonaccorsi, J. P. (1988). Errors in variables with systematic biases. *Communications in Statistics, Theory and Methods*, 18(3), 1001- 1021.

Buonaccorsi, J. P. (1990a). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075–1082.

Buonaccorsi, J. P. (1990b). Double sampling for exact values in the normal discriminant model with application to binary regression. *Communications in Statistics, Series A*, 19, 4569–4586.

Buonaccorsi, J. P. (1991) Measurement error, linear calibration and inferences for means. *Computational Statistics and Data Analysis*, 11, 239-257.

Buonaccorsi, J. P. (1993). Linear measurement error in the response in longitudinal/repeated measures studies. Preprint.

Buonaccorsi, J. P. & Tosterson, T. (1993). Correcting for nonlinear measurement error in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods*, 22, 2687–2702.

Burr, D. (1988). On errors–in–variables in binary regression – Berkson case. *Journal of the American Statistical Association*, 83, 739–743.

Buzas, J. S. & Stefanski, L. A. (1995). A note on corrected score estimation. *Statistics & Probability Letters*, in press.

Buzas, J. S. & Stefanski, L. A. (1996a). Instrumental variable estimation

in a probit measurement error model. Preprint.

Buzas, J. S. & Stefanski, L. A. (1996b). Instrumental variables estimation in generalized linear measurement error models. Preprint.

Cain, K. C. & Breslow, N. E. (1988). Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology*, 128, 1198–1206.

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

Carroll, R. J., Eltinge, J. L. & Ruppert, D. (1993). Robust linear regression in replicated measurement error models. *Letters in Statistics & Probability*, 16, 169–175.

Carroll, R. J., Gail, M. H., & Lubin, J. H. (1993). Case–control studies with errors in predictors. *Journal of the American Statistical Association*, 88, 177–191.

Carroll, R. J. & 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. & Gallo, P. P. (1984). Comparisons between maximum likelihood and method of moments in a linear errors-in-variables regression model. *Design of Experiment: Ranking and Selection*, T. J Santner and A. C. Tamhane, eds., Marcel Dekker, New York.

Carroll, R. J., Gallo, P. P. & 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.

Carroll, R. J. & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184–1186.

Carroll, R. J., Knickerbocker, R. H., & Wang, C. Y. (1995). Dimension reduction in semiparametric measurement error models. *Annals of Statistics*, in press.

Carroll, R. J. & 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., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, in press.

Carroll, R. J. & Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, London.

Carroll, R. J. & Ruppert, D. (1991). Prediction and tolerance intervals with transformation and/or weighting. *Technometrics*, 33, 197–210.

Carroll, R. J. & Spiegelman, C. H. (1986). The effect of small measurement error on precision instrument calibration. *Journal of Quality*

Technology, 18, 170–173.

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

Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T. & Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika, 71*, 19–26.

Carroll, R. J. & Stefanski, L. A. (1990) Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652–663.

Carroll, R. J. & Stefanski, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13, 1265–1282.

Carroll, R. J. & Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, 53, 573–585.

Carroll, R. J., Wang, S. & Wang, C. Y. (1995). Asymptotics for prospective analysis of stratified logistic case–control studies. *Journal of the American Statistical Association*, 90, 157–169.

Casella, G. & Berger, R. L. (1990). *Statistical Inference*. Wadsworth & Cole, Pacific Grove, CA.

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *American Staistician*, 46, 167–174.

Chan, N. N. & Mack, T. K. (1984). Heteroscedastic errors in a linear functional relationship. *Biometrika*, 71, 212-215.

Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, 8, 1095–1106.

Chen, T. T. (1992). Reply to Ekholm. *Statistics in Medicine*, 11, 271–275.

Cheng, C. L. & van Ness, J. W. (1988). Generalized M–estimators for errors in variables regression. *Annals of Statistics*, 20, 385–397.

Cheng, C. L. & Tsai, C. L. (1992). Diagnostics in measurement error models. Preprint.

Clark, R. (1982). Logistic regresison with measurement error in predictors. Unpublished Ph.D. dissertation, Department of Biostatistics, University of North Carolina.

Clayton, D. G. (1991). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health*, Dwyer, J. H., Feinleib, M., Lipsert, P., et al., editors, pages 301-331. Oxford University Press, New York.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W. & Devlin, S. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596–610.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, 10, 637-666.

Cook, J. & Stefanski, L. A. (1995). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314–1328.

Copas, J. B. (1972). The likelihood surface in the linear functional relationship problem. *Journal of the Royal Statistical Society, Series B*, 34, 190-202.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federal Proceedings*, 21, 58–61.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman & Hall, London.

Crouch, E. A. & Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: applications to logistic–normal models. *Journal of the American Statistical Association*, 85, 464–467.

Davidian, M. & Carroll R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079-1091.

Davidian, M. & Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80, 475–488.

DeGracie, J. S. & 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.

DeGruttola, V. & Tu, X. M. (1991). Modeling the relationship between progression of CD4–lymphocyte count and survival time. In *AIDS Epidemiology: Methodological Issues*, N. P. Jewell, K. Dietz & V. T. Farewell, editors, pp 275–296. Birkhäuser, Boston.

Desmond, A. F. (1989), Estimating Equations, Theory of, In *Encyclopedia of Statistical Sciences,* S. Kotz, and N. L. Johnson, editors, pp 56–59, John Wiley & Sons, New York.

Dosemeci, M., Wacholder, S. & Lubin, J. H. (1990). Does nondifferential misclassification of exposure always bias a true effect towards the null value? *American Journal of Epidemiology*, 132, 746–748.

Drum, M. & McCullagh, P. (1993). Comment on the paper by Fitzmaurice, Laird & Rotnitzky. *Statistical Science*, 8, 300–301.

Duan, N. & Li, K. C. (1991). Slicing regression: a link–free regression method. *Annals of Statistics*, 19, 505–530.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling*

Plans. SIAM: Philadelphia.

Efron, B. (1994). Missing data, imputation and the bootstrap. *Journal of the American Statistical Association*, 463–475.

Efron, B. & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expectedFisher information. it Biometrika, 65, 457–487.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

Ekholm, A. (1991). Algorithms versus models for analyzing data that contain misclassification errors. *Biometrics*, 47, 1171–1182.

Ekholm, A. (1992). Letter to the editor concerning the paper by Chen. *Statistics in Medicine*, 11, 271–275.

Ekholm, A., Green, M. & Palmgren, J. (1986). Fitting exponential family nonlinear models in GLIM 3.77. *GLIM Newsletter*, 13, 4–13.

Ekholm, A. & Palmgren, J. (1987). Correction for misclassifiction using doubly sampled data. *Journal of Official Statistics*, 3, 419–429.

Espeland, M. A. & Hui, S. L. (1987). A general approach to analyzing epidemiologic data that contains misclassification errors. *Biometrics*, 43, 1001–1012.

Espeland, M. A. & 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.

Fan, J. (1991a). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19, 1257-1272.

Fan, J. (1991b). Asymptotic normality for deconvolving kernel density estimators, *Sankhyā, Series A*, 53, 97-110.

Fan, J. (1991c). Global behavior of deconvolution kernel estimates. *Statistica Sinica*, 1, 541-551.

Fan, J. (1992a). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20, 23-37.

Fan, J. & Masry, E. (1993). Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes *Journal of Multivariate Analysis*, 43, 237–271.

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

Fan, J., Truong, Y. K. & Wang, Y. (1991). Nonparametric function estimation involving errors-in-variables. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.), 613-627.

Flanders, W. D. & Greenland, S. (1991). Analytic methods for two stage case-control studies and other stratified designs. *Statistics in Medicine*, 10, 739–747.

Fleming, T., Prentice, R., Pepe, M. & Glidden, D. (1993). Surrogate and auxiliary endpoints in clinical trials, with potential applications

in cancer and AIDS research. *Statistics in Medicine.*

Forbes, A. B. & Santner, T. J. (1994). Estimators of odds ratio regression parameters in matched case–control studies with covariate measurement error.

Freedman, L. S., Carroll, R. J. & 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.

Freedman, L., Schatzkin, A. & Wax, Y. (1990). The effect of dietary measurement error on the sample size of a cohort study. *American Journal of Epidemiology*, 132, 1185–1195.

Friedman, J. & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76, 817–823.

Fuller, W. A. (1980). Properties of some estimators for the errors in variables model. *Annals of Statistics*, 8, 407-422.

Fuller, W. A. (1987). *Measurement Error Models.* John Wiley & Sons, New York.

Gail, M. H., Tan, W. Y. & Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75, 57–64.

Gail, M. H., Wieand, S. & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71, 431–444.

Gallo, P. P. (1982). Consistency of some regression estimates when some variables are subject to error. *Communications in Statistics, Series A*, 11, 973-983.

Ganse, R. A., Amemiya, Y. & Fuller, W. A. (1983). Prediction when both variables are subject to error, with application to earthquake magnitude. *Journal of the American Statistical Association*, 78, 761-765.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling–based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. & Rubin, D. R. (1992). Infertence from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

Geman, S. & Geman D. (1984). Stochastic relaxation,, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57, 1317–1339.

Geyer, C. J. (1992). Practical Markov chain Monte–Carlo. *Statistical Science*, 7, 473–511.

Geyer, C. J. & Thompson, E. A. (1992). Constrined Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the*

Royal Statistical Society, Series B, 54, 657–700.

Gleser, L. J. (1981). Estimation in a multivariate errors in variables regression model: large sample results. *Annals of Statistics*, 9, 24–44.

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*, P. J. Brown and W. A. Fuller, editors. American Mathematics Society, Providence.

Gleser, L. J., Carroll, R. J. & Gallo, P. P. (1987). The limiting distribution of least squares in an errors-in-variables linear regression model. *Annals of Statistics*, 15, 220–233.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208–1211.

Godambe, V. P. (1991), *Estimating functions*, Oxford: Clarendon Press, New York.

Green, P., and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, New York.

Greenland, S. (1988a). Statistical uncertainty due to misclassification: implications for validation substudies. *Journal of Clinical Epidemiology*, 41, 1167–1174.

Greenland, S. (1988b). On sample size and power calculations for studies using confidence intervals. *American Journal of Epidemiology*, 128, 231–236.

Greenland, S. (1988c). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, 7, 745–757.

Greenland, S. & Kleinbaum, D. G. (1983). Correcting for misclassification in two–way tables and pair–matched studies. *International Journal of Epidemiology*, 12, 93–97.

Griffiths, P. & Hill, I. D. (1985). *Applied Statistics Algorithms*. Horwood, London.

Hall, P. (1989). On projection pursuit regression. *Annals of Statistics*, 17, 573-588.

Hall, P. G. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, New York.

Härdle, W. & Stoker, T. M. (1989). Investigating smooth multipleregression by the method of average derivatives. *Journal of the American Statistical Association*, 84, 986-995.

Hasenabeldy, N., Fuller, W. A. & Ware, J. (1988). Indoor air pollution and pulmonary performance: investigating errors in exposure assessment. *Statistics in Medicine*, 8, 1109–1126.

Hastie, T. & Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84, 502–516.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall, New York.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Henderson, M. M., Kushi, L. H., Thompson, D. J., et al. (1990). Feasibility of a randomized trial of a low–fat diet for the prevention of breast cancer: dietary compliance in the Women's Health Trial Vanguard Study. *Preventive Medicine*, 19, 115–133.

Hildesheim, A., Mann, V., Brinton, L. A., Szklo, M., Reeves, W. C. & 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.

Horowitz, J. L. & Markatou, M. (1993). Semiparametic estimation of regression models for panel data. Preprint.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, 11, 271–283.

Hsieh, D. A., Manski, C. F. & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, 80, 651–662.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium*, 1, 221-233.

Hughes, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, 49, 1056–1066.

Hunter, W. G. & Lamboy, W. F. (1981). A Bayesian analysis of the linear calibration problem. *Technometrics*, 23, 323-328.

Hwang, J. T. (1986). Multiplicative errors in variables models with applications to the recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association*, 81, 680-688.

Hwang, J. T. and Stefanski, L. A. (1994). Monotonicity of regression functions in structural measurement error models. *Statistics & Probability Letters*, 20, 113-116.

Johnson, N. L. & Kotz, S. (1970). *Distributions in Statistics*, Vol. 2. Boston, Houghton-Mifflin.

Jones, D. Y., Schatzkin, A., Green, S. B., Block, G. Brinton, L. A., Ziegler, R. G., Hoover, R. & 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.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler,

J., Hulley, S. B. & 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.

Kelly, G. (1984). The influence function in the errors in variables problem. *Annals of Statistics*, 12, 87-100.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69, 19-27.

Ketellapper, R. H. & Ronner, R. E. (1984). Are robust estimation methods useful in the structural errors in variables model? *Metrika*, 31, 33-41.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 886-906.

Küchenhoff, H. (1990). *Logit- und Probitregression mit Fehlen in den Variabeln.* Anton Hain, Frankfurt am Main.

Küchenhoff, H. & Carroll, R. J. (1995). Biases in segmented regression with errors in predictors. Preprint.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805–811.

Landin, R., Carroll, R. J. & 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*, in press.

Lee, L. F. & Sepanski, J. H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data, *Journal of the American Statistical Association*, 90, 130–140.

Lesperance, M. L. (1989). Mixture models as applied to models involving many incidental parameters. unpublished Ph.D. dissertation, University of Waterloo, Dept. of Statistics and Actuarial Science.

Lesperance, M. L. & Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, 87, 120-126.

Li, B. & McCullagh, P. (1994). Potential functions and conservative estimating functions. *Annals of Statistics*, 22, 340-356.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342.

Liang, K. Y. & Liu, X. H. (1991). Estimating equations in generalized linear models with measurement error. In *Estimating Functions*, V. P. Godambe, editor. Clarendon Press, Oxford.

Lindley, D. V. (1953). Estimation of a functional relationship. *Biometrika*, 40, 47-49.

Lindley, D. V. & El Sayyad, G. M. (1968). The Bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society, Series B*, 30, 190-202.

Lindsay, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Annals of Statistics*, 13, 914–32.

Lindsay, B. G. (1983). The geometry of mixture likelihoods, Part I: A general theory. *Annals of Statistics*, 11,86-94.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data* John Wiley & Sons, New York.

Liu, M. C. & Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17, 399–410.

Liu, M. C. & Taylor, R. L. (1990). Simulation and computation of a nonparametric density estimator for the deconvolution problem. *Statistical Computation and Simulation*, 35, 145–167.

Liu, K., Stamler, J., Dyer, A., McKeever, J. & McKeever, P. (1978). Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. *Journal of Chronic Diseases*, 31, 399-418.

Liu, X. & Liang, K. Y. (1992). Efficacy of repeated measures in regression models with measurement error. *Biometrics*, 48, 645–654.

Lord, F. M. (1960). Large sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, 55, 307-321.

MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A., & Stamler, J. (1990). Blood pressure, stroke and coronary heart disease: Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, 335, 765–774.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54, 173-205.

Mallick, B. K. & Gelfand, A. E. (1995). Semiparametric errors-in-variables models: a Bayesian approach. *Journal of Statistical Planning and Inference*, in press.

Manski, C. F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Jounral of Econometrics*, 27, 313- 333.

Manski, C. F. & Thompson, T. S. (1986). Operational characteristics of maximum score estimation *Jounral of Econometrics*, 28, 85–108.

Marazzi, A. (1980). ROBETH, a subroutine library for robust statistical procedures. COMPSTAT 1980, Proceedings in Computational Statistics, Physica, Vienna.

Masry, E. & Rice, J. A. (1992). Gaussian deconvolution via differentiation. *Canadian Journal of Statistics*, 20, 9–21.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models, second edition.* Chapman & Hall, London.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 239–265.

McLeish, D. L., & Small, C. G. (1988), *The Theory and Applications of Statistical Inference Functions.* Springer-Verlag, New York.

Michalek, J. E. & Tripathi, R. C. (1980). The effect of errors in diagnosis and measurement on the probability of an event. *Journal of the American Statistical Association*, 75, 713–721.

Monahan, J. & Stefanski, L. A. (1992). Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. In *Handbook of the Logistic Distribution*, pp 529–540, N. Balakrishnan, editor. Marcel Dekker, New York.

Müller, H-G. (1988). *Nonparametric Analysis of Longitudinal Data.* Springer-Verlag, Berlin.

Müller, P. & Roeder, K. (1995). A Bayesian semiparametric model for case-control studies with errors in variables. Preprint.

Nakamura, T. (1990). Corrected score functions for errors–in–variables models: methodology and application to generalized linear models. *Biometrika*, 77, 127–137.

Nakamura, T. (1992). Proportional hazards models with covariates subject to measurement error. *Biometrics*, 48, 829–838.

Newey, W. K. (1991). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5, 99–135.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1–32.

Palmgren, J. (1987). Precision of double sampling estimators for comparing two probabilities. *Biometrika*, 74, 687–694.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, 79, 355–365.

Pepe, M. S. & 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.

Pepe, M. S., Reilly, M. & Fleming, T. R. (1994). Auxilliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42, 137–160.

Pepe, M. S., Self, S. G. & Prentice, R. L. (1989). Further results in covariate measurement errors in cohort studies with time to response

data. *Statistics in Medicine*, 8, 1167–1178.

Pierce, D. A., Stram, D. O., Vaeth, M., Schafer, D. (1992). Some insights into the errors in variables problem provided by consideration of radiation dose–response analyses for the A–bomo survivors. *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. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69, 331-342.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8, 431–440.

Prentice, R. L., Pepe, M. & Self, S. G. (1989). Dietary fat and breast cancer: a review of the literature and a discussion of methodologic issues. *Cancer Research*, 49, 3147–3156.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika*, 66, 403–411.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353–360.

Racine–Poon, A., Weihs, C. & Smith, A. F. M. (1991). Estimation of relative potency with sequential dilution errors in radioimmunoassay. *Biometrics*, 47, 1235–1246.

Rao, C. R., (1947). Large-sample test of statistical hypotheses concerning several parameters with applications to problems of estimation, *Proceedings Cambridge Philosophical Society*, 44, 50–57.

Reilly, M. & Pepe, M. S. (1994). The mean score and hot deck methods for missing and surrogate covariate data. Preprint.

Richardson, S. & Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138, 430–442.

Richardson, S. & Gilks, W. R. (1994). Conditional independence models for epidemiological studies with covariate measurement error. *Biometrics*.

Ritter, C. & Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy Gibbs stopper. *Journal of the American Statistical Association*, 87, 861–868.

Robins, J. M., Hsieh, F. & Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B*, 57, 409–424.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Roeder, K., Carroll, R. J. & Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, in press.

Ronchetti, E. (1982). Robust testing in linear models: the infinitesimal approach. Ph.D. Thesis. ETH, Zurich.

Rosner, B., Spiegelman, D. & 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. & 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. & Streibig, J. C. (1989). Random effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45, 349-362.

Ruppert, D. (1985). M-estimators, In *Encyclopedia of Statistical Sciences, vol. 5*, S. Kotz and N. L. Johnson, editors, pp 443–449. John Wiley & Sons, New York.

Ruppert, D., A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics*, 45, 637–362.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, 1346–1370.

Satten, G. A. & 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. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74, 385-391.

Schafer, D. (1992). Replacement methods for measurement error models. Preprint.

Schafer, D. (1993). Likelihood analysis for probit regression with measurement errors. *Biometrika*, 80, 899–904.

Schafer, D. & James, I. R. (1991). Weibull regression with covariate measurement errors and assessment of unemployment duration dependence. Preprint.

Schmid, C. H. & Rosner, B. (1993). A Bayesian approach to logistic regression models having measurement error following a mixture distribution. *Statistics in Medicine*, 12, 1141–1153.

Schrader, R. M. & Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood criterion. *Biometrika*, 67, 93-101.

Scott, A. J. & Wild, C. J. (1991). Fitting logistic regression models in stratified case–control studies. *Biometrics*, 47, 497–510.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Compo-*

nents, John Wiley & Sons, New York.

Sepanski, J. H. (1992). Score tests in a generalized linear model with surrogate covariates, *Statistics & Probability Letters*, 15, 1–10.

Sepanski, J. H. & Carroll, R. J. (1993). Semiparametric quasilikelihood and variance function estimation in measurement error models. *Journal of Econometrics*, 58, 226–253.

Sepanski, J. H. & Carroll, R. J. & Knickerbocker, R. (1994). A semiparametric correction for attenuation. *Journal of the American Statistical Association*, 89, 1366–1373.

Sepanski, J. H. & Lee, L. F. (1992). Semiparametric estimation of nonlinear errors–in–variables models with a validation study. Preprint.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London and New York.

Smith, A. F. M. & Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling–resampling perspective. *American Statistician*, 46, 84–88.

Spiegelman, C. H. (1986). Two pitfalls of using standard regression diagnostics when both X and Y have measurement error. *The American Statisticain*, 40, 245-248.

Spiegelman, D. (1994). Cost-efficient study designs for relative risk modeling with covariate measurement error. *Journal of Statistical Planning and Inference*, 42, 187–208..

Spiegelman, D. & Gray, R.(1991). Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics*, 47, 851–869.

Sposto, R., Preston, D. L., Shimizu, Y. & Mabuchi, K. (1992). The effect of diagnostic misclassification on non–cancer and cancer mortality dose response in A–bomb survivors. *Biometrics*, 48, 605–618.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, 72, 583-592.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, 18, 4335–4358.

Stefanski, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics & Probability Letters*, 9, 229–235.

Stefanski, L. A. & Buzas, J. S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association*, in press.

Stefanski, L. A. & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics*, 13, 1335–1351.

Stefanski, L. A. & Carroll, R. J. (1986). Deconvoluting kernel density estimators. Technical report.

Stefanski, L. A. & Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74, 703-716.

Stefanski, L. A. & Carroll, R. J. (1990a). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society, Series B*, 52, 345–359.

Stefanski, L. A. & Carroll, R. J. (1990b). Structural logistic regression measurement error models. *Proceedings of the Conference on Measurement Error Models*, P. J. Brown & W. A. Fuller, editors.

Stefanski, L. A. & Carroll, R. J. (1990c). Deconvoluting kernel density estimators. *Statistics*, 21, 165–184.

Stefanski, L. A. & Carroll, R. J. (1991). Deconvolution based score tests in measurement error models. *Annals of Statistics*, 19, 249–259.

Stefanski, L. A. & Cook, J. (1996). Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, in press.

Stephens, D. A. & Dellaportas, P. (1992). Bayesian Analysis of Generalized Linear Models with Covariate Measurement Error. In *Bayesian Statistics 4*, Bernado, J. M., Berger, J. O., Dawid, A. p. and Smith, A. F. M. (Eds.), pp 813–820, Oxford University Press.

Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, second edition.* Springer–Verlag, New York.

Thomas, D. C., Gauderman, J. & Kerber, R. (1993). A nonparametric Monte-Carlo approach to adjustment for covariate measurement errors in regression analysis. Preprint.

Thomas, D., Stram, D. & Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annu. Rev. Publ. Health*, 14, 69–93.

Thompson, F. E., Sowers, M. F., Frongillo, E. A. & Parpia, B. J. (1992). Sources of fiber and fat in diets of U.S. women aged 19–50: implications for nutrition education and policy. *American Journal of Public Health*, 82, 695–718.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York.

Tosteson, T., Stefanski, L. A. & Schafer D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine*, 8, 1139-1147.

Tosteson, T. & Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates.

Biometrika, 75, 507-514.

Tosteson, T. D. & Ware, J. H. (1990). Designing a logistic regression study using surrogate measures of exposure and outcome. *Biometrika*, 77, 11–20.

Tsiatis, A. A., DeGruttola, V., Strawderman, R. L., Dafni, U., Propert, K. J. & Wulfsohn, M. (1992). The relationship of CD4 counts over time to survival in patients with AIDS: is CD4 a good surrogate marker? In *AIDS Epidemiology: Methodological Issues*, N. P. Jewell, K. Dietz & V. T. Farewell, editors. Birkhäuser, Boston.

Tukey, J. (1958), "Bias and Confidence in Not Quite Large Samples," *Annals of Mathematical Statistics*, 29, 614.

van der Vaart, A. (1988). Estimating a real parameter in a class of semiparametric models. *Annals of Statistics*, 16, 1450–1474.

Ulm, K. (1991). A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine*, 10, 341–349.

Wang, C. Y., Wang, S. & Carroll, R. J. (1995). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*, in press.

Wang, N., Carroll, R. J. & Liang, K. Y. (1995). Quasilikelihood and variance functions in measurement error models with replicates. *Biometrics*, in press.

Weinberg, C. R., Umbach, D. M. & Greenland, S. (1993). When will nondifferential misclassification preserve the direction of a trend? Preprint.

Weinberg, C. R. & Wacholder, S. (1993). Prospective analysis of case–control data under general multiplicative–intercept models. *Biometrika*, 80, 461–465.

Whittemore, A. S. (1989). Errors in variables regression using Stein estimates. *American Statistician*, 43, 226–228.

Whittemore, A. S. & Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates. *Applied Statistics*, 40, 81–93.

Whittemore, A. S. & Keller, J. B. (1988). Approximations for regression with covariate measurement error. *Journal of the American Statistical Association*, 83, 1057-1066.

Wild, C. J. (1991). Fitting prospective regression models to case–control data. *Biometrika*, 78, 705–717.

Willett, W. C. (1989). An overview of issues related to the correction of non–differential exposure measurement error in epidemiologic studies. *Statistics in Medicine*, 8, 1031–1040.

Wittes, J., Lakatos, E. & Probstfield, J. (1989). Surrogate endpoints in clinical trials: cardiovascular trials. *Statistics in Medicine*, 8, 415–425.

Willett, W. C., Meir, J. S., Colditz, G. A., Rosner, B. A., Hennekens, C.

H. & Speizer, F. E. (1987). Dietary fat and the risk of breast cancer. *New England Journal of Medicine*, 316, 22–25.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H. & Speizer, F. E. (1985). Reproducibility and validity of a semiquantitative food frequency questionairre. *American Journal of Epidemiology*, 122, 51-65.

Wolfe, D. A. (1976). On testing equality of related correlation coefficients. *Biometrika*, 63, 214–215.

Wu, M. L., Whittemore, A. S. & Jung, D. L. (1986). Errors in reported dietary intakes. *American Journal of Epidemiology*, 124, 826-835.

Zamar, R. H. (1988). Orthogonal regression M-estimators. *Biometrika*, 76, 149–154.

Zamar, R. H. (1992). Bias–robust estimation in the errors in variables model. *Annals of Statistics*, 20, 1875–1888.

Zeger, S. L. & Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Annals of Statistics*, 18, 806–831.

Zhao, L. P. & Lipsitz, S. (1992). Designs and analysis of two–stage studies. *Statistics in Medicine*, 11, 769–782.

Zhao, L. P., Lipsitz, S. & Lew, D. (1994). Regression analysis with missing covariate data using estimating equations. Preprint.

Zhao, L. P., Prentice, R. L. & Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B*, 54, 805–812.

Author index

Abbott, R. A., ix, 65, 88, 141
Amemiya, Y., ix, 37, 69, 108
Armstrong, B., ix, 40, 245
Bailey, K., ix, 65, 88, 141
Baker, S. G., 253,
Beaton, G., 2, 43
Berger, R. L., 165
Berkson, J., 150
Bickel, P. J., 138
Boggs, P., 37
Box, G., 165
Breslow, N., 185, 253
Brown, B., 143, 162
Buonaccorsi, J. P., 233, 245, 252
Burr, D., 150
Buzas, J., 108, 135, 138
Byar, D., ix
Cain, K. B., 185, 253
Carroll, R. J., ix, 11, 24, 35, 50, 65, 79, 88, 97, 108, 122, 136, 160, 173, 201, 210, 238, 244, 261, 269
Casella, G., 166, 271
Chen, T. T., 253
Cheng, C. L., 37
Clayton, D. G., 40, 253, 254
Cleveland, W., 224
Cochran, W. G., 37
Cook, J., 79, 83, 87, 99
Cox, D. R., 254, 261
Cressie, N. A. C., 56
Crouch, E. A. C., 64, 141
Davidian, M., 148, 271
Delaportas, A. P., 167
Desmond, A. F., 261
Devlin, S., 224
Dosemeci, M., 36
Drum, M., 272
Duan, N., 199
Dwyer, J., 247
Efron, B., 87, 258, 274, 277, 279
Ekholm, A., 253
Eltinge, J. L., 37
Espeland, M. A., 253
Fan, J., 216, 218
Flanders, W. D., 185
Fleming, T., 238
Freedman, L. F., xiii, 2, 8, 24, 202
Friedman, J., 199
Fuller, W. A., ix, xii, 21, 31, 33, 37, 40, 69, 108, 143, 145
Gail, M. H., ix, 173, 208, 245, 251
Gallant, R., 148
Gallo, P. P., 37
Ganse, R. A., 37
Gauderman, J., 247
Gelfand, A., 166, 167, 169, 178
Gelman, A., 169
Geman, D., 166
Geman, S., 166
George, E. I., 166
Geyer, C., 153, 169
Gilks, W. R., 167
Gleser, L. J., 26, 37, 40, 47, 122

Godambe, V. P., 261
Gong, G., 141
Gray, R., 252
Greenland, S., 185, 252, 253
Griffiths, P., 266
Härdle, W., 199, 223
Hall, P., 199, 216, 218, 274, 279
Hasenabeldy, N., 37
Hastie, T., 223
Hastings, W. K., 166
Henderson, M. M., 24
Hettmansperger, T. P., 265
Hill, I. D., 266
Hinkley, D., 258, 261
Holubkov, R., 185
Hotelling, H., 71
Hsieh, D. A., 185
Hsieh, F., 187, 191, 195
Huber, P. J., 261
Hughes, M. D., 254
Hui, S. L., 253
Hwang, J. T., 35, 37
Johnson, N. L., 64
Jones, D. Y., 1, 42
Kalbfleisch, J. D., 248
Kannel, W., 4, 11
Keller, J. B., 69
Kent, J. T., 265
Kerber, R., 247
Ketellapper, R. H., 37
Knickerbocker, R. H., 186
Kotz, S., 64
Küchenhoff, H., 79, 83, 97, 101, 141, 144, 160
Kupper, L., 16, 153, 250
Laird, N., 247
Lakatos, E., 229
Lan, K. K. G., ix, 65, 88, 141
Landin, R., 18
Lesperance, M. L., 248
Lew, D., 192, 247
Li, B., 265
Li, K. C., 199, 201, 204
Liang, K. Y., 5, 14, 47, 64, 141
Lindley, D. V., 37, 167
Lindsay, B. G., 138, 247, 248
Lipsitz, S., 185, 192, 247, 253
Little, R. J. A., 18, 169
Liu, M. C., 216
Liu, X., 5, 47, 64
Lombard, F., 79, 83, 101
Lord, F. M., 37
Lubin, J., 36, 173, 245, 251
MacMahon, S., 5, 11, 35
Madansky, A., 37
Makov, U. E., 247
Mallick, B. K., 167
Manski, C., 185, 199
Marazzi, A., 266
Masry, E., 216
McCullagh, P., xii, 126, 136, 265, 269, 272
McCulloch, C. E., 271
McFadden, D., 161, 185
McLeish, D. L., 261
Michalek, J. E., 245
Monahan, J., 64, 163
Müller, H. G., 167, 223
Nakamura, T., 131, 139, 254
Nelder, J. A., xii, 126, 136, 269
Newey, W., 187, 188, 189, 191, 192, 193, 195
Odoroff, C. L., 253
Palmgren, J., 253
Patterson, B. H., 253
Pepe, M., 187, 238, 239, 240, 241, 254
Piantadosi, S., 208
Pierce, D. A., 5
Prentice, R. L., ix, 40, 43, 149, 235, 244, 254, 255
Probstfield, J., 229
Pyke, R., 244
Quenouille, M. H., 87
Racine–Poon, A., 69
Rao, C. R., 260

Rice, J., 216
Richardson, S., 167
Ritov, Y., 138
Ritter, C., 169
Robins, J., 167, 187, 191, 192, 193, 195, 246, 247, 251
Ronchetti, E., 265
Ronner, R. E., 37
Rosner, B., 3, 32, 40, 46, 63, 167
Rotnitzky, A., 191, 246
Rubin, D. B., 18, 169
Rudemo, M., 3, 9, 55, 69, 144, 271
Ruppert, D., 3, 9, 54, 69, 136, 144, 211, 224, 261, 269
Satten, G., 16, 153, 250
Sayyad, G. M., 167
Schafer, D., 48, 141
Schmid, C. H., 167
Schrader, R. M., 265
Scott, A. J., 185
Searle, S. R., 271
Self, S., 149, 254
Sepanski, J. H., 214
Serfling, R. J., 258
Silverman, B. W., 43
Small, C. G., 261
Smith, A. F. M., 69, 166, 169, 178, 247
Spiegelman, C. H., ix, 37, 65, 88, 141
Spiegelman, D., 3, 32, 40, 46, 51, 63, 64, 141, 251, 252
Stefanski, L. A., ix, 11, 35, 40, 47, 50, 53, 64, 79, 83, 87, 99, 108, 122, 163, 210, 214
Stephens, M. A., 167
Stoker, T., 199
Stram, D., 247
Streibig, J., 3, 9, 55, 69, 144, 271
Stuetzle, W., 199, 224
Tan, W. Y., 208
Tanner, M. A., 153, 166, 169
Taylor, R. L., 216
Thomas, D., 247
Thompson, E., 2, 44, 69, 153
Thompson, S., 199
Tiao, G., 165
Tibshirani, R., 223, 274, 279
Titterington, D. M., 247, 249
Tosteson, T. T., 4, 10, 11, 210, 213, 233, 237, 253
Tripathi, R. C., 245
Troung, Y., 216
Tsai, C. L., 37
Tsiatis, A. A., 210, 213
Tukey, J., 87
Ulm, K., 156
van Ness, J., 37
van der Vaart, A., 138
Wacholder, S., 36, 244
Wand, M. P., 224, 238
Wang, C. Y., 186, 244
Wang, N., 14, 47, 141
Wang, S., 244
Ware, J. H., 37, 237, 253
Wax, Y., 24, 253
Weihs, C., 69
Weinberg, C., 35, 244
Whittemore, A., 50, 63, 69, 141
Wieand, S., 208
Wild, C. J., 185
Willett, W.. 3, 32, 40, 43, 46, 63
Wittes, J., 229
Wolfe, D., 71
Wu, M. L., 2, 43
Zamar, R., 37
Zhao, L. P., 149, 185, 191, 192, 246, 247, 253

Subject index

Applications,
  Bioassay, 3, 9, 55-60,
  Bronchitis, 156-160
  Cervical cancer, 173-175
  Cholesterol, 250
  Chronic graft versus host disease, 241
  Framingham Heart Study, 4, 11, 16, 88-94, 111-112, 125, 154-155, 175-181
  Lung functioning, 4
  Nutrition, 1-3, 13, 16, 24, 32, 42-44, 69-71, 209-210, 222, 250
  Urinary sodium chloride, 5, 14
Attenuation, 21-35
  see also Bias caused by measurement error, Correction for attenuation, Regression calibration
Bayesian methods, 165-181
  Credible sets, 165
  Functional methods, 166-167
  Gibbs sampling, 168-170, 175-181
  Importance sampling, 171-173, 175-181
  Weighted bootstrap, 167-170
Berkson model, see Measurement error models
Bias caused by measurement error, 18-36, 230
Bias versus variance in correcting for measurement error, 32-34
Binary regression: see Logistic regression, Probit regression
Bootstrap, 41, 44, 50, 86, 124, 130, 142, 186, 273-279
Box-Cox transformation models: see Transform-Both-Sides model
Calibration, 233
Case-control studies, 16, 243-245
Choice-based sampling, see Case-control studies
Complete data analysis, 182-184, 194-196, 237, 242
Conditional score methods
  Computation, 129-130
  Estimation, 123-130, 244, 248
  Inference, 130-131
  Theory, 139
Contingency tables, 253
Corrected score methods
  Comparison with Conditional score methods, 137-139
  Estimation, 131-133
  Series expansions, 134-137
Correction for attenuation, 27-34, 42-47
  see also Regression calibration
Correlated errors, 6, 14, 24-25, 69-72, 148
  see also Measurement error

models
Deconvolution,
nonparametric, 215-222
parametric, 219
Density estimation, see Deconvolution
Design of validation studies, 182-184, 251-253
Differential measurement error, see Measurement error models
Discrete variables, see Misclassification
Discriminant analysis, 245
Distribution function estimation, 219-220
Efficient score test, 210-214
Equation error, 30-31, 229
see also Orthogonal regression
Error calibration model, 8, 12, 28
Error models, see Measurement error models, Response error
Estimating equations, see Unbiased estimating equations
Exponential family, 273
Free lunch, see identifiability, orthogonal regression
Functional modeling
maximum likelihood estimates, 122-123, 247
models, 6, 29, 122
see also Conditional score methods, Corrected score methods, Instrumental variables, Nonparametric mixture methods, Regression Calibration, SIMEX, Unknown link functions
Gamma model, 124, 126, 131, 270
Generalized linear models, 50, 72-77, 102, 108, 116, 126, 167, 206, 257, 269-273
see also Variance function models
Heteroscedastic regression: see Variance function models
Hypothesis testing, 18, 43, 206-220, 259
Identifiability, 143, 151, 231-236
Ignorable measurement error, 229
Importance sampling, 171-173
see also Bayesian methods
Instrumental variables
Estimation, 12-13, 46, 54, 70, 107-115
Inference, 116-120
Large sample theory, 50, 72-77, 97-106, 116-121, 193-196, 203-205, 262-271
Latent variables: see Maximum likelihood
Likelihood, 18, 57, 60, 123, 141-160, 235, 257-260
see also Maximum likelihood
Linear regression, 17-32, 41, 51-52, 62-67, 80, 95, 102, 122-126, 130, 143-145, 160-162, 245, 264, 270
see also Method of moments
Logistic model: polytomous, 236
Logistic-normal integrals, 64
Logistic regression, 5, 41, 51, 63-66, 102, 125, 128, 135, 142-143, 153, 163, 184, 226-228, 236-238, 242-245, 270, 277
Loglinear mean models 40, 52, 66-68, 95, 123-124, 131
Maximum likelihood,
Comparison with functional methods, 141-142, 144-146
Computation, 142, 153, 161-162
Contingency tables, 253
Identifiability, 143, 151
Inference, 142, 259
Likelihood functions, 57, 60, 71, 123, 141-160, 182-188, 215, 244-248, 257-259

Response error, 235-238
Use in measurement error models, 141-143
When X is partially observed, 152
When X is unobserved, 146-151, 156-160
Measurement error models,
Additive, 8, 22, 27, 47, 69-71, 80-82, 87, 122-127, 133, 138, 141, 147, 154-159, 230, 233
Berkson, 3, 9, 52-56, 67, 143, 150-151, 255
Differential, 16-17, 245-247
General, 2, 87-89, 147
Multiplicative, 48-49, 80, 87, 141, 147, 150, 230, 233
Nondifferential, 16-17, 35
Random coefficient models, 151
Variance estimation, 47-49, 69-71, 84, 133-134
see also Correlated errors, Replicates in additive error models, Response error
Method of moments, 27-30, 40, 124, 245
see also Correction for attenuation, Regression calibration
Misclassification, 44, 142, 253
Missing data, 18, 141, 144-145, 152, 182-185
Mixture methods, 247-251
see also Semiparametric methods
Model robustness, see Functional modeling, Unknown link functions
Naive test, see Hypothesis Testing
Nonignorable measurement error, 299
Nonparametric mixture modeling, 247-251
Nonparametric regression, 215, 223-228
Offset, 238
Optimal estimators, 138
Orthogonal regression, 28-31
Poisson model, 123, 129, 137, 141, 270
Power transformation models: see Transform-Both-Sides model
Prediction in the presence of measurement error, 18
Probit regression, 141, 145, 163
Projection pursuit, see Unknown link functions
Quadratic regression, 52, 68, 78, 96
Quasilikelihood, see Variance function models
Regression calibration,
Accuracy, 61-69
Best linear approximations, 47
Estimation, 41-50, 87, 158, 203
Expanded models, 41, 51-62, 67-68
Inferences, 50-51, 72-77
James-Stein methods, 50
Models, 8-9, 40-44, 53, 61-68, 79, 89, 122, 192, 207, 215, 225, 245, 248, 254-256
Nonparametric regression, 225-228
Regression models: see linear, logistic, loglinear, Poisson, quadratic, segmented models
Regression: nonparametric, 223-228, 240
Regression to the mean, see Attenuation
Reliability ratio, 22, 27
Replicates in additive error models, 5, 17, 47, 70, 87-88
Response error, 229-242
Additive, 230-235

Biased, 233, 236
Likelihoods, 235-238
Semiparametrics 238-240
Robust covariance estimator: see Sandwich method
Sandwich method for standard errors, 41, 72-77, 86, 89, 101, 124, 142, 232-233, 259-265
Segmented regression, 97, 145, 156-160
Semiparametric methods,
As functional methods, 182
Efficient methods, see moments methods
Mean score method, 187-188
Modified pseudolikelihood, 241-242
Moments methods, 188-198
Pseudolikelihood, 185-187, 193-194, 238
Regression calibration, 192-194
Using complete data only, 183-184, 188, 193-196
see also Functional modeling, SIMEX, Conditional scores, Corrected scores, Unknown link functions
SIMEX,
Basic idea, 79-80
Biases, 95-97
Extrapolation step, 83-85
Method, 82, 244
Nonadditive errors, 87-88
Nonparametric regression, 224-228
Relation to the jackknife, 86-87, 98-99
Simulation step, 82-83
Standard errors, 86, 89-105, 122
Sliced inverse regression, 201-202
see also Unknown link functions
Small error approximations, 6, 18, 69, 123, 144-145, 182
Structural modeling, 6, 18, 46, 123, 144-145, 182
see also Bayesian methods, Maximum likelihood
Surrogate, 16
response, 235
See also Measurement error models: nondifferential
Survival analysis, 40, 254-256
Testing, see Hypothesis testing
Transform-Both-Sides regression model, 56, 87, 154-156
Transportability, 10, 148
Transportability: dangers of, 11
Two-stage studies, 246
See also Design, Types of data
Types of data,
External data, 12, 150
Instrumental data, 107-108
Internal data, 12, 46, 182
Predictors without error, 1, 147, 251
Replication data, 12-13, 50-54, 143, 182, 251
True but fallible predictors, 14
Validation data, 12, 46, 54, 143, 182, 236, 238, 240-241, 251
Unbiased estimating equations, 101-102, 124, 261-268
Stacking method, 268-269
Unknown link functions, 199-202
see also Functional modeling
Validation studies, see Design of validation studies, Maximum likelihood, Types of data, Semiparametric methods
Variance function models, 50-56, 62, 67, 72-77, 102, 108, 116, 128, 136, 141, 160, 167, 185-188, 206-207, 226, 230-234, 269-276

Power-of-mean, 231-232
see also Generalized linear models