Lecture Notes in Statistics 211
Proceedings

Brajendra C. Sutradhar *Editor*

# ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers

Springer

# Lecture Notes in Statistics

## Proceedings

## Volume 211



For further volumes:
http://www.springer.com/series/8440

Brajendra C. Sutradhar
Editor

# ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers

Springer

*Editor*
Brajendra C. Sutradhar
Department of Mathematics & Statistics
Memorial University of Newfoundland
St. John's, NL, Canada

ISSN 0930-0325
ISBN 978-1-4614-6870-7 ISBN 978-1-4614-6871-4 (eBook)
DOI 10.1007/978-1-4614-6871-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013941150



Printed on acid-free paper

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

*To*
*Bhagawan Sri Sathya Sai Baba*

# Preface

This special proceedings volume contains nine selected papers that were presented in the *International Symposium in Statistics (ISS) on Longitudinal Data Analysis Subject to Outliers, Measurement Errors, and/or Missing Values*, held at Memorial University, Canada from July 16–18, 2012. Three years ago the ISS-2009 was organized focussing *on inferences in generalized linear longitudinal mixed models (GLLMMs)*, and a special issue of the Canadian Journal of Statistics (2010, Vol. 38, June issue, Wiley) was published with seven selected papers from this symposium. These seven papers from ISS-2009 dealt with progress and challenges in the area of discrete longitudinal data analysis. As a reflection of the theme of the ISS-2012, the papers in the present volume deal with inferences for longitudinal data with additional practical issues such as measurement errors, missing values, and/or outliers. The inferences for this type of complex longitudinal data become more challenging than the inferences for standard longitudinal data following generalized linear longitudinal models (GLLMs). The present volume with nine papers makes a significant contribution toward such challenging inferences. To make it as precise and clear as possible, the papers are grouped into three parts along the theme of the symposium. Part I contains four papers in longitudinal data analysis subject to measurement errors, similarly Part II also contains four papers but they are in longitudinal data analysis subject to missing values, and Part III has one paper dealing with inferences for longitudinal data subject to outliers.

In a longitudinal setup, repeated responses along with a set of multidimensional time-dependent covariates are collected over a small period of time. There are situations where it is realized that the observed multidimensional covariates at a given time point differ from the corresponding true covariates by some measurement errors, but it is of interest to find the regression effects of the true covariates on the repeated responses. The first paper in Part I, by B. C. Sutradhar, begins with a discussion on this measurement error problem for scalar responses. This setup with scalar responses is referred to as the independent setup. In the first part of the paper, the author considers the independent setup and provides an overview of the existing vast literature on inferences for various bias correction approaches. In the longitudinal setup, repeated responses are, however, likely to follow a true

time-dependent covariates-based correlation structure. Because the true covariates are unobserved, this involvement of the true covariates in the correlation structure makes the bias correction to the observed covariates-based regression estimator very difficult, specially for longitudinal discrete such as count and binary data. In the second part of the paper, the author gives a brief discussion on the existing bias corrected generalized method of moments (BCGMM) and generalized quasi-likelihood (BCGQL) approaches in the linear longitudinal models (LLMs) setup. The author then discusses the progress and challenges in obtaining bias corrected inferences in the generalized LLMs (GLLMs) setup, mainly for repeated count and binary data. An overview is given on how to develop a BCGQL inference approach for longitudinal count data when corresponding covariates are subject to measurement errors. The bias correction inference for longitudinal binary data appears to be extremely difficult. The author has discussed the progress and emphasized on further investigations to resolve this challenging problem. In the second paper of Part I, Laine Thomas, Leonard A. Stefanski, and Marie Davidian consider a binary logistic regression model in independent setup where, on top of baseline covariates, the binary responses are also influenced by the mean and variance parameters of a scalar covariate which is repeatedly measured over a period of time. The authors have used a moment approach for the prediction of the variance components involved in the linear regression measurement error models for the repeated values of the covariate, and these predicted variances are used in turn in a conditional scores-based bias correction approach for the estimation of the main regression parameters of the binary outcome model. As opposed to the longitudinal setup, the third paper in Part I, by John P. Buonaccorsi, deals with measurement errors in time series. The author assumes that the true time series follows a dynamic model of interest but the series itself is unobserved. Instead, a series with measurement error is observed. Thus, the observed response at a given time is not necessarily the true response, rather, it follows a suitable distribution with its conditional mean as a function of the true response. The author has discussed various bias correction approaches including moments and likelihood methods for the estimation of the parameters of the dynamic model for the true but unobserved time series. In the fourth paper of Part I, Erik Meijer, Laura Spierdijk, and Tom Wansbeek consider a linear dynamic mixed model in panel data setup, but the true responses satisfying the underlying model are not observed. The observed responses, which are subject to measurement errors, are used to obtain consistent and efficient estimators for the parameters of the model for the true responses. Thus, this paper considers the measurement error in responses, whereas the first and the second paper in this part considered the measurement errors in covariates.

In many biomedical, clinical, and socioeconomic studies, a response and its corresponding multidimensional covariates are collected repeatedly over time from a large number of independent individuals. In this setup, it is assumed that the repeated responses from an individual marginally follow a linear or nonlinear regression model and jointly they follow a longitudinal correlation structure. It is of interest to estimate the regression effects of the time-dependent covariates on the repeated responses. For varieties of reasons it may, however, happen that a portion

of responses are missing from some individuals. In practice, in general, there are three types of missing mechanism such as missing completely at random (MCAR), missing at random (MAR), or missing non-ignorably. Further the nonresponse may occur in a monotonic fashion or they may be intermittent. The analysis of this type of incomplete longitudinal data, specially the inferences for the regression effects by using incomplete data, is complicated. This is because, to develop proper inferences, one requires to accommodate both missing mechanism and the correlation structure for the available longitudinal responses. The first paper in Part II, by B. C. Sutradhar, provides an overview on incomplete data analysis both in independent and aforementioned longitudinal setup. In the independent setup, attempts are made to collect multidimensional responses from a large number of independent individuals, but it may happen that a small portion of individuals do not provide complete multidimensional responses leading to incomplete data. The author first discusses some of the widely used existing estimation methods including the imputation technique for such incomplete data analysis in the independent setup. The author then discusses the progress made and the difficulties encountered, by the existing inference techniques, which do not appear to accommodate the missing mechanism and longitudinal correlations properly. Details are given for some remedies to overcome this anomaly in order to develop proper estimating equations for the regression effects. An unconditional as well as a conditional approach is discussed to develop estimating equations for consistent regression estimates. In the second paper of Part II, Taslim Mallick, Patrick Farrell, and B. C. Sutradhar proposed a GQL approach along the lines of the first paper by B. C. Sutradhar that provides consistent regression estimates. When the responses are MAR, the authors have further demonstrated that the existing generalized estimating equations (GEE) approach encounters serious convergence problems specially when missing proportion is large. This breakdown shows the inconsistency of the GEE-based approaches, whereas the proposed GQL approach does not encounter such convergence problems unless the missing proportion is unreasonably high, and it produces almost unbiased regression estimates with smaller standard errors. In the third paper of Part II, Paul S. Albert, Rajeshwari Sundaram, and Alexander C. McLain discuss a random effects approach to analyze longitudinal data subject to missing. The authors introduce suitable random effects and assume that they cause the correlations among repeated data and also determine the missing mechanism. More specifically, they assume that conditional on the same random effects, the responses do not dependent on the missing data status, also the repeated responses are independent. This yields a simple probability model for the observed random variables, that is, responses and missing indicators, which in turn leads to fairly simple likelihood and/or conditional likelihood inferences for the regression parameters. In the last paper of Part II, Michael A. McIsaac and Richard J. Cook consider a two-phase sampling-based dropout model, where in the first phase a vector of clustered or repeated responses along with a vector of multidimensional auxiliary covariates are collected from a large number of independent individuals. However, in the second phase, a vector of multidimensional expensive covariates are collected only from a portion of individuals. It is of interest to examine the effects

of both auxiliary and expensive covariates on the clustered responses. The authors discuss the incomplete data-based likelihood, mean score, and weighted pseudo-likelihood methods for the estimation of such regression effects.

Part III of the volume contains one paper by B. C. Sutradhar, on the inferences for longitudinal data subject to outliers. It is known in the independent setup that a few outlying responses mainly caused by the associated contaminated covariates may adversely influence the valid inferences for the regression effects. The author first gives an overview of the existing robust approaches in the independent setup for the estimation of the regression effects in linear, count, and binary data models. These approaches include a recently developed fully standardized Mallow's type quasi-likelihood (FSMQL) method that provides almost unbiased regression estimates. The author then extends the overview to the longitudinal setup. The robust inferences for longitudinal binary and count data are, however, not adequately discussed in the literature. The author discusses a robust GQL approach for unbiased regression estimation for count and binary longitudinal data models.

St. John's, NL, Canada Brajendra C. Sutradhar


**ISS-2012 Delegates**

**ISS-2012 Welcome Address by Brajendra C. Sutradhar (Organizer)**

With the name of Lord, we welcome all of you, to Memorial University, the host for the International Symposium in Statistics, 2012 (ISS-2012) on Longitudinal Data Analysis Subject to Outliers, Measurement Errors, and/or Missing Values. It gives us a pleasure to note that we have been able to keep up the spirit of the first symposium (ISS-2009) that took place here in Memorial University, in organizing the present symposium covering extended and more challenging research areas in the longitudinal setup, mainly for discrete data such as count and binary data.We thank all of you for your interest and response to this symposium that has attempted to attract the researchers deeply involved in the inferences for longitudinal data those encounter practical difficulties due to measurement errors, nonresponse, and/or outliers. We hope that you will find the symposium stimulating and will derive spirits for doing more and more quality research in these challenging areas as a service to the society and mankind at large. We also hope that the symposium generates and enhances the spirit of collaborative research among the participants which also might reconfirm our sense of achievement in a greater horizon of life, as the proverb goes: "Life is a march from I to We to He (Sri Sathya Sai Baba, India)". It is indeed a pleasure to note that we have delegates in this specialized symposium from many countries such as Australia, Bangladesh, Brazil, Canada, Mauritius, the Netherlands, Saudi Arabia, Spain, and the USA covering a large part of the globe. We extend our hearty welcome to all of you.

We also welcome you to St. John's, the oldest city of North America, known as the City of Legends, where you can view icebergs, watch whales, and experience Newfoundland and Labrador's unique culture. It is a progressive city and is the site of many world class facilities. A mosaic of fishing villages, cultural festivals, and wildlife tours bring variety to the city. Also, the Cape Spear, the most easterly point of North America, is not far from the city, where one can experience the unique beauty of sunrise. We hope that you have planned for an extended stay in St. John's following the symposium to enjoy these and other endless options!

St. John's, NL, Canada Brajendra C. Sutradhar

# Acknowledgments

This proceedings volume (lecture note) is a collection of selected papers that were presented in ISS-12 (International Symposium in Statistics, 2012) held at Memorial University from July 16–18, 2012. Organizing this symposium would not, however, have been possible without the generous contributions from Memorial University and the Atlantic Association for Research in Mathematical Sciences. I wish to express my special thanks to these two institutes for their support.

All papers in this volume were refereed. Prior to the symposium, the papers were sent to the referees who were also supposed to be present during the presentation. The authors were also benefitted from the warm discussion by the audience of the symposium and prepared the revision of the paper by addressing all suggestions and comments from the referees and the audience. My sincere thanks go to the delegates and referees to make the symposium and this volume a grand success. Some of the contributed papers were also considered for their publication in this volume. A special thanks go to Dr. Alwell Oyet for his warm service in processing all contributed papers for the symposium. It has been a pleasure to work with Marc Strauss, Hannah Bracken, Mary Helena, and Lesley Poliner of Springer-Verlag in preparing this volume.

# Contents

## Part III Longitudinal Data Analysis Subject to Outliers

# Contributors

**Paul S. Albert** *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD, USA

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

**Richard J. Cook** Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

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

**Patrick J. Farrell** School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

**Taslim S. Mallick** School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

**Michael A. McIsaac** Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

**Alexander C. McLain** University of South Carolina, Columbia, SC, USA

**Erik Meijer** RAND Corporation, Santa Monica, CA, USA

**Laura Spierdijk** University of Groningen, Groningen, The Netherlands

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

**Rajeshwari Sundaram** *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD, USA

**Brajendra C. Sutradhar** Memorial University, St. John's, NL, Canada

**Laine Thomas** Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

**Tom Wansbeek** University of Groningen, Groningen, The Netherlands

# Part I
# Longitudinal Data Analysis Subject to Measurement Error

# Measurement Error Analysis from Independent to Longitudinal Setup

**Brajendra C. Sutradhar**

**Abstract** In a generalized linear models (GLMs) setup, when scalar responses along with multidimensional covariates are collected from a selected sample of independent individuals, there are situations where it is realized that the observed covariates differ from the corresponding true covariates by some measurement error, but it is of interest to find the regression effects of the true covariates on the scalar responses. Further it may happen that the true covariates may be fixed but unknown or they may be random. It is understandable that when observed covariates are used for either likelihood or quasi-likelihood-based inferences, the naive regression estimates would be biased and hence inconsistent for the true regression parameters. Over the last three decades there have been a significant number of studies dealing with this bias correction problem for the regression estimation due to the presence of measurement error. In general these bias correction inferences are relatively easier for the linear and count response models, whereas the inferences are complex for the logistic binary models. In the first part of the paper, we review some of the widely used bias correction inferences in the GLMs setup and highlight their advantages and drawbacks where appropriate. As opposed to the independent setup, the bias correction inferences for clustered (longitudinal) data are, however, not adequately addressed in the literature. To be a bit more specific, some attention has been given to deal with bias correction in linear longitudinal setup (also called panel data setup) only. Bias corrected generalized method of moments (BCGMM) and bias corrected generalized quasi-likelihood (BCGQL) approaches are introduced and discussed. In the second part of this paper, we review these BCGMM and BCGQL approaches along with their advantages and drawbacks. The bias correction inferences for count and binary data are, however, more complex, because of the fact that apart from the mean functions, the variance and covariance functions of the clustered responses also involve time-dependent covariates. This makes the bias correction

B.C. Sutradhar (✉)
Memorial University, St. John's, NF, Canada, A1C5S7
e-mail: bsutradh@mun.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_1, 

difficult. However, following some recent works, in the second part of the paper, we also discuss a BCGQL approach for longitudinal models for count data subject to measurement error in covariates. Developing a similar bias correction approach for longitudinal binary data appears to be difficult and it requires further in-depth investigations.

## 1 Introduction

When responses along with covariates are collected from a group of independent individuals in a generalized linear model (GLM) setup, in some practical situations the observed covariates may be subject to measurement errors differing from the true covariates values. These imprecise observed covariates, when used directly, the standard statistical methods such as likelihood and quasi-likelihood methods yield biased and hence inconsistent regression estimates. Bias corrected estimation for the regression effects involved in generalized linear measurement-error models (GLMEMs) with normal measurement errors in covariates has been studied extensively in the literature. See, for example, Fuller (1987), Carroll et al. (2006), and Buonaccorsi (2010), and the references therein. In general, this type of bias correction is studied under two scenarios. First, if for a sample of observed responses and covariates, namely, $\{(y_i, x_i)(i = 1, \ldots, K)\}$, the true covariates $\{z_i\}$ are independent and identically distributed random vectors from some unknown distribution, *a structural error-in-variables model* is obtained; second if $\{z_i\}$ are unknown constants, *a functional error-in-variables model* is obtained (Kendall and Stuart 1979, Chap. 29; Stefanski and Carroll 1987; also known as Berkson error model). Note that the second scenario is more challenging technically because unknown fixed $\{z_i\}$ makes a large set of parameters and direct estimation or prediction of each of them may be impossible, specially when $K$ is large.

For discussions on structural models, especially for inferences, in addition to the aforementioned references, namely, Fuller (1987), Carroll et al. (2006), and Buonaccorsi (2010), one may consult, for example, an instrumental variable technique to obtain bias corrected estimates for regression parameters in GLMs, studied by Buzas and Stefanski (1996a) (see also Stefanski and Buzas 1995; Buzas and Stefanski 1996b; Amemiya 1990; Carroll and Stefanski 1990), among others. In this paper, we, however, mainly deal with functional models, and among many existing studies based on such functional models, we, for example, refer to Stefanski and Carroll (1985), Stefanski (1985), Armstrong (1985), Stefanski and Carroll (1987), Carroll and Stefanski (1990), Nakamura (1990), Carroll and Wand (1991), Liang and Liu (1991), Stefanski (2000), and Carroll et al. (2006). Some of these studies address measurement error problems in various complicated situations such as when the data also contain outliers, and regression function is partly specified. But, they are confined to the independent setup.

As opposed to the independent setup, not much attention is given to the measurement error models for longitudinal count and binary data. Sutradhar and Rao (1996)

have developed a bias correction approach as an extension of Stefanski (1985) for the longitudinal binary data with covariates subject to measurement errors. To be specific, these authors have used a small measurement error variance asymptotic approach to achieve the bias correction, which works well if the measurement error variance is small or moderately large. Wang et al. (1996) considered a measurement error model in a generalized linear regression setup where covariates are replicated and the measurement errors for replicated covariates are assumed to be correlated with a stationary correlation structure such as Gaussian auto-regressive of order 1 (AR(1)) structure. As far as the responses are concerned, they were assumed to be independent, collected at a cross-sectional level from a large number of independent individuals. Thus this study does not address the measurement error problems in the longitudinal setup where responses are collected repeatedly from a large number of independent individuals. With regard to the correlations for the repeated responses, there, however, exit some studies for continuous responses subject to measurement error, in time series setup. For example, we refer to the study by Staudenmayer and Buonaccorsi (2005), where time series responses are assumed to follow the Gaussian auto-regressive order 1 (AR(1)) correlation process subject to measurement error. But, these studies are not applicable to the longitudinal setup, especially for discrete longitudinal data such as for repeated count and binary data with covariates subject to measurement error.

In this paper, first we review some of the widely used inference approaches in the GLMs setup for independent responses, for the estimation of the regression effects on such responses when associated covariates are subject to mainly functional measurement error. The structural measurement error models are discussed in Sect. 2.1.2. The advantages and drawbacks of each approach are highlighted.

As pointed out above, the measurement error analysis is not so developed in the longitudinal setup specially for binary and count data. For linear longitudinal measurement error models, there exist some studies with concentration on econometric data analysis. For example, Wansbeek (2001) (see also Wansbeek and Meijer 2000) considered a measurement error model for linear panel data, where on top of the fixed true covariates $z_i$, some of the other covariates are strictly exogenous. To be more specific, Wansbeek (2001) developed necessary moment conditions to form bias corrected method of moments (BCMM) estimating equations in order to obtain consistent generalized method of moments (GMM) estimates for the regression parameters involved including the effect of the exogenous covariates. More recently, Xiao et al. (2007) studied the efficiency properties of the BCGMM (bias corrected generalized method of moments) approach considered by Wansbeek (2001). Note that the derivation of the efficient BCGMM estimators by Xiao et al. (2007) may be considered as the generalization of the GMM approach of Hansen (1982) to the measurement error models. In studying the efficiency of the BCGMM approach, Xiao et al. (2007), however, assumed that the model errors $\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}$ are independent to each other. Also they assume that the measurement errors collected over times are serially correlated. Recently Fan et al. (2012) have developed a bias corrected generalized quasi-likelihood

(BCGQL) approach that produces more efficient estimates than the BCGMM approach.

As far as the measurement error models for longitudinal count and binary data are concerned, in developing a bias correction method, one has to accommodate both longitudinal correlations of the repeated responses and the measurement errors in covariates. Recently, Sutradhar et al. (2012) have developed a BCGQL approach so that the BCGQL estimating function is unbiased for the GQL estimating function involving the true covariates. They then solved the BCGQL estimating equation to obtain bias corrected regression estimates. These estimates are also efficient. We describe this BCGQL approach in brief from Sutradhar et al. (2012). As opposed to the small measurement error variance-based estimating equation approach (Sutradhar and Rao 1996), developing a similar BCGQL estimating equation for regression effects involved in longitudinal binary data models does not appear to be easy. This would require further in-depth investigations.

## 2 Measurement Error Analysis in Independent Setup

For $i = 1,\ldots,K$, let $Y_i$ denote the binary or count response variable for the $i$th individual and $x_i = (x_{i1},\ldots,x_{ip})'$ be the associated $p$-dimensional covariate vector subject to normal measurement errors. Let $z_i = (z_{i1},\ldots,z_{ip})'$ be the unobserved true covariate vector which may be fixed constant or random and $\beta$ be the regression effect of $z_i$ on $y_i$. For discrete responses, such as for count and binary data, by using exponential family density for $y_i$ given $z_i$, the GLMEM is written as

$$f(y_i;z_i) = \exp[\{y_i\theta_i(z_i) - a(\theta_i(z_i))\} + b(y_i)] \tag{1}$$

$$x_i = z_i + \delta v_i \text{ with } v_i \sim N_p(0, \Lambda = \text{diag}[\sigma_1^2,\ldots,\sigma_p^2]), \tag{2}$$

where $\theta_i(z_i) = h(z_i'\beta)$, with $a(\cdot), b(\cdot)$, and $h(\cdot)$ being known functional form, yielding the first and second derivatives, $a'(\theta_i(z_i))$ and $a''(\theta_i(z_i))$, as the mean and variance of $y_i$, respectively; $v_i$ is a random measurement error vector and $\delta^2$ is a scalar parameter. Note that if for a sample $(y_i,x_i)(i = 1,\ldots,K)$ the covariates $\{z_i\}$ are unknown constants, a functional error-in-variables model (also known as Berkson error model) is obtained; if $\{z_i\}$ are independent and identically distributed random vectors from some unknown distribution, a structural error-in-variables model is obtained (Kendall and Stuart 1979, Chap. 29; Stefanski and Carroll 1987).

Under the functional model, Nakamura (1990) has proposed a corrected score (CS) estimation approach, where for given $z_i$, the log likelihood function for $\beta$ is written by (1) as

$$\ell(\beta;y,z) = \sum_{i=1}^{K}[\{y_i\theta_i(z_i) - a(\theta_i(z_i))\} + b(y_i)],$$

and observed covariates $x_i$-based corrected log likelihood function $\ell^*(\beta;y,x)$ is written such that $E_x[\ell^*(\beta;y,x)] = \ell(\beta;y,z)$. The corrected score estimate of $\beta$, say $\hat{\beta}_{\mathrm{CS}}$, is then obtained by solving the corrected score equation

$$U^*(\beta;y,x) = \frac{\partial \ell^*(\beta;y,x)}{\partial \beta} = 0. \tag{3}$$

This corrected score approach provides closed form estimating equation for $\beta$ for the Poisson regression model, but, the binary logistic regression model does not yield a corrected score function which is a limitation to this approach.

Stefanski and Carroll (1987) proposed a method based on conditional scores (CNS). In this approach, unbiased score equations are obtained by conditioning on certain parameter-dependent sufficient statistics for the true covariates $z$, and the authors have developed the approach in both functional and structural setups. The conditional score equations have a closed form for GLMs such as for normal, Poisson, and binary logistic models. Obtaining a closed form unbiased equation for logistic regression parameter by this conditional approach is an advantage over the direct corrected score approach (Nakamura 1990) which does not yield corrected score function. To elaborate a little more on the conditional score approach, consider, for example, the functional version of the logistic measurement error model with scalar predictor $z_i$ so that the measurement error $v_i$ in (2) follows $N_1(0,\sigma_1^2)$ (Stefanski 2000, Sect. 4.1). For convenience, consider $\delta = 1$ in (2). In this case, the density of $(y_i;x_i)$ is given by

$$f(y_i,x_i;\beta,z_i) = [\frac{\exp(z_i'\beta)}{1+\exp(z_i'\beta)}]^{y_i}[\frac{1}{1+\exp(z_i'\beta)}]^{1-y_i}\frac{1}{\sigma_1}\phi(\frac{x_i-z_i}{\sigma_1}),$$

where $\phi(.)$ is the standard normal density function. The estimation of $\beta$ also requires the estimation of the nuisance parameters $z_i$ or some functions of $z_i$'s for $i = 1,\ldots,K$. However, Stefanski and Carroll (1987) have demonstrated that the parameter-dependent statistic $\lambda_i = x_i + y_i\sigma_1^2\beta$ is sufficient for unknown $z_i$ in the sense that the conditional distribution of $(y_i,x_i)$ given $\lambda_i$ does not depend on the nuisance parameter $z_i$. This fact was exploited to obtain unbiased estimating equation for $\beta$ using either conditional likelihood method or mean variance function models (based on conditional density of $y_i$ given $\lambda_i$) and quasi-likelihood methods. For the scalar regression parameter $\beta$, the unbiased estimating equation has the form (Sutradhar and Rao 1996, Eq. (2.10))

$$\sum_{i=1}^{K}(\lambda_i - \sigma_1^2\beta)(y_i - \tilde{p}_i) = 0, \tag{4}$$

where $\tilde{p}_i = F[\{\lambda_i - (\sigma_1^2/2)\beta\}\beta]$ with $F(t) = 1/[1+\exp(-t)]$. Let $\hat{\beta}_{\mathrm{CNS}}$ denote the solution of (4) for $\beta$.

In structural error-in-variables setup, there exists an instrumental variable technique to obtain bias corrected estimates for regression parameters in GLMs. For this, for example, we refer to Buzas and Stefanski (1996a) (see also Stefanski and Buzas 1995; Buzas and Stefanski 1996a; Amemiya 1990; Fuller 1987). We do not discuss about this technique any further in this paper as our purpose is to deal with functional models as opposed to the structural models.

Note that as in the absence of measurement errors, regression parameters involved in GLMs such as for count and binary models, may be estimated consistently and efficiently by using the first two moments-based quasi-likelihood (QL) approach (Wedderburn 1974), there has been a considerable attention to modify the naive QL (NQL) approach (that directly uses observed covariates ignoring measurement errors) in order to accommodate measurement errors in covariates and obtain bias corrected QL (BCQL) estimates. Some of these BCQL approaches are developed for both structural and functional models, some are developed for the functional models and others are more appropriate for structural models only. Stefanski (1985) proposed a small measurement error variance-based BCQL approach for structural models, Carroll and Stefanski (1990) have used a similar small measurement error variance-based QL approach which is developed to accommodate either of the structural or functional models or both. Liang and Liu (1991) have discussed a BCQL approach for structural model, which was later on generalized by Wang et al. (1996) to accommodate correlated replicates in covariates. Sutradhar and Rao (1996) have used Stefanski's (1985) small measurement error-based BCQL approach for the longitudinal binary data, independent setup being a special case, under functional model only. In the next section, we provide a brief review of some of these existing simpler BCQL approaches which are suitable for functional models.

In Sect. 2.1, we provide an alternative BCQL approach which yields the same corrected regression estimates as the corrected score estimates (Nakamura 1990) for the Poisson model in functional setup. In the binary case, the proposed alternative approach provides a first order approximate BCQL regression estimates.

## 2.1 BCQL Estimation

Note that if $z_i$ were known, then one would have obtained a consistent estimator of $\beta$ by solving the so-called quasi-likelihood (QL) estimating equation

$$\sum_{i=1}^{K}[\frac{\partial a'(\theta_i(z_i))}{\partial \beta}\frac{(y_i - a'(\theta_i(z_i)))}{a''(\theta_i(z_i))}] = \sum_{i=1}^{K}\psi_i(y_i, z_i, \beta) = 0 \tag{5}$$

(Wedderburn 1974), where for $\theta_i(z_i) = h(z_i'\beta)$, both $a'(\theta_i(z_i))$ and $a''(\theta_i(z_i))$ are functions of $\beta$. For example, for the Poisson and binary data $h(\cdot) = 1$, and $a'(\theta_i(z_i)) = \exp(z_i'\beta)$ for the Poisson data, and $a'(\theta_i(z_i)) = \frac{\exp(z_i'\beta)}{1+\exp(z_i'\beta)}$ for the binary

data. Thus, for both Poisson and binary models, the QL estimating (5) reduces to $\sum_{i=1}^{K} z_i(y_i - a'(z_i'\beta)) = 0$, where $a'(z_i'\beta) = \mu_{iz}$ is the mean of $y_i$. Note that this QL estimating equation is also a likelihood estimating equation. However, because the true covariate $z_i$ is not observed, one cannot use the estimating (5) for the estimation of $\beta$.

### 2.1.1 Small Measurement Error Variance-Based QL (SVQL) Approach

Suppose that by replacing $z_i$ with $x_i$ in (5), one constructs a NQL estimating equation, namely

$$\sum_{i=1}^{K} \psi_i(y_i, x_i, \beta) = \sum_{i=1}^{K} w_i[y_i - a'(h(x_i'\beta))]h'(x_i'\beta)x_i$$
$$= \sum_{i=1}^{K} g_i(x_i'\beta)x_i = 0, \tag{6}$$

which is the naive version of the Eq. (10) in Stefanski (1985, 588), where $w_i x_i = \frac{\partial a'(h(x_i'\beta))/\partial\beta}{a''(h(x_i'\beta))}$. Let $\hat{\beta}$ be the solution of this NQL estimating (6). But, because the NQL estimating function in the left-hand side of (6) is a function of $x_i x_i'$ and because $x_i = z_i + \delta v_i$ with $E[x_i x_i'] = z_i z_i' + \delta^2\Lambda$ in the functional setup, $\hat{\beta}$ obtained from (6) cannot converge to $\beta$, it rather converges to a different parameter say $\beta(\delta\Lambda)$. Thus, the naive estimator $\hat{\beta}$ is biased and hence inconsistent for $\beta$. As a remedy, assuming that $\delta$ is small, by expanding the expected function

$$E_x \sum_{i=1}^{K} \psi_i(y_i, x_i, \beta) = E_x \sum_{i=1}^{K} g_i(x_i'\beta)x_i = \sum_{i=1}^{K} \psi_i^*(y_i, z_i, \beta(\delta\Lambda)), \text{ (say)}, \tag{7}$$

about $\delta = 0$, and then equating the expanded function to zero followed by replacing $z_i$ with $x_i$ and $\beta$ with $\hat{\beta}$, Stefanski (1985) obtained a SVQL estimator of $\beta$ as a function of $\delta$ as

$$\hat{\beta}_{\text{SVQL}}(\delta) = \hat{\beta} + \frac{1}{2}\delta^2 \left[\sum_{i=1}^{K} g_i'(x_i'\hat{\beta})x_i x_i'\right]^{-1}$$
$$\times \left[\sum_{i=1}^{K} g_i''(x_i'\hat{\beta})\hat{\beta}'\Lambda\hat{\beta}x_i + 2g_i'(x_i'\hat{\beta})\Lambda\hat{\beta}\right], \tag{8}$$

where $g_i'(\eta_i) = \frac{\partial g_i(\eta_i)}{\partial \eta_i}$, and similarly $g_i''(\eta_i) = \frac{\partial^2 g_i(\eta_i)}{\partial \eta_i^2}$. Note that because in the present independent setup, the mean and variance functions-based QL estimating (5) is the same as the likelihood estimating equation based on GLM (1), Stefanski's

(1985) small variance-based bias correction to naive likelihood estimates is quite flexible. See also Whittemore and Keller (1988) for a similar QL-based modification to the NQL or likelihood estimates. Armstrong (1985) (see also Schafer 1987) also has used QL approach but solved for bias corrected estimates numerically as opposed to obtaining SVQL estimates. Based on small $\delta^2$ approach, Carroll and Stefanski (1990) have developed an approximate SVQL approach in a general framework which can accommodate either structural or functional model or both. In this paper, we, however, concentrate on the functional model only.

Note that as in the count data case $g_i(x_i'\beta) = y_i - \mu_{ix} = y_i - \exp(x_i'\beta)$, the SVQL estimator of $\beta$ by (8) has the formula

$$\text{Poisson case:}\quad \hat{\beta}_{\text{SVQL}}(\delta) = \hat{\beta} + \frac{1}{2}\delta^2 \left[ -\sum_{i=1}^{K} \hat{\mu}_{ix} x_i x_i' \right]^{-1} \times \left[ \sum_{i=1}^{K} (-1)\hat{\mu}_{ix} \hat{\beta}' \Lambda \hat{\beta} x_i - 2\hat{\mu}_{ix} \Lambda \hat{\beta} \right], \tag{9}$$

where $\hat{\mu}_{ix} = \exp(x_i'\hat{\beta})$. Similarly, for the binary data case with $\hat{\mu}_{ix} = \hat{p}_{ix} = \exp(x_i'\hat{\beta})/[1+\exp(x_i'\hat{\beta})]$, the SVQL estimator of $\beta$ has the formula

$$\text{Binary case: } \hat{\beta}_{\text{SVQL}}(\delta) = \hat{\beta} + \frac{1}{2}\delta^2 \left[ -\sum_{i=1}^{K} \hat{p}_{ix} x_i x_i' \right]^{-1} \times \left[ \sum_{i=1}^{K} \hat{p}_{ix}\hat{q}_{ix}\{1-\hat{q}_{ix}\} \hat{\beta}' \Lambda \hat{\beta} x_i - 2\hat{p}_{ix}\hat{q}_{ix} \Lambda \hat{\beta} \right], \tag{10}$$

(see also Sutradhar and Rao 1996, Eq. (2.2), p. 181), where $\hat{q}_{ix} = 1 - \hat{p}_{ix}$.

### 2.1.2 Conditional QL (CNQL) Estimation

In structural setup, there exists a QL approach, developed conditional on $x_i$. Let the true covariate vector $z_i$ be a stochastic variable, distributed as

$$z_i \sim N_p(m, V).$$

Next because $x_i = z_i + \delta v_i$ by (2), it then follows that conditional on $z_i$, $x_i$ has the conditional normal distribution

$$x_i | z_i \sim N_p(z_i, \delta^2 \Lambda).$$

Unconditionally $x_i$ has the normal distribution given by

$$\begin{aligned} x_i &\sim N_p[E(z_i),\ E(\delta^2\Lambda) + \mathrm{var}(z_i)] \\ &\equiv N_p[m, \delta^2\Lambda + V]. \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathrm{cov}(x_i, z_i) &= E_z[\mathrm{cov}((x_i, z_i)|z_i)] + \mathrm{cov}_z[E(x_i|z_i), E(z_i|z_i)] \\ &= \mathrm{cov}_z[z_i, z_i] = V. \end{aligned}$$

It then follows that $z_i$ and $x_i$ have the $2p$-dimensional joint normal distribution given as

$$\begin{pmatrix} z_i \\ x_i \end{pmatrix} \sim N_{2p}\left[\begin{pmatrix} m \\ m \end{pmatrix}, \begin{pmatrix} V & V \\ V & \delta^2\Lambda + V \end{pmatrix}\right], \tag{11}$$

yielding the conditional distribution of $z_i$ given $x_i$ as

$$\begin{aligned} z_i|x_i &\sim N_p[m + V(\delta^2\Lambda + V)^{-1}(x_i - m),\ V - V(\delta^2\Lambda + V)^{-1}V] \\ &\equiv N_p[\{I_p - V(\delta^2\Lambda + V)^{-1}\}m \\ &\qquad + V(\delta^2\Lambda + V)^{-1}x_i,\ \{I_p - V(\delta^2\Lambda + V)^{-1}\}V] \\ &\equiv N_p[\eta_{z|x},\ V_{11.2}]. \end{aligned} \tag{12}$$

The CNQL estimate of $\beta$, say $\hat{\beta}_{\mathrm{CNQL}}$ is then obtained by solving the QL estimating equation

$$\sum_{i=1}^{K} \frac{\partial\{E[Y_i|x_i]\}}{\partial\beta}[\mathrm{var}(Y_i|x_i)]^{-1}(y_i - E[Y_i|x_i]) = 0, \tag{13}$$

(Liang and Liu 1991, Eq. (4.11), p. 51), where by applying (12), the conditional expectation and covariance matrix may be computed by using the formulas

$$\begin{aligned} E[Y_i|x_i] &= E_{z_i|x_i}[Y_i|z_i] = E_{z_i|x_i}[a'(z_i'\beta)], \\ \mathrm{var}[Y_i|x_i] &= E_{z_i|x_i}[\mathrm{var}(Y_i|z_i)] + \mathrm{var}_{z_i|x_i}[E(Y_i|z_i)] \\ &= E_{z_i|x_i}[a''(z_i'\beta)] + \mathrm{var}_{z_i|x_i}[a'(z_i'\beta)]. \end{aligned}$$

Note that in this structural setup, Wang et al. (1996) have used a naive mean and variance-based QL approach where QL estimating equation for $\beta$ is constructed by replacing the observed covariate vector $x_i$ with its mean obtained from a repeated sampling. In fact this type of repeated samples is usually employed to estimate the measurement error variances. Their approximate QL estimating equation has the form

$$\sum_{i=1}^{K}\left[\frac{\partial\{a'(x_i'\beta)\}}{\partial\beta}[a''(x_i'\beta)]^{-1}(y_i-a'(x_i'\beta))\right]_{|x_i=\tilde{x}_i}=0,$$

where $\tilde{x}_i$ is the mean computed from the replicates of $x_i$. The relative performance of this approximate QL approach with other existing approaches is, however, not known.

Turning back to the functional setup, the CNQL estimating (12) may be modified by using fixed $z_i$ and its relationship to $x_i$ given in (2), that is, $x_i = z_i + \delta v_i$. It follows in this case that one may still solve the CNQL (12) for $\beta$, but the conditional expectation and variance are computed as

$$\begin{aligned} E[Y_i|x_i] &= E_{v_i}[\{a'(z_i'\beta)\}_{|z_i=x_i-\delta v_i}] \\ \text{var}[Y_i|x_i] &= E_{v_i}[\{a''(z_i'\beta)\}_{|z_i=x_i-\delta v_i}] + \text{var}_{v_i}[\{a'(z_i'\beta)\}_{|z_i=x_i-\delta v_i}], \end{aligned} \tag{14}$$

where $v_i \sim N_p[0, \Lambda = \text{diag}(\sigma_1^2,\ldots,\sigma_u^2,\ldots,\sigma_p^2)]$.

### 2.1.3 An Approximate BCQL Approach Using Corrected Estimating Function

We propose a bias correction approach along the lines of Nakamura (1990). The difference between Nakamura's and our approach is that Nakamura (1990) developed a corrected score function $\ell^*(\beta;y,x)$ such that its expectation is the true but unknown score function, that is, $E_x[\ell^*(\beta;y,x)] = \ell(\beta;y,z)$, and then solved the corrected score (3) for $\beta$, whereas in our approach we develop a corrected quasi-likelihood function, say $Q^*(y,x,\beta)$, such that

$$E_x[Q^*(y,x,\beta)] = \psi(y,z,\beta), \tag{15}$$

where by (5), $\psi(y,z,\beta) = \sum_{i=1}^{K}\psi_i(y_i,z_i,\beta)$ is the true QL function in unknown covariates $z_i$, and solve the corrected QL equation, that is, $Q^*(\beta,y,x) = 0$ for $\beta$. Also, this bias correction approach is different than the SVQL approach of Stefanski (1985) as it does not require any small variance assumption to hold.

#### Poisson Regression Model

If the true covariates $z_i$ were known, then for the Poisson regression model it follows from (5) that the QL estimating equation would have the form

$$\psi(y,z,\beta) = \sum_{i=1}^{K}\psi_i(y_i,z_i,\beta) = \sum_{i=1}^{K} z_i(y_i-\mu_{iz}) = 0, \tag{16}$$

with $\mu_{iz} = \exp(z_i'\beta)$. For the purpose of developing a corrected QL function $Q^*(\beta, y, x)$, by replacing $z_i$ with $x_i$, we first write the NQL estimating equation as

$$Q(y,x,\beta) = \sum_{i=1}^{K} Q_i(y_i, x_i, \beta) = \sum_{i=1}^{K} x_i(y_i - \mu_{ix}) = 0, \tag{17}$$

where $\mu_{ix} = \exp(x_i'\beta)$. Under the measurement error model (2), that is, when $x_i = z_i + \delta v_i$, it is clear that NQL function $Q(y,x,\beta)$ is not unbiased for the true QL function $\psi(y,z,\beta)$. That is,

$$E_x[Q(y,x,\beta)] = E_x \sum_{i=1}^{K} x_i(y_i - \mu_{ix}) \neq \psi(y,z,\beta) = \sum_{i=1}^{K} z_i(y_i - \mu_{iz}).$$

Note, however, that under the Gaussian measurement error model (2), that is when $x_i \sim N_p(z_i, \delta^2\Lambda)$, one obtains $E_x[\exp(x_i'\beta)|z_i] = \exp(z_i'\beta + \xi) = \mu_{iz}\exp(\xi)$, where $\xi = \frac{\delta^2}{2}\beta'\Lambda\beta$, yielding

$$E_x \mu_{ix} \exp(-\xi) = \mu_{iz}. \tag{18}$$

Further it may be shown that $E_x[x_i \exp(x_i'\beta)|z_i] = [z_i + \delta^2\Lambda\beta]\mu_{iz}\exp(\xi)$ (Nakamura 1990), yielding

$$E_x[x_i \mu_{ix} \exp(-\xi)] = z_i\mu_{iz} + \delta^2\Lambda\beta\mu_{iz}. \tag{19}$$

Now by using (18), it follows from (19) that

$$E_x[\{x_i - \delta^2\Lambda\beta\}\mu_{ix}\exp(-\xi)] = z_i\mu_{iz}. \tag{20}$$

Consequently, one obtains the BCQL function

$$Q^*(y,x,\beta) = \sum_{i=1}^{K}[x_i y_i - \{(x_i - \delta^2\Lambda\beta)\mu_{ix}\exp(-\xi)\}] \tag{21}$$

which satisfies

$$E_x[Q^*(y,x,\beta)] = \sum_{i=1}^{K} z_i(y_i - \mu_i z), \tag{22}$$

yielding the BCQL estimating equation for $\beta$ in the Poisson model as

$$\sum_{i=1}^{K}[x_i y_i - \{(x_i - \delta^2\Lambda\beta)\mu_{ix}\exp(-\xi)\}] = 0. \tag{23}$$

We denote the solution of (23) by $\hat{\beta}_{\text{BCQL}}$. This estimator is consistent for $\beta$. Remark that this BCQL estimating (23) is the same as the corrected score equation derived by Nakamura (1990, Sect. 4.3). Thus, in the Poisson measurement model setup, the BCQL approach provides the same regression estimate as the bias corrected likelihood approach.

Binary Regression Model

In the binary regression case, the true but unknown mean function is given by $\mu_{iz} = \exp(z_i'\beta)/[1+\exp(z_i'\beta)]$, whereas in the Poisson case $\mu_{iz} = \exp(z_i'\beta)$. This makes it difficult to find a corrected QL function $\tilde{Q}(y,x,\beta)$ such that

$$E_x[\tilde{Q}(y,x,\beta)] = \sum_{i=1}^{K} z_i[y_i - \frac{\exp(z_i'\beta)}{1+\exp(z_i'\beta)}] = \tilde{\psi}(y,z,\beta) \tag{24}$$

in the binary case. However, a softer, that is, a first order approximate BCQL (SBCQL) estimating function may be developed as follows. We denote this SBCQL function as $\tilde{Q}_S(y,x,\beta)$ which will be approximately unbiased for $\tilde{\psi}(y,z,\beta)$, that is,

$$E_x[\tilde{Q}_S(y,x,\beta)] \simeq \tilde{\psi}(y,z,\beta).$$

Recall from (18) and (20) that

$$E_x[\exp(x_i'\beta - \xi)] = \exp(z_i'\beta), \tag{25}$$

$$E_x[\{x_i - \delta^2\Lambda\beta\}\exp(x_i'\beta - \xi)] = z_i\exp(z_i'\beta), \tag{26}$$

where $\xi = \frac{\delta^2}{2}\beta'\Lambda\beta$. It then follows that

$$E_x\left[\frac{\{x_i - \delta^2\Lambda\beta\}\exp(x_i'\beta - \xi)}{1+\exp(x_i'\beta - \xi)}\right] \simeq \frac{z_i\exp(z_i'\beta)}{1+\exp(z_i'\beta)}. \tag{27}$$

Next because the true QL function has the form

$$\tilde{\psi}(y,z,\beta) = \sum_{i=1}^{K} z_i y_i - \sum_{i=1}^{K}[\frac{z_i\exp(z_i'\beta)}{1+\exp(z_i'\beta)}],$$

by using (27), one may write a softer BCQL (SBCQL) estimating equation as

$$\sum_{i=1}^{K}\left[x_i y_i - \frac{\{x_i - \delta^2\Lambda\beta\}\exp(x_i'\beta - \xi)}{1+\exp(x_i'\beta - \xi)}\right] = 0. \tag{28}$$

We denote the solution of the SBCQL estimating (28) by $\hat{\beta}_{\mathrm{SBCQL}}$. Note that this estimator may still be biased and on a more serious note it may not even converge to $\beta$. This is because the expectation shown in (27) may differ to a great extent from the actual expectation. However exploiting a better approximation for the expectation as follows may remove the convergence problem and also may yield estimates with smaller bias.

For the purpose, rewrite the expectation in (27) as

$$E_x\left[\frac{\{x_i-\delta^2\Lambda\beta\}\exp(x_i'\beta-\xi)}{1+\exp(x_i'\beta-\xi)}\right]\simeq\frac{z_i\exp(z_i'\beta)}{1+\exp(z_i'\beta)}=\frac{\mu_{W_{z,N}}}{\mu_{W_{z,D}}},\tag{29}$$

and improve the expectation as follows. To be specific, we first compute an improved expectation as

$$\begin{aligned}E_x\left[\frac{\{x_i-\delta^2\Lambda\beta\}\exp(x_i'\beta-\xi)}{1+\exp(x_i'\beta-\xi)}\right]&=E_x\left[\frac{W_{x,N}}{W_{x,D}}\right]\\&\simeq\frac{\mu_{W_{z,N}}}{\mu_{W_{z,D}}}-\frac{\hat{\mathrm{cov}}[W_{x,N},W_{x,D}]}{\hat{\mu}^2_{W_{z,D}}}+\frac{\hat{\mu}_{W_{z,N}}}{\hat{\mu}^3_{W_{z,D}}}\hat{\mathrm{var}}[W_{x,D}],\end{aligned}\tag{30}$$

where we use

$$\begin{aligned}\hat{\mu}_{W_{z,N}}&=\frac{1}{K}\sum_{i=1}^{K}[\{x_i-\delta^2\Lambda\beta\}\exp(x_i'\beta-\xi)]\\\hat{\mu}_{W_{z,D}}&=\frac{1}{K}\sum_{i=1}^{K}[1+\exp(x_i'\beta-\xi)]\\\hat{\mathrm{var}}[W_{x,D}]&=\frac{1}{K}\sum_{i=1}^{K}[1+\exp(x_i'\beta-\xi)]^2-\hat{\mu}^2_{W_{z,D}}\\\hat{\mathrm{cov}}[W_{x,N},W_{x,D}]&=\frac{1}{K}\sum_{i=1}^{K}\left[\{(x_i-\delta^2\Lambda\beta)\exp(x_i'\beta-\xi)\}\{1+\exp(x_i'\beta-\xi)\}\right]\\&\quad-\hat{\mu}_{W_{z,N}}\hat{\mu}_{W_{z,D}}\end{aligned}\tag{31}$$

We then rewrite (30) as

$$E_x\left[\frac{\{x_i-\delta^2\Lambda\beta\}\exp(x_i'\beta-\xi)}{1+\exp(x_i'\beta-\xi)}+t_c\right]=\frac{\mu_{W_{z,N}}}{\mu_{W_{z,D}}},\tag{32}$$

where

$$t_c=\frac{\hat{\mathrm{cov}}[W_{x,N},W_{x,D}]}{\hat{\mu}^2_{W_{z,D}}}-\frac{\hat{\mu}_{W_{z,N}}}{\hat{\mu}^3_{W_{z,D}}}\hat{\mathrm{var}}[W_{x,D}].$$

Thus, instead of (28), we now solve the improved SBCQL estimating equation given by

$$\sum_{i=1}^{K}\left[x_i y_i - \frac{\{x_i - \delta^2 \Lambda\beta\}\exp(x_i'\beta - \xi)}{1+\exp(x_i'\beta-\xi)} - t_c\right] = 0. \tag{33}$$

## 3 Measurement Error Analysis in Longitudinal Setup

With regard to the correlations for the repeated responses, not much attention is paid to model such correlations, where the associated covariates are subject to measurement error. However, in time series setup, there exist some studies for continuous responses subject to measurement error. For example, we refer to the study by Staudenmayer and Buonaccorsi (2005), where time series responses are assumed to follow the Gaussian auto-regressive order 1 (AR(1)) correlation process subject to measurement errors. But, these studies are not applicable to the longitudinal setup, especially for discrete longitudinal data such as for repeated count data with covariates subject to measurement error.

In longitudinal setup, both repeated responses and measurement errors in covariates are likely to be correlated. Because the repeated measurement errors usually share a common instrument/machine/individual effect, in this study we assume that this type of errors follow a familial correlation structure such as mixed model-based equi-correlation structure. As far as the repeated responses are concerned, it is likely that they will follow a dynamic relationship causing certain auto-correlations among them as time effects. Thus, similar to Sutradhar (2011), in this study we assume that the repeated responses will follow a general class of auto-correlation structures. It is, however, known that the repeated linear, count, and binary data exhibit similar but different auto-correlation structures especially when the covariates are time dependent (nonstationary). For this reason, in this section, we deal with the measurement error models for these three types of response data separately and discuss them in sequence in the following three subsections.

### *3.1 Linear Auto-correlation Models with Measurement Error in Covariates*

In this section, we consider functional error-in-variables models for continuous (linear) panel data. Let

$$\begin{aligned} y_{it} &= z_{it}'\beta + w_i\gamma_i^* + \varepsilon_{it}, \text{ for } t = 1,\ldots,T_i, \\ x_{it} &= z_{it} + v_{it}, \end{aligned} \tag{34}$$

represent such a measurement error model, where $y_{it}$ denotes a continuous response for the $i$th $(i = 1,\ldots,K)$ individual recorded at time $t$ $(t = 1,\ldots,T_i)$ with $2 \le T_i \le T$, $z_{it} = (z_{it1}\ldots,z_{itu},\ldots,z_{itp})'$ be the $p \times 1$ true but unobserved time-dependent covariate vector, $\beta = (\beta_1,\ldots,\beta_u,\ldots,\beta_p)'$ be the $p \times 1$ vector of regression parameters, $\gamma_i^*$ is the $i$th individual random effect with $\gamma_i^* \overset{iid}{\sim} (0,\sigma_\gamma^2)$, and $w_i$ is a known additional covariate for the $i$th individual on top of the fixed covariates $z_{it}$. Furthermore, $\varepsilon_{it}$ in (34) is the model error such that marginally $\varepsilon_{it} \sim (0,\sigma_\varepsilon^2)$, but jointly $\varepsilon_{i1},\ldots,\varepsilon_{it},\ldots,\varepsilon_{iT_i}$ follow a serially correlated such as AR(1) (auto-regressive order 1) or MA(1) (moving average order 1) process. Furthermore, in (34),

$$x_{it} = (x_{it1}\ldots,x_{itu},\ldots,x_{itp})', \text{ and } v_{it} = (v_{it1}\ldots,v_{itu},\ldots,v_{itp})',$$

with

$$v_{itu} \sim (0,\sigma_u^2), \text{ for } u = 1,\ldots,p$$

at any time point $t = 1,\ldots,T_i$. Here, as in Sect. 1, $\sigma_u^2$ is known as the measurement error variance for the $u$th covariate. Because the measurement errors $v_{i1u},\ldots,v_{itu},\ldots,v_{iT_iu}$ for measuring the same $u$th covariate values at different times are likely to be correlated due to a common instrumental random effect $m_{i|u}$, (say), we consider

$$v_{itu} = m_{i|u} + a_{itu}, \text{ for } t = 1,\ldots,T_i \tag{35}$$

and assume that $m_{i|u} \overset{iid}{\sim} (0,\tilde{\sigma}_u^2)$ and $a_{itu} \overset{iid}{\sim} (0,\sigma_a^2)$, and $m_{i|u}$ and $a_{itu}$ are independent. It is then clear from (35) that the variance of $v_{itu}$ and the correlation between $v_{isu}$ and $v_{itu}$ are given by

$$\text{var}(v_{itu}) = \sigma_u^2 = \tilde{\sigma}_u^2 + \sigma_a^2, \text{ and } \text{corr}(v_{isu},v_{itu}) = \phi_u = \frac{\tilde{\sigma}_u^2}{\tilde{\sigma}_u^2 + \sigma_a^2}, \tag{36}$$

for all $s \ne t, s,t = 1,\ldots,T_i$.

By writing $Z_i = [z_{i(1)},\ldots,z_{i(u)},\ldots,z_{i(p)}] : T_i \times p$, with $z_{i(u)} = (z_{i1u},\ldots,z_{itu},\ldots,z_{iT_iu})'$; $X_i = [x_{i(1)},\ldots,x_{i(u)},\ldots,x_{i(p)}] : T_i \times p$, with $x_{i(u)} = (x_{i1u},\ldots,x_{itu},\ldots,x_{iT_iu})'$; and $V_i = [v_{i(1)},\ldots,v_{i(u)},\ldots,v_{i(p)}] : T_i \times p$, with $v_{i(u)} = (v_{i1u},\ldots,v_{itu},\ldots,v_{iT_iu})'$, and expressing the measurement error model (34) in matrix notation as

$$y_i = Z_i\beta + 1_{T_i} w_i \gamma_i^* + \varepsilon_i \tag{37}$$

$$X_i = Z_i + V_i \tag{38}$$

with $y_i = (y_{i1},\ldots,y_{iT_i})'$, $\varepsilon_i = (\varepsilon_{i1},\ldots,\varepsilon_{iT_i})'$, and $1_{T_i}$ as the $T_i$-dimensional unit vector, one can first write the so-called naive MM (NMM) estimating equation for $\beta$ as

$$\psi^* = \sum_{i=1}^{K} X_i'(y_i - X_i\beta) = 0, \tag{39}$$

but its solution would produce biased and hence inconsistent estimate for $\beta$, because $E_{x|y}[\sum_{i=1}^{K} X_i'(y_i - X_i\beta)] \neq \sum_{i=1}^{K} Z_i'(y_i - Z_i\beta)$, due to the fact that in the present measurement error setup $E[V_i'V_i] \neq 0$ even though $E[V_i] = 0$. As a remedy, by exploiting

$$\begin{aligned} E[V_i'V_i] &= T_i \text{diag}[\tilde{\sigma}_1^2 + \sigma_a^2, \ldots, \tilde{\sigma}_u^2 + \sigma_a^2, \ldots, \tilde{\sigma}_p^2 + \sigma_a^2] \\ &= T_i \text{diag}[\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2] \\ &= T_i \Lambda(\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2), (\text{say}) \end{aligned} \tag{40}$$

that is,

$$E[X_i'X_i] = Z_i'Z_i + E[V_i'V_i] = Z_i'Z_i + T_i\Lambda(\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2),$$

one may obtain a BCMM estimator for $\beta$ by solving the BCMM estimating equation

$$\begin{aligned} \psi(x, y; \beta, \sigma_1^2, \ldots, \sigma_p^2) &= \sum_{i=1}^{K} X_i'y_i - [\sum_{i=1}^{K} \{X_i'X_i - T_i\Lambda(\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2)\}]\beta \\ &= \sum_{i=1}^{K} \psi_i(x_i, y_i; \beta, \sigma_1^2, \ldots, \sigma_p^2) \end{aligned} \tag{41}$$

(Griliches and Hausman 1986) yielding the BCMM estimator as

$$\hat{\beta}_{BCMM} = \left[\sum_{i=1}^{K} \{X_i'X_i - T_i\Lambda(\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2)\}\right]^{-1} \sum_{i=1}^{K} X_i'y_i. \tag{42}$$

This BCMM estimator is consistent for $\beta$ but can be inefficient.

Recently, some authors such as Wansbeek (2001) (see also Wansbeek and Meijer 2000) considered a slightly different model than (37)–(38) by also involving certain strictly exogenous explanatory variables (in addition to $Z_i$) and by absorbing the random effects $\gamma_i^*$ into the error vector $\varepsilon_i$ that avoids the estimation of the variance component of the random effects $\sigma_\gamma^2$. Wansbeek (2001) developed necessary moment conditions to form BCMM estimating equations in order to obtain consistent GMM estimates for the regression parameters involved including the effect of the exogenous covariates. More recently, Xiao et al. (2007) studied the efficiency properties of the BCGMM approach considered by Wansbeek (2001). Note that the derivation of the efficient BCGMM estimators by Xiao et al. (2007) may be considered as the generalization of the GMM approach of Hansen (1982) to the measurement error models. In studying the efficiency of the BCGMM

approach, Xiao et al. (2007), however, assumed that the model errors $\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}$ are independent to each other. Also they assume that the measurement errors $v_{i1u}, \ldots, v_{iT_iu}$ in (38) (see also (34)) are serially correlated.

More recently, by treating the model errors $\varepsilon_{i1}, \ldots, \varepsilon_{iT_i}$ as serially correlated with a general auto-correlation structure

$$C_i(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T_i-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T_i-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{T_i-1} & \rho_{T_i-2} & \rho_{T_i-3} & \cdots & 1 \end{bmatrix}, \tag{43}$$

(Sutradhar 2003) and by considering a more practical familial type equi-correlation structure (36) for the measurement errors, that is,

$$E[v_{i(u)}v'_{i(u)}] = \sigma_u^2[\phi_u 1_{T_i}1'_{T_i} + (1-\phi)I_{T_i}], \tag{44}$$

Fan et al. (2012) compared the efficiency of the BCGMM estimator with a new BCGQL (also referred to as BCGLS) approach, the latter being more efficient. These two approaches are briefly described in the following two sub-sections.

### 3.1.1 BCGMM Estimation for Regression Effects

Note that the BCMM estimating (41) is an unbiased estimating equation because of the fact that

$$E_y E_{x|y}\psi(x,y;\beta,\sigma_1^2,\ldots,\sigma_p^2) = \sum_{i=1}^{K} E_{y_i}E_{x_i|y_i}\psi_i(x_i,y_i;\beta,\sigma_1^2,\ldots,\sigma_p^2) = 0.$$

Consequently, the BCMM estimator for $\beta$ in (42) was obtained by solving

$$\psi(x,y;\beta,\sigma_1^2,\ldots,\sigma_p^2) = 0,$$

but this estimator can be inefficient. As a remedy, following Hansen (1982) (see also Xiao et al. 2007, Eq. (2.4)), Fan et al. (2012) discuss a BCGMM approach, where one estimates $\beta$ by minimizing the quadratic form

$$Q = \psi' C \psi \tag{45}$$

for a suitable $p \times p$, positive definite matrix $C$, with $C = [\text{cov}(\psi)]^{-1}$ as an optimal choice. In (45), $\psi$ is an unbiased moment function given by (41). Note that since the computation of the $\text{cov}(\psi)$ matrix requires the formulas for the third and fourth order moments of $\{x_{itu}\}$ as well, one cannot compute such a covariance matrix provided the measurement error distributions for the model (34)

are known. However, as argued in the independent setup, it is reasonable for many practical situations that measurement errors are normally distributed. As far as their covariance structure is concerned we assume that they follow the structure in (36). Based on this normality assumption for the measurement error, we reexpress the $C$ matrix in (45) as $C_N$ and obtain the BCGMM estimator for $\beta$ by solving the estimating equation

$$\frac{\partial \psi'}{\partial \beta} C_N \psi = 0, \tag{46}$$

where by (41)

$$\frac{\partial \psi'}{\partial \beta} = [\sum_{i=1}^{K} \{X_i' X_i - T_i \Lambda(\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2)\}].$$

It then follows that the solution of (46), i.e., the BCGMM estimator of $\beta$ is given by

$$\hat{\beta}_{BCGMM} = \left[\frac{\partial \psi'}{\partial \beta} C_N \frac{\partial \psi}{\partial \beta'}\right]^{-1} \left[\frac{\partial \psi'}{\partial \beta} C_N \sum_{i=1}^{K} X_i' y_i\right], \tag{47}$$

with its variance as

$$\begin{aligned} \text{var}(\hat{\beta}_{BCGMM}) = & \left[\frac{\partial \psi'}{\partial \beta} C_N \frac{\partial \psi}{\partial \beta'}\right]^{-1} \\ & \times \left[\frac{\partial \psi'}{\partial \beta} C_N \sum_{i=1}^{K} \text{var}(X_i' y_i) C_N \frac{\partial \psi}{\partial \beta}\right] \left[\frac{\partial \psi'}{\partial \beta} C_N \frac{\partial \psi}{\partial \beta'}\right]^{-1}. \end{aligned} \tag{48}$$

Construction of $C_N$ Matrix

Note that $C_N = [\text{var}(\psi)]^{-1}$ under the assumption that the measurement errors $\{v_{itu}\}$ and hence observed covariates $\{x_{itu}\}$ are normally distributed. For the purpose, we first compute $\text{var}(\psi)$ as follows where $\psi$ is given as in (41):

$$\begin{aligned} \text{var}(\psi) &= \text{var}[\sum_{i=1}^{K} X_i' y_i - \{\sum_{i=1}^{K} X_i' X_i\}\beta] \\ &= \sum_{i=1}^{K} [\text{var}\{X_i' y_i - X_i' X_i \beta\}] \\ &= \sum_{i=1}^{K} [\text{var}\{X_i' y_i\} + \text{var}\{X_i' X_i \beta\} - 2\text{cov}\{X_i' y_i, X_i' X_i \beta\}], \end{aligned} \tag{49}$$

which, in addition to the formulas for the covariance matrix of $y_i$, requires the formulas for all possible second, third, and fourth order moments of $\{x_{itu}\}$. The following two lemmas will be useful in computing the covariance matrices in (49).

**Lemma 3.1.** *Under the measurement error model (34)–(37), let* $var(Y_i) = \Sigma_i = w_i^2\sigma_\gamma^2 J_{T_i} + \sigma_\varepsilon^2 R_i = (\sigma_{i\ell m})$ *denote the* $T_i \times T_i$ *covariance matrix of the response vector* $y_i$*, where* $J_{T_i}$ *is the* $T_i \times T_i$ *unit matrix and* $R_i = (\rho_{i\ell m})$ *is the* $T_i \times T_i$ *correlation matrix for the components of* $\varepsilon_i$ *such as for AR(1) process* $\rho_{i\ell m} = \rho^{|\ell - m|}$*,* $\rho$ *being the correlation index parameter. It then follows that*

$$\sigma_{i\ell m} = cov[Y_{i\ell}, Y_{im}] = \begin{cases} \sigma_i^{*2} & \text{for } \ell = m = 1, \ldots, T_i \\ \sigma_i^{*2}[\theta_i + (1-\theta_i)\rho_{i\ell m}] & \text{for } \ell \neq m, \end{cases} \tag{50}$$

*where* $\sigma_i^{*2} = w_i^2\sigma_\gamma^2 + \sigma_\varepsilon^2$*, and* $\theta_i = \frac{w_i^2\sigma_\gamma^2}{w_i^2\sigma_\gamma^2 + \sigma_\varepsilon^2}$.

**Lemma 3.2.** *Let* $\Delta_{i(u)} = (\delta_{i(uu)\ell m})$ *denote the* $T_i \times T_i$ *covariance matrix of* $x_{i(u)} = (x_{i1u}, \ldots, x_{itu}, \ldots, x_{iT_iu})'$*, where by (36)*

$$cov[x_{i\ell u}, x_{imu}] = \delta_{i(uu)\ell m} = \begin{cases} \sigma_u^2 = \tilde{\sigma}_u^2 + \sigma_a^2 \text{ for } \ell = m = 1, \ldots, T_i \\ \tilde{\sigma}_u^2 = \phi_u \sigma_u^2 \quad \text{for } \ell \neq m. \end{cases} \tag{51}$$

Under the assumption that $v_{i(u)}$ or $x_{i(u)}$ in (38) follows the $T_i$-dimensional normal distribution with covariance matrix $\Delta_{i(u)}$ as in Lemma 3.2, the third and fourth order corrected product moments for the components of $x_{i(u)}$ are given by

$$\eta_{i\ell mt} = E\left[(x_{i\ell u} - z_{i\ell u})(x_{imu} - z_{imu})(x_{itu} - z_{itu})\right] = 0, \tag{52}$$

and

$$\begin{aligned} \xi_{i\ell mst} &= E\left[(x_{i\ell u} - z_{i\ell u})(x_{imu} - z_{imu})(x_{isu} - z_{isu})(x_{itu} - z_{itu})\right] \\ &= \delta_{i(uu)\ell m}\delta_{i(uu)st} + \delta_{i(uu)\ell s}\delta_{i(uu)mt} + \delta_{i(uu)\ell t}\delta_{i(uu)ms}, \end{aligned} \tag{53}$$

respectively.

By applying the Lemmas 3.1 and 3.2, one may compute the covariance matrices in (49). For example, by writing the $p \times 1$ vector $X_i'y_i$ as $X_i'y_i = [\sum_{t=1}^{T_i} x_{it1}y_{it}, \ldots, \sum_{t=1}^{T_i} x_{itu}y_{it}, \ldots, \sum_{t=1}^{T_i} x_{itp}y_{it}]'$, one may compute its $p \times p$ covariance matrix as

$$\text{var}[X_i'y_i] = \begin{cases} \text{var}[\sum_{t=1}^{T_i} x_{itu}y_{it}] & \text{for } u = 1, \ldots, p \\ \text{cov}[\sum_{t=1}^{T_i} x_{itu}y_{it}, \sum_{t=1}^{T_i} x_{itr}y_{it}] & \text{for } u \neq r,\ u, r = 1, \ldots, p, \end{cases} \tag{54}$$

where

$$\begin{aligned}\text{var}[\sum_{t=1}^{T_i} x_{itu}y_{it}] &= \text{var}_y E_x[\sum_{t=1}^{T_i} x_{itu}y_{it}|y] + E_y\text{var}_x[\sum_{t=1}^{T_i} x_{itu}y_{it}|y] \\ &= \text{var}_y[\sum_{t=1}^{T_i} z_{itu}y_{it}] + E_y[\sum_{t=1}^{T_i}\sum_{m=1}^{T_i} \delta_{i(uu)tm}y_{it}y_{im}] \\ &= \sum_{t=1}^{T_i}\sum_{m=1}^{T_i} z_{itu}z_{imu}\sigma_{itm} \\ &+ \sum_{t=1}^{T_i}\sum_{m=1}^{T_i} \delta_{i(uu)tm}[\sigma_{itm} + \beta' z_{it}z'_{im}\beta] \end{aligned} \tag{55}$$

and

$$\begin{aligned}\text{cov}[\sum_{t=1}^{T_i} x_{itu}y_{it}, \sum_{t=1}^{T_i} x_{itr}y_{it}] &= \text{cov}_y[E_x\{\sum_{t=1}^{T_i} x_{itu}y_{it}|y\}, E_x\{\sum_{t=1}^{T_i} x_{itr}y_{it}|y\}] \\ &+ E_y\text{cov}_x[\{\sum_{t=1}^{T_i} x_{itu}y_{it}, \sum_{t=1}^{T_i} x_{itr}y_{it}\}|y] \\ &= \text{cov}_y[\sum_{t=1}^{T_i} z_{itu}y_{it}, \sum_{t=1}^{T_i} z_{itr}y_{it}] + E_y[\sum_{t=1}^{T_i}\sum_{m=1}^{T_i} y_{it}y_{im}\delta_{i(ur)tm}|y] \\ &= \sum_{t=1}^{T_i}\sum_{m=1}^{T_i} z_{itu}z_{imu}\sigma_{itm}, \end{aligned} \tag{56}$$

because two covariates ($u \neq r$) are always independent, i.e., $\delta_{i(ur)tm} = 0$ irrespective of the time points of their measurements. The remaining two covariance matrices in (49) may be computed similarly.

### 3.1.2 BCGQL Estimation for Regression Effects

In this approach, by pretending that the model (37)–(38) does not contain any measurement error, we first write the naive generalized quasi-likelihood (NGQL) estimating equation

$$\Psi^* = \sum_{i=1}^{K} X_i'\Sigma_i^{-1}[y_i - X_i\beta] = 0, \tag{57}$$

where $\Sigma_i$ is the covariance matrix of $y_i$. Note that the estimating function $\Psi^*$ is similar but different than the MM estimating function $\psi^*$ given in (39). The solution of (57) yields an NGQL estimator for $\beta$ as

$$\hat{\beta}_{NGQL} = \left[\sum_{i=1}^{K} X_i' \Sigma_i^{-1} X_i\right]^{-1} \sum_{i=1}^{K} X_i' \Sigma_i^{-1} y_i, \tag{58}$$

which is also familiar as the generalized least squares (GLS) estimator for $\beta$. Note that this NGQL estimator $\hat{\beta}_{NGQL} = \hat{\beta}_{GLS}$ is not consistent for $\beta$. This is because, $\Psi^*$ in (57) is not an unbiased function under the true model (37)–(38), that is, $E(\Psi^*) \neq 0$.

Now to obtain an unbiased and hence consistent estimator for $\beta$, it is necessary to consider an unbiased GQL function under the present model. This would be a generalization of finding the moment conditions for MM studied by Wansbeek (2001) to the actual correlation setup for the panel data.

In order to obtain an unbiased function from the $\Psi_i^*$ function in (57), we first note that in probability $(\to_p)$, $X_i' \Sigma_i^{-1} X_i$ converges as

$$X_i' \Sigma_i^{-1} X_i \to_p \ [Z_i' \Sigma_i^{-1} Z_i + \text{diag}\{tr(\Sigma_i^{-1} \Delta_{i(1)}), \ldots, tr(\Sigma_i^{-1} \Delta_{i(p)})\}], \tag{59}$$

where for $u = 1, \ldots, p$, $\Delta_{i(u)}$ is given in Lemma 3.2 (see also (36)). Now by using (59), we may modify (57) to obtain an unbiased estimating function given by

$$\begin{aligned} \Psi = \sum_{i=1}^{K} X_i' \Sigma_i^{-1} y_i - [\sum_{i=1}^{K} \{X_i' \Sigma_i^{-1} X_i \\ - \text{diag}[tr(\Sigma_i^{-1} \Delta_{i(1)}), \ldots, \ldots, tr(\Sigma_i^{-1} \Delta_{i(p)})]\}]\beta, \end{aligned} \tag{60}$$

that is, $E[\Psi] = 0$ under the model (37)–(38). Consequently, for known measurement error variances, it is now clear from (60) that one may obtain the BCGQL estimator given by

$$\begin{aligned} \hat{\beta}_{BCGQL} = \Bigg[\sum_{i=1}^{K} \{X_i' \Sigma_i^{-1} X_i \\ - \text{diag}[tr(\Sigma_i^{-1} \Delta_{i(1)}), \ldots, tr(\Sigma_i^{-1} \Delta_{i(p)})]\}\Bigg]^{-1} \sum_{i=1}^{K} X_i' \Sigma_i^{-1} y_i, \end{aligned} \tag{61}$$

which is consistent for $\beta$. Also, this BCGQL estimator would be more efficient than the BCMM estimator given in (42). This is because, unlike the BCMM estimator, the BCGQL estimator is constructed by using the covariance matrix $\Sigma_i$ of $y_i$ as the weight matrix in the estimating equation. In fact, in view of the comparative results for GQL and GMM estimators in the linear panel data setup (Rao et al. 2012), this BCGQL estimator (61) may also be more efficient than the BCGMM estimator obtained in (47).

Note that the asymptotic variance of $\hat{\beta}_{BCGQL}$ may be estimated as follows. By writing

$$P_{i,x} = X_i' \Sigma_i^{-1} X_i - \text{diag}[tr(\Sigma_i^{-1} \Delta_{i(1)}), \ldots, tr(\Sigma_i^{-1} \Delta_{i(p)})],$$

and because

$$E_y E_{x|y}[X_i' \Sigma_i^{-1} y_i] = [Z_i' \Sigma_i^{-1} Z_i] \beta$$

is estimated by $P_{i,x}\beta$, one obtains a moment estimator of $\text{var}[\hat{\beta}_{BCGQL}]$ as

$$\hat{\text{var}}[\hat{\beta}_{BCGQL}] = [\sum_{i=1}^{K} P_{i,x}]^{-1} \sum_{i=1}^{K} [X_i' \Sigma_i^{-1} y_i - P_{i,x} \hat{\beta}_{BCGQL}] \times [X_i' \Sigma_i^{-1} y_i - P_{i,x} \hat{\beta}_{BCGQL}]' [\sum_{i=1}^{K} P_{i,x}]^{-1}. \quad (62)$$

A Two-Stage BCGQL (BCGQL2) Estimation of $\beta$

Instead of solving the first stage estimating (60) for BCGQL estimator, similar to the BCGMM estimation (46), Fan et al. (2012) have solved the second stage estimating equation

$$\sum_{i=1}^{K} \left[ \frac{\partial \Psi_i'}{\partial \beta} D_{iN}^{-1} \Psi_i \right] = 0, \quad (63)$$

where, for $\Psi = \sum_{i=1}^{K} \Psi_i$ (60), with $\Psi_i = X_i' \Sigma_i^{-1} y_i - P_{i,x}\beta$,

$$D_{iN} = \text{cov}[\Psi_i]$$

under the assumption of multivariate normality for the random covariates $x_{i(u)} = [x_{i1u}, \ldots, x_{itu}, \ldots, x_{iT_iu}]'$. It then follows that the solution of (63), i.e., the two stage BCGQL BCGQL2 estimator of $\beta$ is given by

$$\hat{\beta}_{BCGQL2} = \left[ \sum_{i=1}^{K} \frac{\partial \Psi_i'}{\partial \beta} D_{iN}^{-1} \frac{\partial \Psi_i}{\partial \beta'} \right]^{-1} \sum_{i=1}^{K} \left[ \frac{\partial \Psi_i'}{\partial \beta} D_{iN}^{-1} X_i' \Sigma_i^{-1} y_i \right], \quad (64)$$

with its variance as

$$\text{var}[\hat{\beta}_{BCGQL2}] = \left[ \sum_{i=1}^{K} \frac{\partial \Psi_i'}{\partial \beta} D_{iN}^{-1} \frac{\partial \Psi_i}{\partial \beta'} \right]^{-1} \sum_{i=1}^{K} \left[ \frac{\partial \Psi_i'}{\partial \beta} D_{iN}^{-1} \text{var}(X_i' \Sigma_i^{-1} y_i) D_{iN}^{-1} \frac{\partial \Psi_i}{\partial \beta'} \right] \times \left[ \sum_{i=1}^{K} \frac{\partial \Psi_i'}{\partial \beta} D_{iN}^{-1} \frac{\partial \Psi_i}{\partial \beta'} \right]^{-1}, \quad (65)$$

where $\text{var}(X_i'\Sigma_i^{-1}y_i)$ may be computed similar to that of $\text{var}(X_i'y_i)$ in (54). Further, the covariance matrix $D_{iN}$ can be computed in the fashion similar to that of $C_N$ in Sect. 3.1.1.

## 3.2 Longitudinal Count Data Models with Measurement Error in Covariates

When compared to the linear measurement error model for correlated data (34), in the present case, one has to deal with a correlation model for repeated count data $y_{i1},\ldots,y_{it},\ldots,y_{iT}$, where $y_{it}$ marginally, as in Sect. 2.1.3, follows a count data distribution such as Poisson distribution with mean $\mu_{iz} = \exp(z_{it}'\beta)$. However, as far as the measurement errors are concerned, they arise through the same relationship $x_{it} = z_{it} + v_{it}$, as in the correlated linear model setup.

For the correlation structure for count data, we consider a practically important AR(1) model following Sutradhar (2010) (see also Sutradhar 2011). The model is written such that conditional on the true covariate vector $z_{it}$, the marginal means and variances satisfy the Poisson distribution-based relationship

$$E(Y_{it}|z_{it}) = \text{var}(Y_{it}|z_{it}) = \mu_{iz,t} = \exp(z_{it}'\beta), \tag{66}$$

for all $t = 1,\ldots,T$. Note that these two moments are nonstationary as they depend on the time-dependent covariates $z_{it}$. As far as the AR(1) correlations among repeated counts are concerned, they arise from the following dynamic relationships:

$$\begin{aligned} y_{i1} &\sim Poi(\mu_{iz,1}) \\ y_{it} &= \rho * y_{i,t-1} + d_{it} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho) + d_{it},\ t = 2,\ldots,T, \end{aligned} \tag{67}$$

where for given counts $y_{i,t-1}$ at time point $t-1$, $\sum_{j=1}^{y_{i,t-1}} b_j(\rho)$ denotes the sum of $y_{i,t-1}$ independent binary values with $Pr[b_j(\rho) = 1] = \rho$ and $Pr[b_j(\rho) = 0] = 1-\rho$, $\rho$ being the longitudinal correlation index parameter. Now under the assumptions that $y_{i,t-1} \sim Poi(\mu_{iz,t-1})$, $d_{it} \sim Poi(\mu_{iz,t} - \rho\mu_{iz,t-1})$, for $t = 2,\ldots,T$, and $d_{it}$ and $y_{i,t-1}$ are independent, it follows from (67) that $y_{ir}$ and $y_{it}$ have nonstationary lag $t-r$ correlations given by

$$\text{corr}(Y_{ir}, Y_{it}) = c_{iz,rt} = \begin{cases} \rho^{t-r}[\mu_{iz,r}\mu_{iz,t}^{-1}]^{\frac{1}{2}}, \text{ for } r < t \\ \rho^{r-t}[\mu_{iz,t}\mu_{ir,t}^{-1}]^{\frac{1}{2}}, \text{ for } r > t. \end{cases} \tag{68}$$

Note that the lag correlations given by (68) are nonstationary by nature as they depend on the time-dependent variances through the covariates $z_{it}$ and $z_{iu}$, whereas

in the stationary case when $z_{it} = z_{iu}$ for all $u \neq t$, they reduce to $\rho^{t-u}$, a Gaussian-type $AR(1)$ correlation structure satisfying (43). Further note that because $E[Y_{it}] = \mu_{iz,t} = \exp(z'_{it}\beta)$ by (67), the regression parameters vector $\beta$ measures the effects of $z_{it}$ on $y_{it}$ for all $t = 1, \ldots, T$. But in the present setup, $z_{it}$'s are unobservable, and hence they cannot be used to estimate $\beta$. Instead, one must use the observed covariates $x_{it}$, which are, however, subject to measurement error explained through the relationship

$$x_{it} = z_{it} + v_{it},$$

with $v_{it} = (v_{it1}, \ldots, v_{itu}, \ldots, v_{itp})'$ satisfying the following assumptions:

1. $v_{it} \sim N(0, \Lambda = \text{diag}[\sigma_1^2, \ldots, \sigma_u^2, \ldots, \sigma_p^2])$ for all $t = 1, \ldots, T$.
2. Also,

$$\text{corr}[v_{iru}, v_{itm}] = \begin{cases} \phi_u, \text{ for } m = u; r \neq t, r, t = 1, \ldots, T \\ 0, \quad \text{for } m \neq u; r, t = 1, \ldots, T. \end{cases}$$

These two assumptions imply that the $u$th covariate has the measurement error variance $\sigma_u^2$ for $u = 1, \ldots, p$, at a given time $t$ for all $t = 1, \ldots, T$. Also, the covariate values for the same $u$th covariate recorded at two different times $r$ and $t$ are equally correlated with correlation $\phi_u$ for all $r \neq t$. This correlation assumption is similar to that of the time-dependent covariates considered by Wang et al. (1996). One may also consider other correlation structures such as AR(1) among the repeated values for the same covariate. More specifically, the above assumptions is equivalent to writing

$$\begin{pmatrix} x_{ir} \\ x_{it} \end{pmatrix} \sim N_{2p} \left[ \begin{pmatrix} z_{ir} \\ z_{it} \end{pmatrix}, \begin{pmatrix} \Lambda & \Lambda_\phi \\ \Lambda_\phi & \Lambda \end{pmatrix} \right], \tag{69}$$

where $\Lambda_\phi = \text{cov}(v_{ir}, v'_{it}) = \text{diag}[\phi_1 \sigma_1^2, \ldots, \phi_u \sigma_u^2, \ldots, \phi_p \sigma_p^2]$.

### 3.2.1 Bias Corrected GQL Estimation

Suppose that by using the observed covariates one writes a NGQL estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial \mu'_{ix}}{\partial \beta} \Sigma_{ix}^{-1} (y_i - \mu_{ix}) = 0, \tag{70}$$

where

$$\mu_{ix,t} = \mu_{iz,t}|_{z_{it}=x_{it}}, \text{ and } \Sigma_{ix} = (\sigma_{ix,rt}) = \Sigma_{iz}|_{z=x} = ([c_{iz,rt}\sqrt{\mu_{iz,r}\mu_{iz,t}}]|_{z_{it}=x_{it}}).$$

But, this NGQL estimating (70) will yield biased and hence inconsistent estimate for $\beta$. This is because the NGQL estimating function in the left-hand side of the (70) is not unbiased for the true covariates-based GQL estimating function. That is,

$$E_x\left[\sum_{i=1}^{K}\frac{\partial\mu_{ix}'}{\partial\beta}\Sigma_{ix}^{-1}(y_i-\mu_{ix})\right]\neq\sum_{i=1}^{K}\frac{\partial\mu_{iz}'}{\partial\beta}\Sigma_{iz}^{-1}(y_i-\mu_{iz}). \tag{71}$$

Recently, Sutradhar et al. (2012) have proposed a bias correction to the NGQL estimating function and developed a BCGQL estimating function which is unbiased for the true covariates-based estimating function $\sum_{i=1}^{K}\frac{\partial\mu_{iz}'}{\partial\beta}\Sigma_{iz}^{-1}(y_i-\mu_{iz})$. This provides the BCGQL estimating equation as

$$\begin{aligned}g_x(x,\beta,\rho,\Lambda,\phi_1,\ldots,\phi_p|y)=\sum_{i=1}^{K}\Big[&\{M_{1\phi}X_i'-M_{1\phi}B_{1\phi}(\beta\otimes 1_T')\}\\&\times\{A_{ix}^{\frac{1}{2}}\tilde{Q}_{ix}(\rho)A_{ix}^{-\frac{1}{2}}\}y_i-\{M_{2\phi}X_i'-M_{2\phi}B_{2\phi}(\beta\otimes 1_T')\}\\&\times\{A_{ix}^{\frac{1}{2}}\tilde{Q}_{ix}(\rho)A_{ix}^{-\frac{1}{2}}\}\mu_{ix}\Big]=0,\end{aligned} \tag{72}$$

where $y_i=(y_{i1},\ldots,y_{it},\ldots,y_{iT})'$ is the $T\times 1$ vector of repeated count responses, with its mean $\mu_{ix}=\exp(x_{it}'\beta)$ in observed covariates; $X_i'=(x_{i1},\ldots,x_{it},\ldots,x_{iT})$ is the $p\times T$ observed covariates matrix; $A_{ix}=\text{diag}[\mu_{ix,1},\ldots,\mu_{ix,t},\ldots,\mu_{ix,T}]$; $1_T'=(1\ldots,1)$ is the $1\times T$ vector of unity, $\otimes$ denotes the well-known Kronecker or direct product, so that $\beta\otimes 1_T'$ is the $p\times T$ matrix containing $\beta=(\beta_1\ldots,\beta_p)'$ in each column of the matrix; and

$$\begin{aligned}B_{1\phi}&=\frac{1}{2}(\Lambda-\Lambda_\phi),B_{2\phi}=\frac{1}{2}(\Lambda+\Lambda_\phi),\\M_{1\phi}&=\text{diag}[m_1,\ldots,m_1]:p\times p;M_{2\phi}=\text{diag}[m_2,\ldots,m_2]:p\times p,\end{aligned}$$

with

$$m_1=\exp\{-\frac{1}{4}\beta'(\Lambda-\Lambda_\phi)\beta\},\text{ and }m_2=\exp\{-\frac{1}{4}\beta'(\Lambda+\Lambda_\phi)\beta\}.$$

Furthermore, in (72), $\tilde{Q}_{ix}(\rho)=\tilde{C}_{ix}^{-1}(\rho)$, with $C_{ix}(\rho)=(\tilde{c}_{ix,rt})$ as an unbiased correlation matrix for the AR(1) correlation matrix in true covariates, namely $C_{iz}(\rho)=(c_{iz,rt})$. The formula for the $(r,t)$-th element of the unbiased correlation matrix is given by

$$\tilde{c}_{ix,rt}=\rho^{t-r}[\exp(x_{ir}-x_{it})'\frac{\beta}{2}-\frac{1}{4}\beta'(\Lambda-\Lambda_\phi)\beta] \tag{73}$$

satisfying

$$E_x[\tilde{c}_{ix,rt}] = c_{iz,rt} = \rho^{t-r}\{\mu_{iz,r}/\mu_{iz,t}\}^{\frac{1}{2}}. \tag{74}$$

We reexpress the BCGQL estimating (72) as

$$g_x(x,\beta,\rho,\Lambda,\phi_1,\ldots,\phi_p|y) = \sum_{i=1}^{K}[D_{i1}(x)y_i - D_{i2}(x)\mu_{ix}] = 0, \tag{75}$$

where $D_{i1}(x)$ and $D_{i2}(x)$ are $p \times T$ matrix functions of observed covariates. Let $\hat{\beta}_{BCGQL}$ be the solution of (73). Now conditional on the observed covariates $x_i$, solving this equation for $\beta$ is equivalent to use the iterative equation,

$$\begin{aligned}\hat{\beta}_{BCGQL}(r+1) \;=\; \hat{\beta}_{BCGQL}(r) - \Bigg[\{\hat{E}_y(\frac{\partial g_x(x,\beta,\rho,\Lambda,\phi_1,\ldots,\phi_p|y)}{\partial\beta'})\}^{-1} \\ \times \sum_{i=1}^{K}\{D_{i1}(x)y_i - D_{i2}(x)\mu_{ix}\}\Bigg]_{\hat{\beta}_{BCGQL}(r)},\end{aligned} \tag{76}$$

where $\hat{\beta}_{BCGQL}(r)$ denote the $\beta$ estimate at the r-th iteration. Note that under the true model involving covariates $z_i$,

$$y_i \sim [\mu_{iz}, \Sigma_{iz} = (\sigma_{itm})],$$

where for $t < m$, $\sigma_{itm} = \text{cov}(y_{it}, y_{im}) = (\rho^{m-t}\mu_{iz,t})$, and $y_1,\ldots,y_i,\ldots,y_K$ are $T$-dimensional independent vectors. Thus, under some mild moment conditions, by using Lindeberg-Feller central limit theorem (Amemiya 1985, Theorem 3.3.6, p. 92), it follows from (76) that as $K \to \infty$, $\hat{\beta}_{BCGQL} \sim N_p(\beta, V^*)$, where

$$\begin{aligned}V^* = \Bigg[\hat{E}_y(\frac{\partial g_x(x,\beta,\rho,\Lambda,\phi_1,\ldots,\phi_p|y)}{\partial\beta'})\Bigg]^{-1}\sum_{i=1}^{K}D_{i1}(x)\Sigma_{iz}D_{i1}'(x) \\ \times \Bigg[\hat{E}_y(\frac{\partial g_x(x,\beta,\rho,\Lambda,\phi_1,\ldots,\phi_p|y)}{\partial\beta'})\Bigg]^{-1},\end{aligned} \tag{77}$$

which may be consistently estimated by using the moment estimate for $\Sigma_{iz}$ in (77). For this moment estimate, when $\rho$ is known, one estimates the $(t,m)$th element $(t < m)$ of this matrix by using

$$\hat{\sigma}_{itm} = \rho^{m-t}\hat{\mu}_{iz,t} = \rho^{m-t}\mu_{ix,t} = \rho^{m-t}[exp\,(x_{it}'\beta - \frac{1}{2}\beta'\Lambda\beta)]_{|\beta=\hat{\beta}_{BCGQL}}.$$

### 3.2.2 A Simulation-Based Numerical Illustration

We consider two $(p = 2)$ covariates with measurement error variances $\sigma_1^2$ and $\sigma_2^2$, respectively. It is expected that these measurement error variances are small in

practice. We, however, consider them ranging from 0.1 to 0.3 for $\sigma_1^2$; and from 0.1 to 0.8 for $\sigma_2^2$. Note that these ranges are quite large, whereas in the independence ($\rho = 0.0$) setup and for one covariate case ($p = 1$), Nakamura (1990) examined the performance of the bias corrected score estimator for $\sigma_1^2$ up to 0.1. We have also included the independence case but for larger measurement error variances as compared to that of Nakamura (1990).

The main purpose of this section is to illustrate the performance of the proposed BCGQL estimator obtained from (72) (see also (76)) when AR(1) count responses are generated with some positive correlation index, where the covariates are subject to measurement error with variances $\sigma_1^2$ and $\sigma_2^2$ for the two covariate case. We consider $\rho = 0.5$. As mentioned above we also include the independence case ($\rho = 0.0$). In all these cases, we first show that if measurement errors are not adjusted, the so-called NGQL approach (70) produces highly biased estimates and the correction by using BCGQL approach performs well.

We consider 500 simulations and generate correlated count data following the AR(1) Poisson model (67)–(68) for $K = 100$ individuals over a period of $T = 4$ time points. The true covariates $z_{it1}$ and $z_{it2}$ were generated as

$$z_{it1} \overset{iid}{\sim} N(0,1), \text{ and } z_{it2} \overset{iid}{\sim} \frac{\chi_4^2 - 4}{\sqrt{8}}$$

with their effects $\beta_1 = 0.3$ and $\beta_2 = 0.1$, respectively, on the repeated response $y_{it}$. Note that even though the true covariates $z_{itu}$ are generated following the standard normal and standardized $\chi^2$ distribution, these values are treated as fixed under all simulations. Further note that these true covariates are unobserved in the present setup, instead $x_{it1}$ and $x_{it2}$ are observed. We generate the observed covariates following the relationship

$$x_{itu} = z_{itu} + v_{itu},\ u = 1, \ldots, p,$$

where $v_{itu}$'s are generated by using a random effect model given by

$$v_{itu} = k_u + e_{itu}, \text{ with } k_u \overset{iid}{\sim} N(0, \sigma_u^{*2}) \text{ and } e_{itu} \overset{iid}{\sim} N(0, \sigma_e^{*2}), \tag{78}$$

yielding

$$\begin{aligned} \text{var}(v_{itu}) &= \sigma_u^{*2} + \sigma_e^{*2} = \sigma_u^2 \\ \text{corr}(v_{itu}, v_{iru}) &= \frac{\sigma_u^{*2}}{\sigma_u^{*2} + \sigma_e^{*2}} = \phi_u, \end{aligned} \tag{79}$$

where $\sigma_u^2$ is the measurement error variance for the $u$th ($u = 1, 2$). Notice that $\phi_u$ represents the equi-correlations among the repeated values of the same covariate. Thus, $\phi_u = 1$ would represent the situation where covariate values are same over

**Table 1** Simulated regression estimates, and their standard errors (SSEs), with true regression parameters $\beta_1 = 0.3$, $\beta_2 = 0.1$, under AR(1) count data model for selected response correlation $\rho$, measurement error variances $\sigma_1^2$, $\sigma_2^2$, with $K = 100$; $T = 4$; and measurement error correlations $\phi_1$ and $\phi_2$; and true covariate values $Z_1 \sim N(0,1)$ and $Z_2 \sim \frac{\chi_4^2 - 4}{\sqrt{8}}$

| | | | | | Estimates | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | NGQL | | BCGQL | |
| $\rho$ | $\phi_1$ | $\phi_2$ | $\sigma_1^2$ | $\sigma_2^2$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ |
| 0.0 | 1.0 | 1.0 | 0.1 | 0.3 | 0.2683 | 0.0849 | 0.3025 | 0.1026 |
| | | | | | (0.0501) | (0.0380) | (0.0583) | (0.0448) |
| | | | 0.3 | 0.3 | 0.2274 | 0.0800 | 0.3052 | 0.1033 |
| | | | | | (0.0462) | (0.0383) | (0.0680) | (0.0468) |
| | | | 0.3 | 0.8 | 0.2221 | 0.0652 | 0.3068 | 0.1052 |
| | | | | | (0.0450) | (0.0338) | (0.0701) | (0.0525) |
| 0.5 | 1.0 | 1.0 | 0.1 | 0.3 | 0.2688 | 0.0900 | 0.3036 | 0.1085 |
| | | | | | (0.0689) | (0.0535) | (0.0803) | (0.0634) |
| | | | 0.3 | 0.3 | 0.2286 | 0.0854 | 0.3069 | 0.1099 |
| | | | | | (0.0640) | (0.0536) | (0.0920) | (0.0662) |
| | | | 0.3 | 0.8 | 0.2232 | 0.0707 | 0.3100 | 0.1138 |
| | | | | | (0.0623) | (0.0493) | (0.0979) | (0.0786) |
| 0.0 | 0.25 | 0.50 | 0.1 | 0.3 | 0.2680 | 0.0842 | 0.2772 | 0.0914 |
| | | | | | (0.0502) | (0.0372) | (0.0523) | (0.0401) |
| | | | 0.3 | 0.3 | 0.2270 | 0.0793 | 0.2435 | 0.0871 |
| | | | | | (0.0461) | (0.0372) | (0.0499) | (0.0402) |
| | | | 0.3 | 0.8 | 0.2218 | 0.0643 | 0.2404 | 0.0779 |
| | | | | | (0.0450) | (0.0329) | (0.0495) | (0.0392) |
| 0.5 | 0.25 | 0.50 | 0.1 | 0.3 | 0.2338 | 0.0751 | 0.2800 | 0.1050 |
| | | | | | (0.0617) | (0.0484) | (0.0764) | (0.0652) |
| | | | 0.3 | 0.3 | 0.1687 | 0.0683 | 0.2525 | 0.1148 |
| | | | | | (0.0521) | (0.0476) | (0.0943) | (0.1065) |
| | | | 0.3 | 0.8 | 0.1642 | 0.0492 | 0.2488 | 0.1097 |
| | | | | | (0.0506) | (0.0406) | (0.1314) | (0.1579) |

time and in this case we consider $x_{itu} = k_u$, which yields $\text{corr}(x_{itu}, x_{iru}) = \phi_u = 1.0$. But it does not mean though responses are same, rather responses follow the AR(1) correlation structure. In the simulation study, we, however, consider both situations where $\phi_u = 1.0$ for $u = 1, 2$, in one situation; and in the other situation $\phi_1 = 0.25$ and $\phi_2 = 0.5$.

The simulated estimates along with their standard errors are presented in Table 1 for all selected values of the parameters. As expected, the NGQL estimates appear to be highly biased. For example, when $\phi_1 = \phi_2 = 1.0$, the response correlation index is 0.5, and measure error variances are $\sigma_1^2 = 0.3$, $\sigma_2^2 = 0.8$, the NGQL approach produces the estimates of $\beta_1 = 0.3$ and $\beta_2 = 0.1$ as 0.22 and 0.07, whereas the BCGQL approach yields almost unbiased estimates as 0.31 and 0.11, respectively. When $\phi_1 = 0.25$ and $\phi_2 = 0.5$, for this set of large measurement error variances,

the NGQL approach produces useless estimates, 0.16 for $\beta_1 = 0.30$, and 0.05 for $\beta_2 = 0.10$. In this case, BCGQL approach still appears to produce reasonably good estimates, 0.25 for $\beta_1 = 0.30$, and 0.11 for $\beta_2 = 0.10$. The BCGQL estimates for $\beta_2$ appears to be unbiased in all selected situations. As far as the independence case $\rho = 0.0$ is concerned, the BCGQL approach works similarly to the correlation case with $\rho = 0.5$.

**Acknowledgment** The author fondly acknowledges the stimulating discussion by the audience of the symposium and wishes to thank for their comments and suggestions.

## References

Amemiya, T.: Advanced Econometrics. Harvard University Press, Cambridge (1985)
Amemiya, Y.: Two-stage instrumental variable estimator for the non-linear errors-in-variables model. J. Econom. **44**, 311–332 (1990)
Armstrong, B.: Measurement error in the generalized linear model. Comm. Stat. Simulat. Comput. **14**, 529–544 (1985)
Buonaccorsi, J.P.: Measurement Error: Models, Methods, and Applications. Chapman and Hall/CRC, London (2010)
Buzas, J.S., Stefanski, L.A.: Instrumental variable estimation in generalized linear measurement error models. J. Am. Stat. Assoc. **91**, 999–1006 (1996a)
Buzas, J.S., Stefanski, L.A.: Instrumental variable estimation in a probit measurement error model. J. Stat. Plann. Infer. **44**, 47–62 (1996b)
Carroll, R.J., Ruppert, D., Stefanski, L., Crainiceanunn, C.M.: Measurement Error in Nonlinear Models–A Modern Perspective. Chapman & Hall/CRC, London (2006)
Carroll, R.J., Stefanski, L.A.: Approximate quasi-likelihood estimation in models with surrogate predictors. J. Am. Stat. Assoc. **85**, 652–663 (1990)
Carroll, R.J., Wand, M.P.: Semiparametric estimation in logistic measurement error models. J. R. Stat. Soc. B **53**, 573–585 (1991)
Fan, Z., Sutradhar, B., Rao, R.P.: Bias corrected generalised method of moments and generalised quasi-likelihood inferences in linear models for panel data with measurement error. Sankhya B **74**, 126–148 (2012)
Fuller, W.A.: Measurement Error Models. Wiley, New York (1987)
Griliches, Z., Hausman, J.A.: Errors in variables in panel data. J. Econometrics **32**, 93–118 (1986)
Hansen, L.-P.: Large-sample properties of generalized method of moment estimators. Econometrica **50**, 1029–1054 (1982)
Kendall, M.G., Stuart, A.: The Advanced Theory of Statistics, vol. 2. Griffin, London (1979)
Liang, K.-Y., Liu, X.: Estimating equations in generalized linear models with measurement error. In: Godambe, V.P. (ed.) Estimating Functions, pp. 47–63. Oxford University Press, New York (1991)
Nakamura, T.: Corrected score function for errors-in-variables models: methodology and application to generalized linear models. Biometrika **77**, 127–137 (1990)
Rao, R.P., Sutradhar, B.C., Pandit, V.N.: GMM versus GQL inferences in linear dynamic panel data models. Braz. J. Probab. Stat. **26**, 167–177 (2012)
Schafer, D.W.: Covariate measurement errors in generalized linear models. Biometrika **74**, 385–391 (1987)
Stefanski, L.A.: The effects of measurement error on parameter estimation. Biometrika **72**, 583–592 (1985)
Stefanski, L.A.: Measurement error models. J. Am. Stat. Assoc. **95**, 1353–1358 (2000)

Stefanski, L.A., Buzas, J.S.: Instrumental variable estimation in binary regression measurement error variables. J. Am. Stat. Assoc. **87**, 87–90 (1995)
Stefanski, L.A., Carroll, R.J.: Covariate measurement error in logistic regression. Ann. Stat. **13**, 1335–1351 (1985)
Stefanski, L.A., Carroll, R.J.: Conditional scores and optimal scores for generalized linear measurement-error models. Biometrika **74**, 703–716 (1987)
Staudenmayer, J., Buonaccorsi, J.P.: Measurement error in linear autoregressive models. J. Am. Stat. Assoc. **100**, 841–852 (2005)
Sutradhar, B.C.: An review on regression models for discrete longitudinal responses. Stat. Sci. **18**, 377–393 (2003)
Sutradhar, B.C.: Inferences in generalized linear longitudinal mixed models. Can. J. Stat. **38**, 174–196 (2010)
Sutradhar, B.C.: Dynamic Mixed Models for Familial Longitudinal Data. Springer, New York (2011)
Sutradhar, B.C., Rao, J.N.K.: Estimation of regression parameters in generalized linear models for cluster correlated data with measurement error. Can. J. Stat. **24**, 177–192 (1996)
Sutradhar, B.C., Rao, R.P., Chowdhury, R.I.: Inferences for longitudinal count data with measurement error in covariates. Technical Report, Department of Mathematics and Statistics, Memorial University, Canada (2012)
Wang, N., Carroll, R.J., Liang, K.-Y.: Quasilikelihood estimation in measurement error models with correlated replicates. Biometrics **52**, 401–411 (1996)
Wansbeek, T.J.: GMM estimation in panel data models with measurement error. J. Econometrics **104**, 259–268 (2001)
Wansbeek, T.J., Meijer, E.: Measurement Error and Latent Variables in Econometrics. North-Holland, Amsterdam (2000)
Wedderburn, R.: Quasilikelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika **61**, 439–447 (1974)
Whittemore, A.S., Keller, J.B.: Approximations for regression with covariate measurement error. J. Am. Stat. Assoc. **83**, 1057–1066 (1988)
Xiao, Z., Shao, J, Xu, R., Palta, M.: Efficiency of GMM estimation in panel data models with measurement error. Sankhya **69**, 101–118 (2007)

# Bias Reduction in Logistic Regression with Estimated Variance Predictors

**Laine Thomas, Leonard A. Stefanski, and Marie Davidian**

**Abstract** We study the problem of modeling a response as a function of baseline covariates and a primary predictor of interest that is a noisy measurement of a subject-specific variance. The problem arises naturally in biostatistical joint models wherein the subjects' primary endpoints are related to the features of subject-specific longitudinal risk processes or profiles. Often the longitudinal process features of interest are parameters of a longitudinal mean function. However, there is a relatively recent and growing interest in relating primary endpoints to longitudinal process variances. In the application motivating our work longitudinal processes consist of 30-day blood pressure trajectories measured between 91 and 120 days post dialysis therapy, with the primary endpoints being short-term mortality. Often the longitudinal risk processes are adequately characterized in terms of trends such as the slopes and intercepts identified with the subject-specific biomarker processes. Modeling of the trend lines results in subject-specific estimated intercepts and slopes, thus inducing a heteroscedastic measurement-error model structure where the estimated trend parameters play the role of measurements of the "true" subject-specific trend parameters that appear as predictors in the primary endpoint model. Our interest lies in models in which the residual variances of the longitudinal processes feed into the model for the primary endpoint. These subject-specific variance parameters are estimated in the course of trend-line fitting creating a measurement error model scenario where variances are predictors and mean squared errors are their noisy measurements. Background literature is reviewed and several methodological approaches for addressing the resulting errors-in-variances problem are studied.

L. Thomas
Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA
e-mail: laine.thomas@duke.edu

L.A. Stefanski (✉) • M. Davidian
Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA
e-mail: stefansk@ncsu.edu; davidian@ncsu.edu

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_2, 

## 1 Introduction

In biological studies of health effects a primary endpoint is sometimes related to the longitudinal profiles of a continuous secondary response (e.g., systolic blood pressure). The subject-specific random effects that describe each individual's longitudinal trajectory are often included as covariates in an outcome model. It is well known that using these estimates of these subject-specific random effects results in biased inference. A variety of methods exist to correct this bias (see, for example, Li et al. 2004). Longitudinal measurements also allow for assessment of variability over time. Recent studies seek to quantify the relationship between a disease outcome and subject-specific variability (Yang et al. 2007; Havlik et al. 2002). Substantial bias may occur when variance estimates are imputed for true variances as covariates in the outcome model (Lyles et al. 1999). In the most general setting both subject-specific random effects estimates and variance estimates can enter into the primary-endpoint model. In this case, coefficient estimation should account for both sources of measurement error. We focus on the case that only estimated variances enter the model (along with error-free baseline covariates), while explaining how to generalize our methods to the more general modeling setting. In particular we study three relatively simple approaches for the special case of logistic regression when the subject-specific variances enter the model as *precisions*, i.e., as inverse variances. In this case the conditional score method (Stefanski and Carroll 1987; Carroll et al. 2006) provides a nice solution to the problem that allows for models with interactions between the variance predictor and baseline predictors. We compare the conditional score estimator to two simple data-adjustment methods that counter the measurement-error induced attenuation by shrinking the noisy predictors. The first can be viewed as either a variant of regression calibration (Carroll and Stefanski 1990; Rosner et al. 1989) or a simple linear measurement error model correction for attenuation. We call it *attenuation-corrected calibration*. The second is an extension of the first that imputes variance estimates adjusted to have means, variances, and covariances (with baseline predictors) that are asymptotically correct (Tukey 1974; Bay 1997; Freedman et al. 2004; Elliott 2009). Both attenuation-corrected calibration (ACCal) and moment adjusted imputation are attractive because of their simplicity—post adjustment, the adjusted values are used in standard software without further modification to obtain parameter estimates (accurate standard error estimation is more involved but can usually be done by the bootstrap). We describe the methods and compare them via simulation.

Health outcomes are frequently modeled by logistic regression where the covariates may include subject-specific random effects and/or residual variances which characterize the longitudinal features of a continuous response. In this paper, we investigate the relationship between longitudinal blood pressure variability and short-term mortality in hemodialysis (HD) patients. In healthy patients blood pressure variability is often measured using 24-hour ambulatory BP monitoring. However describing variability in HD patients is more complex since their treatment may induce fluctuations over longer periods than 24 hours. For this reason, in a retrospective analysis of the Accelerated Mortality on Renal Replacement (ArMORR) cohort,

Yang et al. 2007 measured blood pressure variability over a three-month exposure window (90–180 days) before a six-month follow-up period (181–365 days) to asses mortality. Thus they focused on relatively long-term variability and mortality among patients who had survived at least 180 days after the introduction of HD treatment. Blood pressure was not measured until 90 days in order to allow subjects to "acclimate" to dialysis therapy. The ArMORR study includes adults ($> 18$ years) who were incident to hemodialysis between June 2004 and August 2005. Blood pressure measurements were obtained from 90 to 120 days, and patient follow-up began at 120 days. The average number of replicates observed during this period is 12 with standard deviation of 1.5. Consequently the subject-specific estimates are subject to substantial variability about the corresponding unknown parameters.

This problem has been thoroughly studied for outcome models which include subject-specific random effects like longitudinal slope and intercept parameters. Wang et al. (2000) characterize these joint models and show that a naive approach that simply substitutes ordinary least squares estimates for the random effects leads to biased inference. To adjust for this measurement error bias, they propose a variety of structural methods including regression calibration (Carroll et al. 2006, Chap. 4) and a psuedo-expected estimating equation approach. These methods rely on normality of the underlying random effects. In order to avoid parametric assumptions on the random effects parameters Li et al. (2004) use a conditional-score approach for generalized linear models. They follow the strategy of Stefanski and Carroll (1987) and derive unbiased estimating equations by conditioning on sufficient statistics for the unknown random effects parameters. The common feature in these joint models is that the estimated random effect parameters have a normal distribution, conditional on the true random effects. In other words, the measurement error is normally distributed. This is suitable for joint modeling applications where interest focuses on longitudinal slope and intercept parameters.

Many recent studies define joint models where the primary outcome model depends on subject-specific variance components. Yang et al. (2007) identified an association between hemoglobin variability and mortality in patients with end stage renal disease, even after controlling for the absolute level and temporal trend in hemoglobin. Havlik et al. (2002) reported increased risk of late-life white matter brain lesions for patients with higher variability in systolic blood pressure during mid-life. Similar work is abundant (Brunelli et al. 2008; Liu et al. 1978; Grove et al. 1997; Iribarren et al. 1995). These outcome models often include both random effects and variance components to describe the longitudinal data. Coefficient estimation should account for both sources of measurement error, which are correlated. Moreover, the measurement error in the variance estimates is clearly not normal (under normality assumptions for the longitudinal model the variance estimates, appropriately scaled, have a chi-square distribution).

It seems that the problem of measurement error is not well appreciated in this context since the aforementioned studies all take a naive approach to model fitting. This is not surprising since, to the best of our knowledge, there is very little methodology to address this problem. A key exception is a maximum likelihood method proposed by Lyles et al. 1999. They note that variance estimates have

high variability relative to estimates of a mean from a similar sample size, so the problem of measurement error can be even more pronounced in the case of estimated variances. They study a continuous outcome, decline in CD4 counts after seroconversion, modeled by linear regression that depends on both the subject-specific mean and variance of longitudinal CD4 counts measured before seroconversion. They obtain a full likelihood for the observed data by assuming that the subject-specific means are normally distributed and independent of log-normally distributed subject-specific variances. Based on these assumptions they fit models by maximum likelihood as well as a regression calibration-type approach. The methods we study make less demanding assumptions about the distribution of the true subject variances and therefore provide an alternatively to heavily parameterized modeling.

In Sect. 2 we describe a general joint-effects model where primary modeling can depend on both subject-specific longitudinal mean parameters and longitudinal variance parameters. Then we narrow the focus to the case that the primary model is logistic regression and that in addition to baseline covariates only subject-specific variances enter the primary model. Estimators for that special case are derived in Sect. 3. We derive the conditional score estimator and also the necessary adjustment formulas for the attenuation-corrected calibration and moment adjusted imputation estimators. Section 4 presents results from a simulation study. Section 5 presents some extensions and their limitations. Section 6 concludes with summary remarks and notes on possible future research.

## 2 Joint Model with Variance Predictors

In the general joint model of interest data $(\mathbf{Z}_i, \mathbf{S}_i, \mathbf{T}_i, Y_i)$ are recorded for each of $i = 1, \ldots, n$ subjects. The outcome variable is $Y_i$ which we assume is binary; baseline covariates are codified in $\mathbf{Z}_i$; the vector $\mathbf{S}_i$ contains repeated measurements $\mathbf{S}_i = (S_{it}, \ldots, S_{ir_i})^T$ collected at times $\mathbf{T}_i = (t_{i1}, \ldots, t_{ir_i})^T$. In the application motivating our model the $S_{it}$ are blood pressure measurements taken post dialysis after a stabilization period, and $Y_i$ is an indicator of mortality in specified follow-up period.

The longitudinal component of our joint model is a linear random-coefficient model

$$\mathbf{S}_i = \mathbf{D}_i \boldsymbol{\gamma}_i + \sigma_i \mathbf{U}_i, \qquad \mathbf{U}_i(\mathbf{0}, \mathbf{I}_{r_i}), \tag{1}$$

where $\mathbf{D}_i$ has $j^{\text{th}}$ row $(1, t_{ij})$, $j = 1, \ldots, r_i$ and $\boldsymbol{\gamma}_i = (\gamma_{0i}, \gamma_{1i})^T$. The model has subject-specific intercepts $\gamma_{0,1}, \ldots, \gamma_{0,n}$, subject-specific slopes, $\gamma_{1,1}, \ldots, \gamma_{1,n}$, *and* in its most general form, subject-specific variances $\sigma_1^2, \ldots, \sigma_n^2$. We do not assume that the $\boldsymbol{\gamma}_i$ or the $\sigma_i^2$ are random, choosing rather to work under the more general framework that these are fixed unknown parameters. The latter assumption corresponds to the so-called *functional* measurement error model, and thus our methods are *functional methods* in the terminology of Carroll et al. (2006).

We study the case in which the outcome is binary and is modeled via logistic regression. The models studied in this paper are all submodels of the general form

$$P(Y_i = 1 | \mathbf{Z}_i, \boldsymbol{\gamma}_i, \sigma_i^2; \boldsymbol{\beta}) = H\left(\beta_0 + \mathbf{Z}_i^T \boldsymbol{\beta}_Z + \boldsymbol{\gamma}_i^T \boldsymbol{\beta}_\gamma + \beta_\sigma T(\sigma_i^2) + \mathbf{Z}_{1i}^T \boldsymbol{\beta}_{Z_2\sigma} T(\sigma_i^2)\right), \tag{2}$$

where $\mathbf{Z}_{1i}$ is a subvector of $\mathbf{Z}_i$, $H(t) = 1/(1+e^{-t})$, and $T()$ is a transformation applied to the subject-specific variance, e.g., identity, square root, logarithm, reciprocal. In the models above $\boldsymbol{\beta} = \left(\beta_0,\ \boldsymbol{\beta}_Z^T,\ \boldsymbol{\beta}_\gamma^T,\ 0\beta_\sigma,\ \boldsymbol{\beta}_{Z_2\sigma}\right)^T$.

### 2.1 Longitudinal Model Summary Statistics

Assuming that the subject-specific longitudinal models are fit via ordinary least squares we have that:

$$\hat{\boldsymbol{\gamma}}_i = (\mathbf{D}_i^T \mathbf{D}_i)^{-1} \mathbf{D}_i^T \mathbf{S}_i \tag{3}$$

and

$$\hat{\sigma}_i^2 = \frac{(\mathbf{S}_i - \mathbf{D}_i \hat{\boldsymbol{\gamma}}_i)^T (\mathbf{S}_i - \mathbf{D}_i \hat{\boldsymbol{\gamma}}_i)}{\nu_i}. \tag{4}$$

If we add to our the model the assumption that

$$\mathbf{U}_i \sim N(\mathbf{0}, \mathbf{I}_{r_i}), \quad \text{for} \quad i = 1, \ldots, n, \tag{5}$$

then it follows that conditioned on $(\boldsymbol{\gamma}_i, \sigma_i^2)$ the estimators $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\sigma}_i^2$ are independently distributed as

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_i &\sim \mathrm{N}\{\boldsymbol{\gamma}_i, \sigma_i^2 (\mathbf{D}_i^T \mathbf{D}_i)^{-1}\} \\ \frac{\nu_i}{\sigma_i^2} \hat{\sigma}_i^2 &\sim \chi^2(\nu_i), \end{aligned} \tag{6}$$

for $i = 1, \ldots, n$.

## 3 Outcome Model Methods of Analysis

### 3.1 Simple Substitution (aka the "Naive" Method)

In the case that $\nu_i = r_i - 2$ are large for all $i$, so that the estimation variation in $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\sigma}_i^2$ is small relative to the estimation variability inherent in the outcome model, the parameter $\boldsymbol{\beta} = \left(\beta_0,\ \boldsymbol{\beta}_Z^T,\ \boldsymbol{\beta}_\gamma^T,\ \beta_\sigma,\ \boldsymbol{\beta}_{Z_2\sigma}\right)^T$ in the outcome model in (2)

is estimated by logistic regression of $Y_i$ on $\left(\mathbf{Z}_i,\ \hat{\gamma}_i,\ \hat{\sigma}_i^2\right)$. The validity of the *simple substitution* method depends crucially on the aforementioned negligibility of the estimation variability in the subject-specific estimated regression parameters and variance parameters.

In many applications the subject-specific information is not large enough to justify the use of the simple substitution method. Rather the estimation variability in $\hat{\boldsymbol{\gamma}}_i$ and $\hat{\sigma}_i^2$ creates a measurement error model problem:

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_i &= \boldsymbol{\gamma}_i + \boldsymbol{\Upsilon}_{\gamma,i}, \\ \hat{\sigma}_i^2 &= \sigma_i^2 + \boldsymbol{\Upsilon}_{\sigma,i}, \end{aligned} \tag{7}$$

where

$$\begin{pmatrix} \boldsymbol{\Upsilon}_{\gamma,i} \\ \boldsymbol{\Upsilon}_{\sigma,i} \end{pmatrix} \sim \left\{ \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{V}_{\gamma,i} & \mathbf{0} \\ \mathbf{0} & V_{\sigma,i} \end{pmatrix} \right\}, \tag{8}$$

with $\mathbf{V}_{\gamma,i} = \sigma_i^2 \left(\mathbf{D}_i^T \mathbf{D}_i\right)^{-1}$ and $V_{\sigma,i} = 2\sigma_i^4/\nu_i$ for $i = 1, \ldots, n$. This measurement error model is nonstandard in the sense that the measurement errors are heteroscedastic (with possibly large differences in variation), and the usual normal error model is clearly not appropriate for the errors in the variance estimates. The most natural parametric model for the variance estimators is a scaled chi-squared, consistent with the usual linear model assumptions.

## 3.2 Longitudinal Variance with Baseline Interactions Model

We focus on an analysis of a specialized submodel of (2) that includes only the subject-specific variances and baseline covariates. Because the model does not include the longitudinal regression parameters $\boldsymbol{\gamma}_i$, it is limited to assessing the marginal effects of variance $\sigma_i^2$ on the response, albeit with appropriate adjustment for baseline covariates *and* for the estimation error variability in $\hat{\sigma}_i^2$. A further restriction is to the case where $T(\sigma^2) = 1/\sigma^2$ so that the resulting model has logits that are linear in the subject-specific *precisions* $\sigma_i^{-2}$:

$$P(Y_i = 1 | \ \mathbf{Z}_i, \sigma_i^2; \boldsymbol{\beta}) = H\left(\beta_0 + \mathbf{Z}_i^T \boldsymbol{\beta}_Z + \beta_\sigma (1/\sigma_i^2) + \mathbf{Z}_{1i}^T \boldsymbol{\beta}_{Z_2\sigma} (1/\sigma_i^2)\right), \tag{9}$$

where for this submodel $\boldsymbol{\beta} = (\beta_0,\ \boldsymbol{\beta}_Z^T,\ \beta_\sigma,\ \boldsymbol{\beta}_{Z_2\sigma})^T$.

We also assume that the observed variances $\sigma_1^2, \ldots, \sigma_n^2$ are independent, with $\nu_i \hat{\sigma}_i^2/\sigma_i^2 \sim \chi_{\nu_i}^2$. With this assumption, the logistic outcome model coupled with the $\chi^2$ error model possesses the exponential-family structure required to adapt the conditional-score estimating approach of Stefanski and Carroll (1987) (see also Carroll et al. 2006, Chap. 7, Wang et al. 2000; Li et al. 2004). The conditional score is widely used to obtain consistent estimators for the logistic regression

model parameters in the presence of measurement error. The sufficient statistics are identified by factorization of the observed data likelihood. This involves the measurement error distribution, which, in the present case, is not normally distributed but chi-square. However, the general approach laid out in Stefanski and Carroll (1987) applies. We present the extended version of the conditional score developed by Dagalp (2001) for models with interactions between the error-prone and error-free predictors.

In the exponential family form, the corresponding density of $Y$ for a given true predictor $\sigma^2$ is

$$f_Y(y;\sigma^2,\boldsymbol{\beta}) = \exp\left[y\eta + \ln\{1-H(\eta)\}\right], \tag{10}$$

where $\eta = \beta_0 + \mathbf{Z}^T\boldsymbol{\beta}_Z + \beta_\sigma(1/\sigma^2) + \mathbf{Z}_2^T\boldsymbol{\beta}_{Z_2\sigma}(1/\sigma^2)$. The observed variance estimates $\hat{\sigma}^2$ have conditional density

$$f_{\hat{\sigma}^2}(\hat{\sigma}^2;\sigma^2) = \frac{(\nu/\sigma^2)}{2^{\nu/2}\Gamma(\nu/2)}\left(\frac{\nu\hat{\sigma}^2}{\sigma^2}\right)^{\nu/2-1}\exp\left(\frac{-\nu\hat{\sigma}^2}{2\sigma^2}\right)I(\hat{\sigma}^2>0). \tag{11}$$

Under the surrogacy assumption that $Y$ and $\hat{\sigma}^2$ are conditionally independent given $\sigma^2$, the joint density of the observed pair $(Y,\hat{\sigma}^2)$ given $\sigma^2$ and $\boldsymbol{\beta}$ is the product of (21) and (11),

$$\begin{aligned} f_{Y,\hat{\sigma}^2}(y,\hat{\sigma}^2;\sigma^2,\boldsymbol{\beta}) = {} & \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\exp\left\{y\beta_0 + (\nu/2)\ln(\nu) + (\nu/2-1)\ln(\hat{\sigma}^2)\right\}\times \\ & \exp\left[\left(\frac{2y(\beta_\sigma + \mathbf{Z}_2^T\boldsymbol{\beta}_{Z_2\sigma})}{\nu} - \hat{\sigma}^2\right)\left(\frac{\nu}{2\sigma^2}\right)\right. \\ & \left. + \ln\left\{1 - F\left(\beta_0 + \frac{\beta_\sigma}{\sigma^2}\right)\right\} - \nu\ln(\sigma)\right]. \end{aligned} \tag{12}$$

If $\sigma^2$ is viewed as a parameter in density (12), and $\beta_\sigma$ is regarded as known, then the "statistic"

$$\Delta = \Delta(Y,\hat{\sigma}^2,\beta_\sigma) = \hat{\sigma}^2 - 2Y\left(\beta_\sigma + \mathbf{Z}_2^T\boldsymbol{\beta}_{Z_2\sigma}\right)/\nu$$

is complete and sufficient for $\sigma^2$. Of course, $\Delta$ is not statistic in the strict sense as it depends on unknown parameters. However, it is true that the conditional distribution of $Y|\Delta$ does not depend on the unknown $\sigma^2$, and thus it can be used to construct unbiased estimating equations for the unknown logistic regression parameters that are free of the unknown nuisance parameters $\sigma_i^2$ as described in Stefanski and Carroll (1987) and Carroll et al. (2006). Because $Y$ is binary in our case we need to only calculate $\Pr(Y=1 \mid \Delta=\delta)$. Straightforward manipulations reveal that

$$\Pr(Y=1|\Delta=\delta) = \frac{1}{1+M}, \tag{13}$$

where

$$M = \frac{\left\{1 - P(Y_i = 1 | \mathbf{Z}_i, \sigma_i^2; \boldsymbol{\beta})\right\} f_{\hat{\sigma}^2}(\delta; \sigma^2)}{P(Y_i = 1 | \mathbf{Z}_i, \sigma_i^2; \boldsymbol{\beta}) f_{\hat{\sigma}^2}\left(\delta + 2\left(\beta_\sigma + \mathbf{Z}_2^T \boldsymbol{\beta}_{Z_2\sigma}\right)/\nu; \sigma^2\right)}. \tag{14}$$

Substituting the expressions in (9) and (11) for the corresponding probabilities and densities in the expression above results in $\Pr(Y = 1 | \Delta = \delta) = Q(\delta, \boldsymbol{\beta})$, where

$$Q(\delta, \boldsymbol{\beta}) = \frac{I_1}{1 + I_2 \exp(-\mathbf{Z}^T \boldsymbol{\beta}_Z) R^{(\nu/2-1)}}, \tag{15}$$

with

$$\begin{aligned} I_1 &= I\left\{\delta + 2\left(\beta_\sigma + \mathbf{Z}_2^T \boldsymbol{\beta}_{Z_2\sigma}\right)/\nu > 0\right\}, \\ I_2 &= I\{\delta > 0\}, \\ R &= \frac{\delta}{\delta + 2\left(\beta_\sigma + \mathbf{Z}_2^T \boldsymbol{\beta}_{Z_2\sigma}\right)/\nu}. \end{aligned} \tag{16}$$

It follows that the score function defined as

$$\psi_{\text{Cond}}(Y, \mathbf{Z}, \hat{\sigma}^2, \nu, \boldsymbol{\beta}) = \left[ \{Y - Q(\delta, \boldsymbol{\beta})\} \begin{pmatrix} 1 \\ \mathbf{Z} \\ 1/\delta \\ \mathbf{Z}_1/\delta \end{pmatrix} \right]_{\delta = \hat{\sigma}^2 - \frac{2Y}{\nu}\left(\beta_\sigma + \mathbf{Z}_2^T \boldsymbol{\beta}_{Z_2\sigma}\right)}, \tag{17}$$

is unbiased in the sense that

$$E\left\{\psi_{\text{Cond}}(Y, \mathbf{Z}, \hat{\sigma}^2, \nu, \boldsymbol{\beta})\right\} = E\left[E\left\{\psi_{\text{Cond}}(Y, \mathbf{Z}, \hat{\sigma}^2, \nu, \boldsymbol{\beta}) | \Delta\right\}\right] = 0.$$

We define the *conditional score* estimator, $\hat{\boldsymbol{\beta}}_{\text{Cond}}$, as the solution to the conditional score equations

$$\sum_{i=1}^{n} \psi_{\text{Cond}}(Y_i, \mathbf{Z}_i, \hat{\sigma}_i^2, \nu_i, \boldsymbol{\beta}) = \mathbf{0}. \tag{18}$$

The conditional score function is conditionally unbiased for all $\nu = 1, 2, \ldots$; however, it is not smooth for small $\nu$. This is best exemplified in the case $\nu = 2$, the conditional probability $Q(\delta, \boldsymbol{\beta})$ defined in (15) reduces to

$$Q(\Delta, \boldsymbol{\beta}) = \frac{I\left\{\hat{\sigma}^2 > (Y - 1)\left(\beta_\sigma + \mathbf{Z}_2^T \boldsymbol{\beta}_{Z_2\sigma}\right)\right\}}{1 + I\left\{\hat{\sigma}^2 > Y\left(\beta_\sigma + \mathbf{Z}_2^T \boldsymbol{\beta}_{Z_2\sigma}\right)\right\} \exp(-\mathbf{Z}^T \boldsymbol{\beta}_Z)}, \tag{19}$$

from which it is apparent that $Q(\Delta, \boldsymbol{\beta})$ is not a continuous function. At first blush, it is not readily evident that in this case that $E\{Y - Q(\Delta, \boldsymbol{\beta})\} = 0$. However, this follows in just a few steps upon invoking the fact the $\chi^2$ distribution with two degrees of freedom is exponential. The more important point for application is that the estimating equations are not easily solved when all of the $\nu_i$ are small, because the estimating equations can be discontinuous. In fact, there may not be an exact solution to the conditional estimating equations (18) because of discontinuities. In such cases it is necessary to redefine $\hat{\boldsymbol{\beta}}_{\text{Cond}}$ as

$$\hat{\boldsymbol{\beta}}_{\text{Cond}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\| \sum_{i=1}^{n} \boldsymbol{\psi}_{\text{Cond}}(Y_i, \mathbf{Z}_i, \hat{\sigma}_i^2, \nu_i, \boldsymbol{\beta}) \right\|^2. \tag{20}$$

The applications motivating our work generally have $\nu$ large enough that the nonsmoothness for small $\nu$ is not a problem. More to the point, it is likely the case that if most estimated variances are based on just two or three or four measurements, then no method of analysis is likely to perform well in sample sizes typically encountered in biomedical studies. If such applications are encountered a possible beneficial modification to the conditional score estimator would be to smooth the indicator functions appearing in the expression for $Q(\delta, \boldsymbol{\beta})$ in (15).

Because $\hat{\boldsymbol{\beta}}_{\text{Cond}}$ is an m-estimator, large-sample approximate variances and standard errors can be derived using the either (conditional) model-based variance formulas or the empirical sandwich formula variance estimator, see, for example, Stefanski and Boos (2002). In the Monte Carlo study reported in the next section sandwich formula variance estimates were studied and found to be adequate for sample sizes common in large studies.

### *3.3 Attenuation-corrected Calibration/Moment Matching*

In linear regression measurement error models much of the attenuation in regression coefficients is due to the increased variability of the measured-with-error predictor, which in our cases are the subject-specific estimated variances. A relative simple and often versatile method is obtained by simply "shrinking" the observed variances to obtain $\hat{\sigma}^2_{\text{s},i}$ in such a way that the $\hat{\sigma}^2_{\text{s},i}$ have sample mean and sample variance converging to the mean and variance of the so-called true $\sigma_i^2$. That is

$$\overline{\hat{\sigma}}^2_{\text{s},\cdot} = \frac{1}{n-1} \sum_{i=1}^{n} \hat{\sigma}^2_{\text{s},i} \xrightarrow{P} E(\sigma^2),$$

and

$$\frac{1}{n-1} \sum_{i=1}^{n} \left( \hat{\sigma}^2_{\text{s},i} - \overline{\hat{\sigma}}^2_{\text{s},\cdot} \right)^2 \xrightarrow{P} \text{Var}\left(\sigma^2\right).$$

For observed variances $\sigma_1^2, \ldots, \sigma_n^2$ that are independent, with $\nu_i \hat{\sigma}_i^2 / \sigma_i^2 \sim \chi^2_{\nu_i}$, the moment adjusted estimates are defined as

$$\sigma^2_{\mathrm{s},i} = \hat{\alpha}\hat{\sigma}_i^2 + (1-\hat{\alpha})\overline{\hat{\sigma}^2_{\cdot}},$$

where

$$\hat{\alpha} = \sqrt{\frac{S^2_{\hat{\sigma}^2} - 2\widetilde{\hat{\sigma}^4_{\cdot}}}{S^2_{\hat{\sigma}^2}}}$$

with

$$\overline{\hat{\sigma}^2_{\cdot}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}_i^2,$$

$$S^2_{\hat{\sigma}^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\sigma}_i^2 - \overline{\hat{\sigma}^2_{\cdot}}\right)^2,$$

and

$$\widetilde{\hat{\sigma}^4_{\cdot}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}_i^4/(\nu_i+2).$$

Note the average adjusted estimate satisfies

$$\overline{\sigma^2_{\mathrm{s},\cdot}} = \frac{1}{n}\sum_{i=1}^{n} = \overline{\hat{\sigma}^2_{\cdot}} \xrightarrow{P} E(\sigma^2).$$

Now consider the sample variance of the adjusted estimates

$$S^2_{\hat{\sigma}^2_S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\sigma}^2_{\mathrm{s},i} - \hat{\sigma}^2_{\mathrm{s},\cdot}\right)^2 = \hat{\alpha}^2 S^2_{\hat{\sigma}^2} = S^2_{\hat{\sigma}^2} - 2\widetilde{\hat{\sigma}^4_{\cdot}}$$

Using the variance decomposition identity

$$\mathrm{Var}\left(\hat{\sigma}^2\right) = E\left(\mathrm{Var}(\hat{\sigma}^2|\sigma^2)\right) + \mathrm{Var}\left(E(\hat{\sigma}^2|\sigma^2)\right)$$

and moments from the $\chi^2$ distribution it can be shown that

$$S^2_{\hat{\sigma}^2} - 2\widetilde{\hat{\sigma}^4_{\cdot}} \xrightarrow{P} \mathrm{Var}(\sigma^2).$$

Thus asymptotically the adjusted estimates $\hat{\sigma}^2_{\mathrm{s},i}$ have the same first- and second-moments of the true subject-specific variances. If employed in a simple linear measurement error regression model with $\sigma_i^2$ playing the role of the true predictor,

the resulting estimator would be a heteroscedastic measurement error variance version of the correction-for-attenuation estimator.

Once the adjusted estimates are calculated one proceeds acting as though they are equal to the true subject-specific variance. That is, one fits the logistic regression model (9) to the (imputed) data $\left\{\left(\mathbf{Z}_i,\ \hat{\sigma}^2_{S,i},\ Y_i\right)\right\}_1^n$. At the expense of detracting from the simplicity of the plug-in nature of data imputation, variance estimates, and standard errors from the usual logistic regression model fit can be replaced by estimates that properly account for the variability in the imputation step. In the simulation study in Sect. 4 the usual standard errors were calculated and proved to adequately reflect variability. The method described above does not account for covariates $\mathbf{Z}$, a deficiency that is addressed in the next section.

### *3.4 Moment Adjusted Imputation*

A refinement of the approach in the previous section calculates adjusted measured values that correct not only for the moments of the observed variances but also for their covariances with other covariates ($\mathbf{Z}$) in the logistic model.

Consider adjusted variance estimators of the form

$$\hat{\sigma}^2_{i,\text{MAI}} = \gamma_1 + \gamma_\sigma \hat{\sigma}^2_i + \boldsymbol{\gamma}_Z^T \mathbf{Z}_i.$$

The parameter $\boldsymbol{\gamma} = (\gamma_1, \gamma_\sigma, \boldsymbol{\gamma}_Z^T)^T$ is chosen so that the following moment equalities are satisfied:

$$\overline{\sigma^2_{\text{MAI},\cdot}} = \frac{1}{n}\sum_{i=1}^n \hat{\sigma}^2_{i,\text{MAI}} = \overline{\hat{\sigma}^2_\cdot},$$

$$S^2_{\hat{\sigma}^2_{\text{MAI}}} = \frac{1}{n-1}\sum_{i=1}^n \left(\hat{\sigma}^2_{i,\text{MAI}} - \overline{\sigma^2_{\text{MAI},\cdot}}\right)^2 = S^2_{\hat{\sigma}^2_S},$$

and

$$\mathbf{S}_{Z\hat{\sigma}^2_{\text{MAI}}} = \frac{1}{n-1}\sum_{i=1}^n \hat{\sigma}^2_{i,\text{MAI}}\left(\mathbf{Z}_i - \overline{\mathbf{Z}}_\cdot\right) = \frac{1}{n-1}\sum_{i=1}^n \hat{\sigma}^2_i\left(\mathbf{Z}_i - \overline{\mathbf{Z}}_\cdot\right).$$

So constructed, the adjusted subject-specific variances $\hat{\sigma}^2_{i,\text{MAI}}$ asymptotically have the same first-order moment properties (including covariances) as the true subject-specific variances $\sigma^2_i$. Adjusted estimators of this form have a long history in statistics, starting with Tukey's so-called *named and faceless values* Tukey (1974). They appear in the Bayesian literature as modified empirical Bayes estimators Louis (1984) and they have been studied (and extended) for measurement error model applications by Bay (1997), Freedman et al. (2004), and Elliott (2009). Additional variations on this theme are possible including use of an estimated best

linear predictor of $\sigma_i^2$ derived from estimating the mean and covariance matrix of $\left(\sigma_i^2,\ \hat{\sigma}_i^2,\ \mathbf{Z}_i\right),\ i=1,\ldots,n$.

As with the simple moment adjusted estimators of the previous subsection, once the adjusted estimates are calculated one proceeds acting as though they are equal to the true subject-specific variance. In the simulation study in Sect. 4 the usual standard errors were calculated and proved to adequately reflect variability.

## 4 Simulation Results

We now report on a simulation study designed to investigate the performance of the conditional score estimator for models with and without interactions. Data for the simulation study were generated as follows.

1. Baseline covariate: $Z_{1,i}$ and $Z_{2,i}$ are independent with $Z_{1,i} \sim \mathrm{N}(0,1)$ and $Z_{2,i} \sim \mathrm{Bin}(1,.5)$.
2. True subject-specific variances: $\sigma_i^2 \sim .25 + .25Z_{2,i} + \zeta_i^2$ where the $\zeta_i$ are $\mathrm{N}(0,1)$;
3. Response: $Y_i \sim \mathrm{Bin}(1,p_i)$ with $p_i$ given by the logistic model (9) with $\beta_{Z_1} = 1$, $\beta_{Z_2} = -1$, $\beta_\sigma = 1$, and for cases with an interaction term (Tables 1 and 2), $\beta_{\sigma Z_2} = .5$. The intercept was adjusted to achieve marginal mean responses of $E(Y) = .5$ (Tables 1 and 3) and $E(Y) = .25$ (Tables 2 and 4).
4. Estimated variances: $\hat{\sigma}_i^2 \,|\, \sigma_i^2 \overset{D}{=} \sigma_i^2 \chi^2_{\nu_i}/\nu_i$ where the degrees of freedom $\nu_i$ were generated independently as $\nu_i \overset{D}{=} 9 + \mathrm{Round}(20U_i)$ with $U_i \sim \mathrm{Unif}(0,1)$.

Sample sizes of $n = 1,000$ and 5,000 were studied. The number of simulated data sets was 10,000 in all cases. Results for the case $n = 1,000$ are reported in Tables 1–4. The results for $n = 5,000$ are qualitatively similar and are not reported here.

All three measurement error estimation methods reduce bias substantially relative to the naive estimator. However, only the conditional score estimator reduces bias to levels comparable to biases in the true-data estimator. The statistically significant biases revealed by the large Monte Carlo sample size (10,000 replications) of the conditional score estimator are also manifest in the true-data estimator. These biases are attenuated much more so in the $n = 5,000$ results not shown here. Thus the biases in the conditional-score and true-data estimators are consistent with the conclusion that they are finite-sample/nonlinear estimator biases, rather than biases inherent with the method of analysis.

The table entries labeled AvgSD/MCSD are ratios of Monte Carlo mean standard deviations calculated from m-estimator, sandwich-formula variance estimators to the Monte Carlo standard deviation of the estimators. None of the ratios are sufficiently less than one to be troublesome. Thus the m-estimator, sandwich-formula variance estimators yield reasonable unbiased standard errors.

Confidence intervals were calculated using the usual large-sample

$$\text{estimate} \pm 1.96 \times (\text{standard error})$$

**Table 1** Est, estimator; Int, intercept; $\mathbf{Z}_1$, coefficient for $\mathbf{Z}_1$; $\mathbf{Z}_2$, coefficient for $\mathbf{Z}_2$; $1/\sigma^2$, coefficient for $1/\sigma^2$; $H_0$ : Bias $= 0$, $t$-statistic; AvgSE/MCSD, average standard error from sandwich variance estimate divided by the Monte Carlo Standard deviation; 95% CI CP and Length; coverage probability and length of nominal large-sample 95% confidence interval

Case 1: Marginal mean $E(Y) = .5$; No-interaction model; $n = 1,000$.

| | Est | Int | $\mathbf{Z}_1$ | $\mathbf{Z}_2$ | $1/\sigma^2$ |
|---|---|---|---|---|---|
| Bias | | | | | |
| | True | 0.00 | 0.01 | −0.01 | 0.01 |
| | Naive | 0.44 | −0.04 | −0.12 | −0.36 |
| | Cond | −0.01 | 0.01 | −0.01 | 0.01 |
| | ACCal | 0.17 | −0.03 | −0.15 | −0.01 |
| | MAI | 0.15 | −0.03 | −0.10 | −0.01 |
| AvgSE/MCSD | | | | | |
| | True | 0.99 | 0.99 | 0.98 | 1.00 |
| | Naive | 0.98 | 0.99 | 0.98 | 0.98 |
| | Cond | 0.99 | 0.99 | 0.98 | 0.98 |
| | ACCal | 0.98 | 0.99 | 0.98 | 0.97 |
| | MAI | 0.98 | 0.98 | 0.98 | 0.97 |
| 95% CI CP | | | | | |
| | True | 0.95 | 0.95 | 0.95 | 0.95 |
| | Naive | 0.24 | 0.92 | 0.88 | 0.02 |
| | Cond | 0.95 | 0.95 | 0.95 | 0.95 |
| | ACCal | 0.83 | 0.93 | 0.84 | 0.94 |
| | MAI | 0.86 | 0.93 | 0.89 | 0.94 |
| 95% CI Length | | | | | |
| | True | 0.67 | 0.36 | 0.62 | 0.38 |
| | Naive | 0.64 | 0.34 | 0.60 | 0.33 |
| | Cond | 0.79 | 0.37 | 0.63 | 0.51 |
| | ACCal | 0.68 | 0.35 | 0.61 | 0.42 |
| | MAI | 0.68 | 0.35 | 0.61 | 0.42 |

formula. Coverage probabilities for the true-data estimator and the conditional-score estimator are acceptably close to the nominal 95%. The lower coverages of the other estimators (Naive, ACCal, and MAI) are due largely to their biases rather than underestimation of variability. The near-nominal coverage probabilities of the conditional-score estimator come at the price of increased confidence-interval length of between 25% and 70%.

Finally we note that the computational algorithm for the conditional-score estimator was very reliable. An internal check on (fast) convergence of the algorithm was made for each data set. If that test failed, then that particular data set was bypassed in the simulation study. Failure occurred in only six of 80,006 data sets generated to obtain the results for both sample sizes ($n = 1,000$, $n = 5,000$). All six

**Table 2** Est, estimator; Int, intercept; $\mathbf{Z}_1$, coefficient for $\mathbf{Z}_1$; $\mathbf{Z}_2$, coefficient for $\mathbf{Z}_2$; $1/\sigma^2$, coefficient for $1/\sigma^2$; $H_0$ : Bias = 0, $t$-statistic; AvgSE/MCSD, average standard error from sandwich variance estimate divided by the Monte Carlo Standard deviation; 95% CI CP and Length; coverage probability and length of nominal large-sample 95% confidence interval

Case 2: Marginal mean $E(Y) = .25$; No-interaction model; $n = 1,000$.

| | Est | Int | $\mathbf{Z}_1$ | $\mathbf{Z}_2$ | $1/\sigma^2$ |
|---|---|---|---|---|---|
| Bias | | | | | |
| | True | −0.02 | 0.01 | −0.01 | 0.01 |
| | Naive | 0.68 | −0.06 | −0.26 | −0.42 |
| | Cond | −0.03 | 0.01 | −0.01 | 0.02 |
| | ACCal | 0.25 | −0.05 | −0.28 | 0.02 |
| | MAI | 0.25 | −0.05 | −0.22 | 0.00 |
| AvgSE/MCSD | | | | | |
| | True | 1.00 | 1.00 | 1.01 | 0.99 |
| | Naive | 0.98 | 1.00 | 1.00 | 0.96 |
| | Cond | 0.98 | 0.99 | 1.01 | 0.97 |
| | ACCal | 0.98 | 1.00 | 1.01 | 0.95 |
| | MAI | 0.98 | 0.99 | 1.00 | 0.94 |
| 95% CI CP | | | | | |
| | True | 0.95 | 0.95 | 0.95 | 0.95 |
| | Naive | 0.06 | 0.89 | 0.73 | 0.00 |
| | Cond | 0.95 | 0.95 | 0.95 | 0.94 |
| | ACCal | 0.74 | 0.91 | 0.70 | 0.94 |
| | MAI | 0.75 | 0.92 | 0.80 | 0.93 |
| 95% CI Length | | | | | |
| | True | 0.85 | 0.41 | 0.76 | 0.37 |
| | Naive | 0.72 | 0.39 | 0.73 | 0.30 |
| | Cond | 1.11 | 0.44 | 0.81 | 0.53 |
| | ACCal | 0.80 | 0.39 | 0.73 | 0.42 |
| | MAI | 0.81 | 0.39 | 0.73 | 0.42 |

non-rapid convergent cases were for the case of $n = 1,000$ and the marginal mean $E(Y) = 0.25$. In all cases the estimation algorithms used the naive estimate as a starting value (other starting values were not tried for the six exceptional cases).

## 5 Extensions and Limitations

A not-so-appealing feature of the conditional score method is the manner in which it extends to models that include estimated longitudinal process mean parameters in addition to variance parameters. The conditional score approach relies on obtaining sufficient statistics for mismeasured covariates. For the model studied in this

**Table 3** Est, estimator; Int, intercept; $\mathbf{Z}_1$, coefficient for $\mathbf{Z}_1$; $\mathbf{Z}_2$, coefficient for $\mathbf{Z}_2$; $1/\sigma^2$, coefficient for $1/\sigma^2$; $\mathbf{Z}_2/\sigma^2$, coefficient for interaction term; $H_0$ : Bias = 0, $t$-statistic; AvgSE/MCSD, average standard error from sandwich variance estimate divided by the Monte Carlo Standard deviation; 95% CI CP and Length; coverage probability and length of nominal large-sample 95% confidence interval

Case 3: Marginal mean $E(Y) = .5$; Interaction model; $n = 1,000$.

| | Est | Int | $\mathbf{Z}_1$ | $\mathbf{Z}_2$ | $1/\sigma^2$ | $\mathbf{Z}_2/\sigma^2$ |
|---|---|---|---|---|---|---|
| Bias | | | | | | |
| | True | −0.01 | 0.01 | 0.00 | 0.01 | 0.00 |
| | Naive | 0.35 | −0.06 | 0.37 | −0.29 | −0.43 |
| | Cond | −0.02 | 0.01 | −0.01 | 0.02 | 0.01 |
| | ACCal | −0.01 | −0.04 | 0.44 | 0.14 | −0.50 |
| | MAI | 0.01 | −0.04 | 0.39 | 0.10 | −0.41 |
| AvgSE/MCSD | | | | | | |
| | True | 1.01 | 0.99 | 1.00 | 1.00 | 0.99 |
| | Naive | 0.98 | 0.99 | 0.98 | 0.97 | 0.97 |
| | Cond | 0.99 | 0.99 | 0.99 | 0.97 | 0.98 |
| | ACCal | 0.99 | 0.99 | 0.99 | 0.97 | 0.99 |
| | MAI | 0.99 | 0.99 | 0.99 | 0.97 | 0.97 |
| 95% CI CP | | | | | | |
| | True | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 |
| | Naive | 0.54 | 0.88 | 0.72 | 0.25 | 0.35 |
| | Cond | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| | ACCal | 0.95 | 0.91 | 0.69 | 0.84 | 0.40 |
| | MAI | 0.95 | 0.91 | 0.74 | 0.89 | 0.55 |
| 95% CI Length | | | | | | |
| | True | 0.74 | 0.36 | 1.20 | 0.42 | 0.84 |
| | Naive | 0.75 | 0.34 | 1.09 | 0.42 | 0.72 |
| | Cond | 0.87 | 0.38 | 1.55 | 0.54 | 1.19 |
| | ACCal | 0.81 | 0.34 | 1.19 | 0.54 | 0.89 |
| | MAI | 0.80 | 0.34 | 1.19 | 0.53 | 0.90 |

paper this means modeling in terms of precision rather than variance, an arguable reasonable strategy. However, we now show that for models that include estimated longitudinal process mean parameters in addition to variance parameters, things are not so palatable.

We consider logistic models, but for simplicity we assume that there are no covariates $\mathbf{Z}$ (and thus no interaction term either). Assume that $\Pr(Y = 1|\sigma^2; \boldsymbol{\beta}) = F\{\beta_0 + \beta_{\gamma_0}(\gamma_0/\sigma^2) + \beta_\sigma(1/\sigma^2)\}$ for $\boldsymbol{\beta} = (\beta_0, \beta_{\gamma_0}, \beta_\sigma)$. The corresponding density of $Y$ is

$$f_Y(y; \gamma_0, \sigma^2, \boldsymbol{\beta}) =$$
$$\exp\left[y\left(\beta_0 + \beta_{\gamma_0}\frac{\gamma_0}{\sigma^2} + \beta_\sigma\frac{1}{\sigma^2}\right) + \log\left\{1 - F\left(\beta_0 + \beta_{\gamma_0}\frac{\gamma_0}{\sigma^2} + \beta_\sigma\frac{1}{\sigma^2}\right)\right\}\right], \quad (21)$$

**Table 4** Est, estimator; Int, intercept; $\mathbf{Z}_1$, coefficient for $\mathbf{Z}_1$; $\mathbf{Z}_2$, coefficient for $\mathbf{Z}_2$; $1/\sigma^2$, coefficient for $1/\sigma^2$; $\mathbf{Z}_2/\sigma^2$, coefficient for interaction term; $H_0$ : Bias = 0, $t$-statistic; AvgSE/MCSD, average standard error from sandwich variance estimate divided by the Monte Carlo Standard deviation; 95% CI CP and Length; coverage probability and length of nominal large-sample 95% confidence interval

Case 4: Marginal mean $E(Y) = .25$; Interaction model; $n = 1,000$.

| | Est | Int | $\mathbf{Z}_1$ | $\mathbf{Z}_2$ | $1/\sigma^2$ | $\mathbf{Z}_2/\sigma^2$ |
|---|---|---|---|---|---|---|
| Bias | | | | | | |
| | True | −0.02 | 0.01 | −0.02 | 0.01 | 0.01 |
| | Naive | 0.66 | −0.07 | 0.36 | −0.39 | −0.48 |
| | Cond | −0.04 | 0.02 | −0.06 | 0.02 | 0.04 |
| | ACCal | 0.12 | −0.05 | 0.53 | 0.11 | −0.60 |
| | MAI | 0.16 | −0.05 | 0.47 | 0.06 | −0.51 |
| AvgSE/MCSD | | | | | | |
| | True | 0.99 | 0.98 | 0.98 | 0.99 | 0.98 |
| | Naive | 0.97 | 0.98 | 0.97 | 0.97 | 0.96 |
| | Cond | 0.98 | 0.98 | 0.97 | 0.97 | 0.97 |
| | ACCal | 0.97 | 0.98 | 0.98 | 0.95 | 0.99 |
| | MAI | 0.97 | 0.97 | 0.98 | 0.94 | 0.97 |
| 95% CI CP | | | | | | |
| | True | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 |
| | Naive | 0.15 | 0.86 | 0.77 | 0.02 | 0.21 |
| | Cond | 0.95 | 0.95 | 0.94 | 0.94 | 0.93 |
| | ACCal | 0.90 | 0.90 | 0.67 | 0.87 | 0.27 |
| | MAI | 0.87 | 0.90 | 0.73 | 0.92 | 0.42 |
| 95% CI Length | | | | | | |
| | True | 0.95 | 0.40 | 1.69 | 0.40 | 1.04 |
| | Naive | 0.83 | 0.38 | 1.22 | 0.36 | 0.67 |
| | Cond | 1.22 | 0.45 | 2.52 | 0.56 | 1.71 |
| | ACCal | 0.96 | 0.38 | 1.42 | 0.51 | 0.91 |
| | MAI | 0.95 | 0.38 | 1.42 | 0.49 | 0.93 |

where $\gamma_0$ and $\sigma^2$ are regarded as unknown constants. The observed variance estimators $\hat{\sigma}^2$ have density

$$f_{\hat{\sigma}^2}(\hat{\sigma}^2;\sigma^2) = \frac{\left(\frac{\nu}{\sigma^2}\right)}{2^{\nu/2}\Gamma(\nu/2)}\left(\frac{\nu\hat{\sigma}^2}{\sigma^2}\right)^{\nu/2-1}\exp\left(\frac{-\nu\hat{\sigma}^2}{2\sigma^2}\right)I(\hat{\sigma}^2 > 0).$$

The estimator $\hat{\gamma}_0$ has density

$$f_{\hat{\gamma}_0}(\hat{\gamma}_0;\gamma_0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 c}}\exp\left\{-\frac{1}{2\sigma^2 c}(\hat{\gamma}_0-\gamma_0)^2\right\}.$$

and thus

$$f_{Y,\hat{\sigma}^2,\hat{\gamma}_0}(y,\hat{\sigma}^2,\hat{\gamma}_0;\gamma_0,\sigma^2,\boldsymbol{\beta}) = g(Y,\hat{\sigma}^2,\hat{\gamma}_0)h(\sigma^2,\gamma_0)\times$$
$$\exp\left[\frac{-1}{2\sigma^2}\left\{(\gamma_0 - y\beta_{\gamma_0} - \hat{\gamma}_0)^2 - (y\beta_{\gamma_0} + \hat{\gamma}_0)^2 + \hat{\gamma}_0^2 + \nu\hat{\sigma}^2 - 2y\beta_\sigma\right\}\right].$$

By the Factorization Theorem, the two-dimensional statistic $\{Y\beta_{\gamma_0} + \hat{\gamma}_0, -(Y\beta_{\gamma_0} + \hat{\gamma}_0)^2 + \hat{\gamma}_0^2 + \nu\hat{\sigma}^2 - 2Y\beta_\sigma\}$ is sufficient for $(\gamma_0, \sigma^2)$.

This factorization relies critically on the specific form of the logistic model

$$\Pr(Y = 1|\,\sigma^2;\boldsymbol{\beta}) = F\{\beta_0 + \beta_{\gamma_0}\left(\gamma_0/\sigma^2\right) + \beta_\sigma(1/\sigma^2)\}.$$

Otherwise, we are unable to factor the joint density of observed data. When $\gamma_0$ is not in the model, i.e., $\beta_{\gamma_0} = 0$, the resulting model in terms of the precision $1/\sigma^2$ is defensibly reasonable. However, with $\gamma_0$ in the model the conditional score approach requires modeling in terms of the precision $1/\sigma^2$ and the precision-scaled mean $\gamma_0/\sigma^2$. Whether such transformations are useful for applied work will require experience with many data sets for which subject-specific means and variances are used as predictors.

For a logistic model that includes both a longitudinal intercept and slope parameter $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^T$ as covariates, a similar factorization reveals that the requisite logistic model has the form

$$\Pr(Y = 1|\,\boldsymbol{\gamma}, \sigma^2;\boldsymbol{\beta}) = F\left(\beta_0 + (1/\sigma^2)\boldsymbol{\gamma}^T\mathbf{D}^T\mathbf{D}\boldsymbol{\beta}_\gamma + \beta_\sigma(1/\sigma^2)\right).$$

In this case, the model involves linear combinations of $\gamma_0/\sigma^2$ and $\gamma_1/\sigma^2$, and also the design matrix $\mathbf{D}$, which renders the model untenable. In future work we will study moment-adjusted imputation methods for the full logistic model 2, as well as hybrid methods that use moment-adjustment to account for measurement error variability in some parameters and conditioning on sufficient statistics to account for errors in the remaining parameters.

## 6 Summary

The conditional score method is an attractive approach that performed well for the case studied in this paper, namely the model with the only error-prone predictor being the longitudinal process variance parameter. Depending on one's point of view, modeling in terms of precision rather than variance may be positive or negative. In applications, this matter has to be considered in light of the available data—a heavily left-skewed distribution of variance estimates can produce a heavily right-skewed distribution of precisions. The precision-based model (9) should not be adopted solely for its tractability and good performance with simulated data sets. Ultimately

the decision to transform variances will be data dependent and thus general-purpose methods such as moment adjusted imputation are attractive because in principle they should work reasonably well for different model/transformations and thus are useful for facilitating such modeling decisions. However, the possibility (and likelihood) of transformation raises the possibility of doing moment-adjustment *on the transformed scale*. For example, for the precision-based model studied in this paper, moment-based adjustments could be made for the estimated precisions using inverse $\chi^2$ moments. Moment adjustment could also be done on the standard deviation scale as well. Such transformation-specific moment adjustment negates some of the generality-grounded appeal of moment adjustment. Regardless, it may prove useful in application. Finally, we showed in Sect. 5 that extensions of the conditional score method to models that include estimated longitudinal process mean parameters in addition to variance parameters is possible only for models of questionable practical use. A strategy for dealing with models that include estimated longitudinal process mean parameters in addition to variance parameters that is worthy of future study is the approximate best linear predictor approach described at the end of Sect. 3.4 wherein an estimated best linear predictor of $\sigma_i^2$ derived from estimating the mean and covariance matrix of $\left(\sigma_i^2,\ \hat{\sigma}^2{}_i,\ \mathbf{Z}_i\right),\ i = 1, \ldots, n$.

**Acknowledgments** The authors are grateful for the opportunity to present their work at the *2012 International Symposium in Statistics (ISS) on Longitudinal Data Analysis Subject to Outliers, Measurement Errors, and/or Missing Values, St. John's, NL CA*. Special thanks go to Professor Brajendra Sutradhar for organizing the conference, to members of the symposium audience for insightful discussion of our presentation, and to two anonymous referees for their thoughtful comments on our manuscript. The authors' research was supported by NIH grants R01CA085848 and P01CA142538 and NSF grant DMS 0906421.

## References

Bay, J.: Adjuting data for measurement error. Unpublished dissertation thesis, North Carolina State University (1997)

Brunelli, S.M., Thadhani, R.I., Lynch, K.E., Ankers E.D., Joe, M.M., Boston, R., Chang, Y., Feldman, H.I.: Association between long-term blood pressure variability and mortality among incident hemodialysis patients. Am. J. Kidney Dis. **52**, 716–726 (2008)

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edn. Chapman and Hall, Boca Raton (2006)

Carroll, R.J., Stefanski, L.A.: Approximate quasi-likelihood estimation in models with surrogate predictors. J. Am. Stat. Assoc. **85**, 652–663 (1990)

Dagalp R.: Estimators For generalized linear measurement error models with interaction terms. Unpublished dissertation thesis, North Carolina State University (2001)

Elliott, L.: Adjusting for measurement error. Unpublished dissertation thesis, North Carolina State University (2009)

Freedman, L.S., Fainberg, V., Kipnis, V., Midthune, D., Carroll, R.J.: A new method for dealing with measurement error in explanatory variables of regression models. Biometrics **60**, 172–181 (2004)

Grove, J.S., Reed, D.M., Tano, K., Hwang, L.J.: Variability in systolic blood pressure—a risk factor for coronary heart disease? Am. J. Epidemiol. **145**, 771–776 (1997)

Havlik, R.J., Foley, D.J., Sayer, B., Masaki, K., White, L., Launer, L.J.: Variability in midlife systolic blood pressure is related to late-life brain white matter lesions. Stroke **33**, 26–30 (2002)

Iribarren, C.I., Sharp, D.S., Burchel, C.M., Petrovitch, H.P.: Association of weight loss and weight fluctuation with mortality among Japanese American men. New Engl. J. Med. **333**, 686–692 (1995)

Li, E., Zhang, D., Davidian, M.: Conditional estimation for generalized linear models when covariate are subject-specific parameters in a mixed model for longitudinal parameters. Biometrics **60**, 1–7 (2004)

Liu, K., Stammler, J., Dyer, A., Mc Keever, J., McKeever, P.: Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. J. Chron. Dis. **31**, 399–418 (1978)

Louis, T.A.: Estimating a population of parameter values using Bayes and empirical Bayes methods. J. Am. Stat. Assoc. **79**, 393–398 (1984)

Lyles, R.H., Munoz, A., Xu, J., Taylor, J.M.G., Chmiel, J.S.: Adjusting for measurement error to asses health effects of variability in biomarkers. Stat. Med. **18**, 1069–1086 (1999)

Yang, W., Israni, R.K., Brunelli, S.M., Joe, M.M., Fishbance, S., Feldman, H.I.: Hemoglobin variability and mortality in ESRD. J. Am. Soc. Nephrol. **18**, 3164–3170 (2007)

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

Stefanski, L.A., Boos, D.D.: The calculus of M-estimation. Am. Statist. **56**, 29–38 (2002)

Stefanski, L.A., Carroll, R.J.: Conditional scores and optimal scores for generalized linear measurement error models. Biometrika **74**, 703–716 (1987)

Tukey, J.W.: Named and faceless values: an initial exploration in memory of Prasanta C. Mahalanobis. Sankhya A. **36**, 125–176 (1974)

Wang, C.Y., Wang, N., Wang, S.: Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. Biometrics **56**, 487–495 (2000)

# Measurement Error in Dynamic Models

**John P. Buonaccorsi**

**Abstract** Many time series contain measurement (often sampling) error and the problem of assessing the impacts of such errors and accounting for them has been receiving increasing attention of late. This paper provides a survey of this problem with an emphasis on estimating the coefficients of the underlying dynamic model, primarily in the context of fitting linear and nonlinear autoregressive models. An overview is provided of the biases induced by ignoring the measurement error and of methods that have been proposed to correct for it, and remaining inferential challenges are outlined.

## 1 Introduction

Measurement error is a commonly occurring problem and is especially prominent in many time series, where the variable of interest often has to be estimated rather than observed exactly. There is a fairly diverse statistical literature which has addressed the problem of measurement error in time series as well as a burgeoning ecological literature, where the problem of modeling population dynamics in the presence of the so-called observation error has garnered considerable attention. Included among the many papers addressing this problem with real data are ones that account for errors in series involving population abundances of waterfowl (Lillegard et al. 2008; Saether et al. 2008; Viljugrein et al. 2005), voles (Stenseth et al. 2003), grouse (Ives et al. 2003) as well as labor force statistics (Pfeffermann et al. (1998), retail sales (Bell and Wilcox 1993), the number of households in Canada (Feder 2001), and disease rates (Burr and Chowell 2006).

The main ingredients here are a dynamic model for the true (but unobserved) values and a measurement error model. The time series of interest is denoted by

J.P. Buonaccorsi (✉)
University of Massachusetts, Amherst, MA, USA
e-mail: johnpb@math.umass.edu

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_3, 

$\{Y_t, t = 1, \ldots, T\}$, where $Y_t$ is random and $t$ indexes time. The realized true value at time $t$ is denoted by $y_t$. Measurement error occurs where instead of $y_t$ we observe the outcome of $W_t$, where $W_t$ is an estimator or general index of $y_t$. A particular point of emphasis in our coverage is to allow the behavior of the measurement error to depend on the underlying true value and/or sampling effort, as commonly occurs in practice.

The two main questions are 1. What happens if we ignore the measurement error? and 2. How can we correct for the measurement error? Of course, there are many possible objectives of a time series analysis and we need to limit our scope. The primary focus in this paper is on estimation of the parameters in the dynamic model, autoregressive models in particular. This is a logical first step as these parameters provide the building blocks for other objectives including forecasting or estimating probabilities about the process in the future. Also, in many of the ecological problems we discuss the estimation of the coefficients of the underlying dynamic process is the main thing of interest (and for this reason our discussion has a heavy ecological orientation to it).

Because of space limitations, there are a number of important related problems that we mostly ignore. These include repeated sample surveys where the main objective is updated estimation of the current true value, predicting and/or filtering in the presence of measurement error (e.g., Berliner (1991); Tripodis and Buonaccorsi (2009)), direct estimation of trends, model identification and problems where there are other variables in addition to the dynamic model. There are many examples of the latter; see, e.g., Ives et al. (2003), De Valpine and Hilborn (2005), and references therein for access to an extensive fisheries literature, Burr and Chowell (2006) in fitting SIR model to disease dynamics and Schmid et al. (1994).

Within the above stated focus, the objective is to provide a broad survey of modeling considerations, the effects of, and ways to correct for, measurement error and some of the challenges in carrying out estimation and inference. The intent is not to look at any particular model in great detail, although we do illustrate some key concepts with linear autoregressive models and the Ricker model and present a few new results. Dynamic models for the true values and measurement error models are discussed in Sects. 2.1 and 2.2, respectively. This is followed by a discussion about the performance of naive analyses that ignore measurement error in Sect. 3 and then a survey of correction methods in Sect. 4. Concluding remarks appear in Sect. 5.

## 2 Models

### *2.1 Dynamic Models for True Values*

There is of course a very rich class of dynamic models that can be used for time series. As noted in the introduction the main interest here is in autoregressive

**Table 1** Some dynamic models

| Name | Model |
|---|---|
| AR(p) | $Y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t$ |
| Ricker | $Y_t = \phi_0 + y_{t-1} + \phi_1 e^{y_{t-1}} + \varepsilon_t$ |
| Beaverton-Holt | $Y_t = y_{t-1} + log(\phi_0) - log(1 + \phi_1 n_{t-1}) + \varepsilon_t$ |
| Theta-logistic | $Y_t = y_{t-1} + \phi_0 + (\phi_1 n_{t-1})^{\theta} + \varepsilon_t$ |
| Logistic | $Y_t = y_{t-1} + \phi_0 + log(1 - e^{yt-1}) + \varepsilon_t$. |

models with $E(Y_t|\mathbf{y}_{t-1}) = m(\mathbf{y}_{t-1}, \boldsymbol{\phi})$ and $V(Y_t|\mathbf{y}_{t-1}) = v(\mathbf{y}_{t-1}, \boldsymbol{\phi}, \boldsymbol{\sigma})$, where $\mathbf{y}_{t-1} = (\ldots, y_{t-2}, y_{t-1})$ indicates past values and $\boldsymbol{\sigma}$ contains additional variance parameters. Alternatively we can write $Y_t|\mathbf{y}_{t-1} = m(\mathbf{y}_{t-1}, \boldsymbol{\phi}) + \varepsilon_t$, where $E(\varepsilon_t) = 0$ and $V(\varepsilon_t) = v(\mathbf{y}_{t-1}, \boldsymbol{\phi}, \boldsymbol{\sigma})$.

We allow the mean function to be linear or nonlinear in $Y$ with some examples given in Table 1. All of these models, with the last four being nonlinear in $Y$, arise in population ecology with $Y_t = log(N_t)$ where $N_t$ is abundance, or density, at time $t$. The $\varepsilon_t$ is typically taken to be normal with mean 0 and $V(\varepsilon_t) = \sigma^2$, referred to as process error. There are, of course, numerous other models that can be considered that we won't discuss in any detail. For example, as noted briefly later, measurement error has received some attention in an autoregressive integrated moving average (ARIMA) and basic structural models (BSMs) used to model labor variables, medical indices, and other variables over time. Another important model in the economics literature is to extend the linear autoregressive model in Table 1 to allow conditional heteroscedasticity in the process errors; the so-called ARCH model.

With $Y_t = log(N_t)$ where $N_t$ is a count or scaled count, the usual assumption on $\varepsilon_t$ may not be appropriate and an alternate nonlinear models arise by working explicitly if we work with $N_t$. These lead to another set of mean-variance models. (It is worth noting that one can only move directly from a model for $N_t|\mathbf{n}_{t-1}$ to one for $Y_t|\mathbf{y}_{t-1}$ in certain special cases, e.g., when $\delta_t$ is log-normal.) For example with the Ricker model, $N_t = n_{t-1} e^{(\phi_0 + \phi_1 n_{t-1})} \delta_t$. If we assume $\delta_t$ has mean 1 and constant variance $\sigma_\delta^2$, then $E(N_t|n_{t-1}) = m(n_{t-1}, \boldsymbol{\phi}) = n_{t-1} e^{(\phi_0 + \phi_1 n_{t-1})}$ and $V(N_t|n_{t-1}) = m(n_{t-1}, \boldsymbol{\phi})^2 \sigma_\delta^2$. For the multiplicative version of the AR(1) model we might just work with $E(N_t|n_{t-1}) = e^{\phi_0} n_{t-1}^{\phi_1}$. In general, if $N_t|n_{t-1}$ is distributed Poisson, then $V(N_t|n_{t-1}) = E(N_t|n_{t-1})$. For a full discussion of fitting dynamic models with count data, possibly with the inclusion of time varying covariates, see Mallick and Sutradhar (2008).

Turning to the linear $AR(p)$ models we will assume the model is stationary (e.g., Box et al. 1994, Chap. 3), which for AR(1) model means $|\phi_1| < 1$. In general for stationary models we denote $E(Y_t) = \mu_Y$ and $V(Y_t) = \sigma_Y^2$, both constant in $t$, and $Cov(Y_t, Y_{t+k}) = \gamma_k$, a function only of the lag $k$. In the population ecology literature the AR models are referred to as the Gompertz model. There, based on the multiplicative versions for the AR(1) and AR(2), the primary objective is estimation of $\phi_1$ and/or $\phi_2$, usually interpreted as measures of density dependence and delayed density dependence, respectively; see Stenseth et al. (2003); Solow (2001), and

references therein. The AR(p) models can also be extended to multivariate autoregressive (MAR) models (also called vector autoregressive models or VAR); see Ives et al. (2003) and Aigner et al. (1984), both which accommodate measurement error.

All of the models in Table 1 have the random walk model $Y_t = y_{t-1} + \mu + \varepsilon_t$ as a special case. If the $\varepsilon_t$'s are assumed to be independent and identically distributed (iid) with mean 0 and variance $\sigma^2$, this is a moving average of order 1, which is non-stationary. In the ecological context, once again with $Y = log(N)$, it represents a density independent model and is commonly used in population viability analysis and related trend analysis (Morris and Doak (2002)). There is a fairly large literature on measurement error in this problem. While there are some methodological connections to our discussion here we will not discuss it in detail both for space reasons and the fact that with the difference there is no dynamic piece left. See Buonaccorsi and Staudenmayer (2009) for a comprehensive treatment and references.

For likelihood-based approaches, the joint density of $Y_1, \ldots, Y_T$ is denoted $f(\mathbf{y}; \boldsymbol{\phi}, \boldsymbol{\sigma})$, with the term density applying for either the continuous or discrete case. For conditional likelihood approaches, assuming an autoregressive model (see Box et al. (1994, Chap. 7) for extension) suppose the distribution of $Y_t|\mathbf{y}_{t-1}$ depends on the past $p$ values. Then partition $\mathbf{y}$ into $\mathbf{y}' = (\mathbf{y}_1^{*'}, \mathbf{y}_2^{*'})$, where $\mathbf{y}_1^* = (y_1, \ldots, y_p)'$ and $\mathbf{y}_2^* = (y_{p+1}, \ldots, y_T)'$. The conditional density of $\mathbf{Y}_2^*$ given $\mathbf{y}_1^*$ is

$$f_2(\mathbf{y}_2^*; \boldsymbol{\phi}, \boldsymbol{\sigma}, \mathbf{y}_1^*) = \prod_{t=p+1}^{T} f(y_t|\mathbf{y}_{t-1}, \boldsymbol{\phi}, \boldsymbol{\sigma}), \tag{1}$$

where $f(y_t|\mathbf{y}_{t-1}, \boldsymbol{\phi}, \boldsymbol{\sigma})$ is the density of $\mathbf{Y}_t$ given $\mathbf{y}_{t-1}$. Here, $\mathbf{y}_1^*$ is treated as fixed.

Notice that we can also write $f(\mathbf{y}; \boldsymbol{\phi}, \boldsymbol{\sigma}) = \int_{\mathbf{y}_1^*} f_2(\mathbf{y}_2^*; \boldsymbol{\phi}, \boldsymbol{\sigma}, \mathbf{y}_1^*) f(\mathbf{y}_1^*) d\mathbf{y}_1^*$, where $f(\mathbf{y}_1^*)$ is the density of $\mathbf{Y}_1^*$, which can depend on some parameters (suppressed in the notation). If the $Y$'s are discrete, then integration is replaced by summation above.

## 2.2 Measurement Error Models

The measurement error model describes the conditional behavior of the observable random variables $\mathbf{W} = (W_1, \ldots, W_T)'$ given $\mathbf{y} = (y_1, \ldots, y_T)'$. Given the huge number of sampling methods that can be used to estimate the $y_t$'s, there are many measurement error models that can be entertained here. The bulk of the literature assumes that the measurement errors are conditionally uncorrelated and, frequently, additive. The assumption of conditionally uncorrelated measurement errors is reasonable when there is independent sampling at each time point. There are settings, however, where some common sampling units occur over time, leading to correlated measurement errors. This can occur, for example, in biological and climatological monitoring and is also a key feature in large national repeated sample surveys which use block resampling, where fairly general dynamic models have

been used for the measurement error itself; see, for example, Pfeffermann et al. (1998); Feder (2001), and references therein.

The conditional mean and covariance are $E(\mathbf{W}|\mathbf{y})$ and $Cov(\mathbf{W}|\mathbf{y})$, which usually depend on unknown parameters, contained in $\boldsymbol{\theta}$. A more convenient representation is

$$\mathbf{W} = \mathbf{y} + \mathbf{u}, \quad E(\mathbf{u}|\mathbf{y}) = \mathbf{B}_c, \quad \text{and} \quad Cov(\mathbf{u}|\mathbf{y}) = Cov(\mathbf{W}|\mathbf{y}) = \boldsymbol{\Sigma}_{uc}. \tag{2}$$

The $\mathbf{B}_c' = (\theta_{1c}, \ldots, \theta_{Tc})$ contains conditional biases, while $\boldsymbol{\Sigma}_{uc}$ is the conditional covariance matrix. The $c$ here is a reminder that these are conditional on $\mathbf{y}$ and this conditional behavior could be a function of the underlying true values or sampling effort/design. This is elaborated on in more detail below when discussing the measurement error variances.

The frequently used additive model assumes $E(\mathbf{W}|\mathbf{y}) = \mathbf{y}$ or $E(\mathbf{u}|\mathbf{y}) = \mathbf{0}$, but models with bias have been considered including constant bias, $E(W_t|y_t) = y_t + \boldsymbol{\theta}$, and proportional bias, $E(W_t|y_t) = \theta y_t$. The constant bias model arises in various ways. For example, if $\hat{N}_t$ is the estimated abundance at time $t$, $W_t = log(\hat{N}_t)$, $Y_t = log(N_t)$, and $E(\hat{N}_t|n_t) = an_t$ with constant coefficient of variation, then $E(W_t|y_t) = y_t + \theta$, exactly or approximately (see Buonaccorsi et al. 2006). Often, a constant bias term can be easily absorbed. The proportional bias model arises from Poisson-type sampling where $y_t$ is abundance, $W_t$ is a count, adjusted for sampling effort, and $\theta$ is an unknown representing "catchability"; e.g., Stenseth et al. (2003). There are certainly even richer bias models that can be considered. For example, in calibrating aerial counts $W$ versus ground counts $Y$ of waterfowl, Lillegard et al. (2008) build a model where $W_t^{1/2} \sim N(\theta_0 + \theta_1 y_t^{1/2}, \tau^2)$, leading to $E(W_t|y_t) = \tau^2 + \theta_0^2 + 2\theta_0\theta_1 y_t^{1/2} + \theta_1^2 y_t^2$.

Much of our coverage is around settings assuming conditionally uncorrelated measurement errors with

$$Cov(\mathbf{W}|\mathbf{y}) = \boldsymbol{\Sigma}_{uc} = diag(\sigma_{u1c}^2, \ldots, \sigma_{uTc}^2), \tag{3}$$

a diagonal matrix with $(t,t)$ element $V(W_t|y_t) = \sigma_{utc}^2$. Note that this allows for heteroscedastic measurement errors where the conditional variance, $\sigma_{utc}^2$, possibly depending on $y_t$, or on sampling effort, although the functional nature of that dependence need not be specified. The unconditional (over random $Y$) variance is denoted by $\sigma_{ut}^2$. Suppose, for example, that $\sigma_{utc}^2 = h(y_t, \boldsymbol{\theta})$ and there is additive error or constant bias. Then, unconditionally, $\sigma_{ut}^2 = V(u_t) = E[V(u_t|Y_t)] + V[E(u_t|Y_t)] = E[h(Y_t, \boldsymbol{\theta})]$. An important point here is that if the conditional variance only changes over $t$ as a result of $y_t$ and the process is stationary, then $h(Y_t, \boldsymbol{\theta})$ is stationary and unconditionally $\sigma_{ut}^2 = \sigma_u^2$. Hence, we can have conditional heteroscedasticity but unconditional homoscedasticity. As noted above, however, the conditional and unconditional variance may also change as a function of sampling effort.

For assessing the properties of naive estimators and corrected estimators under the additive model and (3) we assume that

$$Lim_{T\to\infty}\frac{\sum_{t=1}^{T}\sigma_{ut}^2}{T}=\sigma_u^2 \tag{4}$$

exists, where the limit is in probability if $\sigma_{utc}^2$ depends on $Y_t$.

Returning to the general setting, the unconditional moments of the observable $\mathbf{W}$ are given by $E(\mathbf{W})=E[E(\mathbf{W}|\mathbf{Y}))]$ and $Cov(\mathbf{W})=E[Cov(\mathbf{W}|\mathbf{Y}))]+Cov[E(\mathbf{W}|\mathbf{Y})]$. For additive measurement error this becomes $E(\mathbf{W})=E(\mathbf{Y})(=\mu\mathbf{1}$ under stationarity) and $Cov(\mathbf{W})=E[\boldsymbol{\Sigma}_{uc}]+Cov[\mathbf{Y}]=\boldsymbol{\Sigma}_u+\boldsymbol{\Sigma}_Y$.

For likelihood methods $\mathbf{W}|\mathbf{y}$ is assumed to have density $f(\mathbf{w}|\mathbf{y},\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ includes any measurement error parameters. When $\boldsymbol{\theta}$ is a parameter vector of fixed length (e.g., $\sigma_u^2$, or $(\theta_0,\theta_1,\tau^2)$ in the model of Lillegard et al. (2008)) there is no difficulty in interpreting and using this density in standard fashion. The handling of $\boldsymbol{\theta}$ is more delicate if we allow the conditional measurement error parameters to change over time in some unspecified manner. For example, if we assume $W_t|y_t\sim N(y_t,\sigma_{utc}^2)$ where no structure is given to $\sigma_{utc}^2$, then $\boldsymbol{\theta}=(\sigma_{u1c}^2,\ldots,\sigma_{uTc}^2)'$, which increases in size with $T$. Further, if $\sigma_{utc}^2$ involves $y_t$, then unconditionally $\boldsymbol{\theta}$ is random. There are two densities for $\mathbf{W}$ that will be used later,

$$f_W(\mathbf{w};\phi,\sigma,\boldsymbol{\theta})=\int_y f(\mathbf{w}|\mathbf{y},\boldsymbol{\theta})f(\mathbf{y};\boldsymbol{\phi},\boldsymbol{\sigma})d\mathbf{y} \tag{5}$$

and

$$f_W^*(\mathbf{w};\phi,\sigma,\boldsymbol{\theta},\mathbf{y}_1^*)=\int_{\mathbf{y}_2^*} f(\mathbf{w}|\mathbf{y},\boldsymbol{\theta})f_2(\mathbf{y}_2^*;\boldsymbol{\phi},\boldsymbol{\sigma},\mathbf{y}_1^*)d\mathbf{y}_2^*, \tag{6}$$

where the second conditions on $\mathbf{y}_1^*$ with $f_2(\mathbf{y}_2^*;\boldsymbol{\phi},\boldsymbol{\sigma},\mathbf{y}_1^*)$ as given in (1). As before, integrals are replaced by sums for the discrete case.

## 3 Performance of Naive Estimators

An important question, especially in the absence of any specific information about the measurement error, is the performance of the so-called naive analyses, which ignore the measurement error and treat $W_t$ as if it is $Y_t$. Based on what we know from regression models it is not surprising that the measurement error will lead to biases in estimated coefficients as well as in the process variance parameters. However, we will see that obtaining analytical expressions for asymptotic or approximate biases is generally difficult here.

Of course the first question to ask here is "which naive analysis?", since there are a plethora of approaches to estimation in time series. For stationary normal models, maximum likelihood (ML) and (more recently) restricted maximum likelihood (REML) are the most popular. Conditional maximum likelihood (CML), which maximizes $L(\boldsymbol{\phi},\boldsymbol{\sigma},\mathbf{y}_1^*|\mathbf{y})$ arising from (1), is another choice. This drops making an assumption about the marginal distribution of $\mathbf{Y}_1^*$ and usually leads to easier

**Table 2** Selected papers assessing the performance of naive estimators via simulation

| Author(s) | Dynamic Model | ME Model | Method |
|---|---|---|---|
| Solow (1998) | Logistic | Poisson | LS |
| De Valpine and Hastings (2002) | Ricker and | $N(0,\sigma_u^2)$ | ML |
| De Valpine (2002) | Beaverton-Holt | 〃 〃 | 〃 〃 |
| SB (2005) | AR(1) | $N(0,\sigma_{utc}^2)$ | ML |
| Lele (2006) | AR(1) | Poisson | ML |
| Hovestadt and Nowicki (2008) | Modified Ricker | $N(0,\sigma_u^2)$ | LS |
| Barker and Sibly (2008) | theta-logistic | $N(0,\sigma_u^2)$ | LS |
| Ives et al. (2010) | ARMA | $N(0,\sigma_u^2)$ | REML |
| Resendes (2011) | Ricker | $N(0,\sigma_u^2)$ | LS |

computing. For a normal autoregressive model with the $\varepsilon_t$ assumed iid $N(0,\sigma^2)$ this leads to least squares, where $\hat{\boldsymbol{\phi}}_{CLS}$ minimizes $\sum_{t=p+1}^{T}(y_t - m(\mathbf{y}_{t-1},\boldsymbol{\phi}))^2$, and $p$ is the number of previous $y$ terms in the $m$ function. Obviously if $m(\mathbf{y}_{t-1},\boldsymbol{\phi})$ is linear in the $\phi$'s, then this leads to simple least squares. More generally these lead to nonlinear least squares with estimating equations of the form

$$\sum_{t=p+1}^{t}(y_t - m(\mathbf{y}_{t-1},\boldsymbol{\phi}))\Delta(\mathbf{y}_{t-1},\boldsymbol{\phi}) = \mathbf{0}, \tag{7}$$

where the $jth$ element of the $\Delta(\mathbf{y}_{t-1},\boldsymbol{\phi})$ is $\partial m(\mathbf{y}_{t_1},\boldsymbol{\phi})/\partial\phi_j$. Similar estimating equations can arise from conditional maximum likelihood in other situations. For example, if $Y_t|y_{t-1}$ is assumed to be Poisson with mean $e^{\phi_0+\phi_1 log(y_{t-1})}$, then CML leads to (7) with $\Delta(\mathbf{y}_{t-1},\boldsymbol{\phi})' = [1, log(y_{t-1})]$.

A number of general strategies have been tried to investigate the performance of naive estimators in the presences of measurement error. These include:

1. Use an explicit expression for the estimators, e.g., $\hat{\boldsymbol{\phi}} = g(\mathbf{W})$), and determine the limiting or approximate properties analytically. This is mainly used for the linear autoregressive models, discussed in the next section.
2. View the naive estimators as solutions to estimating equations $S(\mathbf{W},\boldsymbol{\phi},\boldsymbol{\sigma}) = \mathbf{0}$. In this case the naive estimators (under some conditions) will converge to $\boldsymbol{\phi}^*$ and $\boldsymbol{\sigma}^*$ which satisfy $E_{\boldsymbol{\phi},\boldsymbol{\sigma}}[S(\mathbf{W},\boldsymbol{\phi}^*,\boldsymbol{\sigma}^*)]/T \to \mathbf{0}$.
3. Find an "induced" model, or an approximation, for the behavior of $W_t|\mathbf{w}_{t-1}$. If this is in the same class as the original with $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}$ replaced by $\phi^*$ and $\boldsymbol{\sigma}^*$, then the naive methods are consistent (or approximately consistent) for $\phi^*$ and $\boldsymbol{\sigma}^*$.
4. Simulate the performance of the naive estimators. While it is difficult to gain a good understanding of the nature of the biases from simulations alone, this is often the only option. Even when analytical asymptotic properties are available, as in linear AR models, simulations are needed to assess "small" sample behavior; see the next section. Table 2 summarizes some of the papers that have utilized simulation. SB (2005) refers to Staudenmayer and Buonaccorsi (2005), an abbreviation used throughout.

**Table 3** Simulation means for naive least squares estimators from the Ricker model with $\varepsilon_t \sim N(0, \sigma^2)$ with $K = 100$ and $\sigma = 0.2$. From Resendes (2011), with permission

| | True $\phi_0$ | | | | | |
|---|---|---|---|---|---|---|
| | 0.2 | 0.75 | 1.5 | 2.4 | 2.6 | $\sigma_u^2$ |
| $\hat{\phi}_0$ | 0.413 | 0.822 | 1.444 | 2.381 | 2.575 | 0.05 |
| | 0.593 | 0.939 | 1.266 | 2.206 | 2.427 | 0.2 |
| | 0.729 | 0.975 | 1.182 | 1.988 | 2.231 | 0.3 |
| | 0.813 | 0.983 | 1.053 | 1.628 | 1.855 | 0.5 |
| $\hat{K}$ | 97.537 | 100.093 | 100.007 | 100.188 | 100.035 | 0.05 |
| | 94.771 | 102.044 | 102.324 | 101.516 | 101.668 | 0.2 |
| | 96.631 | 104.456 | 104.028 | 102.375 | 104.261 | 0.3 |
| | 100.686 | 113.058 | 113.294 | 111.585 | 112.367 | 0.5 |
| $\hat{\sigma}$ | 0.202 | 0.206 | 0.205 | 0.226 | 0.231 | 0.05 |
| | 0.301 | 0.291 | 0.294 | 0.456 | 0.536 | 0.2 |
| | 0.387 | 0.371 | 0.379 | 0.617 | 0.720 | 0.3 |
| | 0.586 | 0.566 | 0.573 | 0.906 | 1.069 | 0.5 |

While methods 2 and 3 have proved fruitful in many regression problems (see Carroll et al. (2006) and Buonaccorsi (2010)) they are less useful in dynamic situations. A simple example that illustrates the shortcomings of method 3 is the linear AR(p) model with additive normal measurement error with constant variance. This leads to an induced model which is an Autoregressive Moving Average (ARMA) (p,p) (see Sect. 4.2.1), so the problem is one of model misspecification. Method 2, an approach first used in regression contexts by Stefanski (1985), is the only recourse for assessing bias analytically when the estimators do not have a closed form. Expanding the estimating equations leads to $\boldsymbol{\phi}^* \approx \boldsymbol{\phi} - (Lim_{T\to\infty} \sum_t E[\dot{S}(W_t, \boldsymbol{\phi})]/T)^{-1} Lim_{t\to T} \sum_t E(S(W_t, \boldsymbol{\phi}))/T$, assuming the limits exist and where $\dot{S}$ denotes partial derivatives with respect to the parameters. The problem here is finding the expected values and limits. The best potential for this method is with using conditional ML/LS approaches leading to estimating equations as given in (7), but even there, the analysis is not straightforward. To illustrate we consider the Ricker model, which was investigated by Resendes (2011). Although nonlinear in $Y$, it is linear in the parameters, leading to simple linear least squares. Defining, $D_t = W_{t+1} - W_t$ and $C_t = e^{W_t}$ $S_{DC} = \sum_{t=1}^{T-1}(D_t - \bar{D})(C_t - \bar{C})/T$ and $S_{CC} = \sum_{t=1}^{T-1}(C_t - \bar{C})^2/T$, then the naive estimators of $\phi_1$ is $\hat{\phi}_{1,naive} = S_{DC}/S_{CC}$ and asymptotically $\hat{\phi}_{1,naive} \Rightarrow LimE(S_{DC})/LimE(S_{CC})$, provided the two limits exists. We faced two problems here; the first in finding the expected values involving nonlinear functions, the second in determining whether the sums converge, and to what. If the process is assumed stationary, then these limits generally exist but one needs to determine the stationary moments for the two series involved. Resendes (2011) examined the biases of the naive estimators in the Ricker model extensively, both via simulation and via the estimating equation just described (on both additive and multiplicative scales). While clean final expressions for the approximate biases proved elusive, his approximations did show how the direction of the bias can change with the values of the parameters. This is also seen in his simulations, a very small portion of which appear in Table 3 and highlight his main

conclusions. These are based on parameter values similar to those in De Valpine and Hastings (2002) but with some larger measurement error variances. The magnitude and direction of the biases depend heavily on $\phi_0$ with naive estimators of $\phi_0$ tending to be overestimates at small $\phi_0$ and underestimates as $\phi_0$ increases. The same thing happens for the naive estimator of the carrying capacity $K = -\phi_0/\phi_1$ (the Ricker model can be parameterized instead in terms of $\phi_0$ and $K$), but it is less sensitive to the measurement error than naive estimators of $\phi_0$ are. It also shows overestimation of $\sigma$, although modest at small measurement error variances.

## *3.1 Linear Autoregressive Models*

Here we summarize and illustrate some results for the stationary linear AR models under the additive error model with (2). This is one of the simpler settings, allowing exactly analytical bias expressions, and is useful for illustrating some key points. We concentrate on the coefficients, but it also can be shown that $\sigma^2$ is overestimated. The standard method of analysis here is traditionally maximum likelihood, but there is increasing support for the use of REML given that it reduces small sample bias (e.g., Cheang and Reinsel (2000)), REML estimators for the AR(1) model can be easily obtained using most mixed models software. SB (2005) extended earlier work and developed a general expression for the limiting values of the Yule-Walker (YW) estimates allowing changing measurement error variances and under the assumption in (4). For the AR(1) model this leads to $\hat{\phi}_{1,naive} \Rightarrow \kappa\phi_1$, where $\Rightarrow$ denotes convergence in probability and

$$\kappa = \sigma_Y^2/(\sigma_Y^2 + \sigma_u^2). \tag{8}$$

This shows asymptotic attenuation (bias towards 0) in the estimator of $\phi_1$. If, as in ecological applications, the focus is on $\beta_1 = \phi_1 - 1$, then $\hat{\beta}_{1,naive} \Rightarrow \kappa\beta_1 + (\kappa - 1)$, which is greater than $\beta_1$ if $\beta_1 < -1$ but less than $\beta_1$ if $\beta_1 > -1$.

For the AR(2), we have

$$\begin{pmatrix} \hat{\phi}_{1,naive} \\ \hat{\phi}_{2,naive} \end{pmatrix} \Rightarrow \begin{pmatrix} (\kappa\rho_1 - \kappa^2\rho_1\rho_2)/(1 - \kappa^2\rho_1^2) \\ (\kappa\rho_2 - \kappa^2\rho_1^2)/(1 - \kappa^2\rho_1^2) \end{pmatrix},$$

where $\rho_j = \gamma_j/\sigma_X^2$ for $j = 1$ and 2 with $\gamma_1 = (\phi_1\sigma_X^2)/(1 - \phi_2)$ and $\gamma_2 = (\phi_1^2 + \phi_2 - \phi_2^2)\sigma_X^2/(1 - \phi_2)$. This leads to some more interesting results, as the bias in either element of $\hat{\boldsymbol{\phi}}_{naive}$ can be either attenuating or accentuating (larger in absolute value), depending on both the amount of measurement error and the other parameters.

It is well known that YW and ML/REML estimators have the same asymptotic properties without measurement error. With measurement error the asymptotic properties of the ML (equivalently REML) can be examined through their associated estimating equations, see the Appendix. This leads to the same asymptotic behavior

**Table 4** Performance of maximum likelihood (ML), Yule-Walker (YW), and restricted maximum likelihood (REML) estimators of $\phi_1$, using true and mismeasured values in the AR(1) model. $\phi_1^* = \kappa\phi_1$ = limiting value of naive estimator

| | | | True (using $Y$) | | | | Naive (using $W$) | | |
|---|---|---|---|---|---|---|---|---|---|
| $\phi_1$ | $\sigma_u^2$ | $n$ | YW | ML | REML | $\phi_1^*$ | YW | ML | REML |
| 0.2 | 0.15 | 10 | 0.023 | 0.029 | 0.176 | 0.169 | 0.006 | 0.009 | 0.152 |
| 0.2 | 0.15 | 30 | 0.140 | 0.146 | 0.187 | 0.169 | 0.116 | 0.120 | 0.161 |
| 0.2 | 0.15 | 50 | 0.161 | 0.164 | 0.188 | 0.169 | 0.136 | 0.139 | 0.163 |
| 0.2 | 0.15 | 100 | 0.179 | 0.180 | 0.192 | 0.169 | 0.149 | 0.150 | 0.162 |
| 0.2 | 0.6 | 10 | 0.031 | 0.040 | 0.185 | 0.116 | −0.022 | −0.020 | 0.120 |
| 0.2 | 0.6 | 30 | 0.151 | 0.157 | 0.199 | 0.116 | 0.074 | 0.077 | 0.116 |
| 0.2 | 0.6 | 50 | 0.171 | 0.174 | 0.198 | 0.116 | 0.090 | 0.092 | 0.115 |
| 0.2 | 0.6 | 100 | 0.185 | 0.187 | 0.199 | 0.116 | 0.101 | 0.102 | 0.113 |
| 0.8 | 0.15 | 10 | 0.356 | 0.435 | 0.623 | 0.749 | 0.290 | 0.354 | 0.536 |
| 0.8 | 0.15 | 30 | 0.648 | 0.686 | 0.750 | 0.749 | 0.586 | 0.619 | 0.681 |
| 0.8 | 0.15 | 50 | 0.710 | 0.731 | 0.768 | 0.749 | 0.650 | 0.669 | 0.705 |
| 0.8 | 0.15 | 100 | 0.756 | 0.765 | 0.783 | 0.749 | 0.701 | 0.709 | 0.727 |
| 0.8 | 0.6 | 10 | 0.357 | 0.444 | 0.641 | 0.630 | 0.195 | 0.235 | 0.410 |
| 0.8 | 0.6 | 30 | 0.652 | 0.689 | 0.753 | 0.630 | 0.463 | 0.487 | 0.542 |
| 0.8 | 0.6 | 50 | 0.709 | 0.729 | 0.766 | 0.630 | 0.526 | 0.540 | 0.572 |
| 0.8 | 0.6 | 100 | 0.756 | 0.766 | 0.785 | 0.630 | 0.574 | 0.582 | 0.598 |

as the YW estimators when $\boldsymbol{\Sigma}_u = \sigma_u^2\mathbf{I}$. Note that the $\boldsymbol{\Sigma}_u$ is the unconditional covariance of $\mathbf{u}$ and can have conditional heteroscedasticity arising through $Y_t$; see the discussion in Sect. 2.2.

One question is how useful the bias expressions are for "small" samples. Even without measurement error, the issue of bias in small samples is an important one with time series. To illustrate data was generated from the AR(1) model with $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$ and $W_t = y_t + u_t$, where the $\varepsilon_t$ are iid $N(0, \sigma^2)$ and the $\varepsilon_t$ are iid $N(0, \sigma_u^2)$. The process variance was held to $\sigma^2 = .8$, while $\phi_1 = .2, .5$, or .8 and $\sigma_u^2 = .15, .4$ and .8. The case with $\phi_1 = .5$ and $\sigma_u^2 = .15$ is based roughly on an analysis of mouse dynamics given in Buonaccorsi (2010, Chap. 12). For each combination 1,000 simulations were run and YW, ML, and REML estimators obtained, using the true $Y$'s and the error prone $W$'s. Partial results appear in Table 4. The analysis with true values is given for two reasons. First it shows the clear superiority of REML to ML estimation, especially at small sample sizes, with the ML only being modestly better than the YW estimator. Notice that even using true values all of the estimators are attenuated towards zero, sometimes dramatically with $n = 10$. Second it gives a baseline to compare the performance of the naive estimators to. The measurement error leads to further attenuation, increasing in $\sigma_u^2$, as it should. The REML estimator obviously provides some extra protection against measurement error compared to ML and YW, especially at small sample sizes. The variable $\phi_1^*$ is the limiting value of the naive estimator (whether YW, ML, or REML). The asymptotic bias associated with this limiting value can be significantly off compared to the simulated bias, sometimes even with samples of size 50.

## 4 Correcting for Measurement Error

There are three general contexts within which measurement error corrections are carried out. 1. Using only the observed $W_1, \ldots, W_T$; 2. Using the observed $W$'s and estimated measurement error parameters leading to pseudo-methods. 3. Using richer data than just $W_t$' from each time point t. We first comment on each of these three.

1. Unlike many other measurement error problems it has been shown that with some assumptions on the measurement error (e.g., that the errors are independent with mean 0 and common variance) all of the parameters can be identified (and estimated) from the $\mathbf{W}$ data alone for a variety of dynamic models; see, e.g., some of the references in Sect. 4.2.1 and Aigner et al. (1984). The majority of the work on correcting for measurement error in dynamic settings has attacked the problem from this perspective. In the likelihood context these are state-space models and the approach is standard in principle but can face computational challenges as discussed later. The shortcoming of this approach is the potential restrictive nature of the measurement error model and the fact that identifiability does not guarantee good estimators. However, without any information about the measurement error process the only option is to use this approach to estimate the dynamic and measurement error parameters simultaneously.
2. There is often data that allows for estimating the measurement error parameters, contained in $\hat{\boldsymbol{\theta}}$, say. This may include estimated variances as well as biases and/or correlations if they are part of the model. With additive uncorrelated measurement errors, allowing changing variance over time then $\hat{\boldsymbol{\theta}}$ contains an estimate of measurement error variances, $\hat{\sigma}^2_{ut}$, at each time point $t$, typically arising from the same data that produces $W_t$. The pseudo-methods set $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and then estimate the parameters in the dynamic model. While it seems natural to exploit estimates about the measurement error parameters, this strategy has been seriously underutilized. Note that $\hat{\boldsymbol{\theta}}$ may be of fixed size or contain separate estimates at each time point, depending on the assumed structure of $\boldsymbol{\theta}$; see the discussion in Sect. 2.2. We subsume under pseudo-methods approaches that simply assume that the measurement error parameters are known. The difference between viewing $\hat{\boldsymbol{\theta}}$ as estimated or known and fixed will come in trying to account for uncertainty $\hat{\boldsymbol{\theta}}$; see Sects. 4.1 and 4.2.2.
3. The third approach has some richer "data," denoted $\mathbf{Q}_t$, at time $t$ with a model for $\mathbf{Q}_t$ given $y_t$. The simplest example is where $\mathbf{Q}_t$ contains replicate measures of $y_t$; e.g, Wong et al. (2001); Dennis et al. (2010), and Knape et al. (2011). The analyses here connect to the previous two approaches. If $\mathbf{Q}_t|\mathbf{y}_t$ depends on a finite collection of parameters $\boldsymbol{\theta}$, then this is like approach 1 but with $W_t$ replaced by $\mathbf{Q}_t$. Or, the richer data can be used to first estimate the measurement error parameters, in which case this reduces to a pseudo-method. In some cases those two strategies concur. With these connections we won't mention this method in detail in the later sections.We do note that one advantage of using the full $\mathbf{Q}_t$ is that the uncertainty from estimating the measurement error parameters is accounted for. It does, however, require that $\boldsymbol{\theta}$ be fixed and that a distribution is

**Table 5** Selected papers correcting for measurement error

| Methods using W only | | | |
|---|---|---|---|
| Author(s) | Model for $Y$ | ME Model | Method |
| De Valpine and Hilborn (2005) | General,AR(1), | $N(0,\sigma_u^2)$ | ML |
| De Valpine and Hastings (2002) | Ricker and | | |
| De Valpine (2002) | Beaverton-Holt | | |
| Calder et al. (2003) | Gen | Gen | Bayesian |
| Clark and Bjornstad (2004) | Gen | Gen | Bayesian |
| SB (2005) | AR(1) | $N(0,\sigma_u^2)$ | ML/ARMA |
| | AR(p) | $(0,\sigma_{ut}^2)$ | modified YW |
| Wang et al. (2006) | RW, AR(1), AR(2) | $N(0,\sigma_u^2)$ | ML |
| Dennis et al. (2006) | RW, AR(1) | $N(0,\sigma_u^2)$ | ML/REML |
| Lele (2006) | Gen | Gen | ComML |
| Knape (2008) | AR(1) | $N(0,\sigma_u^2)$ | ML |
| Ponciano et al. (2009) | Gen | Gen | ML |
| Knape et al. (2011) | AR(1) | varied | ML |
| Pseudo-methods | | | |
| Solow (1998) | "logistic" | Poisson | SIMEX |
| Williams et al. (2003) | AR(1)+ trend | $N(0,\sigma_{ut}^2)$ | ML |
| Ives et al. (2003) | MAR(1) | $N(0,\sigma_u^2)$ | ML |
| Clark and Bjornstad (2004) | Gen | $(0,\sigma_{ut}^2)$ | Bayesian |
| SB (2005) | AR(p) | $(0,\sigma_{ut}^2)$ | CEE/ML |
| Wang (2007) | Theta-logistic | $N(0,\sigma_u^2)$ | ML/Bayesian |
| Lillegard et al. (2008) | MAR(1) | (* below) | Bayesian |
| Ives et al. (2010) | ARMA | $N(0,\sigma_u^2)$ | ML/REML |
| Dennis et al. (2010) | AR(1) | $N(0,\sigma_u^2)$(reps) | REML |
| Knape et al. (2011) | AR(1) | varied | ML |
| Resendes (2011) | Ricker | varied | SIMEX, MEE |

$* \sqrt{(W_t)} \sim N(a + b\sqrt{exp(Y_t)}, \sigma_u^2)$

specified for the within time data. In many applications the within time sampling can be very complex and the results will be reduced to $W_t$ and $\hat{\sigma}_{ut}^2$, leading us back to the pseudo-methods.

Table 5 contains a partial listing of papers addressing connection techniques, many containing simulations evaluating the methods. A number of these will be referred to in the later discussion, along with additional papers.

Within each of the three contexts described above, moment-based, maximum likelihood, and Bayesian methods are all options. The pseudo-methods also open the door for other methods including Simex, modifying the estimating equations and what is known as "regression calibration" all of which have received limited, or no, attention, in dynamic settings. Our survey below is categorized by the correction technique (moment, likelihood, Simex, etc.) with the last subsection addressing the use of bootstrapping.

## 4.1 Moment Methods

Moment-based corrections are for the most part limited to linear problems including the random walk model and linear autoregressive models, which we concentrate on here. Working only with the $W_t$'s, if the measurement error is additive and conditionally uncorrelated but with possibly changing variances, modified Yule-Walker estimators for $\boldsymbol{\phi}$ are available. These take advantage of the fact that the lag covariances, which involve the $\phi$'s, are estimated consistently; see Walker (1960); Sakai et al. (1979); Chanda (1996), and comments in SB (2005). While these estimators are consistent and robust to changing measurement error variances, in practice they are often very ill behaved even for moderate sample sizes and cannot be recommended.

For a pseudo approach with uncorrelated additive measurement errors, but possibly unequal variances, define

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_0 & \cdots & \gamma_{p-1} \\ \vdots & \ddots & \vdots \\ \gamma_{p-1} & \cdots & \gamma_0 \end{pmatrix} \text{ and } \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \dots \\ \gamma_p \end{bmatrix},$$

where $\gamma_k$ is the lag $k$ covariance and $\gamma_0 = \sigma_Y^2$. SB (2005) proposed the simple estimator $\hat{\boldsymbol{\phi}}_{CEE} = (\hat{\boldsymbol{\Gamma}}_W - \hat{\sigma}_u^2 \mathbf{I})^{-1} \hat{\boldsymbol{\gamma}}_W$, where $\hat{\boldsymbol{\Gamma}}_W$ and $\hat{\boldsymbol{\gamma}}_W$ are naive estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{\gamma}$ using the sample variances and covariances of the observed $W_t$'s and $\hat{\sigma}_u^2 = \sum_{t=1}^T \hat{\sigma}_{ut}^2/T$. The estimator can be viewed as arising from either correcting the naive estimating equations so they have mean 0 (hence CEE for corrected estimating equation) or from a simple correction based on the fact that the sample variance of the observed $W_t$'s estimates $\sigma_Y^2 + \sigma_u^2$. They show that $\hat{\boldsymbol{\phi}}_{CEE}$ is consistent as long as (4) holds and in addition $\hat{\sigma}_u^2$ converges in probability to $\sigma_u^2$. They also establish asymptotic normality and provide the asymptotic covariance of $\hat{\boldsymbol{\phi}}_{CEE}$ but only under certain assumptions. An extension of their result is the following:

**Proposition 1.** *Let* $G_1 = \partial \hat{\boldsymbol{\phi}}_{CEE}/\partial \hat{\boldsymbol{\gamma}}_W|_*$, $G_2 = \partial \hat{\boldsymbol{\phi}}_{CEE}/\partial \hat{\sigma}_u^2|_*$, *(with* $|_*$ *denoting evaluation at* $\hat{\gamma}_{Wj} = \gamma_{Wj}$ *and* $\hat{\sigma}_u^2 = \sigma_u^2$*),* $\mathbf{X}_t' = [W_t^2, W_t W_{t+1}, \dots, W_t W_{t+p}]$, *and* $\bar{\mathbf{X}} = \sum_t \mathbf{X}_t/T$. *Assuming the following limits exist,* $\mathbf{Q}_1 = LimCov(\bar{\mathbf{X}})T$, $Q_{12} = LimCov(\bar{\mathbf{X}}, \hat{\sigma}_u^2)T$ *and* $Q_{22} = LimV(\hat{\sigma}_u^2)T$, *then the asymptotic/approximate covariance of* $\hat{\boldsymbol{\phi}}_{CEE}$ *is*

$$ACov(\hat{\boldsymbol{\phi}}_{CEE}) = (G_1 Q_1 G_1 + + G_1 Q_{12} G_2' + G_1' Q_{12}' G_2)/T + G_2 G_2' V(\hat{\sigma}_u^2). \quad (9)$$

The first term $G_1 Q_1 G_1/T = C_K$, say, is the approximate covariance of $\hat{\boldsymbol{\phi}}_{CEE}$ if the $\hat{\sigma}_{ut}^2$'s are treated as known. The terms involved in (9) get complicated, as does estimation of them. We omit details (given in Buonaccorsi and Staudenmayer (2012)) but the general result is useful for highlighting the difficulty in accounting for the uncertainty in the measurement error parameters by needing to handle the terms involving $\hat{\sigma}_{ut}^2$. The problem simplifies considerably when the $\hat{\sigma}_{ut}^2$'s are

assumed independent of the $W_t$'s, as would hold if the $W_t$ and $\hat{\sigma}^2_{ut}$ arise from using normal replicate measures at time $t$. In this case $ACov(\hat{\boldsymbol{\phi}}_{CEE}) = C_K + G_2 G_2' V(\hat{\sigma}^2_u)$ (where $C_k$ was defined above) so only $\eta_4 = \sum_t E(u_t^4)/T (= 3\sigma^4_{ut}$ if $u_t$ is normal) and $V(\hat{\sigma}^2_u)$ and estimates of them are needed. To illustrate (see SB (2005)), for the AR(1) model, the approximate variance of $\hat{\phi}_{1,CEE}$ is

$$\frac{1}{T}\left[\frac{(1-\phi_1^2)(2-\kappa)}{\kappa} + \frac{(1-\phi_1^2)(\sum_t \sigma^4_{ut}/T) + \phi_1^2\eta_4}{\sigma^2_Y}\right] + \frac{\phi_1^2 V(\hat{\sigma}^2_u)}{\sigma^2_Y},$$

where $\kappa$ is given in (8).

More broadly, the challenge is to accommodate the case where the conditional variance (or higher moments) of $u_t|y_t$ may depend on $y_t$. Unconditionally this leads to the $\hat{\sigma}^2_{ut}$'s being correlated with the $W_t$'s and expressions and estimates for $Q_1$ and $Q_{12}$ are needed. It appears impossible to do this robustly using just the $W_t$ and $\hat{\sigma}^2_{ut}$ in the most general setting, without some assumptions on the measurement error variances. Buonaccorsi and Staudenmayer (2012) develop a strategy for estimating the covariance matrix under the assumptions that either i) the sampling effort is constant so the heteroscedasticity arises through the $Y_t$ only or ii) the measurement error variance is inversely proportional to known sampling effort. The methodology used considers the joint series $(W_t, \hat{\sigma}^2_{ut})$ and exploits time series methods for estimating covariance structures robustly using consistent estimators of the spectral density of a multivariate stationary process (e.g., the modified Bartlett kernel estimator); see, e.g., Fuller (1996, Chap. 7).

## 4.2 Likelihood Methods

Likelihood and related Bayesian methods have dominated the correction approaches. There are two-likelihoods of interest, the full and conditional likelihoods $L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}|\mathbf{w}) = f(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta})$ and $L^*(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \mathbf{y}_1^*|\mathbf{w}) = f^*(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \mathbf{y}_1^*)$, based on the densities in (5) and (6), respectively. For stationary normal linear models, an alternative is to use the REML likelihood in place of $L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}|\mathbf{w})$.

### 4.2.1 Using $W$ Values Only

Assuming $\boldsymbol{\theta}$ is of fixed size, these methods maximize either the full or conditional likelihood, or their REML versions. Of course, they are only used when all of the parameters are identifiable. This is a classical state-space formulation and there is a fairly large literature on fitting these models, whether linear or nonlinear using maximum likelihood, and in a few cases, REML. While this is straightforward in principle, there are a number of challenges in using these techniques, including needing to dealing with local maxima and/or the maximum occurring on the boundary of the parameter space and, general difficulty in computing the likelihood

function and an associated covariance matrix of the estimates for use in inference. However, as discussed below, there have been recent advances tackling some of the computational challenges.

*Normal linear stationary models:* Assuming $\mathbf{u}|\mathbf{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ where $\boldsymbol{\Sigma}_u$ depends on a fixed number of parameters, the full likelihood approach can be used based on $\mathbf{W} \sim N(\mu\mathbf{1}, \boldsymbol{\Sigma}_Y + \boldsymbol{\Sigma}_u)$. (It is important to note that if the measurement error variances (or covariances if present) are changing with $\mathbf{y}$, then even if $\mathbf{u}$ is conditionally normal and unconditionally $Cov(\mathbf{u}) = \boldsymbol{\Sigma}_u$, $\mathbf{W} = \mathbf{Y} + \mathbf{u}$ is not normal since it involves mixtures of normals with changing variances. This same comment applies for the pseudo likelihood methods in the next section.) Often $\boldsymbol{\Sigma}_u$ is taken to be $\sigma_u^2\mathbf{I}$, while $\boldsymbol{\Sigma}_y$ is a function of the parameters, depending on the specific model for the true values. Computational methods here typically use the Kalman filter or some variation on it; see, e.g., Harvey (1990), Brockwell and Davis (2002), Ives et al. (2010, 2003) and, for the AR(1) models, Dennis et al. (2006) and Knape (2008). These last two papers touch on the important problem that even in the simple AR(1) model there may be issues with local maxima and/or the maximum occurring on the boundary. Knape (2008) also zeros in more on the fact that, not surprisingly, it can be difficult to separate the process error variance and the measurement error variance. These models can also be cast as mixed models which may employ other computational techniques (e.g., the EM algorithm and modifications of it), although only the AR(1) model is typically available in canned mixed model routines (e.g, proc Mixed in SAS and lmm in R).

For normal autoregressive (and more generally ARMA models) there is another quick and easy option. If the model for true values is ARMA(p,p) and the measurement error model is an $MA(q)$ process ($q = 0$ corresponding to $u_t$'s being iid $N(0, \sigma_u^2)$), then it is well known that the model for $W$ is an ARMA(p,p + q) and the autoregressive parameters are unchanged (Box et al. 1994, Sect. A4.3, Ives et al. 2010). This means we could estimate the autoregressive coefficients in $\boldsymbol{\phi}$ by simply fitting an ARMA$(p,q)$ model. This is the hybrid approach of Wong and Miller (1990) used in an ARIMA setting. For the $AR(p)$ models with measurement errors being iid $N(0, \sigma_u^2)$, this means we can estimate the autoregressive coefficients in simply fitting an ARMA$(p,p)$ model. In the case of $p = 1$ this yields the ML estimator of $\boldsymbol{\phi}_1$ and using results on fitting ARMA models (see Brockwell and Davis (2002)) an exact, but somewhat complicated, expression for the asymptotic variance of $\hat{\phi}_{1,ML}$ can be obtained; see SB (2005). When $p > 1$ this does not yield the ML estimate for $\boldsymbol{\phi}$, since there are restrictions involving the moving average parameters. There is an older literature that attacked this problem by first fitting the ARMA model and then bringing in the restrictions in various ways to get approximate MLEs (e.g., Lee and Shin (1997); Pagano (1974)).

*General Models.* The computational challenges are more severe for nonlinear models both in the integration required to obtain the likelihood, and in getting the information matrix or other quantities for use in inference. Note that the full likelihood function $L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}|\mathbf{w})$ also requires specifying a marginal distribution for $Y_1$, while in working with the conditional likelihood $\mathbf{y}_1^*$ is treated as

another parameter. This would seem to argue more for the use of the conditional approach. For overviews and references, see De Valpine and Hastings (2002), De Valpine (2002), De Valpine and Hilborn (2005), and Wang (2007) (who assumes the measurement error parameters are known but the basic algorithms are the same). Recently, some new methods have been developed to try and overcome some of the computational challenges. These include the Monte Carlo Kernel Likelihood (MCKL) (De Valpine (2004)), a method called data cloning, which borrows from Bayesian computing (Lele et al. (2007); Ponciano et al. (2009)) and composite maximum likelihood estimation (ComML) (Lele 2006).

### 4.2.2 Pseudo Likelihood Methods

The pseudo ML estimates maximize either $L(\boldsymbol{\phi},\boldsymbol{\sigma},\hat{\boldsymbol{\theta}}|\mathbf{w})$ or $L_2(\boldsymbol{\phi},\boldsymbol{\sigma},\hat{\boldsymbol{\theta}},\mathbf{y}_1^*|\mathbf{w})$, or an REML modified version, where $\hat{\boldsymbol{\theta}}$ contains estimated, or assumed known, measurement error parameters. These methods also have a long history of use in modeling with repeated samples surveys where the true values are ARIMA (and special cases thereof) or follow a basic structural model (BSM) and the sampling error also may follow a dynamic model (AR, MA, etc.), which is first estimated and then held fixed. See Bell and Wilcox (1993); Koons and Foutz (1990); Wong and Miller (1990); Miazaki and Dorea (1993); Lee and Shin (1997); Feder (2001), and references therein. Many pseudo-approaches treat stationary normal models assuming $\mathbf{W} \sim N(\mu\mathbf{1}, \boldsymbol{\Sigma}_Y + \hat{\boldsymbol{\Sigma}}_u)$, but see the caution in the previous section about non-normality of $\mathbf{W}$ if the measurement error covariance change with $Y_t$.

For the most part, the pseudo-likelihood approaches face the same computational demands as the non-pseudo-likelihood methods. An added challenge lies in finding the covariance matrix for the estimated dynamic parameters which accounts for uncertainty in $\hat{\boldsymbol{\theta}}$. To illustrate, let $\omega' = (\boldsymbol{\phi}', \boldsymbol{\sigma}')$ for the collection of dynamic parameters and $I(\boldsymbol{\omega})$ the corresponding information matrix, with submatrices $I_\phi$, etc. If $\hat{\boldsymbol{\theta}}$ is treated as fixed, the asymptotic covariance matrix of $\hat{\boldsymbol{\omega}}_{ML}$ is $I(\boldsymbol{\omega})^{-1}$, leading to an asymptotic covariance matrix of $\hat{\boldsymbol{\phi}}_{ML}$ of $Acov(\hat{\boldsymbol{\phi}}_{ML,K}) = (I_\phi - I_{\phi,\sigma} I_\sigma^{-1} I'_{\phi,\sigma})^{-1}$, $K$ for known. This covariance can be estimated in standard fashion, computational issues aside.

What about accounting for the uncertainty in $\hat{\boldsymbol{\theta}}$, say with covariance $\boldsymbol{\Sigma}_{\hat{\theta}}$? This part has been essentially ignored, an exception being SB (2005). If $\boldsymbol{\theta}$ is of fixed dimension and $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal with covariance matrix $\boldsymbol{\Sigma}_{\hat{\theta}}$, then as shown by Parke (1986), often

$$Acov(\hat{\boldsymbol{\omega}}_{ML}) = I(\boldsymbol{\omega})^{-1} + I(\boldsymbol{\omega})^{-1} I_{\omega,\theta} \boldsymbol{\Sigma}_{\hat{\theta}} I'_{\omega,\theta} I(\boldsymbol{\omega})^{-1}. \tag{10}$$

Hence, $Acov(\hat{\boldsymbol{\phi}}_{ML}) = Acov(\hat{\boldsymbol{\phi}}_{ML,K}) + \mathbf{Q}$, where $\mathbf{Q}$ is the upper left $p \times p$ block of the second matrix in (10) and $p$ is the size of $\boldsymbol{\phi}$. If $\hat{\theta} = \hat{\sigma}_u^2$, then $\boldsymbol{\Sigma}_{\hat{\theta}}$ is the exact or approximate variance of $\hat{\sigma}_u^2$. SB (2005) used this result for the AR(1) model with additive constant measurement error variance and compared the asymptotic variance

of the pseudo-MLE and that of the CEE estimator in Sect. 4.1. As expected, under normality the pseudo-MLE is more efficient but in many cases the moment-based CEE did not lose much and in finite sample simulations the performances were similar.

The expression in (10) depends on $\hat{\boldsymbol{\theta}}$ being asymptotically uncorrelated with $\hat{\boldsymbol{\omega}}$ (if computed at the true $\boldsymbol{\theta}$). This certainly holds if $\hat{\boldsymbol{\theta}}$ is independent of $\mathbf{W}$. Besides possibly violating these conditions on $\hat{\boldsymbol{\theta}}$, we also need to worry about the case where the pseudo method uses the individual measurement error parameters (e.g., the $\hat{\sigma}^2_{ut}$'s) and not a simple function of them, such as the mean, so $\hat{\boldsymbol{\theta}}$ increases in dimension as $T$ increases. Suppose that $\hat{\boldsymbol{\theta}}$ can be written as solving equations $S_2(\boldsymbol{\theta}) = \mathbf{0}$ and the $\boldsymbol{\omega}$ arises from solving $S_1(\boldsymbol{\omega}, \hat{\boldsymbol{\theta}}) = \mathbf{0}$; e.g, score equations. Note that both $S_1$ and $S_2$ depend on random quantities which have been suppressed in the notation. Use of a standard first order expansion of the estimating equations and results on the inverse of a partitioned matrix lead to an approximate covariance matrix: $Cov(\hat{\boldsymbol{\omega}}) \approx Acov(\hat{\boldsymbol{\omega}}_{ML,K}) + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\mathbf{H}_{12}'\mathbf{H}_{11}^{-1'} - \mathbf{P}$, where $\mathbf{H}_{11} = E(\partial \mathbf{S}_1(\boldsymbol{\omega}, \boldsymbol{\theta})/\partial \boldsymbol{\omega})$, $\mathbf{H}_{12} = E(\partial \mathbf{S}_1(\boldsymbol{\omega}, \boldsymbol{\theta})/\partial \boldsymbol{\theta})$, $\mathbf{H}_{22} = E(\partial \mathbf{S}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta})$, $\mathbf{C}_{12} = E(\mathbf{S}_1\mathbf{S}_2')$, and $\mathbf{P} = \mathbf{H}_{11}^{-1}(\mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{C}_{12}' + \mathbf{C}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{12}')\mathbf{H}_{11}^{-1}$. This is a bit daunting and we are faced with many of the same issues faced in using the CEE estimator in the linear autoregressive models; see Sect. 4.1. This is a case where bootstrapping will help. Also, the approximate covariance above comes from simply using a first order approximation to the estimating equations. Work remains to be done, however, to carefully examine asymptotics in this setting where the size of $\hat{\boldsymbol{\theta}}$ is increasing with $T$. Notice that treating the asymptotics for the CEE estimator was easier since only the average estimated measurement variance was used.

## 4.3 Bayesian Methods

Bayesian methods begin with the same structure as the likelihood methods above but utilize priors for the parameters ($\boldsymbol{\phi}, \boldsymbol{\sigma}$, and $\boldsymbol{\theta}$) and base inferences on the posterior distribution of the parameters with the main focus being on estimating $\boldsymbol{\phi}$. The formulas are standard so we won't repeat them here. Some of the computational challenges are similar to those in the likelihood setting, but there are some formulations that can be easily fit using Winbugs (see, e.g., Bolker (2008, Section 11.6.2) and Viljugrein et al. (2005)) General discussion of the Bayesian approach can be found in Calder et al. (2003); Clark and Bjornstad (2004); Wang (2007), and Jungbacker and Koopman (2007). Viljugrein et al. (2005) use a pseudo-approach by using estimated measurement error variances from each point in time. Clark and Bjornstad (2004) is notable for its treatment of unequal measurement error variances by incorporating different priors on each of the $\sigma^2_{utc}$ to reflect the amount of information about them. Additional applications of the Bayesian method can be found in Stenseth et al. (2003); Saether et al. (2008), and Lillegard et al. (2008), among others.

## 4.4 SIMEX, MEE, and RC

Here we briefly discuss three other correction methods, Simex (simulation-extrapolation), modifying estimating equations and regression calibration, all of which have been successful in treating standard regression settings; see Carroll et al. (2006) and Buonaccorsi (2010) for background and details. All of three use information about the measurement error parameters and are designed to both ease the computational burden and, more importantly, relax some of the distributional assumptions underlying likelihood and Bayesian techniques.

*SIMEX.* Originally due to Stefanski and Cook (1995), it has been used mainly with additive measurement error but can also accommodate multiplicative errors (e.g, Solow (1998); Resendes (2011)). Briefly it proceeds by simulating different amounts of additional measurement error to the observed $W$'s, estimating the mean behavior of the naive approach at each of these levels of measurement error (by simulating multiple samples at that level of measurement error), then fitting a curve relating the mean behavior to the level of measurement error and projecting back to the case of no measurement error. It has seen some, but rather, limited use in dynamic settings; see Solow (1998); Ellner et al. (2002), and Bolker (2008, Chap. 11). While certainly not bullet proof, Simex has proven itself to perform quite well across a variety of regression models. Its great advantage is the need to only have to be able to fit the naive estimator. Getting analytical standard errors is more challenging and has not been examined in dynamic contexts. This is another place where the bootstrap will come in handy. Resendes (2011) evaluated the performance of SIMEX in fitting the Ricker model via simulation and obtained bootstrap standard errors and confidence intervals. He found that except for large (and unreasonable) levels of measurement error, SIMEX was quite successful in removing bias and bootstrap-based inferences performed fairly well. One problem with SIMEX in combination with the bootstrap was the huge number of fits that need to be done. This was easy for the version of the Ricker model that leads to linear least squares but was more problematic when needing to use root finding methods for solving nonlinear equations for the multiplicative version.

*Modified estimating equations.* This is closely related to finding corrected scores and is also motivated by minimizing distributional assumptions. The idea is to use the estimated measurement error parameters and try to modify the naive estimating equations so the corrected equations have asymptotic mean $\mathbf{0}$. For linear autoregressive models modifying the Yule-Walker equation leads to the CEE estimator while under normality modifying the score equations leads to pseudo-ML estimators. For many other cases however, it is difficult to implement this method for the same reasons associated with using the estimating equations to assess bias (Sect. 3), although approximate corrections can be found for least squares-type estimators. An advantage of an explicit set of corrected estimating equations is the ability to build off of them to get analytical expressions for the approximate covariance matrix of the estimators. While a promising approach in

general, an extensive investigation into its use for the Ricker model (the easiest of the "nonlinear" models) by Resendes (2011) found that the resulting estimator could be erratic, was outperformed by SIMEX and it was difficult to get good standard errors, either analytically or via the bootstrap. Further fine-tuning of the correction term might alleviate some of these issues.

*Regression calibration.* Finally we briefly speculate on regression calibration, an extremely popular method in regression contexts, which has not yet explored at all in the dynamic contexts. For an autoregressive model, consider

$$\begin{aligned} E(W_t|\mathbf{W}_{t-1}) &= E[E(W_t|\mathbf{W}_{t-1},\mathbf{Y}_{t-1}] = E[E(W_t|\mathbf{Y}_{t-1}|\mathbf{W}_{t-1})] \\ &= E[m(\boldsymbol{\phi},\mathbf{Y}_{t-1})|\mathbf{W}_{t-1}] \approx m(\boldsymbol{\phi},E(\mathbf{Y}_{t-1}|\mathbf{W}_{t-1})), \end{aligned}$$

where the approximation is exact if the model is linear in the $Y$'s. (The last step of running the expectation through the $m$ function is what motivated RC methods in regression.) This suggests finding an estimate $\hat{\mathbf{Y}}_{t-1}$ of $E(\mathbf{Y}_{t-1}|\mathbf{W}_{t-1})$ and then estimating $\boldsymbol{\phi}$ by regressing $W_t$ on $\hat{\mathbf{Y}}_{t-1}$. Notice that this is *not* the same as running the usual naive analysis but replacing $W_t$ with $\hat{Y}_t$ since we are leaving $W_t$ as is when it is the "outcome" but modifying "predictors" by using $\hat{\mathbf{Y}}_{t-1}$ in place of $\mathbf{W}_{t-1}$. For normal stationary models and using estimated best linear predictors, it can be shown that this leads essentially to the CEE estimator in Sect. 4.1. A fruitful line of future work would be to examine the procedure above for nonlinear but stationary models and also to consider modifications to handle non-stationary models without an explicit expression for $E(Y_t)$ and $V(Y_t)$ which enter the best linear predictor of $Y_t$.

## 4.5 Bootstrapping

The preceding discussions provide a number of reasons why the bootstrap will be useful, both for getting standard errors and for assessing bias. The parametric bootstrap, based on an assumed distribution for both the measurement errors and the true values is relatively easy to implement. For an autoregressive model depending on the past $p$ values, for each bootstrap sample $b$ (= 1 to $B$), we can set $y_{b1} = (y_{b1},\ldots,y_{bp})'$, where $y_{bj} = W_j$ for $j = 1$ to $p$. We would then generate (sequentially) $Y_{bt} = m(\hat{\boldsymbol{\phi}},\mathbf{y}_{b,t-1}) + e_{bt}$, where $e_{bt}$ is based on the distribution of $\varepsilon_t$ with estimated parameters. Measurement error is then added to generate $\mathbf{W}_b$ from $\mathbf{y}_b$ according to the estimated model for $\mathbf{W}|\mathbf{y}$. If a pseudo-method is being used and we want to account for uncertainty from estimating the measurement error parameters, then we also would generate $\hat{\boldsymbol{\theta}}_b$, based on a distributional assumption. For each bootstrap sample, the corrected estimators are obtained and standard bootstrap inferences obtained using the $B$ bootstrap values.

Notice that we need to resample from the dynamic model explicitly. (There are some methods that bootstrap via block resampling, e.g., but these are of limited

value with short series and their use needs to be carefully considered if there are measurement errors present with changing properties over time). The difficulty in using a nonparametric bootstrap is getting an estimate of the process error distribution; i.e., the distribution of the $\varepsilon_t$. To see the problem suppose we knew the dynamic parameters exactly and examine residuals $r_t = W_t - m(\boldsymbol{\phi}, W_{t-1}) = Y_t + u_t - m(\boldsymbol{\phi}, \mathbf{Y}_{t-1}) + (m(\boldsymbol{\phi}, \mathbf{Y}_{t-1}) - m(\boldsymbol{\phi}, W_{t-1}) = \varepsilon_t + u_t + \varepsilon_t + (m(\boldsymbol{\phi}, \mathbf{Y}_{t-1}) - m(\boldsymbol{\phi}, W_{t-1})$. This is contaminated by the measurement errors and some type of "unmixing" (related to deconvolution) is needed to get a nonparametric estimate of the distribution of $\varepsilon_t$. This is not an easy problem and is especially challenging with short series. The problem is even more difficult if the last term is nonlinear in $Y$ and exacerbated further, of course, when an estimated $\hat{\boldsymbol{\phi}}$ is used. There is no work on unmixing in this particular context and so the nonparametric bootstrap remains undeveloped here, as it still does for many regression problems with measurement error.

## 5 Discussion

The main goal here was a broad overview of modeling and methodological issues when accounting for measurement error in fitting dynamic models. Much of the work in this area has tended to focus on likelihood methods involving distributional assumptions under fairly limited measurement error models. While these may provide good approximations in some settings, in general methods that drop the distributional assumptions and/or allow for richer measurement error models are often required; as are methods that explicitly exploit estimated measurement error parameters, which may be changing with time. While some important strides have been made in addressing these problems, the only problem with a somewhat complete solution is for additive errors in the linear autoregressive models. Of course one question still to be answered thoroughly is whether using the estimated measurement error parameters always improves the situation. More generally, a number of possible approaches to correcting for measurement error were described in Sect. 4. While a few papers have compared a couple of techniques (e.g., SB (2005) and Resendes (2011)) the majority of papers assess the performance of a single method often in comparison with a naive approach which ignores measurement error. A more comprehensive understanding of the pros and cons of the different methods is still needed. Future work is needed to explore the performance and robustness of the various procedures under a variety of assumptions; in particular the robustness of likelihood-based methods to distributional violations.

We identified a number of other problems that need further attention. One is in the treatment of nonlinear models without distributional assumptions. This turns out to be a challenge both for assessing bias and correcting for measurement error, even with fairly simple measurement error models; see Sects. 3 and 4.4 and Resendes (2011). Simex in combination with the bootstrap appears to be the best option here,

but as noted in Sect. 4.4 currently only the parametric bootstrap is available. The development of the nonparametric bootstrap would be helpful, although this will be very difficult with short series. Modifying the estimating equations does not seem all that promising here, but the regression calibration approach is worthy of investigation.

Two other areas needing attention, even when working under distributional assumptions, are allowing for richer measurement error models (such as letting the measurement error variance be different at each point in time) and accounting from uncertainty from the estimated measurement error parameters. The latter is of added difficulty when the number of estimated measurement error parameters is changing in time. In that context some avenues worth pursuing are just using the average measurement error variance for each $t$ (there is theoretical justification for this if the process is stationary and the variance is changing as just a function of $Y_t$) or smoothing the variances in some way.

Lastly we note that in addition to other problems put aside for space reasons in the introduction, there is the problem of simultaneously fitting models to multiple series. This is an important topic where the use of series from many different locations which exploit the spatial structure or other assumptions about common dynamic parameters can help with the ever present problem of short series. There has been some work on measurement error in these contexts (e.g., Lillegard et al. (2008), Ives et al. (2003)), but a number of the issues raised above in treating a single series remain of interest.

## Appendix

*Assessing bias via estimating equations.* Suppose $\mathbf{Y} \sim N(\mu\mathbf{1}, \boldsymbol{\Sigma}_Y)$. The estimating equations for the ML estimators of the parameters in $\boldsymbol{\Sigma}_Y$ (say $\psi_1, \ldots, \psi_J$) can be written McCulloch et al. (2008, p. 165) as $tr(\boldsymbol{\Sigma}_Y^{-1}\mathbf{G}_j) - (\mathbf{y} - \mu\mathbf{1})'\boldsymbol{\Sigma}_Y^{-1}\mathbf{G}_j\boldsymbol{\Sigma}_Y^{-1}(\mathbf{y} - \mu\mathbf{1}) = 0$, for $j = 1$ to $J$, where $J$ is the number of parameters in $\boldsymbol{\Sigma}_Y$ and $\mathbf{G}_j = \partial\boldsymbol{\Sigma}_Y/\partial\psi_j$. Replacing $\mathbf{y}$ with $\mathbf{W}$ and taking the expected value, but denoting the arguments of the estimating equations denoted with a $*$ leads to an expected value of the $jth$ estimating equation of $E_j = tr(\boldsymbol{\Sigma}_Y^{*-1}\mathbf{G}_j^*) - tr(\boldsymbol{\Sigma}_Y^{*-1}\mathbf{G}_j^*\boldsymbol{\Sigma}_Y^{*-1}\boldsymbol{\Sigma}_W)$. If we can find $\Sigma_Y^*$ of the same form as $\boldsymbol{\Sigma}_Y$ so that each $E_j$ is 0, then the naive estimators of the parameters in $\boldsymbol{\Sigma}_Y$ are consistent for the parameters in $\boldsymbol{\Sigma}_Y^*$. Obviously $E_j$ is 0 if $\boldsymbol{\Sigma}_Y^* = \boldsymbol{\Sigma}_W$ but this only provides the asymptotic bias immediately if $\boldsymbol{\Sigma}_W$ is of the same form as $\boldsymbol{\Sigma}_Y$. If the measurement errors are additive with constant (unconditional) variance $\sigma_u^2$, then $\boldsymbol{\Sigma}_W = \boldsymbol{\Sigma}_Y + \sigma_u^2\mathbf{I}$, and we can take $\boldsymbol{\Sigma}_Y^* = \boldsymbol{\Sigma}_W$. This means the naive estimator or $\sigma_Y^2$ asymptotically estimates $\sigma_Y^2 + \sigma_u^2$ while the naive estimators of the off-diagonal covariance terms in $Y$ are correct and the ML estimators are asymptotically like the YW estimators.

Allowing unequal unconditional variances (as can occur with changing sampling effort) $\boldsymbol{\Sigma}_W = \boldsymbol{\Sigma}_Y + Diag(\sigma_{u1}^2, \ldots, \sigma_{uT}^2)$ the question, which we have not investigated, is whether $E_j \to 0$ as $T$ increases if we take $\boldsymbol{\Sigma}_Y^* = \boldsymbol{\Sigma}_Y + \sigma_u^2$, where $\sigma_u^2 = \sum_{t=1}^T \sigma_{ut}^2/T$.

**Acknowledgments** I wish to express my appreciation to Brajendra Sutradhar for organizing the International Symposium in Statistics on Longitudinal Data Analysis and for serving as editor for the written contributions to this volume. I would also like to thank the members of the audience at the conference for helpful feedback and particularly thank the two referees whose comments and corrections helped improve the initial draft of this paper. Portions of this work grew out of earlier joint work with John Staudenmayer, which was supported by NSF-DMS-0306227, and with David Resendes. I am grateful to David for allowing the use of some results here from his dissertation. This paper also grew out of a seminar presented at the Center for Ecological and Evolutionary Synthesis (CEES), at the University of Oslo, in spring of 2010. I am very grateful to Nils Stenseth and CEES for support during my stay there.

## References

Aigner, D., Hsiao, C., Kapteyn, A., Wansbeek, T.: Latent variable models in econometrics. In: Griliches, Z., Intriligator, M.D. (eds.) Handbook of Econometrics, pp. 1321–1393. Elsevier, Amsterdam (1984)

Barker, D., Sibly, R.M.: The effects of environmental perturbation and measurement error on estimates of the shape parameter in the theta-logistic model of population regulation. Ecol. Model. **219**, 170–177 (2008)

Bell, W.R., Wilcox, D.W.: The effect of sampling error on the time series behavior of consumption data. J. Econometrics. **55**, 235–265 (1993)

Berliner L.M.: Likelihood and Bayesian prediction of chaotic systems. J. Am. Stat. Assoc. **86**, 938–952 (1991)

Bolker, B.: Ecological Models and Data in R. Princeton University Press, Princeton, NJ (2008)

Box, G., Jenkins, G., Reinsel, G.: Time Series Analysis, Forecasting and Control, 3rd edn. Prentice Hall, Englewood Cliffs, NJ (1994)

Brockwell, P., Davis, R.: Introduction to Time Series and Forecasting. Springer, New York (2002)

Buonaccorsi, J.P.: Measurement Error, Models, Methods, and Applications. Chapman and Hall, London (2010)

Buonaccorsi, J.P., Staudenmayer, J., Carreras, M.: Modeling observation error and its effects in a random walk/extinction model. Theor. Popul. Biol. **70**, 322–335 (2006)

Buonaccorsi J,P., Staudenmayer, J.: Statistical methods to correct for observation error in a density-independent population model. Ecol. Monogr. **79**, 299–324 (2009)

Buonaccorsi, J.P., Staudenmayer, J.: Measurement error in linear autoregressive models II: further results and inferences. Working Paper, University of Massachusetts (2012)

Burr, T., Chowell, G.: Observation and model error effects on parameter estimates in susceptible-infected-recovered epidemic model. Far East J. Theor. Stat. **19**, 163–183 (2006)

Calder, C., Lavine, M., Muller, P., Clark, J.: Incorporating multiple sources of stochasticity into dynamic population models. Ecology **84**, 1395–1402 (2003)

Carroll, R.J., Stefanski, L.A., Ruppert, D., Crainiceanu C.M.: Measurement Error in Nonlinear Models, 2nd edn. Chapman and Hall, London (2006)

Chanda, K.C.: Asymptotic properties of estimators for autoregressive models with errors in variables. Ann. Stat. **24**, 423–430 (1996)

Cheang, W., Reinsel, G.C.: Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood. J. Am. Stat. Assoc. **95**, 1173–1184 (2000)

Clark, J.S., Bjornstad, O.N.: Population time series, process variability, observation errors, missing values, lags, and hidden states. Ecology. **85**, 3140–3150 (2004)

Dennis, B., Ponciano, J., Lele, S., Taper, M., Staples, D.: Estimating density dependence, process noise, and observation error. Ecol. Monogr. **76**, 323–341 (2006)

Dennis, B., Ponciano, J.M., Taper, M.L.: Replicated sampling increases efficiency in monitoring biological populations. Ecology **91**, 610–620 (2010)

De Valpine, P.: Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear non-Gaussian state-space models. Bull. Mar. Sci. **70**, 455–471 (2002)
De Valpine, P.: Monte-Carlo state-space likelihoods by weighted posterior kernel density estimation. J. Am. Stat. Assoc. **99**, 523–536 (2004)
De Valpine, P., Hastings, A.: Fitting population models incorporating process noise and observation error. Ecol. Monogr. **72**, 57–76 (2002)
De Valpine, P., Hilborn, R.: State-space likelihoods for nonlinear fisheries time series. Can. J. Fish. Aquat. Sci. **62**, 1937–1952 (2005)
Ellner, S., Yodit, S., Smith, R.: Fitting population dynamic models to time-series by gradient matching. Ecology **83**, 2256–2270 (2002)
Feder, M.: Time series analysis of repeated surveys: the state-space approach. Stat. Neerl. **55**, 182–199 (2001)
Harvey, A.C.: Forecasting, Structural Time Series Models, and the Kalman Filter. Cambridge University Press, Cambridge (1990)
Fuller, W.: Time Series Analysis. Wiley, New York (1996)
Hovestadt, T., Nowicki, P.: Process and measurement errors of population size, their mutual effects on precision and bias of estimates for demographic parameters. Biodivers. Conserv. **17**, 3417–3429 (2008)
Ives, A.R., Dennis, B., Cottingham, K.L., Carpenter, S.R.: Estimating community stability and ecological interactions from time-series data. Ecol. Monogr. **73**, 301–330 (2003)
Ives, A.R, Abbott, K., Ziebarth, N.: Analysis of ecological time series with ARMA(p,q) models. Ecololgy **91**, 858–871 (2010)
Jungbacker, B., Koopman, S.J.: Monte Carlo estimation for nonlinear non-Gaussian state space models. Biometrika **94**, 827–839 (2007)
Knape, J.: Estimability of density dependence in models of time series data. Ecology **89**, 2994–3000 (2008)
Knape, J., Jonzén, N., Skold, M.: Observation distributions for state space models of population survey data. J. Anim. Ecol. **80**, 1269–1277 (2011)
Koons, B.K., Foutz, R.V.: Estimating moving average parameters in the presence of measurement error. Comm. Stat. **19**, 3179–3187 (1990)
Lee, J.H., Shin, D.W.: Maximum likelihood estimation for ARMA models in the presence of ARMA errors. Comm. Stat. Theor. Meth. **26**, 1057–1072 (1997)
Lele, S.R.: Sampling variability and estimates of density dependence, a composite-likelihood approach. Ecology **87**, 189–202 (2006)
Lele, S.R., Dennis, B., Lutscher, F.: Data cloning, easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol. Lett. **10**, 551–563 (2007)
Lillegard, M., Engen, S., Saether, B.E., Grotan, V., Drever, M.: Estimation of population parameters from aerial counts of North American mallards: a cautionary tale. Ecol. Appl. **18**, 197–207 (2008)
Mallick, T., Sutradhar, B.: GQL versus conditional GQL inferences for non-stationary time series of counts with overdispersion. J. Time Anal. **29**, 402–420 (2008)
Miazaki, E.S., Dorea, C.C.Y.: Estimation of the parameters of a time series subject to the error of rotation sampling. Commun. Stat. A - Theor. Meth. **22**, 805–825 (1993)
McCulloch, C., Searle, S., Neuhaus, J. : Generalized, Linear, and Mixed Models, 2nd edn. Wiley, New York (2008)
Morris, W.F., Doak, D.F.: Quantitative Conservation Biology: Theory and Practice of Population Variability Analysis. Sinauer Associates, Sunderland, MA (2002)
Pagano, M.: Estimation of models of autoregressive signal plus white noise. Ann. Stat. **2**, 99–108 (1974)
Parke, W.: Pseudo maximum likelihood estimation: the asymptotic distribution. Ann. Stat. **14**, 355–357 (1986)

Pfeffermann, D., Feder, M., Signorelli, D.: Estimation of autocorrelations of survey errors with application to trend estimation in small areas. J. Bus. Econ. Stat. **16**, 339–348 (1998)
Ponciano, J., Taper, M., Dennis, B., Lele, S.: Hierarchical models in ecology, confidence intervals hypothesis testing, and model selection using data cloning. Ecology **90**, 356–362 (2009)
Resendes, D.: Statistical methods for nonlinear dynamic models with measurement error using the Ricker model. Ph.D. thesis, University of Massachusetts, Amherst (2011)
Sakai, H., Soeda, T., Hidekatsu, T.: On the relation between fitting autoregression and periodogram with applications. Ann. Stat. **7**, 96–107 (1979)
Saether, B., Lilligard, M., Grotan, V., Drever, M., Engen, S., Nudds, T., Podruzny, K.: Geographical gradients in the population dynamics of North American prairie ducks. J. Anim. Ecol. **77**, 869–882 (2008)
Schmid, C.H., Segal, M.R., Rosner, B.: Incorporating measurement error in the estimation of autoregressive models for longitudinal data. J. Stat. Plann. Infer. **42**, 1–18 (1994)
Solow, A.R.: On fitting a population model in the presence of observation error. Ecology **79**, 1463–1466 (1998)
Solow, A.R.: Observation error and the detection of delayed density dependence. Ecology **82**, 3263–3264 (2001)
Staudenmayer, J., Buonaccorsi, J.P.: Measurement error in linear autoregressive models. J. Am. Stat. Assoc. **100**, 841–852 (2005)
Stefanski, L.: The effects of measurement error on parameter estimation. Biometrika **72**, 583–592 (1985)
Stefanski, L., Cook, J.: Simulation-extrapolation: the measurement error jackknife. J. Am. Stat. Assoc. **90**, 1247–1256 (1995)
Stenseth, N.C., Viljugrein, H., Saitoh, T., Hansen, T.F., Kittilsen, M.O., Bolviken, E., Glockner, F.: Seasonality, density dependence, and population cycles in Hokkaido voles. Proc. Natl. Acad. Sci. USA. **100**, 11478–11483 (2003)
Tripodis, Y., Buonaccorsi, J.P.: Prediction and forecasting in linear models with measurement error and unknown parameters. J. Stat. Plann. Infer. **139**, 4039–4050 (2009)
Viljugrein, H., Stenseth, N.C., Smith, G.W., Steinbakk, G.H.: Density dependence in North American ducks. Ecology **86**, 245–254 (2005)
Walker, A.M.: Some consequences of superimposed error in time series analysis. Biometrika **47**, 33–43 (1960)
Wang, G.: On the latent state estimation of nonlinear population dynamics using Bayesian and non-Bayesian state-space models. Ecol. Model. **20**, 521–528 (2007)
Wang, G., Hobbs, N.T., Boone, R.B., Illius, A.W., Gordon, I.J., Gross, J.E., Hamlin, K.L.: Spatial and temporal variability modify density dependence in populations of large herbivores. Ecology **87**(1), 95–102 (2006)
Williams, C.K., Ives, A.R., Applegate, R.D.: Population dynamics across geographical ranges, time-series analyses of three small game species. Ecology **84**, 2654–2667 (2003)
Wong, W-K., Miller, R.B.: Repeated time series analysis of ARIMA-noise models. J. Bus. Econ. Stat. **8**, 243–250 (1990)
Wong, W-K., Miller, R.B., Shrestha, K.: Maximum likelihood estimation of ARMA models with error processes for replicated observations. J. Appl. Stat. Sci. **10**, 287–297 (2001)

# Measurement Error in the Linear Dynamic Panel Data Model

**Erik Meijer, Laura Spierdijk, and Tom Wansbeek**

**Abstract** We study measurement error in the simplest dynamic panel data model without covariates. We start by investigating the first-order effects, on the most commonly used estimator, of the presence of measurement error. As was to be expected, measurement error renders this estimator inconsistent. However, with a slight adaptation, the estimator can be made consistent. This approach to consistent estimation is ad hoc and we next develop a systematic approach to consistent estimation. We show how to obtain the most efficient estimator from this class of consistent estimators. We illustrate our findings through an empirical example.

## 1 Introduction

In econometrics, the analysis of panel data is a rapidly expanding research area. Frequently, the models formulated are dynamic models in the sense that the lagged dependent variables is among the regressors. Especially the linear dynamic panel data model (LDPDM) is hugely popular. This model typically has both the lagged dependent variable and an individual effect on the right-hand side. Its estimation is not entirely straightforward since the least-squares estimator is inconsistent.

E. Meijer
RAND Corporation, Santa Monica, CA 90401, USA
e-mail: meijer@rand.org

L. Spierdijk • T. Wansbeek (✉)
University of Groningen, 9700 AV Groningen, The Netherlands
e-mail: l.spierdijk@rug.nl; t.j.wansbeek@rug.nl

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_4, 

Consistent estimation of the LDPDM has inspired many researchers, and the number of publications on the topic is still growing. The leading idea is to transform the model equation into first differences over time, and next use the twice lagged-dependent variable as an instrumental variable (IV). Under the crucial assumption that the error term is not correlated over time, it is easy to see that this approach gives a consistent estimator of the regression coefficient. The idea is due to Anderson and Hsiao (1981, 1982). Arellano and Bond (1991) pointed out that all preceding values of the dependent variable, not just the directly preceding one, can be used as IVs, leading to more IVs and smaller asymptotic variance, if there are more than three periods. The estimator due to Arellano and Bond (1991) has found application on a very large scale.

Little attention has been paid to issues around measurement error in the LDPDM. This comes not entirely as a surprise, as the situation for the much simpler static model is not vastly more favorable. But at least, there is a line of literature for the static model. The pioneering contribution there is Griliches and Hausman (1986). Much of the literature for the static model is reviewed by Meijer et al. (2012), where also a number of new ways are described for consistent estimation.

As to the literature on the LDPDM with measurement error, an early contribution is Wansbeek and Kapteyn (1992). For the model without exogenous regressors, they derive the probability limit of the within estimator and the OLS estimator after first-differencing the model and suggest to use the result to construct a consistent estimator of the autoregressive parameter. In an empirical study on income dynamics, Antman and McKenzie (2007) consider a dynamic panel data model where the current value depends on a cubic function of the lagged value. Their estimator is based on outside information on the reliability of the income variable, that is, on the ratio of the true variance and the observed variance. Chen et al. (2008), in their study of the dynamics of students' test scores, construct consistent estimators through IVs derived from within the model, adapting an approach due to Altonji and Siow (1987) to the dynamic case. Komunjer and Ng (2011) consider a VARX model with all variables contaminated by measurement error and exploit the dynamics of the model for consistent estimation. Biørn (2012) presents a thorough treatment of the topic, with IVs based on the absence of correlation between regressors and disturbances for some combinations of time indices.

In this paper we contribute to the literature on the LDPDM with measurement error in two ways. In the first place we derive, in Sect. 2, the effect of measurement error on the Arellano–Bond (AB) estimator. For the simple case of a panel with three waves, we investigate the inconsistency of this estimator in the presence of measurement error. We provide some interpretation and further elaboration in Sect. 3. Next, we move over to consistent estimation. In Sect. 4 we consider a wide class of estimators that are consistent, and in Sect. 5, we study efficiency within this class. An illustrative example is given in Sect. 6. In Sect. 7 we make some concluding remarks.

## 2 The Effect of Measurement Error

In this section we consider the simplest possible LDPDM and investigate the effect of measurement error when it is estimated in the usual way. The model is represented by the following two equations,

$$\eta_{nt} = \gamma\eta_{n,t-1} + \alpha_n + \varepsilon_{nt} \tag{1}$$

$$y_{nt} = \eta_{nt} + v_{nt}, \tag{2}$$

for $n = 1,\ldots,N$ and $t = 1,\ldots,T$. In this model, $\eta_{nt}$ is an unobserved variable, according to (1) subject to an autoregressive process of order one. The error term in (1) consists of two components, a time-constant one, $\alpha_n$, and a time-varying one, $\varepsilon_{nt}$. The link between the unobserved variable $\eta_{nt}$ and the observed variable $y_{nt}$ is given by the measurement equation (2), where $v_{nt}$ represents the measurement error. All variables have mean zero over $n$, possibly after demeaning per time period thus accounting for fixed time effects. The parameter of interest in the autocorrelation parameter $\gamma$. We restrict ourselves to the case where $-1 < \gamma < 1$.

It is assumed that $\alpha_n$, $\varepsilon_{nt}$, and $v_{nt}$ are uncorrelated over $n$. Moreover, $\varepsilon_{nt}$ and $v_{nt}$ are assumed uncorrelated over $t$. This is quite a simplification but, somewhat surprisingly, these assumptions are commonly made. The various error terms are taken homoskedastic, with means zero and variances $\sigma_\alpha^2$, $\sigma_\varepsilon^2$, and $\sigma_v^2$, respectively. As is usual in econometrics, these parameters, in particular the absolute or relative measurement error variance, are taken to be unknown. Finally, it is assumed that the process has been going on since minus infinity, and $-1 < \gamma < 1$, so that the distributions of all variables are stationary.

We take $N$ to be large relative to $T$ and hence, in our asymptotic results, keep $T$ fixed and let $N$ go to infinity. So our perspective is cross-sectional. For a time-series perspective on measurement error, see Aigner et al. (1984, Sect. 6) and Buonaccorsi (2010, Chap. 12).

Even if $\eta_{nt}$ would be observed, estimation of (1) is not straightforward as (1) implies that $\alpha_n$ is correlated with $\eta_{n\tau}$ for all $\tau$, including the case $\tau = t-1$. So, since $\eta_{n,t-1}$ is the regressor, the regression model (1) has an error term that is correlated with the regressor. Hence least squares gives an inconsistent result for the parameter of interest, $\gamma$.

The Anderson–Hsiao (AH) estimator (Anderson and Hsiao 1981, 1982) first transforms (1) into first differences over time, and next uses $\eta_{n,t-2}$ as an IV. Under the crucial assumption that $\varepsilon_{nt}$ is not correlated over time, it is easy to see that this approach gives a consistent estimator of $\gamma$. The Arellano–Bond (AB) estimator (Arellano and Bond 1991) is a generalization that uses all preceding values of the dependent variable, not just the directly preceding one, as IVs, leading to more instruments and smaller asymptotic variance, if the number of observed periods is larger than three. In the derivations in this section we restrict ourselves, for reasons of tractability and emphasis on the essentials, to three periods, so our estimator is the AH estimator.

This is all about the case where $\eta_{nt}$ is observed and not clouded by measurement error. If there is measurement error, so if (2) enters the stage, the consistency of the AH estimator is at peril. We now have measurement error twice in the model. It evidently enters both the dependent variable and the regressor. As is well known from the measurement error literature, measurement error in the dependent variable does not affect consistency, but measurement error in a regressor does, in most cases in the form of a bias towards zero of the estimator. Here the measurement error in both variables comes from the same source, that is, from (2), and the effect of measurement error is not straightforward.

In order to gain insight into the effect of measurement error on the AH estimator, we transform the model into first differences over time and substitute out the unobserved variable $\eta$ from the model. This gives us

$$y_{nt} - y_{n,t-1} = \gamma(y_{n,t-1} - y_{n,t-2}) + u_{nt}, \tag{3}$$

where the error term $u_{nt}$ is defined as

$$u_{nt} \equiv (\varepsilon_{nt} - \varepsilon_{n,t-1}) + (v_{nt} - v_{n,t-1}) - \gamma(v_{n,t-1} - v_{n,t-2}).$$

The AH estimator of $\gamma$ is obtained by estimating (3) with $y_{n,t-2}$ as the IV. In the presence of measurement error this estimator is not consistent, because the IV is not valid as it is not orthogonal to the error term in (3):

$$\begin{aligned} \mathrm{E}(y_{n,t-2}u_{nt}) &= \mathrm{E}\{(\eta_{n,t-2} + v_{n,t-2})[(\varepsilon_{nt} - \varepsilon_{n,t-1}) \\ &\quad + (v_{nt} - v_{n,t-1}) - \gamma(v_{n,t-1} - v_{n,t-2})]\} \\ &= \gamma\sigma_v^2. \end{aligned}$$

In order to derive the probability limit of the AH estimator, we first notice that, through repeated substitution, the unobserved variable can be expressed as

$$\eta_{nt} = \frac{\alpha_n}{1-\gamma} + \sum_{s=0}^{\infty} \gamma^s \varepsilon_{n,t-s}. \tag{4}$$

Hence

$$\mathrm{E}(\eta_{nt}\eta_{n,t-\tau}) = \frac{\sigma_\alpha^2}{(1-\gamma)^2} + \gamma^\tau \frac{\sigma_\varepsilon^2}{1-\gamma^2}. \tag{5}$$

Consequently,

$$\begin{aligned} \omega_\tau &\equiv \mathrm{E}(y_{nt}y_{n,t-\tau}) \\ &= \frac{\sigma_\alpha^2}{(1-\gamma)^2} + \gamma^\tau \frac{\sigma_\varepsilon^2}{1-\gamma^2} + I(\tau = 0)\sigma_v^2, \end{aligned} \tag{6}$$

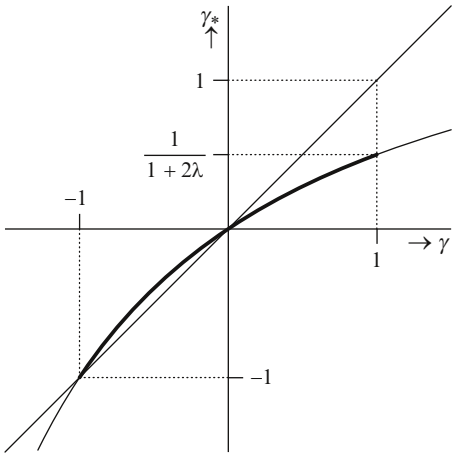


**Fig. 1** Probability limit of the Anderson–Hsiao estimator with $\lambda = \frac{1}{2}$

where $I(\cdot)$ is the indicator function, which is 1 if its argument is true and 0 otherwise. The AH estimator is given by

$$\hat{\gamma} = \frac{\frac{1}{N}\sum_n y_{n,t-2}(y_{nt} - y_{n,t-1})}{\frac{1}{N}\sum_n y_{n,t-2}(y_{n,t-1} - y_{n,t-2})}.$$

Let $\lambda \equiv \sigma_v^2/\sigma_\varepsilon^2$ be the ratio of the measurement error variance to the equation error variance. Under weak assumptions, the probability limit of the AH estimator is

$$\begin{aligned}\gamma_* &\equiv \text{plim}_{N\to\infty}\hat{\gamma} \\ &= \frac{\omega_2 - \omega_1}{\omega_1 - \omega_0} \\ &= \frac{(\gamma^2 - \gamma)\frac{1}{1-\gamma^2}}{(\gamma - 1)\frac{1}{1-\gamma^2} - \lambda} \\ &= \frac{\gamma}{1 + (1+\gamma)\lambda}. \end{aligned} \tag{7}$$

This result is depicted in Fig. 1. Clearly, measurement error causes the estimator to be biased towards zero. The bias towards zero is a well-known phenomenon from the literature on measurement error in a single cross-section.

The figure is made for the case of $\lambda = \frac{1}{2}$, so $\sigma_v^2 = \frac{1}{2}\sigma_\varepsilon^2$. With decreasing measurement error, so with decreasing $\lambda$, the hyperbola will become closer to the 45° line. It is also striking how asymmetric the biasing effect is: the bias is much larger for positive values of $\gamma$ (which are arguably more likely in most applications) than for negative values, both absolutely and in a relative sense.

## 3 Interpretation and Elaboration

Another view of the result (7) can be obtained as follows. Due to the stationarity and the absence of serial correlation in the measurement errors, and with $\Delta$ denoting the first-difference operator, we have

$$\begin{aligned}
\mathrm{E}\sum_n y_{n,t-2}(y_{nt} - y_{n,t-1}) &= \mathrm{E}\sum_n \eta_{n,t-2}(\eta_{nt} - \eta_{n,t-1}) \\
&= \gamma \mathrm{E}\sum_n \eta_{n,t-2}(\eta_{n,t-1} - \eta_{n,t-2}) \\
&= -\tfrac{1}{2}\gamma \sum_n (\eta_{n,t-1} - \eta_{n,t-2})^2 \\
&= -\tfrac{1}{2}\gamma \sigma^2_{\Delta\eta}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\sum_n y_{n,t-2}(y_{n,t-1} - y_{n,t-2}) &= -\frac{1}{2}\mathrm{E}(y_{n,t-1} - y_{n,t-2})^2 \\
&= -\tfrac{1}{2}\sigma^2_{\Delta y}.
\end{aligned}$$

We thus obtain

$$\gamma_* = \frac{\sigma^2_{\Delta\eta}}{\sigma^2_{\Delta y}}\gamma.$$

The bias factor is the "reliability" of $\Delta y$ as a proxy for $\Delta\eta$. The situation closely resembles the situation in the classical measurement error model for a single cross-section, where the same result holds but then in levels, not differences. Also, in that case the reliability does not mathematically depend on $\gamma$, because it is the reliability of the exogenous variable. In the LDPDM, it depends on $\gamma$, causing the curvature depicted in Fig. 1.

As mentioned above, the AH estimator is a special instance of the AB estimator. Nowadays, researchers often use the "systems" generalized method-of-moments (GMM) estimator (Arellano and Bover 1995; Blundell and Bond 1998), which combines the building blocks of the AB estimator with those that can be derived when the correlation between $y_{nt}$ and the individual effect $\alpha_n$ does not depend on $t$. Usually, the systems GMM estimator greatly outperforms the AB estimator. In the present setup we have from (4) that

$$\mathrm{E}(y_{nt}\alpha_n) = \frac{\sigma^2_\alpha}{1-\gamma}, \tag{8}$$

establishing this equicorrelation here. It can be used to estimate $\gamma$ in a way that in a sense is the mirror image of AB. There, previous values of $y$ are used as IV for a

model in first differences. Here, we keep the model in levels but use previous values of $y$ in first-difference form as IV. Then

$$\begin{aligned}\tilde{\gamma} &= \frac{\frac{1}{N}\sum_n (y_{n,t-1} - y_{n,t-2}) y_{nt}}{\frac{1}{N}\sum_n (y_{n,t-1} - y_{n,t-2}) y_{n,t-1}} \\ &\xrightarrow{p} \frac{\omega_1 - \omega_2}{\omega_0 - \omega_1} \\ &= \gamma_*.\end{aligned}$$

Thus, the inconsistency is the same as with the AH estimator.

So, with measurement error, we encounter the inconsistency issue well known from the cross-sectional case. A major difference, though, is that the latter case, in its simplest form of linearity, normality, and independence of observations (LIN), results in an identification problem that precludes the existence of a consistent estimator (e.g., Wansbeek and Meijer 2000, p. 79). In the LDPDM, LIN does not apply and the situation is more favorable. In fact, a consistent estimator is easily found; instead of using $y_{n,t-2}$ as an IV, we can use $y_{n,t-3}$ (assuming $T > 3$). (Bond et al. 2001, make the same observation.) We call this estimator the Anderson–Hsiao lagged (AHL) estimator. Its probability limit is

$$\begin{aligned}\operatorname{plim}_{N\to\infty} \hat{\gamma}_{\text{AHL}} &= \frac{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_n y_{n,t-3}(y_{nt} - y_{n,t-1})}{\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_n y_{n,t-3}(y_{n,t-1} - y_{n,t-2})} \\ &= \frac{\omega_3 - \omega_2}{\omega_2 - \omega_1} \\ &= \frac{(\gamma^3 - \gamma^2)\frac{1}{1-\gamma^2}}{(\gamma^2 - \gamma)\frac{1}{1-\gamma^2}} \\ &= \gamma.\end{aligned}$$

So AHL is a consistent estimator, due to the assumed lack of correlation over time of the measurement error. Analogously, the Arellano–Bond lagged (ABL) estimator is obtained by removing $y_{n,t-2}$ from the list of IVs of the Arellano–Bond estimator, which is also a consistent estimator in our setup. Arellano and Bond (1991) mention this estimator in the context of autocorrelation resulting from a moving average process in the errors, so it serves a dual purpose.

This approach to consistent estimation is of course somewhat ad hoc. Moreover, it breaks down when the measurement errors are autocorrelated. To gain some insight here, we assume that the measurement errors are subject to an autoregressive process of order one, AR(1), so

$$v_{nt} = \rho v_{n,t-1} + w_{nt},$$

with $w_{nt}$ white noise with variance $\sigma_w^2$. Instead of (6), we now have

$$\omega_\tau = \frac{\sigma_\alpha^2}{(1-\gamma)^2} + \gamma^\tau \frac{\sigma_\varepsilon^2}{1-\gamma^2} + \rho^\tau \frac{\sigma_w^2}{1-\rho^2}.$$

With $\lambda$ redefined as $\lambda \equiv \sigma_w^2/\sigma_\varepsilon^2$, the probability limit of the AH estimator now becomes

$$\begin{aligned}\gamma_* &= \frac{(\gamma^2-\gamma)\frac{\sigma_\varepsilon^2}{1-\gamma^2} + (\rho^2-\rho)\frac{\sigma_w^2}{1-\rho^2}}{(\gamma-1)\frac{\sigma_\varepsilon^2}{1-\gamma^2} + (\rho-1)\frac{\sigma_w^2}{1-\rho^2}} \\ &= \frac{\gamma(1+\rho)+\rho(1+\gamma)\lambda}{(1+\rho)+(1+\gamma)\lambda}.\end{aligned}$$

So the effect of measurement error is now more complicated. The measurement error in the dependent variable, which in the classical case has no effect on the estimator of the regression coefficient, now plays a role due to its correlation with the measurement error in the regressor. The estimator is consistent ($\gamma_* = \gamma$) if $\rho = \gamma$ or $\lambda = 0$. The estimator has a positive bias if $\rho > \gamma$ and a negative bias if $\rho < \gamma$. In the most likely case that $0 < \rho < \gamma$, we see the usual attenuation bias towards zero. Note that we can write $\gamma_* = \phi\gamma + (1-\phi)\rho$, with

$$\phi = \frac{\frac{\sigma_\varepsilon^2}{1+\gamma}}{\frac{\sigma_\varepsilon^2}{1+\gamma} + \frac{\sigma_w^2}{1+\rho}},$$

so $\gamma_*$ is a weighted average of $\gamma$ and $\rho$. Although the weights themselves depend on $\gamma$ and $\rho$, they are always between 0 and 1.

Using a similar derivation, it follows that the AHL estimator, or in general using previous values of $y$ as an IV, does not yield a consistent estimator anymore. We now turn to a more systematic approach to consistent estimation, for general values of $T$. We first investigate what consistency implies and derive a class of consistent estimators. We next consider issues of optimality.

## 4 Consistent Estimation

Our approach to consistent estimation extends Wansbeek and Bekker (1996) by taking measurement error into account. They derive an instrumental variable that is linear in the values of the dependent variable across time and that results in an IV estimator that has minimal asymptotic variance. Harris and Mátyás (2000) extend this approach to include exogenous regressors. They compare the estimator thus defined with the Arellano–Bond estimator and some estimators based on Ahn and Schmidt (1995) and find that Wansbeek and Bekker's estimator "generally outperformed all other estimators when $T$ was moderate *in all of the situations that an applied researcher might encounter*" [italics in original]. We adapt the Wansbeek

and Bekker (1996) approach also in another way, in that we assume stationarity throughout.

We now turn to the model and derive our estimator. For general $T$, it is convenient to move over to matrix notation. With observations $y_{nt}$, $n = 1,\ldots,N$, $t = 0,\ldots T$, we define

$$y_n \equiv \begin{pmatrix} y_{n1} \\ \vdots \\ y_{nT} \end{pmatrix} \quad y_{n,-1} \equiv \begin{pmatrix} y_{n0} \\ \vdots \\ y_{n,T-1} \end{pmatrix} \quad y_{n,+} \equiv \begin{pmatrix} y_{n0} \\ \vdots \\ y_{nT} \end{pmatrix}.$$

For $\eta$, $v$, and $\varepsilon$, we use analogous notation. Note that the number of observed periods is $T+1$ now, as opposed to the $T$ used before. The model can now be written as

$$\eta_n = \gamma\eta_{n,-1} + \alpha_n\iota_T + \varepsilon_n, \tag{9}$$

where $\iota_T$ is a $T$-vector of ones, and $\varepsilon_n \sim (0, \sigma_\varepsilon^2 I_T)$. The measurement equation is

$$y_n = \eta_n + v_n, \tag{10}$$

with

$$v_n = \rho v_{n,-1} + w_n,$$

where $w_n \sim (0, \sigma_w^2 I_T)$, thus allowing for measurement errors correlated over time according to an AR(1) process.

One way to estimate the model parameters consistently is through GMM. From (5) we obtain

$$\begin{aligned} \Sigma_\eta &\equiv \mathrm{E}(\eta_{n,+}\eta_{n,+}') \\ &= \frac{\sigma_\alpha^2}{(1-\gamma)^2}\iota_{T+1}\iota_{T+1}' + \frac{\sigma_\varepsilon^2}{1-\gamma^2}V_\gamma, \end{aligned}$$

where $V_\gamma$ is the AR(1) correlation matrix of order $(T+1)\times(T+1)$, that is, the matrix whose $(t,s)$th element is $\gamma^{|t-s|}$. So the second-order implication of the model for the observations, taking the measurement error into account, is

$$\begin{aligned} \Sigma_y &\equiv \mathrm{E}(y_{n,+}y_{n,+}') \\ &= \mathrm{E}\left((\eta_{n,+}+v_{n,+})(\eta_{n,+}+v_{n,+})'\right) \qquad (11) \\ &= \Sigma_\eta + \frac{\sigma_w^2}{1-\rho^2}V_\rho \\ &= \frac{\sigma_\alpha^2}{(1-\gamma)^2}\iota_{T+1}\iota_{T+1}' + \frac{\sigma_\varepsilon^2}{1-\gamma^2}V_\gamma + \frac{\sigma_w^2}{1-\rho^2}V_\rho. \qquad (12) \end{aligned}$$

From this, an essential identification problem with the model is immediately clear. The model is locally identified but not globally. The parameter set $(\sigma_\varepsilon^2, \gamma)$ can be interchanged with the parameter set $(\sigma_w^2, \rho)$ as they play a symmetric role in $\Sigma_y$, and the data do not provide sufficient information to tell which is which. Hence we restrict ourselves to the case where the measurement error has no autocorrelation and have $\rho = 0$ from now on.

The GMM estimator of the parameters is obtained by minimizing the distance between

$$\sigma_y \equiv \text{vech } \Sigma_y,$$

the vector containing the non-redundant elements of $\Sigma_y$, and its sample counterpart

$$s_y \equiv \text{vech } S_y,$$

where $S_y \equiv \sum y_{n,+} y_{n,+}'/N$. For an appropriate choice of the weight matrix in the distance function, the GMM estimator is asymptotically efficient among all estimators based on $S_y$.

A drawback of the GMM estimator in this case is that it will be cumbersome to compute as $\Sigma_y$ depends on the parameter of interest, $\gamma$, in a highly nonlinear way. Hence we consider a simpler way to obtain a consistent estimator, focusing on $\gamma$. The price for this simplicity is that this estimator does not exploit all the structure imposed by the model on $\Sigma_y$ and hence will be asymptotically inefficient.

As a start, we eliminate $\eta_n$ from the model by substitution from (10) into (9) to obtain

$$y_n = \gamma y_{n,-1} + \upsilon_n \tag{13}$$

$$\upsilon_n \equiv \alpha_n \iota_T + \varepsilon_n + v_n - \gamma v_{n,-1}. \tag{14}$$

We consider IV estimation of $\gamma$. As an IV, we consider a general linear function of $y_{n,+}$ of the form $A' y_{n,+}$ for some $(T+1) \times T$-matrix $A$. Below we will also use the form

$$a \equiv \text{vec } A.$$

Given $A$, our IV estimator of $\gamma$ is

$$\begin{aligned} \hat{\gamma} &= \frac{\sum_n y_{n,+}' A y_n}{\sum_n y_{n,+}' A y_{n,-1}} \\ &= \gamma + \frac{\sum_n y_{n,+}' A \upsilon_n}{\sum_n y_{n,+}' A y_{n,-1}}. \end{aligned} \tag{15}$$

We now investigate the conditions under which this estimator exists and, if so, if it is consistent. In order to do so, we need the following notation. Let $C_0' \equiv (I_T, 0_T)$

and let $C_1', \ldots, C_T'$ be a series of matrices of order $T \times (T+1)$, where $C_1' \equiv (0_T, I_T)$, in $C_2'$ the ones are moved one position to the right, and so on, ending with $C_T'$, which is zero, except for its $(1, T+1)$ element. For example, for $T = 3$, we have

$$C_0' = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix};\ C_1' = \begin{pmatrix} 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{pmatrix};\ C_2' = \begin{pmatrix} 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0\,0 \end{pmatrix};\ C_3' = \begin{pmatrix} 0\,0\,0\,1 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{pmatrix}.$$

Next, let

$$C \equiv (\text{vec}\, C_0, \ldots, \text{vec}\, C_T).$$

We now consider the requirements that $A$ has to satisfy.

In the first place, $A$ should be such that $\hat{\gamma}$ exists. More precisely, the expression for $\hat{\gamma}$ should be meaningful in the sense that neither numerator nor denominator is identically equal to zero. For example, if $T = 2$ and

$$A_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ and } A_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ 0 & 0 \end{pmatrix}$$

we have $y_{n,+}' A_1 y_n = 0$ and $y_{n,+}' A_2 y_{n,-1} = 0$. To exclude such cases, consider

$$y_{n,+} \otimes y_{n,+} = D_{T+1}(y_{n,+} \bar{\otimes} y_{n,+}),$$

where the bar over the Kronecker product indicates the omission of duplicate elements. The duplication matrix $D_{T+1}$, of order $(T+1)^2 \times (T+1)(T+2)/2$ restores them (see, e.g., Magnus and Neudecker 1986). Next, let

$$F_\tau \equiv (C_\tau' \otimes I_{T+1}) D_{T+1}, \quad \tau = 0, 1,$$

and note that

$$s_y = \frac{1}{N} \sum_n (y_{n,+} \bar{\otimes} y_{n,+}).$$

Using $y_{n,-1} = C_0' y_{n,+}$ and $y_n = C_1' y_{n,+}$, we can now write the estimator of $\gamma$ as

$$\begin{aligned} \hat{\gamma} &= \frac{a' \sum_n (y_n \otimes y_{n,+})}{a' \sum_n (y_{n,-1} \otimes y_{n,+})} \\ &= \frac{a'(C_1' \otimes I_{T+1}) D_{T+1} \sum_n (y_{n,+} \bar{\otimes} y_{n,+})}{a'(C_0' \otimes I_{T+1}) D_{T+1} \sum_n (y_{n,+} \bar{\otimes} y_{n,+})} \\ &= \frac{a' F_1 s_y}{a' F_0 s_y}. \end{aligned} \tag{16}$$

So for a meaningful estimator, we should have both $F_0'a \neq 0$ and $F_1'a \neq 0$.

We next turn to consistency. From (14) and (15) we see that it requires

$$\begin{aligned} 0 &= \mathrm{E}\left(y'_{n,+}A\upsilon_n\right) \\ &= \mathrm{E}\left(\alpha_n y'_{n,+}A\iota_T\right) + \mathrm{tr}\left(\mathrm{E}\left[(\varepsilon_n + v_n - \gamma v_{n,-1})y'_{n,+}\right]A\right), \end{aligned} \tag{17}$$

where tr indicates the trace. First, we have the equicorrelation property from (8),

$$\mathrm{E}(\alpha_n y_{n,+}) = c \cdot \iota_{T+1}$$

with $c = \sigma_\alpha^2/(1-\gamma)$. So one requirement for consistency is $\iota'_{T+1}A\iota_T = 0$ or

$$\iota'_{T(T+1)}a = 0. \tag{18}$$

Since

$$\mathrm{E}(\varepsilon_n y'_{n,+}) = \sigma_\varepsilon^2 \sum_{\tau=1}^{T} \gamma^{\tau-1} C'_\tau,\ \mathrm{E}(v_n y'_{n,+}) = \sigma_v^2 C'_1,\ \mathrm{E}(v_{n,-1} y'_{n,+}) = \sigma_v^2 C'_0, \tag{19}$$

we conclude from (17) and (19) that consistency is obtained when we let $A$ be such that $\mathrm{tr}(C'_t A) = 0$ or $(\mathrm{vec}\, C_t)'a = 0$ for $t = 0, \ldots, T$. This means that $a$ should satisfy

$$C'a = 0_{T+1}. \tag{20}$$

Any estimator of the form (16) that satisfies (18), (20), and the existence conditions is consistent.

## 5 Efficient Estimation

To find an estimator that is not only consistent but also asymptotically efficient we have to distinguish between two kinds of efficiency, which we may label as local and global. We call $\hat{\gamma}$ locally efficient if it is in the class of estimators defined by $\hat{\gamma}$, with $a$ properly restricted. A globally efficient estimator is as efficient as the GMM estimator discussed at the beginning of the previous section. There we saw that GMM on the covariance matrix was a daunting task. This task is greatly simplified when we already have estimators that are consistent without any further optimality qualities. We can then adapt these estimators such that we get asymptotically efficient estimators in just a single step. This approach is called *linearized GMM* (see, e.g., Wansbeek and Meijer 2000, Sect. 9.3). It requires initial consistent estimators not just of $\gamma$ but also of the other model parameters, $\sigma_\alpha^2$, $\sigma_\varepsilon^2$, and $\sigma_v^2$. Given a consistent estimator of $\gamma$, such estimators can be easily

constructed by employing some appropriately chosen moment conditions implied by the structure of $\Sigma_y$.

We now turn to local efficiency. To that end we write the restrictions on $a$, which are all linear, in the condensed form $a = Qb$, where $Q$ is full column rank, and $b$ can be chosen freely, subject to the numerator and denominator of $\hat{\gamma}$ not becoming identically zero. So we now have

$$\begin{aligned}\hat{\gamma} &= \frac{a'\sum_n (y_n \otimes y_{n,+})}{a'\sum_n (y_{n,-1} \otimes y_{n,+})} \\ &= \frac{b'Q'F_1 s_y}{b'Q'F_0 s_y}.\end{aligned}$$

With

$$\Psi_y \equiv \text{plim}_{N\to\infty} \frac{1}{N}\sum_n \left[(y_{n,+} \bar{\otimes} y_{n,+} - s_y)(y_{n,+} \bar{\otimes} y_{n,+} - s_y)'\right],$$

we obtain

$$\text{AVar}(\hat{\gamma}) = \frac{b'\Phi_y b}{(b'Q'F_0\sigma_y)^2}, \tag{21}$$

where $\Phi_y \equiv Q'F_1\Psi_y F_1'Q$. If $\Phi_y$ were nonsingular, we could use the Cauchy–Schwarz inequality to derive the lower bound $(\sigma_y'F_0'Q\Phi_y^{-1}Q'F_0\sigma_y)^{-1}$ of the asymptotic variance, with equality for $b = \Phi_y^{-1}Q'F_0\sigma_y$. Hence, (16) would become

$$\begin{aligned}\hat{\gamma} &= \frac{\hat{b}'Q'F_1 s_y}{\hat{b}'Q'F_0 s_y} \\ &= \frac{s_y'F_0'Q\hat{\Phi}_y^{-1}Q'F_1 s_y}{s_y'F_0'Q\hat{\Phi}_y^{-1}Q'F_0 s_y},\end{aligned}$$

where $\hat{b}$ denotes $b$ with sample counterparts for $\Psi_y$ and $\sigma_y$ substituted. Unfortunately, however, it turns out that $\Phi_y$ is singular, and it is not immediately clear whether there exists a feasible optimal estimator, and if so, what this estimator would be. We leave this problem for future research. However, from the analysis here, an appealing consistent estimator is obtained by replacing the regular inverse with the Moore–Penrose generalized inverse. Thus, we propose the estimator

$$\begin{aligned}\hat{\gamma}_{\text{MP}} &\equiv \frac{s_y'F_0'Q\hat{\Phi}_y^{+}Q'F_1 s_y}{s_y'F_0'Q\hat{\Phi}_y^{+}Q'F_0 s_y} \\ &= \frac{s_y'F_0'Q(Q'F_1\hat{\Psi}_y F_1'Q)^{+}Q'F_1 s_y}{s_y'F_0'Q(Q'F_1\hat{\Psi}_y F_1'Q)^{+}Q'F_0 s_y},\end{aligned} \tag{22}$$

and we call this the Moore–Penrose (MP) estimator. Its asymptotic variance can be estimated by the sample counterpart of (21), that is

$$\widehat{\mathrm{AVar}}(\hat{\gamma}_{\mathrm{MP}}) = \frac{1}{s_y' F_0' Q (Q' F_1 \hat{\Psi}_y F_1' Q)^{+} Q' F_0 s_y}.$$

## 6 Illustrative Example

To illustrate the application of these estimators, we study the persistence in household wealth in the Health and Retirement Study (HRS; Juster and Suzman 1995). The HRS started in 1992 with a sample of individuals born in 1931–1941 and their spouses and interviewed them biennially afterward. Over time, additional cohorts have been added. We select all individuals who participated in all ten waves from 1992 to 2010 and who were either single across all waves or married to the same spouse across all waves. (We treat cohabitation the same as marriage, as is common in HRS analyses.) Because wealth is reported at the household level, we select only one respondent per household. This leaves us with a sample of 2,668 households.

We use the RAND version of the HRS, version L (St. Clair et al. 2011), including the imputations, and study total household wealth excluding the second home (HwATOTA), because information about the second home is not available in all waves. We compute the inverse hyperbolic sine transform of this variable and then subtract the wave-specific average, which captures macro effects and age effects. We then estimate the simple LDPDM for this transformed variable. We computed the standard Anderson–Hsiao estimator (by 2SLS), the Arellano-Bond estimator (using two-step GMM), the consistent "lagged" versions of these introduced earlier (AHL and ABL), and the MP estimator.

Table 1 shows the results. We clearly see the attenuation in the AH and AB estimators. Unfortunately, the standard errors increase substantially for the AHL and ABL estimators, compared to the AH and AB estimators. Meijer and Wansbeek (2000) showed this phenomenon for a cross-sectional regression model, but here it is even more dramatic. The MP estimate is close to the AH estimate, but its standard error is much smaller, though still almost four times as large as the standard errors of the (inconsistent) AH and AB estimators. Nevertheless, the MP estimate is highly

**Table 1** Estimates of $\gamma$ for transformed household wealth in the HRS

| | AH | AB | AHL | ABL | MP |
|---|---|---|---|---|---|
| $\hat{\gamma}$ | 0.107*** | 0.119*** | 0.424 | 0.264* | 0.417*** |
| (s.e.) | (0.022) | (0.020) | (0.224) | (0.112) | (0.078) |

*AH* Anderson–Hsiao, *AB* Arellano–Bond, *AHL* Anderson–Hsiao lagged, *ABL* Arellano–Bond lagged, *MP* Moore–Penrose estimator (22)
$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

significant and indicates a much stronger persistence in household wealth than we would conclude from the standard AH and AB estimators.

## 7 Discussion

Measurement error is a common problem in economic data, and this may have especially grave consequences in dynamic models. We study this and show the inconsistency of standard estimators for dynamic panel data models. We then develop a characterization of a class of consistent estimators and study efficiency within this class. Based on efficiency considerations, we propose an estimator, the Moore–Penrose (MP) estimator that has attractive statistical properties, although we have not been able to conclude whether it is the most efficient estimator in its class.

We apply the theory to the study of persistence of household wealth. We show that the attenuation bias of estimators that do not take measurement error into account can be quite large, and that our proposed estimator is much more efficient than two consistent estimators that are ad-hoc adaptations of the Anderson–Hsiao and Arellano–Bond estimators.

The results here are still quite limited. The set of model specifications needs to be expanded. Adding exogenous covariates is relatively straightforward, and weakly exogenous covariates can also be accommodated without much trouble. Our derivations thus far assume homoskedasticity, which is too strong in many economic applications. Relaxing this assumption adds restrictions that the estimator must satisfy, but does not conceptually change much. As indicated by Arellano and Bond (1991), a moving average process of the errors can be accommodated by dropping the first few lags of the dependent variable. Within our framework, this translates into additional linear restrictions. Although in the example, our estimator appears to work well, further efficiency gains may be obtained by GMM estimation based on (12). Fan et al. (2012) pursue such an approach for the static panel data model with measurement error and obtain even better results with a generalized quasi-likelihood-based estimator. We leave the development of the specifics to further research.

**Acknowledgments** We would like to thank the symposium organizer, Brajendra Sutradhar, and an anonymous reviewer of an earlier version of this paper for their useful comments and suggestions.

## References

Ahn, S.C., Schmidt, P.: Efficient estimation of models for dynamic panel data. J. Econom. **68**, 5–27 (1995)

Aigner, D.J., Hsiao, C., Kapteyn, A., Wansbeek, T.: Latent variable models in econometrics. In: Griliches, Z., Intriligator, M.D. (eds.) Handbook of Econometrics, vol. 2, pp. 1321–1393. North-Holland, Amsterdam (1984)

Altonji, J.G., Siow, A.: Testing the response of consumption to income changes with (noisy) panel data. Q. J. Econ. **101**, 293–328 (1987)
Anderson, T.W., Hsiao, C.: Estimation of dynamic models with error components. J. Am. Stat. Assoc. **77**, 598–606 (1981)
Anderson, T.W., Hsiao, C.: Formulation and estimation of dynamic models using panel data. J. Econom. **18**, 47–82 (1982)
Antman, F., McKenzie, D.: Poverty traps and nonlinear income dynamics with measurement error and individual heterogeneity. J. Dev. Stud. **43**, 1057–1083 (2007)
Arellano, M., Bond, S.: Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev. Econ. Stud. **58**, 277–297 (1991)
Arellano, M., Bover, O.: Another look at the instrumental variable estimation of error-components models. J. Econom. **68**, 29–51 (1995)
Biørn, E.: The measurement error problem in dynamic panel data analysis: modeling and GMM estimation, Working paper, University of Oslo, 2012
Blundell, R., Bond, S.: Initial conditions and moment restrictions in dynamic panel data models. J. Econom. **87**, 115–143 (1998)
Bond, S., Hoeffler, A., Temple, J.: GMM estimation of empirical growth models, CEPR Working Paper 3048, 2001
Buonaccorsi, J.: Measurement Error, Models, Methods, and Applications. Chapman & Hall, Boca Raton (2010)
Chen, J., Ni, S., Podgursky, M.: Estimating dynamic panel data models with measurement errors with an application to school evaluation based on student test scores. In: Proceedings of the 2008 Meeting of the American Statistical Association, pp. 951–957 (2008)
Fan, Z., Sutradhar, B.C., Rao, R.P.: Bias corrected generalized method of moments and generalized quasi-likelihood inferences in linear models for panel data with measurement error. Sankhyā B. **74**, 126–148 (2012)
Griliches, Z., Hausman, J.A.: Errors in variables in panel data. J. Econom. **31**, 93–118 (1986)
Harris, M.N., Mátyás, L.: Performance of the operational Wansbeek-Bekker estimator for dynamic panel data models. Appl. Econ. Lett. **7**, 149–153 (2000)
Juster, F.T., Suzman, R.: An overview of the Health and Retirement Study. J. Hum. Resour. **30**, S7–S56 (1995)
Komunjer, I., Ng, S.: Measurement errors in dynamic models, Working paper, 2011
Magnus, J.R., Neudecker, H.: Symmetry, 0–1 matrices and Jacobians: A review. Econom. Theory **2**, 157–190 (1986)
Meijer, E., Wansbeek, T.J.: Measurement error in a single regressor. Econ. Lett. **69**, 277–284 (2000)
Meijer, E., Spierdijk, L., Wansbeek, T.J.: Point and set identification in linear panel data models with measurement error. Working Paper WR-941, RAND Corporation, Santa Monica, 2012
St. Clair, P., et al.: RAND HRS Data Documentation, Version L. RAND Corporation, Santa Monica (2011)
Wansbeek, T.J., Bekker, P.A.: On IV, GMM and ML in a dynamic panel data model. Econ. Lett. **51**, 145–152 (1996)
Wansbeek, T.J., Kapteyn, A.: Simple estimators for dynamic panel data models with errors in the variables. In: Bewley, R., Van Hoa, T. (eds.) Contributions to Consumer Demand and Econometrics. McMillan, London (1992)
Wansbeek, T.J., Meijer, E.: Measurement Error and Latent Variables in Econometrics. North-Holland, Amsterdam (2000)

# Part II
# Longitudinal Data Analysis Subject to Missing Values

# Inference Progress in Missing Data Analysis from Independent to Longitudinal Setup

**Brajendra C. Sutradhar**

**Abstract** In the independent setup with multivariate responses, the data become incomplete when partial responses, such as responses on some variables as opposed to all variables, are available from some individuals. The main challenge here is obtaining valid inferences such as unbiased and consistent estimates of mean parameters of all response variables by using available responses. Typically, unbalanced correlation matrices are formed and moments or likelihood analysis based on the available responses are employed for such inferences. Various imputation techniques also have been used. In the longitudinal setup, when a univariate response is repeatedly collected from an individual, these repeated responses become correlated and the responses form a multivariate distribution. In this setup, it may happen that a portion of responses are not available from some individuals under study. These non-responses may be monotonic or intermittent. Also the response may be missing following a mechanism such as missing completely at random (MCAR), missing at random (MAR), or missing non-ignorably. In a longitudinal regression setup, the covariates may also be missing, but typically they are known for all time periods. Obtaining unbiased and consistent regression estimates specially when longitudinal responses are missing following MAR or ignorable mechanism becomes a challenge. This happens because one requires to accommodate both longitudinal correlations and missing mechanism to develop a proper inference tool. Over the last three decades some progress has been made toward this mainly by taking partial care of missing mechanism in developing estimation techniques. But overall, they fall short and may still produce biased and hence inconsistent estimates. The purpose of this paper is to outline these perspectives in a comprehensive manner so that real progress and challenges are understood in order to develop proper inference techniques.

B.C. Sutradhar (✉)
Memorial University, St. John's, Canada, A1C5S7
e-mail: bsutradh@mun.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_5, 

## 1 Introduction

Missing data analysis in the independent setup with multivariate responses has a long history. For example, for an early work, we refer to Lord (1995) who considered a set of incomplete trivariate normal responses collected from $K$ independent individuals. But all components of all three variables were not available from the $K$ individuals. To estimate the mean parameters consistently, instead of dropping out the individuals with incomplete information, Lord (1995) has utilized the available information and constructed unbalanced (bivariate and trivariate) probability functions for individuals toward writing a likelihood function for the desired inference. Note that this technique for consistent estimation of the parameters and other similar inferences by using incomplete data have been used by many researchers over the last six decades. See, for example, Mehta and Gurland (1973), Morrison (1973), Naik (1975), Little (1988), and Krishnamoorthy and Pannala (1999), among others.

In the independent setup, techniques of imputation and multiple imputation (Rubin 1976; Rubin and Schenker 1986; Meng 1994) have also been widely used. Some authors such as Paik (1997) used this imputation technique in repeated measure (longitudinal) setup. The imputation at a given time point is done mainly by averaging over the responses of other individuals at that time who has the same covariates history as that of the individual concerned. Once the missing values are estimated, they are used as data with necessary adjustments to construct complete data based estimating equations for the desired parameters.

In a univariate longitudinal response setup, when $T$ repeated measures are taken they become correlated and hence they jointly follow a $T$-dimensional multivariate distribution. However, unlike in the Gaussian setup for linear data, the multivariate distributions for repeated binary and count data become complex or impractical. However if a portion of individuals do not provide responses for all $T$ time points, then adopting likelihood approach by blending missing mechanism and correlation structure of the repeated data would naturally become extremely complicated or impossible. As a remedy, either imputation or estimating equation approaches became popular which, however, work well if the missing data occur following the simplest MCAR mechanism. When the missing data occur following the MAR mechanism, writing a proper estimating equation by accommodating both longitudinal correlations and missing mechanism becomes difficult. Robins et al. (1995) proposed an inverse probability weights based generalized estimating equations (WGEE) approach as an extension of the GEE approach proposed by Liang and Zeger (1986) to the incomplete setup. Remark that as demonstrated by Sutradhar and Das (1999) and Sutradhar (2010), for example, the GEE approach can produce less efficient regression estimates than the well-known simpler moments or quasi-likelihood (QL) estimates, in the complete data setup. Thus, to be realistic, there is no reason how WGEE approach can be more efficient in the incomplete longitudinal setup as compared to simpler moments and QL estimates. In fact in the incomplete longitudinal setup, the WGEE approach constructed based on working correlations as opposed to the use of

MAR based correlation matrix may yield biased and hence inconsistent regression estimates (Sutradhar and Mallick 2010). Further remark that this inconsistency issue was, however, not adequately addressed in the literature including the studies by Robins et al. (1995), Paik (1997), Rotnitzky et al. (1998), and Birmingham et al. (2003). One of the main reasons is this that none of the studies used any stochastic correlation structure in conjunction with the missing mechanism to model the longitudinal count and binary data in the incomplete longitudinal setup. Details on this inconsistency problem are given in Sect. 3, whereas in Sect. 2 we provide a detailed discussion on missing data analysis in independent setup.

Without realizing the aforementioned inconsistency problems that can be caused because of the use of working correlations in the estimating equations under the MAR based longitudinal setup, some authors such as Wang (1999) and Rotnitzky et al. (1998) used similar estimating equations approach in non-ignorable missing mechanism-based incomplete longitudinal setup. Some authors such as Troxel et al. (1998) (see also Troxel et al. 1997) and Ibrahim et al. (2001) (see also Ibrahim et al. 1999) have used random effects based generalized linear mixed model to accommodate the longitudinal correlations and certain binary logistic models to generate the non-ignorable mechanism based response indicator variables. In general expectation-maximization (EM) techniques are used to estimate the likelihood based parameters. These approaches appear to encounter similar difficulties as the existing MAR based approaches in generating first the response indicator and then the responses so that underlying longitudinal correlation structure is satisfied. Thus the inference validity of these approaches is not yet established. This problem becomes more complicated when longitudinal correlations are not generated through random effects and writing a likelihood such as for repeated count data becomes impossible. For clarity, in this paper we discuss in detail the successes and challenges with the inferences for MAR based incomplete longitudinal models only. The non-ignorable missing data based longitudinal analysis will therefore be beyond the scope of the paper.

## 2 Missing Data Analysis in Independent Setup

Missing data analysis in the independent setup with multivariate responses has a long history. For example, for an early work, we refer to Lord (1995) who considered a set of incomplete trivariate normal responses collected from $K$ independent individuals. To be specific, suppose that $y = (y_1, y_2, y_3)'$ represents a trivariate response, but all components of $y$ were not available from $K$ individuals. Suppose that $y_3$ was recorded from all $K$ individuals, and either $y_1$ or $y_2$ was recorded for all individuals, but not both. For $j = 1, \ldots, 3$, let $K_j$ denote the number of individuals having the response $y_j$. It then follows that

$$K_1 + K_2 = K,\ K3 = K.$$

Further suppose that the $K_1$ individuals for whom $y_1$ is recorded will be denoted collectively as group 1 $(G_1)$; and the $K_2$ individuals with $y_2$ will be denoted as group 2 $(G_2)$. Now because $y_1$ and $y_2$ are correlated, it is obvious that the data for $G_2$ contain some information relevant for estimating the parameters of variable $y_1$, and that the data for $G_1$ contain some information relevant for estimating the parameters of $y_2$. The problem is to use the available data as efficiently as possible for estimating the parameters concerned. Denote the distribution of $y = [y_1, y_2, y_3]'$ as

$$y \sim N(\mu, \Sigma),$$

with $\mu = [\mu_1, \mu_2, \mu_3]'$ and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \rho_{12}[\sigma_{11}\sigma_{22}]^{\frac{1}{2}} & \rho_{13}[\sigma_{11}\sigma_{33}]^{\frac{1}{2}} \\ & \sigma_{22} & \rho_{23}[\sigma_{22}\sigma_{33}]^{\frac{1}{2}} \\ & & \sigma_{33} \end{pmatrix}.$$

Note that in this setup, there are no data available to estimate $\rho_{12}$. For the likelihood estimation of all the other parameters, define

$$\bar{y}_1^* = \frac{1}{K_1}\sum_{i=1}^{K_1} y_{1i},\ \bar{y}_2^* = \frac{1}{K_2}\sum_{i=1}^{K_2} y_{2i},\ \bar{y}_3 = \frac{1}{K}\sum_{i=1}^{K} y_{3i},\ \bar{y}_3^* = \frac{1}{K_1}\sum_{i=1}^{K_1} y_{3i},\ \bar{y}_3^{**} = \frac{1}{K_2}\sum_{i=1}^{K_2} y_{3i}$$

$$s_{11}^* = \frac{1}{K_1}\sum_{i=1}^{K_1}[y_{1i} - \bar{y}_1^*]^2,\ s_{22}^* = \frac{1}{K_2}\sum_{i=1}^{K_2}[y_{2i} - \bar{y}_2^*]^2,\ s_{33} = \frac{1}{K}\sum_{i=1}^{K}[y_{3i} - \bar{y}_3]^2,$$

$$s_{33}^* = \frac{1}{K_1}\sum_{i=1}^{K_1}[y_{3i} - \bar{y}_3^*]^2,\ s_{33}^{**} = \frac{1}{K_2}\sum_{i=1}^{K_2}[y_{3i} - \bar{y}_3^{**}]^2$$

$$r_{13} = \frac{1}{K_1}\sum_{i=1}^{K_1}[(y_{1i} - \bar{y}_1^*)(y_{3i} - \bar{y}_3^*)]/[s_{11}^* s_{33}^*]^{\frac{1}{2}},$$

$$r_{23} = \frac{1}{K_2}\sum_{i=1}^{K_2}[(y_{2i} - \bar{y}_2^*)(y_{3i} - \bar{y}_3^{**})]/[s_{22}^* s_{33}^{**}]^{\frac{1}{2}}. \tag{1}$$

The maximum likelihood estimators for the means are then given by

$$\hat{\mu}_1 = \bar{y}_1^* - b_{13}[\bar{y}_3^* - \bar{y}_3],\ \hat{\mu}_2 = \bar{y}_2^* - b_{23}[\bar{y}_3^{**} - \bar{y}_3],\ \text{and}\ \hat{\mu}_3 = \bar{y}_3, \tag{2}$$

where

$$b_{13} = r_{13}\frac{s_{11}^*}{s_{33}^*},\ \text{and}\ b_{23} = r_{23}\frac{s_{22}^*}{s_{33}^{**}}.$$

These estimators in (2) are unbiased and consistent for $\mu_1$, $\mu_2$, and $\mu_3$, respectively. The remaining parameters may also be estimated similarly (Lord 1995).

Note that the aforementioned technique for consistent estimation of the parameters and for other similar inferences by using incomplete data has been subsequently used by many researchers over the last six decades. See, for example, Mehta and Gurland (1973), Morrison (1973), Naik (1975), Little (1988), and Krishnamoorthy and Pannala (1999), among others. This idea of making inferences about the underlying model parameters such that the missing data (assuming a small proportion of missing) may not to any major extent negatively influence the inferences has also been extended to the analysis of incomplete repeated measure data. For example, one may refer to Little (1995), Robins et al. (1995), and Paik (1997), as some of the early studies. This inference procedure for incomplete longitudinal data is discussed in detail in the next section.

In the independent setup, techniques of imputation and multiple imputation (Rubin 1976; Rubin and Schenker 1986; Meng 1994) have also been widely used. Later on some authors also used this imputation technique in repeated measure (longitudinal) setup. For example, here we illustrate an imputation formula from Paik (1997) in repeated measure setup. The imputation at a given time point is done mainly by averaging over the responses of other individuals at that time who has the same covariates history as that of the individual concerned. Once the missing values are estimated, they are used as data with necessary adjustments to construct complete data based estimating equations for the desired parameters.

In a univariate longitudinal response setup, when $T$ repeated measures are taken they become correlated and hence they jointly follow a $T$-dimensional multivariate distribution. Now suppose that $T_i$ responses are observed for the $i$th $(i = 1, \ldots, K)$ individual. So, one requires to impute $T - T_i$ missing values which may be done following Paik (1997), for example. Interestingly, a unified recursive relation can be developed as follows to obtain the imputed value $\tilde{y}_{i,T_i+k_i}$ at time point $T_i + k_i$ for all $k_i = 1, \ldots, T - T_i$. For this, first define

$$\tilde{y}^{(0)}_{j,T_i+k_i} = y_{j,T_i+k_i} \tag{3}$$

for the $j$th individual where $j \neq i$, $j = 1, \ldots, K$. Also, let $D_{iT_i}$ denote the covariate history up to time point $T_i$ for the $i$th individual, and

$$D^*_{i,T_i+k_i} = (x_{i,T_i+1}, \ldots, x_{i,T_i+k_i})$$

is the covariate information for the $i$th individual from time $T_i + 1$ up to $T_i + k_i$ for $k_i = 1, \ldots, T - T_i$. Further let, $r_{jw} = 1, \text{or}, 0$, for example, indicates the response status of the $j$th individual at $w$th time. One may then obtain $\tilde{y}_{i,T_i+k_i}$ by computing $\tilde{y}^{(k_i)}_{i,T_i+k_i}$, that is, $\tilde{y}_{i,T_i+k_i} \equiv \tilde{y}^{(k_i)}_{i,T_i+k_i}$, where

$$\tilde{y}_{i,T_i+k_i}^{(k_i)} = \left[ \sum_{j=1}^{K} \tilde{y}_{j,T_i+k_i}^{(0)} \Pi_{u=1}^{k_i} r_{j,T_i+u} I(D_{jT_i} = D_{iT_i}, D_{j,T_i+k_i}^{*} = D_{i,T_i+k_i}^{*}) \right.$$

$$+ \sum_{m_i=1}^{k_i-1} \sum_{j=1}^{K} \tilde{y}_{j,T_i+k_i}^{(m_i)} \Pi_{u=k_i-(m_i-1)}^{k_i} (1 - r_{j,T_i+u})$$

$$\left. \times \Pi_{u=1}^{k_i-m_i} r_{j,T_i+u} I(D_{jT_i} = D_{iT_i}, D_{j,T_i+k_i}^{*} = D_{i,T_i+k_i}^{*}) \right]$$

$$\times \left[ \sum_{j=1}^{K} r_{j,T_i+1} I(D_{jT_i} = D_{iT_i}, D_{j,T_i+k_i}^{*} = D_{i,T_i+k_i}^{*}) \right]^{-1}. \tag{4}$$

Note that $\tilde{y}_{i.T_i+k_i} \equiv \tilde{y}_{i,T_i+k_i}^{(k_i)}$ is an unbiased estimate of $\mu_{i,T_i+k_i}$ as the individuals used to impute the missing value of the $i$th subject has the same covariate history up to time point $T_i + k_i$, unlike the covariate history up to time point $T_i$ (Paik 1997).

## 3 Missing Data Models in Longitudinal Setup

Let $Y_{it}$ be the potential response from the $i$th $(i = 1, \ldots, K)$ individual at time point $t$ which may or may not be observed, and $x_{it} = (x_{it1}, \ldots, x_{itp})'$ be the corresponding $p$-dimensional covariate vector which is assumed to be available for all times $t = 1, \ldots, T$. In this setup, $K$ is large $(K \to \infty)$ and $T$ is small such as 3 or 4. Suppose that $\beta = (\beta_1, \ldots, \beta_p)'$ denote the effect of $x_{it}$ on $y_{it}$. Irrespective of the situation whether $Y_{it}$ is observed or not, it is appropriate in the longitudinal setup to assume that the repeated responses follow a correlation model with known functional forms for the mean and the variance, but the correlation structure may be unknown. Recall that in the independent setup, Lord (1995) considered multivariate responses having a correlation structure and incompleteness arose because of missing information on some response variables, whereas in the present longitudinal setup, repeated responses from an individual form a multivariate response with a suitable mean, variance, and correlation structures, but it remains a possibility that one individual may not provide responses for the whole duration of the study. As indicated in the last section, suppose that for the $i$th $(i = 1, \ldots, K)$ individual $T_i$ responses $(1 < T_i \leq T)$ are collected. Also suppose that the remaining $T - T_i$ potential responses are missing and the non-missing responses occur in a monotonic pattern.

As far as the mean, variance, and correlation structure of the potential responses are concerned, it is convenient to define them for the complete data. Let $y_i^c = (y_{i1}, \cdots, y_{it}, \cdots, y_{iT})'$ and $X_i^c = (x_{i1}, \cdots, x_{it}, \cdots, x_{iT})'$ denote the $T \times 1$ complete outcome vector and $T \times p$ covariate matrix, respectively, for the $i$-th $(i = 1, \cdots, K)$ individual over $T$ successive points in time. Also, let

$$E(Y_i^c|x_i^c) = \mu_i^c(\beta) = (\mu_{i1}(\beta), \cdots, \mu_{it}(\beta), \cdots, \mu_{iT}(\beta))' \tag{5}$$

where $\mu_{it}(\beta) = h^{-1}(\eta_{it})$ with $\eta_{it} = x_{it}'\beta$, $h$ being a suitable link function. For example, for linear models, a linear link function is used so that $\mu_{it}(\beta) = x_{it}'\beta$; whereas for the binary data a logistic link function is commonly used so that $\mu_{it}(\beta) = \exp(\eta_{it})/[1 + \exp(\eta_{it})]$, and for count data a log linear link function is used so that $\mu_{it}(\beta) = \exp(\eta_{it})$. Further let

$$\Sigma_i^c(\beta, \rho) = A_i^{c\frac{1}{2}}(\beta)\tilde{C}_i(\rho, x_i^c)A_i^{c\frac{1}{2}}(\beta) \tag{6}$$

be the true covariance matrix of $y_i^c$, where $A_i^c(\beta) = \text{diag}[\sigma_{i11}(\beta), \cdots, \sigma_{itt}(\beta), \cdots, \sigma_{iTT}(\beta)]$ with $\sigma_{itt}(\beta) = var(Y_{it})$, and $\tilde{C}_i(\rho, x_i^c)$ is the correlation matrix for the $i$th individual with $\rho$ as a suitable vector of correlation parameters, for example, $\rho \equiv (\rho_1, \ldots, \rho_\ell, \ldots, \rho_{T-1})'$, where $\rho_\ell$ is known to be the lag $\ell$ auto-correlation. Note that when covariates are time dependent, the true correlation matrix is free from time-dependent covariates in linear longitudinal setup, but it depends on the time-dependent covariates through $X_i^c$ in the discrete longitudinal setup (Sutradhar 2010). In the stationary case, that is, when covariates are time independent, we will denote the correlation matrix by $\tilde{C}(\rho)$ in the complete longitudinal setup, and similar to Sutradhar (2010, 2011), this matrix satisfies the auto-correlation structure given by

$$\tilde{C}(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots & 1 \end{bmatrix}, \tag{7}$$

where for $\ell = 1, \ldots, T$, $\rho_\ell$ is known to be the $\ell$th lag auto-correlation. Note that when this correlation structure (7) will be used in the incomplete longitudinal setup, it would be denoted by $\tilde{C}_i(\rho)$ as it will be constructed for $T_i$ available responses.

As far as the missing mechanism is concerned, it is customary to assume that a longitudinal response may be missing completely at random (MCAR), or missing at random (MAR), or the missing can be non-ignorable. Under the MCAR mechanism, the missing-ness does not depend on any present, past, or future responses. Under the MAR mechanism, the missing-ness depends only on the past responses but not on the present or future responses, whereas under the non-ignorable mechanism the missing-ness depends on the past, present, and future possible responses. In notation, let $R_{it}$ be a response indicator variable at time $t$ $(t = 1, \cdots, T)$ for the $i$-th $(i = 1, \cdots, K)$ individual, so that

$$R_{it} = \begin{cases} 1, \text{ if } Y_{it} \text{ is observed} \\ 0, \text{ otherwise.} \end{cases} \tag{8}$$

Note that all individuals provide the responses at the first time point $t = 1$. Thus, we set $R_{i1} = 1$ with $P(R_{i1} = 1) = 1.0$ for all $i = 1, \cdots, K$. Further we assume that the response indicators satisfy the monotonic relationship

$$R_{i1} \geq R_{i2} \geq \cdots \geq R_{it} \geq \cdots \geq R_{iT}. \tag{9}$$

Next suppose that $r_{it}$ denote the observed value for $R_{it}$. For $t = 2, \ldots, T$, one may then describe the aforementioned three missing mechanisms as

$$\begin{aligned}
\text{MCAR Model} \ : \ & Pr(R_{it} = 1 \mid y_i^c, x_i, r_{i,t-1} = 1) = Pr(R_{it} = 1 \mid r_{i,t-1} = 1) \\
\text{MAR Model} \ : \ & Pr(R_{it} = 1 \mid y_i^c, x_i, r_{i,t-1} = 1) \\
& = Pr(R_{it} = 1 \mid y_{i1}, \cdots, y_{i,t-1}, x_i, r_{i,t-1} = 1) \\
\text{Non-ignorable Model} \ : \ & Pr(R_{it} = 1 \mid y_i^c, x_i, r_{i,t-1} = 1) \\
& = Pr(R_{it} = 1 \mid y_{i1}, \cdots, y_{i,t-1}, y_{it}, \ldots, y_{iT}, x_i, r_{i,t-1} = 1)
\end{aligned}$$

(Little and Rubin 1987; Laird 1988; Fitzmaurice et al. 1996). Furthermore, it follows under the monotonic missing pattern (9) that $Pr(R_{it} = 1 | y_i^c, x_i, r_{i,t-1} = 0) = 0$ irrespective of the missing mechanism. Note that the inferences based on the non-ignorable missing mechanism may be quite complicated, and we do not include this complicated mechanism in the current paper.

## 3.1 *Inferences When Longitudinal Responses Are Subject to MCAR*

When the longitudinal responses are MCAR, $R_{it}$ does not depend on the past, present, or future responses. In such a situation, $R_{it}$ and $Y_{it}$ are independent, implying that

$$E[R_{it}(Y_{it} - \mu_{it}(\beta))] = E[R_{it}]E[Y_{it} - \mu_{it}(\beta)] = 0, \tag{10}$$

because $E[Y_{it} - \mu_{it}(\beta)] = 0$. It is then clear that the inference for $\beta$ involved in $\mu_{it}(\beta)$ is not affected by the MCAR mechanism. Thus, one may estimate the regression effects $\beta$ consistently and efficiently by solving the GQL estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta} \Sigma_i^{-1}(\beta, \hat{\rho})(y_i - \mu_i(\beta)) = 0, \tag{11}$$

where for $T_i$-dimensional observed response vector $y_i = (y_{i1}, \ldots, y_{iT_i})'$,

$$\begin{aligned}
\mu_i(\beta) = E[Y_i] &= (\mu_{i1}(\beta), \cdots, \mu_{it}(\beta), \cdots, \mu_{iT_i}(\beta))' \\
\Sigma_i(\beta, \hat{\rho}) &= A_i^{1/2}(\beta) \tilde{C}_i(\rho, x_i) A_i^{1/2}(\beta),
\end{aligned}$$

with $A_i(\beta) = \text{diag}(\sigma_{i,11}(\beta), \cdots, \sigma_{i,tt}(\beta), \cdots, \sigma_{i,T_iT_i}(\beta))$, where $\sigma_{i,tt}(\beta) = \text{var}[Y_{it}]$. Note that the incomplete data based estimating equation (11) can be written in terms of pretended complete data. To be specific, by using the available responses $y_i = (y_{i1}, \ldots, y_{iT_i})'$ corresponding to the known response indicators

$$R_i^c = r_i^c = \begin{bmatrix} I_{T_i} & 0 \\ 0 & 0, \end{bmatrix}$$

one may write the GQL estimating equation (11) under the MCAR mechanism as

$$\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}(\beta)}{\partial \beta} \left[ \{I - r_i^c\} + r_i^c \Sigma_i^c(\beta, \hat{\rho}) r_i^{c'} \right]^{-1} r_i^c \, (y_i^c - \mu_i^c(\beta)) = 0, \tag{12}$$

where $y_i^c = (y_i^{'}, y_{im}^{'})'$ with $y_{im}$ representing the $T - T_i$ dimensional missing responses which are unobserved but for the computational purpose in the present approach one can use it as a zero vector, for convenience, without any loss of generality. Let $\hat{\beta}_{GQL,MCAR}$ denote the solution of (11) or (12). This estimator is asymptotically unbiased and hence consistent for $\beta$.

Note that the computation of $\tilde{C}_i(\hat{\rho}, x_i)$ matrix in (11) in general, i.e., when covariates are time dependent, depends on the specific correlation structure (Sutradhar 2010). In stationary cases as well as in linear longitudinal model setup, one may, however, compute the stationary correlation matrix $\tilde{C}_i(\hat{\rho})$, by first computing a larger $\tilde{C}(\hat{\rho})$ matrix for $\ell = 1, \ldots, T-1$, and then using the desired part of this large matrix for $t = 1, \ldots, T_i$. Turning back to the computation for the larger matrix with dimension $T = \max_{1 \le i \le K} T_i$ for $T_i \ge 2$, we exploit the observed response indicator $r_{it}$ given by

$$r_{it} = \begin{cases} 1 \text{ if } t \le T_i \\ 0 \text{ if } T_i < t \le T. \end{cases}$$

for all $t = 1, \ldots, T$. For known $\beta$ and $\sigma_{itt}$, the $\ell$th lag correlation estimate $\hat{\rho}_\ell$ for the larger $\tilde{C}(\hat{\rho})$ matrix may be computed as

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K} \sum_{t=1}^{T-\ell} r_{it} r_{i,t+\ell} [\left( \frac{y_{it} - x_{it}'\beta}{\sigma_{itt}} \right) \left( \frac{y_{i,t+\ell} - x_{it,t+\ell}'\beta}{\sigma_{i,t+\ell,t+\ell}} \right)] / \sum_{i=1}^{K} \sum_{t=1}^{T-\ell} r_{it} r_{i,t+\ell}}{\sum_{i=1}^{K} \sum_{t=1}^{T} r_{it} [\frac{y_{it} - x_{it}'\beta}{\sigma_{itt}}]^2 / \sum_{i=1}^{K} r_{it}}, \tag{13}$$

(cf. Sneddon and Sutradhar 2004, eqn. (16)) for $\ell = 1, \ldots, T-1$. Note that as this estimator contains $\hat{\beta}_{GQL,MCAR}$, both (11) and (13) have to be computed iteratively until convergence.

Further note that in the existing GEE approach, instead of (11), one solves the estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta} V_i^{-1}(\beta, \hat{\alpha})(y_i - \mu_i(\beta)) = 0, \tag{14}$$

[Liang and Zeger 1986] where $V_i(\beta,\hat{\alpha}) = A_i^{1/2}(\beta)Q_i(\alpha)A_i^{1/2}(\beta)$, with $Q_i(\alpha)$ as the $T_i \times T_i$ "working" correlation matrix of $y_i$. It is, however, known that this GEE approach may sometimes encounter consistency breakdown (Crowder 1995) because of the difficulty in estimating the "working" correlation or covariance structure, leading to the failure of estimation of $\beta$ or the non-convergence of $\beta$ estimator to $\beta$. Furthermore, even if GEE $\beta$ estimate becomes consistent, it may produce inefficient estimate than simpler independence assumption based moment or quasi-likelihood (QL) estimate (Sutradhar and Das 1999; Sutradhar 2011). Thus, one should be clear from these points that the GEE approach even if corrected for missing mechanism may encounter similar consistency and inefficiency in estimating the regression parameters.

We also remark that even though the non-response probability is not affected by the past history under the MCAR mechanism, the respective efficiency of GQL and GEE estimators will decrease if $T_i$ is very small as compared to the attempted complete duration $T$, that is, if $T - T_i$ is large. As far as the value of $T_i$ is concerned, it depends on the probability, $P[R_{it} = 1]$ which in general decreases due to the monotonic condition (9). This is because under this monotonic property (9) and following MCAR mechanism, one writes

$$\begin{aligned} Pr[R_{it} = 1] &\equiv Pr[R_{i1} = 1, R_{i2} = 1, \ldots, R_{it} = 1] \\ &= \Pi_{j=2}^{t} P[R_{ij} = 1], \end{aligned} \tag{15}$$

which gets smaller as $t$ gets larger, implying that $T_i$ can be small as compared to $T$ if $P[R_{ij} = 1]$ is far away down from 1 such as $P[R_{ij} = 1] = 0.90$, say.

## 3.2 *Inferences When Longitudinal Responses Are Subject to MAR*

Unlike in the MCAR case, $R_{it}$ and $y_{it}$ are not independent under the MAR mechanism. That is

$$E[R_{it}(Y_{it} - \mu_{it}(\beta))] \neq 0 \text{ under MAR.} \tag{16}$$

This is because

$$\begin{aligned} & E\left[R_{it}(Y_{it} - \mu_{it}(\beta)) \mid H_{i,t-1}(y)\right] \\ &= E_{Y_{it}} E\left[R_{it}(Y_{it} - \mu_{it}(\beta)) \mid Y_{it}, H_{i,t-1}(y)\right] \\ &= E_{Y_{it}}\left[\{(Y_{it} - \mu_{it}(\beta))|H_{i,t-1}(y)\} E\{R_{it}|Y_{it}, H_{i,t-1}(y)\}\right] \\ &= E_{Y_{it}}\left[\{(Y_{it} - \mu_{it}(\beta))|H_{i,t-1}(y)\} E\{R_{it}|H_{i,t-1}(y)\}\right] \end{aligned} \tag{17}$$

as $R_{it}$ does not depend on $Y_{it}$ by the definition of MAR.

Next due to the monotonic property (9) of the response indicators

$$\begin{aligned}
&E\left[R_{it} \mid H_{i,t-1}(y)\right] \\
&= P\left[R_{i1}=1, R_{i2}=1, \cdots, R_{i,t-1}=1, R_{it}=1 | H_{i,t-1}(y)\right] \\
&= P(R_{i1}=1) P\left[R_{i2}=1 \mid R_{i1}=1; H_{i1}(y)\right] \cdots \\
&\times P\left[R_{it}=1 \mid R_{i1}=1, \cdots, R_{i,t-1}=1; H_{i,t-1}(y)\right] \\
&= \prod_{j=1}^{t} g_{ij}(y_{i,j-1}, \cdots, y_{i,j-q}; \gamma) \\
&= w_{it}\{H_{i,t-1}(y); \gamma\}, \qquad (18)
\end{aligned}$$

and

$$E_{Y_{it}}\left[(Y_{it}-\mu_{it}(\beta)) | H_{i,t-1}(y)\right] = (\lambda_{it}(H_{i,t-1}(y), \beta, \rho) - \mu_{it}(\beta)), \qquad (19)$$

where $\lambda_{it}(H_{i,t-1}(y), \beta, \rho)$ is the conditional mean of $Y_{it}$. In (18), one may, for example, use $g_{ij}(\gamma)$ as

$$\begin{aligned}
g_{ij}(\gamma) &= Pr[(R_{ij}=1) | R_{i1}=1, \ldots, R_{i,j-1}=1, H_{i,j-1}(y)] \\
&= \frac{\exp(1+\gamma y_{i,j-1})}{1+\exp(1+\gamma y_{i,j-1})}. \qquad (20)
\end{aligned}$$

Now because both $w_{it}\{H_{i,t-1}(y); \gamma\}$ and $\lambda_{it}(H_{i,t-1}(y), \beta, \rho)$ are functions of the past history of responses $H_{i,t-1}(y)$, and because

$$E_{H_{i,t-1}(y)}[\lambda_{it}(H_{i,t-1}(y), \beta, \rho) - \mu_{it}(\beta)] = 0, \qquad (21)$$

it then follows from (17), by (18) and (19), that

$$E[R_{it}(Y_{it}-\mu_{it}(\beta))] = E_{H_{i,t-1}(y)} E\left[R_{it}(Y_{it}-\mu_{it}(\beta)) \mid H_{i,t-1}(y)\right] \neq 0, \qquad (22)$$

unless $w_{it}\{H_{i,t-1}(y); \gamma\}$ is a constant free of $H_{i,t-1}(y)$, which is, however, impossible under MAR missing mechanism as opposed to the MCAR mechanism. Thus, $E[R_{it}\{Y_{it}-\mu_{it}(\beta)\}] \neq 0$.

### 3.2.1 Existing Partially Standardized GEE Estimation for Longitudinal Data Subject to MAR

Note, however, that

$$\begin{aligned}
&E\left\{\frac{R_{it}}{w_{it}\{H_{i,t-1}(y); \gamma\}}(Y_{it}-\mu_{it}(\beta))\right\} \\
&= E_{H_{i,t-1}(y)}\left[(\lambda_{it}(H_{i,t-1}(y), \beta, \rho) - \mu_{it}(\beta))\right] = 0. \qquad (23)
\end{aligned}$$

Now suppose that

$$\Delta_i = \text{diag}[\delta_{i1}, \delta_{i2}, \cdots, \delta_{iT_i}] \text{ with } \delta_{it} = R_{it}/w_{it}\{H_{i,t-1}(y); \gamma\}$$

implying that $E[\Delta_i|H_i(y)] = I_{T_i}$, and where $H_i(y)$ is used to denote appropriate past history showing that the response indicators are generated based on observed responses only.

By observing the unconditional expectation property from (23), in the spirit of GEE [Liang and Zeger 1986], Robins et al. (1995, eqn. (10), p. 109) proposed a conditional inverse weights based PSGEE for the estimation of $\beta$ which has the form

$$\sum_{i=1}^{K} \frac{\partial E_{H_i(y)} E[\{\Delta_i \mu_i(\beta)\}'|H_i(y)]}{\partial \beta} V_i^{-1}(\hat{\alpha})\{\Delta_i(y_i - \mu_i(\beta))|H_i(y)\}$$

$$= \sum_{i=1}^{K} \frac{\partial \{\mu_i(\beta)\}'}{\partial \beta} V_i^{-1}(\hat{\alpha})\{\Delta_i(y_i - \mu_i(\beta))\} = 0, \quad (24)$$

(see also Paik 1997, eqn. (1), p. 1321). Note that we refer to the GEE in (23) as a partly or partially standardized GEE (PSGEE) because $V_i(\alpha) = \hat{\text{cov}}(Y_i)$ used in this GEE is a partial weight matrix which ignores the missing mechanism, whereas $\text{cov}[\Delta_i(y_i - \mu_i(\beta))]$ would be a full weight matrix.

Note that over the last decade many researchers have used this PSWGEE approach for studying various aspects of longitudinal data subject to non-response. See, for example, the studies by Rotnitzky et al. (1998), Preisser et al. (2002), and Birmingham et al. (2003), among others. However, even if the MAR mechanism is accommodated to develop an unbiased estimating function $\Delta_i(y_i - \mu_i(\beta))$ (for 0) to construct the fully standardized GEE (FSGEE), the consistency of the estimator of $\beta$ may break down (see Crowder 1995 for complete longitudinal models) because of the use of "working" covariance matrix $V_i(\alpha)$, whereas the true covariance matrix for $y_i$ is given by $\text{cov}[Y_i] = \Sigma_i(\rho)$. This can happen for those cases where $\alpha$ is not estimable. To be more clear, $V_i(\alpha)$ is simply a "working" covariance matrix of $y_i$, whereas a proper estimating equation must use the correct variance (or its consistent estimate) matrix of $\{\Delta_i(y_i - \mu_i(\beta))\}$.

To understand the roles of both missing mechanism and longitudinal correlation structure in constructing a proper estimating equation, we now provide following three estimating equations for $\beta$. The difficulties and/or advantages encountered by these equations are also indicated.

### 3.2.2 Partially Standardized GQL (PSGQL) Estimation for Longitudinal Data Subject to MAR

When $V_i(\alpha)$ matrix in (24) is replaced with the true $T_i \times T_i$ covariance matrix of the available responses, that is, $\Sigma_i(\rho) = \text{cov}[Y_i]$, one obtains the PSGQL estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial\{\mu_i(\beta)\}'}{\partial\beta} \Sigma_i^{-1}(\hat{\rho})\{\Delta_i(y_i - \mu_i(\beta))\} = 0, \tag{25}$$

which also may produce biased and hence inconsistent estimate. This is because $\Sigma_i(\rho)$ may still be very different than the covariance matrix of the actual variable $\{\Delta_i(y_i - \mu_i(\beta))\}$. Thus, if the proportion of missing values is more, one may not get convergent solution to the estimating equation (25) and the consistency for $\beta$ would break down (Crowder 1995). The convergence problems encountered by (24) would naturally be more severe as even in the complete data case $V_i(\alpha)$ may not be estimable.

### 3.2.3 Partially Standardized Conditional GQL (PSCGQL) Estimation for Longitudinal Data Subject to MAR

Suppose that one uses conditional (on history) variance

$$\text{cov}\{\Delta_i(Y_i - \mu_i(\beta))\}|H_i(y) = \Sigma^*_{ich}(H_i(y), \beta, \rho, \gamma), \tag{26}$$

to construct the estimating equation. Then following (25), one may write the PSCGQL estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial\{\mu_i(\beta)\}'}{\partial\beta} {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)\{\Delta_i(y_i - \mu_i(\beta))\} = 0 \tag{27}$$

It is, however, seen that

$$\begin{aligned} & {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)[\Delta_i(Y_i - \mu_i(\beta))] \\ \to\ & {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)E[\{\Delta_i(Y_i - \mu_i(\beta))\}|H_i(y)] \\ =\ & {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)[\lambda_i(H_i(y)) - \mu_i(\beta)] \end{aligned} \tag{28}$$

But,

$$\begin{aligned} & E_{H_i(y)}\left[\frac{\partial\{\mu_i(\beta)\}'}{\partial\beta} {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)\right. \\ \times\ & \left.[\lambda_i(H_i(y)) - \mu_i(\beta)]\right] \neq 0, \end{aligned} \tag{29}$$

even though

$$E_{H_i(y)}[\lambda_i(H_i(y)) - \mu_i(\beta)] = 0.$$

Thus, the PSCGQL estimating equation (27) is not an unbiased equation for 0, and may produce bias estimate.

*Computational formula for* $\Sigma^*_{ich}(\beta,\rho,\gamma)$

For convenience, we first write

$$\Delta_i = W_i^{-1}R_i, \text{ with}$$
$$W_i = \text{diag}[w_{i1}, w_{i2}, \ldots, w_{iT_i}], \text{ and } R_i = \text{diag}[R_{i1}, \ldots, R_{iT_i}].$$

It then follows that

$$\begin{aligned}\Sigma^*_{ich}(\beta,\rho) &= \text{cov}[\{\Delta_i\{(y_i - \mu_i(\beta))\}\}|H_i(y)] \\ &= W_i^{-1}\text{cov}[\{R_i(y_i - \mu_i(\beta))\}|H_i(y)]W_i^{-1}. \end{aligned} \tag{30}$$

Now to compute the covariance matrix in the middle term in the right-hand side of (30), we first re-express $R_i(y_i - \mu_i(\beta))$ as

$$R_i(y_i - \mu_i(\beta)) = [R_{i1}(y_{i1} - \mu_{i1}), \ldots, R_{it}(y_{it} - \mu_{it}), \ldots, R_{iT_i}(y_{iT_i} - \mu_{iT_i})]',$$

and compute the variances for its components as

$$\text{var}[\{R_{i1}(y_{i1} - \mu_{i1})\}|y_{i1}] = 0, \tag{31}$$

because $R_{i1} = 1$ always and $y_{i1}$ is random . In the Poisson case $\sigma_{i,11} = \mu_{i1}$ and in the binary case $\sigma_{i,11} = \mu_{i1}(1-\mu_{i1})$, with appropriate formula for $\mu_{i1}$ in a given case. Next for $t = 2, \ldots, T_i$,

$$\begin{aligned}&\text{var}[R_{it}(y_{it} - \mu_{it})|H_{i,t-1}(y)] = \text{var}[R_{it}|H_{i,t-1}(y)]\text{var}[y_{it}|H_{i,t-1}(y)] \\ &+ E^2[R_{it}|H_{i,t-1}(y)]\text{var}[(y_{it})|H_{i,t-1}(y)] + \text{var}[R_{it}|H_{i,t-1}(y)]E^2[(y_{it} - \mu_{it})|H_{i,t-1}(y)] \\ &= w_{it}(1-w_{it})\sigma_{ic,tt} + w_{it}^2\sigma_{ic,tt} + w_{it}(1-w_{it})\{\lambda_{it} - \mu_{it}\}^2 \\ &= w_{it}[\sigma_{ic,tt} + (\lambda_{it} - \mu_{it})^2] - w_{it}^2(\lambda_{it} - \mu_{it})^2, \end{aligned} \tag{32}$$

where, given the history, $\lambda_{it}$ and $\sigma_{ic,tt}$ are the conditional mean and variance of $y_{it}$, respectively.

Furthermore, all pairwise covariances conditional on the history $H_{i,t-1}(y)$ may be computed as follows. For $u < t$,

$$\begin{aligned}&\text{cov}[\{R_{iu}(y_{iu} - \mu_{iu}), R_{it}(y_{it} - \mu_{it})\}|H_{i,t-1}(y)] \\ &= E[\{R_{iu}R_{it}(y_{iu} - \mu_{iu})(y_{it} - \mu_{it})\}|H_{i,t-1}(y)] - E[\{R_{iu}(y_{iu} - \mu_{iu})\}|H_{i,t-1}(y)] \\ &\times E[\{R_{it}(y_{it} - \mu_{it})\}|H_{i,t-1}(y)] \\ &= (y_{iu} - \mu_{iu})E[\{R_{it}(y_{it} - \mu_{it})\}|H_{i,t-1}(y)] - [w_{iu}(y_{iu} - \mu_{iu})][w_{it}(\lambda_{it} - \mu_{it})] \\ &= [(y_{iu} - \mu_{iu})(1 - w_{iu})][w_{it}(\lambda_{it} - \mu_{it})] \end{aligned} \tag{33}$$

### 3.2.4 A Fully Standardized GQL (FSGQL) Approach

All three estimating equations, namely PSGEE (24), PSGQL (25), and PSCGQL (27) may produce bias estimates, PSGEE being the worst. The reasons for the poor performance of PSGEE are two fold. This is because it completely ignores the missing mechanism and uses a working correlation matrix to accommodate the longitudinal nature of the available data. As opposed to the PSGEE approach, PSGQL approach uses the true correlation structure under a class of auto-correlations but similar to the PSGEE approach it also ignores the missing mechanism. As far as the PSCGQL approach it uses a correct conditional covariance matrix which accommodates both missing mechanism and correlation structure. However, the resulting estimating equation may not unbiased for zero as the history of the responses involved in covariance matrix make a weighted distance function which is not unbiased.

To remedy the aforementioned problems, it is therefore important to use the correct covariance matrix or its consistent estimate to construct the weight matrix by accommodating both missing mechanism and longitudinal correlations of the repeated data. For this to happen, because the distance function is unconditionally unbiased for zero, i.e.,

$$E_{H_i(y)}E[\{\Delta_i(Y_i-\mu_i(\beta))\}|H_i(y)]=0,$$

one must use the unconditional covariance matrix of $\{\Delta_i(Y_i-\mu_i(\beta))\}$ to compute the incomplete longitudinal weight matrix, for the construction of a desired unbiased estimating equation. Let $\Sigma_i^*(\beta,\rho,\gamma)$ denote this unconditional covariance matrix which is computed by using the formula

$$\begin{aligned}\Sigma_i^*(\beta,\rho,\gamma) &= \text{cov}\{\Delta_i(Y_i-\mu_i(\beta))\} = E_{H_i(y)}[\text{cov}\{\Delta_i(Y_i-\mu_i(\beta))\}|H_i(y)] \\ &\quad + \text{cov}_{H_i(y)}[E\{\Delta_i(Y_i-\mu_i(\beta))\}|H_i(y)]. \qquad (34)\end{aligned}$$

In the spirit of Sutradhar (2003), we propose the FSGQL estimating equation for $\beta$ given by

$$\begin{aligned}&\sum_{i=1}^{K}\frac{\partial E_{H_i(y)}E[\{\Delta_i\mu_i(\beta)\}'|H_i(y)]}{\partial\beta}[\text{cov}\{\Delta_i(y_i-\mu_i)\}]^{-1}\{\Delta_i(y_i-\mu_i(\beta))\} \\ &=\sum_{i=1}^{K}\frac{\partial\mu_i'}{\partial\beta}[\Sigma_i^*(\beta,\rho,\gamma)]^{-1}\{\Delta_i(y_i-\mu_i(\beta))\}=0, \qquad (35)\end{aligned}$$

where $\Sigma_i^*(\beta,\rho,\gamma)$ is yet to be computed. This estimating equation is solved iteratively by using

$$\hat{\beta}_{FSGQL}(m+1) = \hat{\beta}_{FSGQL}(m) + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta}[\Sigma_i^*(\beta,\rho,\gamma)]^{-1}\frac{\partial \mu_i(\beta)}{\partial \beta'}\right]_m^{-1}$$

$$\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{'}}{\partial \beta}[\Sigma_i^*(\beta,\rho,\gamma)]^{-1}\Delta_i(y_i-\mu_i(\beta))\right]_m \tag{36}$$

*Computation of* $\Sigma_i^*(\beta,\rho,\gamma) = cov[\Delta_i(y_i-\mu_i)]$

Rewrite (34) as

$$\begin{aligned}\Sigma_i^*(\beta,\rho,\gamma) &= E_{H_i(y)}[\text{cov}\{\Delta_i(Y_i-\mu_i(\beta))\}|H_i(y)] \\ &+ \text{cov}_{H_i(y)}[E\{\Delta_i(Y_i-\mu_i(\beta))\}|H_i(y)] \\ &= E_{H_i(y)}[\Sigma_{ich}^*(\beta,\rho)] + \text{cov}_{H_i(y)}[E_{ich}(\beta,\rho)],\end{aligned} \tag{37}$$

where $\Sigma_{ich}^*(\beta,\rho)$ is constructed by (30) by using the formulas from (31) to (33), and $E_{ich}(\beta,\rho)$ has the form $E_{ich}(\beta,\rho) = [(y_{i1}-\mu_{i1}),(\lambda_{i2}-\mu_{i2}),\ldots,(\lambda_{iT_i}-\mu_{iT_i})]'$.

It then follows that the components of the $T_i \times T_i$ unconditional covariance matrix $\Sigma_i^*(\beta,\rho,\gamma)$ are given by

$$\text{cov}[\delta_{iu}(y_{iu}-\mu_{iu}),\delta_{it}(y_{it}-\mu_{it})] \tag{38}$$

$$= \begin{cases} \text{var}_{y_{i1}}(y_{i1}-\mu_{i1}) = \sigma_{i11} & \text{for u=t=1} \\ E_{H_i(y)}[w_{it}^{-1}\{\sigma_{ic,tt}+(\lambda_{it}-\mu_{it})^2\}-(\lambda_{it}-\mu_{it})^2]+E_{H_i(y)}(\lambda_{it}-\mu_{it})^2, & \text{for u=t=2,}\ldots \\ E_{H_i(y)}[(y_{i1}-\mu_{i1})(\lambda_{it}-\mu_{it})] & \text{for u=1,t=2,}\ldots \\ E_{H_i(y)}[(w_{iu}^{-1}-1)\{(y_{iu}-\mu_{iu})(\lambda_{it}-\mu_{it})\}] & \\ \quad +E_{H_i(y)}[(\lambda_{iu}-\mu_{iu})(\lambda_{it}-\mu_{it})], & \text{for u=2,}\ldots;\ u<t \end{cases}$$

(a). Example of $\Sigma_i^*(\beta,\rho,\gamma)$ under linear longitudinal models with $T=2$

Note that $R_{i1} = r_{i1} = 1$ always. But $R_{i2}$ can be 1 or 0 and under MAR, its probability depends on $y_{i1}$. Consider

$$Pr[R_{i1}=1] = g_{i1} = w_{i1} = 1.0$$

$$P[R_{i2}=1|r_{i1}=1,y_{i1}] = g_{i2}(\gamma) = \frac{\exp\{1+\gamma y_{i1}\}}{1+\exp\{1+\gamma y_{i1}\}} \tag{39}$$

by (20), yielding

$$w_{i2} = E[R_{i2}|H_{i1(y)}] = P[R_{i1} = 1, R_{i2} = 1|H_{i1}(y)]$$
$$= P[R_{i1} = 1]P[R_{i2} = 1|H_{i1}(y)] = g_{i1}g_{i2}(y_{i1}),$$

(see also (18)). With regard to the longitudinal model for potential responses $y_{i1}, y_{i2}$, along with their non-stationary (time dependent covariates), consider the model as:

$$y_{it} \sim (x'_{it}\beta, \frac{\sigma^2}{1-\rho^2}),\ \text{corr}(Y_{it}, Y_{i,t+\ell}) = \rho^{\ell}. \tag{40}$$

Assuming normal distribution, one may write

$$E[Y_{it}|H_{i,t-1}] = x'_{it}\beta + [\text{cov}(y_{it}, y_{i,t-\ell})][\text{var}(y_{i,t-\ell})]^{-1}(y_{i,t-\ell} - x'_{i,t-\ell}\beta)$$
$$= x'_{it}\beta + [\frac{\sigma^2\rho^{\ell}}{1-\rho^2}] = x'_{it}\beta + \rho^{\ell}[y_{i,t-\ell} - x'_{i,t-\ell}\beta]. \tag{41}$$

When the response $y_{it}$ depends on its immediate history, the conditional mean has the formula

$$E[Y_{it}|y_{i,t-1}] = \lambda_{it} = x'_{it}\beta + \rho(y_{i,t-1} - x'_{i,t-1}\beta),$$

implying that the unconditional mean is given by $\mu_{it} = E[Y_{it}] = x'_{it}\beta$, which is the same as the mean in (40), as expected.

Now following (38), we provide the elements of the $2 \times 2$ matrix $\Sigma_i^*(\beta, \rho, \gamma)$ as

$$\sigma^*_{i11} = \frac{\sigma^2}{1-\rho^2}$$

$$\sigma^*_{i12} = \sigma^*_{i21} = \rho\,\text{var}[Y_{i1} - x_{i11}\beta] = \rho\frac{\sigma^2}{1-\rho^2}$$

$$\sigma^*_{i22} = E_{y_{i1}}[w_{i2}^{-1}\{\text{var}(Y_{i2}|y_{i1}) + (\lambda_{i2} - \mu_{i2})^2\}]$$
$$= E_{y_{i1}}[\{1 + \frac{1}{\exp(1+\gamma y_{i1})}\}\{\sigma^2 + \rho^2(y_{i1} - x_{i11}\beta)^2\}]$$
$$= \frac{\sigma^2}{1-\rho^2} + \sigma^2 E_1 + \rho^2 E_2, \tag{42}$$

where

$$E_1 = \int [\frac{1}{\exp(1+\gamma y_{i1})}] g_N(y_{i1}) dy_{i1},\ \text{and}$$
$$E_2 = \int [\frac{\{y_{i1} - x_{i11}\beta\}^2}{\exp(1+\gamma y_{i1})}] g_N(y_{i1}) dy_{i1},$$

$g_N(y_{i1})$ being the normal (say) density of $y_{i1}$. Thus, $\Sigma_i^*(\beta,\rho,\gamma)$ has the formula

$$\Sigma_i^*(\beta,\rho,\gamma)=\frac{\sigma^2}{1-\rho^2}\begin{bmatrix}1 & \rho\\ \rho & \{1+(1-\rho^2)E_1+\frac{\rho^2(1-\rho^2)}{\sigma^2}E_2\}\end{bmatrix}, \tag{43}$$

Note that in the complete longitudinal case $w_{i2}$ would be 1 and $\sigma_{i22}^*$ would reduce to $\frac{\sigma^2}{1-\rho^2}$, leading to

$$\Sigma_i^*(\beta,\rho,\gamma)=\Sigma_i(\beta,\rho)=\frac{\sigma^2}{1-\rho^2}\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}, \tag{44}$$

which is free from $\beta$ in this linear model case, and the *PSGEE* (24) uses a "working" version of (44), namely

$$V_i(\alpha)=\frac{\sigma^2}{1-\rho^2}\begin{bmatrix}1 & \alpha\\ \alpha & 1\end{bmatrix}, \tag{45}$$

whereas the FSGQL estimating equation (35) would use $\Sigma_i^*(\beta,\rho,\gamma)$ from (43). This shows the effect of missing mechanism in the construction of the weight matrix for the estimating equation.

(b). Example of $\Sigma_i^*(\beta,\rho,\gamma)$ under binary longitudinal AR(1) model with $T=2$

Consider a binary AR(1) model with

$$\lambda_{it}=E[Y_{it}|y_{i,t-1}]=\mu_{it}+\rho(y_{i,t-1}-\mu_{i,t-1}),\ t=2,\ldots,T, \tag{46}$$

where $\mu_{it}=\frac{\exp(x_{it}'\beta)}{1+\exp(x_{it}'\beta)}$, for all $t=1,\ldots,T$.

Now considering $y_{i1}$ as fixed, by using (31)–(33) we first compute the history-dependent conditional covariance matrix $\Sigma_{ich}(\beta,\rho)=\text{cov}[\{\Delta_i(y_i-\mu_i)\}|H_i(y)]$ as:

$$\begin{aligned}&\text{var}[\delta_{i1}(y_{i1}-\mu_{i1})]=0\\ &\text{var}[\{\delta_{i2}(y_{i2}-\mu_{i2})\}|y_{i1}]=\frac{1}{w_{i2}}[\lambda_{i2}(1-\lambda_{i2})+\rho^2(y_{i1}-\mu_{i1})^2]-\rho^2(y_{i1}-\mu_{i1})^2\\ &\text{cov}[\{\delta_{i1}(y_{i1}-\mu_{i1}),\delta_{i2}(y_{i2}-\mu_{i2})\}|y_{i1}]=0,\end{aligned} \tag{47}$$

yielding

$$\begin{aligned}&E_{H_i(y)}[\Sigma_{ich}(\beta,\rho)] \\ &=\begin{cases}E_{y_{i1}}[\sigma_{ich,11}]=E_{y_{i1}}[0]=0\\ E_{y_{i1}}[\sigma_{ich,22}]=E_{y_{i1}}[\frac{1}{w_{i2}}[\lambda_{i2}(1-\lambda_{i2})+\rho^2(y_{i1}-\mu_{i1})^2]-\rho^2(y_{i1}-\mu_{i1})^2]\\ E_{y_{i1}}[\sigma_{ich,12}]=E_{y_{i1}}[0]=0\end{cases}\end{aligned} \tag{48}$$

Next because

$$E_{ich}(\beta,\rho) = E\{\Delta_i(y_i-\mu_i)\}|H_i(y)] = [(y_{i1}-\mu_{i1}),(\lambda_{i2}-\mu_{i2})]',$$

one obtains

$$\begin{aligned} &\mathrm{cov}_{H_i(y)}[E_{ich}(\beta,\rho)] \qquad (49)\\ &= \begin{cases} \mathrm{var}_{y_{i1}}[y_{i1}-\mu_{i1}] = \mu_{i1}[1-\mu_{i1}] \\ \mathrm{cov}_{y_{i1}}[(y_{i1}-\mu_{i1}),(\lambda_{i2}-\mu_{i2})] = \rho\,\mathrm{var}_{y_{i1}}[y_{i1}-\mu_{i1}] = \rho\mu_{i1}[1-\mu_{i1}] \\ \mathrm{var}_{y_{i1}}[\lambda_{i2}-\mu_{i2}] = \mathrm{var}_{y_{i1}}[\rho(y_{i1}-\mu_{i1})] = \rho^2[\mu_{i1}(1-\mu_{i1})]. \end{cases} \end{aligned}$$

By combining (48) and (49), it follows from (38) that the $2\times 2$ unconditional covariance matrix $\Sigma_i^*(\beta,\rho,\gamma)$ has the form

$$\begin{aligned} \mathrm{var}[\delta_{i1}(y_{i1}-\mu_{i1})] &= \mu_{i1}[1-\mu_{i1}] \\ \mathrm{var}[\delta_{i2}(y_{i2}-\mu_{i2})] &= E_{y_{i1}}[\frac{1}{w_{i2}}\{\lambda_{i2}(1-\lambda_{i2})+\rho^2(y_{i1}-\mu_{i1})^2\}] \\ &= [\mu_{i2}(1-\mu_{i2})]E[w_{i2}^{-1}] \qquad (50)\\ &\quad +\rho(1-2\mu_{i2})E[w_{i2}^{-1}(y_{i1}-\mu_{i1})] \\ &= [\mu_{i2}(1-\mu_{i2})]E_{1y}+\rho(1-2\mu_{i2})[E_{2y}-\mu_{i1}E_{1y}] \\ \mathrm{cov}[\delta_{i1}(y_{i1}-\mu_{i1}),\delta_{i2}(y_{i2}-\mu_{i2})] &= \rho[\mu_{i1}\{1-\mu_{i1}\}], \qquad (51) \end{aligned}$$

where

$$\begin{aligned} E_{1y} &= E[w_{i2}^{-1}] = \{1+\exp(-1)+\mu_{i1}\exp(-1)(\exp(-\gamma)-1)\} \\ E_{2y} &= E[\frac{y_{i1}}{w_{i2}}] = \mu_{i1}\{1+\exp(-\gamma-1)\}. \end{aligned}$$

*General formula for $\Sigma_i^*(\beta,\rho,\gamma)$ under the binary AR(1) model*

In general, it follows from (38) that the elements of the $T_i\times T_i$ unconditional covariance matrix $\Sigma_i^*(\beta,\rho,\gamma)$ under AR(1) binary model are given by

$$\begin{aligned} &\mathrm{cov}[\delta_{iu}(y_{iu}-\mu_{iu}),\delta_{it}(y_{it}-\mu_{it})] \qquad (52)\\ &\equiv \begin{cases} \sigma^*_{i,11} = \mu_{i1}[1-\mu_{i1}] \\ \sigma^*_{i,tt} = E_{H_i(y)}[w_{it}^{-1}\{\mu_{it}(1-\mu_{it})+\rho(1-2\mu_{it})(y_{i,t-1}-\mu_{i,t-1})\}, (\text{for } t=2,\ldots,T_i) \\ \sigma*_{i,ut} = \rho\rho^{t-1-u}\mu_{iu}(1-\mu_{iu}),\ (\text{for } u=1<t) \\ \sigma^*_{i,ut} = \rho^2\rho^{t-u}\mu_{i(u-1)}(1-\mu_{i(u-1)}),\ (\text{for } 1<u<t). \end{cases} \end{aligned}$$

### 3.3 An Empirical Illustration

First, to illustrate the performance of the existing PSGEE (24) approach, we refer to some of the simulation results reported by Sutradhar and Mallick (2010). It was shown that this approach may produce highly biased and hence inconsistent regression estimates. In fact these authors also demonstrated that PSGEE(I) (independence assumption based) approach produces less biased estimates than any "working" correlation structures based PSGEE approaches. For example, we consider here their simulation design chosen as

*Simulation Design*

$K = 100$, $T = 4$, $p = 2$, $q = 1$, $\gamma = 4$, $\rho = 0.4, 0.8$, $\beta_1 = \beta_2 = 0$ along with two time-dependent covariates:

$$x_{it1} = \begin{cases} \frac{1}{2} & \text{for } i = 1, \cdots, \frac{K}{4}; t = 1,2 \\ 0 & \text{for } i = 1, \cdots, \frac{K}{4}; t = 3,4 \\ -\frac{1}{2} & \text{for } i = \frac{K}{4}+1, \cdots, \frac{3K}{4}; t = 1 \\ 0 & \text{for } i = \frac{K}{4}+1, \cdots, \frac{3K}{4}; t = 2,3 \\ \frac{1}{2} & \text{for } i = \frac{K}{4}+1, \cdots, \frac{3K}{4}; t = 4 \\ \frac{t}{2T} & \text{for } i = \frac{3K}{4}+1, \cdots, K; t = 1, \cdots, 4 \end{cases}$$

and

$$x_{it2} = \begin{cases} \frac{t-2.5}{2T} & \text{for } i = 1, \cdots, \frac{K}{2}; t = 1, \cdots, 4 \\ 0 & \text{for } i = \frac{K}{2}+1, \cdots, K; t = 1,2 \\ \frac{1}{2} & \text{for } i = \frac{K}{2}+1, \cdots, K; t = 3,4 \end{cases}$$

Details on the MAR based incomplete binary data generation, one may be referred to Sutradhar and Mallick (2010, Sect. 2.1). Based on 1,000 simulations, the PSGEE estimates obtained from (24) and PSGEE (I) obtained from (24) by using zero correlation are displayed in Table 1.

These results show that the PSGEE estimates for $\beta_1 = 0$ and $\beta_2 = 0$ are highly biased. For example, when $\rho = 0.8$, the estimates of $\beta_1$ and $\beta_2$ are $-0.213$ and $-0.553$, respectively. These estimates are inconsistent and unacceptable. Note that these biases are caused by the wrong correlation matrix used to construct the PSGEE (24), whereas this PSGEE provides almost unbiased estimates when data are treated to be independent even if truly they are not so. However the standard errors of the PSGEE(I) estimates appear to be large and hence it may provide inefficient estimates. In fact when the proportion of missing values is large, the PSGEE(I) will also encounter estimation breakdown or it will produce biased estimates. This

**Table 1** Simulated means (SMs), simulated standard errors (SSEs), and simulated mean squared errors (SMSEs) for "working" correlations based PSGEE (24) estimates, when the incomplete longitudinal responses were generated based on MAR mechanism (20) with $\gamma = 4.0$ and a longitudinal AR(1) correlation structure with correlation index parameter $\rho$; $\beta_1 = \beta_2 = 0$; based on 1,000 simulations

| | | Estimation approach | | | | |
|---|---|---|---|---|---|---|
| | | PSGEE(AR(1)) | | | PSGEE (I) | |
| $\rho$ | Statistic | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\rho}$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ |
| 0.4 | SM | −0.076 | −0.224 | 0.404 | 0.015 | 0.015 |
| | SSE | 0.361 | 0.544 | 0.062 | 0.384 | 0.587 |
| | SMSE | 0.136 | 0.346 | 0.004 | 0.148 | 0.344 |
| 0.8 | SM | −0.213 | −0.553 | 0.802 | 0.007 | 0.017 |
| | SSE | 0.257 | 0.381 | 0.038 | 0.378 | 0.614 |
| | SMSE | 0.112 | 0.450 | 0.001 | 0.143 | 0.377 |

is verified by a simulation study reported by Mallick et al. (2013). The reason for this inconsistency encountered by PSGEE and PSGEE(I) is the failure of accommodating MAR mechanism in the covariance matrix used as the longitudinal weights.

As a remedy to this inconsistency, we have developed a FSGQL (35) estimating equation by accommodating both MAR mechanism and longitudinal correlation structure in constructing the weight matrix $\Sigma_i^*(\beta, \rho, \gamma)$. This FSGQL equation would provide consistent and efficient regression estimates. For simplicity, Mallick et al. (2013) have demonstrated through a simulation study that FSGQL(I) approach by using $\rho = 0$ in $\Sigma_i^*(\beta, \rho = 0, \gamma)$ produces almost unbiased estimates with small variances. This provides a guidance that ignoring missing mechanism in constructing the weight matrix would provide detrimental results, whereas ignoring longitudinal correlations does not appear to cause any significant loss.

**Acknowledgment** The author fondly acknowledges the stimulating discussion by the audience of the symposium and wishes to thank for their comments and suggestions.

## References

Birmingham, J., Rotnitzky, A., Fitzmaurice, G.M.: Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. J. R. Stat. Soc. Ser. B **65**, 275–297 (2003)

Crowder, M.: On the use of a working correlation matrix in using generalized linear models for repeated measures. Biometrika **82**, 407–410 (1995)

Fitzmaurice, G.M., Laird, N.M., Zahner, G.E.P.: Multivariate logistic models for incomplete binary responses. J. Am. Stat. Assoc. **91**, 99–108 (1996)

Ibrahim, J.G., Lipsitz, S.R., Chen, M.H.: Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. J. R. Stat. Soc. Ser. B **61**, 173–190 (1999)

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R.: Missing responses in generalized linear mixed models when the missing data mechanism is non-ignorable. Biometrika **88**, 551–564 (2001)

Krishnamoorthy, K., Pannala, M.K.: Confidence estimation of a normal mean vector with incomplete data. Can. J. Stat. **27**, 395–407 (1999)
Laird, N.M.: Missing data in longitudinal studies. Stat. Med. **7**, 305–315 (1988)
Liang, K.-Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
Little, R.J.A.: A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc. **83**, 1198–1202 (1988)
Little, R.J.A.: Modeling the drop-out mechanism in repeated-measures studies. J. Am. Stat. Assoc. **90**, 1112–1121 (1995)
Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
Lord, F.M.: Estimation of parameters from incomplete data. J. Am. Stat. Assoc. **50**, 870–876 (1995)
Mallick, T., Farrell, P.J., Sutradhar, B.C.: Consistent estimation in incomplete longitudinal binary models. In: Sutradhar, B.C. (ed.) ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers. Springer Lecture Notes Series, pp. 125146. Springer, New York, (2013)
Mehta, J.S., Gurland, J.: A test of equality of means in the presence of correlation and missing values. Biometrika **60**, 211–213 (1973)
Meng, X.L.: Multiple-imputation inferences with uncongenial sources of input. Stat. Sci. **9**, 538–573 (1994)
Morrison, D.F.: A test for equality of means of correlated variates with missing data on one response. Biometrika **60**, 101–105 (1973)
Naik, U.D.: On testing equality of means of correlated variables with incomplete data. Biometrika **62**, 615–622 (1975)
Paik, M.C.: The generalized estimating equation approach when data are not missing completely at random. J. Am. Stat. Assoc. **92**, 1320–1329 (1997)
Preisser, J.S., Lohman, K.K., Rathouz, P.J.: Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Stat. Med. **21**, 3035–3054 (2002)
Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data . J. Am. Stat. Assoc. **90**, 106–121 (1995)
Rotnitzky, A., Robins, J.M., Scharfstein, D.O.: Semi-parametric regression for repeated outcomes with nonignorable nonresponse. J. Am. Stat. Assoc. **93**, 1321–1339 (1998)
Rubin, D.B.: Inference and missing data (with discussion). Biometrika **63**, 581–592 (1976)
Rubin, D.B., Schenker, N.: Multiple imputation for interval estimation from simple random sample with ignorable nonresponses. J. Am. Stat. Assoc. **81**, 366–374 (1986)
Sneddon, G., Sutradhar, B.C.: On semi-parametric familial longitudinal models. Statist. Prob. Lett. **69**, 369–379 (2004)
Sutradhar, B.C.: An overview on regression models for discrete longitudinal responses. Stat. Sci. **18**, 377–393 (2003)
Sutradhar, B.C.: Inferences in generalized linear longitudinal mixed models. Can. J. Stat. **38**, 174–196 (2010), Special issue
Sutradhar, B.C.: Dynamic Mixed Models for Familial Longitudinal Data. Springer, New York (2011)
Sutradhar, B.C., Das, K.: On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika **86**, 459–65 (1999)
Sutradhar, B.C., Mallick, T.S.: Modified weights based generalized quasilikelihood inferences in incomplete longitudinal binary models. Can. J. Stat. **38**, 217–231 (2010), Special issue
Troxel, A.B., Lipsitz, S.R., Harrington, D.P.: Marginal models for the analysis of longitudinal measurements subject to non-ignorable and non-monotonic missing data. Biometrika **85**, 661–672 (1988)
Troxel, A.B., Lipsitz, S.R., Brennan, T.A.: Weighted estimating equations with nonignorably missing response data. Biometrics **53**, 857–869 (1997)
Wang, Y.-G.: Estimating equations with nonignorably missing response data. Biometrics **55**, 984–989 (1999)

# Consistent Estimation in Incomplete Longitudinal Binary Models

**Taslim S. Mallick, Patrick J. Farrell, and Brajendra C. Sutradhar**

**Abstract** It is well known that in the complete longitudinal setup, the so-called working correlation-based generalized estimating equations (GEE) approach may yield less efficient regression estimates as compared to the independence assumption-based method of moments and quasi-likelihood (QL) estimates. In the incomplete longitudinal setup, there exist some studies indicating that the use of the same "working" correlation-based GEE approach may provide inconsistent regression estimates especially when the longitudinal responses are at risk of being missing at random (MAR). In this paper, we revisit this inconsistency issue under a longitudinal binary model and empirically examine the relative performance of the existing weighted (by inverse probability weights for the missing indicator) GEE (WGEE), a fully standardized GQL (FSGQL) and conditional GQL (CGQL) approaches. In the comparative study, we consider both stationary and non-stationary covariates, as well as various degrees of missingness and longitudinal correlation in the data.

## 1 Introduction

Consider a longitudinal binary data setup where $y_{it}$ is the Bernoulli response for the $i$-th ($i = 1, \cdots, K$) individual at the $t$-th time point ($t = 1, \cdots, T$) and $x_{it} = (x_{it1}, \cdots, x_{itu}, \cdots, x_{itp})'$ is the associated $p$-dimensional covariate vector. When the longitudinal data are complete (that is, there are no missing responses from any of the individuals in the study), an estimating approach such as generalized quasi-

T.S. Mallick • P.J. Farrell (✉)
School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada
e-mail: pfarrell@math.carleton.ca

B.C. Sutradhar
Memorial University, St. John's, NF, Canada, A1C5S7
e-mail: bsutradh@mun.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_6, 

likelihood (GQL) can be used to obtain an estimate of the regression parameter vector, $\beta$, that is both consistent and efficient, provided that the correlation structure associated with the repeated binary responses is known (see Sutradhar 2003). In order to describe the longitudinal correlation in the data, it seems reasonable to assume deterioration in the association between observations on the same individuals that are further apart in time. Thus, to achieve this, we let $\rho$ be a longitudinal correlation parameter and consider a conditional linear binary dynamic (CLBD) model proposed by Zeger et al. (1985) (see also Qaqish 2003), which is given by

$$P(Y_{i1}=1)=\mu_{i1}, \text{ and}$$
$$P(Y_{it}=1 \mid y_{i,t-1})=\mu_{it}+\rho(y_{i,t-1}-\mu_{i,t-1})=\lambda_{i,t|t-1}(\beta,\rho)=\lambda_{it}, \text{ for } t=2,\cdots,T \tag{1}$$

with $\mu_{it}=exp(x_{it}^{'}\beta)/[1+exp(x_{it}^{'}\beta)]$ for $t=1,\cdots,T$. According to model (1), the marginal means and variances of $y_{it}$ are

$$E(Y_{it})=\mu_{it} \tag{2}$$

and

$$Var(Y_{it})=\sigma_{i,tt}=\mu_{it}(1-\mu_{it}), \tag{3}$$

while the correlations between $Y_{it}$ and $Y_{i,t+l}$ for $l=1,\cdots,T-1, t=1,\cdots,T-l$ are given by

$$corr(Y_{it},Y_{i,t+l})=\rho^{l}\left[\frac{\sigma_{i,tt}}{\sigma_{i,t+l,t+l}}\right]^{1/2}. \tag{4}$$

The means, variances, and covariances defined by (2) through (4) are nonstationary, since they are all functions of time-dependent covariates $\{x_{it}\}$. However, if the $\sigma_{i,tt}$ are not extremely different, the correlations given by (4) assume a behavior that is analogous to an autoregressive process of order one, AR(1). Under the present model, the correlation parameter $\rho$ must satisfy the range restriction

$$max\left[-\frac{\mu_{it}}{1-\mu_{i,t-1}},-\frac{1-\mu_{it}}{\mu_{i,t-1}}\right]\leq\rho\leq min\left[\frac{1-\mu_{it}}{1-\mu_{i,t-1}},\frac{\mu_{it}}{\mu_{i,t-1}}\right]. \tag{5}$$

Suppose that we let $\mu_i$ and $\Sigma_i(\rho)$ represent the mean vector and the covariance matrix of the complete data vector $Y_i$, where $\mu_i=(\mu_{i1},\cdots,\mu_{it},\cdots,\mu_{iT})^{'}$ and $\Sigma_i(\rho)=A_i^{1/2}C_i(\rho)A_i^{1/2}$. Here, $C_i(\rho)$ is the $T\times T$ correlation matrix based on (4), and $A_i=diag(\sigma_{i,11},\cdots,\sigma_{i,tt},\cdots,\sigma_{i,TT})$. An estimator for $\beta$ that is both consistent and highly efficient can be obtained by solving the GQL estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i^{'}}{\partial \beta} [\Sigma_i(\rho)]^{-1} (y_i - \mu_i) = 0, \tag{6}$$

(Sutradhar 2003).

In practice, it is typically the case that some of the responses associated with each of a number of individuals in the study may be missing. To acknowledge this phenomenon during the data collection process, we introduce an indicator variable $R_{it}$, that takes on a value of one if $Y_{it}$ is observed, and zero otherwise. For purposes of our investigation here, we adopt the not-so-unreasonable assumption that all individuals provide a response at the first time point, so that $R_{i1} = 1$ for all $i = 1, \cdots, K$. We also assume monotonic missingness, suggesting that the $R_{it}$ satisfy the inequality $R_{i1} \geq R_{i2} \geq \cdots \geq R_{it} \geq \cdots \geq R_{iT}$. Thus, if responses are no longer observed for the $i$-th individual after the $j$-th time point, for this individual we would have available $y_{it}$ for $t = 1, \cdots, T_i = j$.

Regarding the missing data mechanism, at this time we distinguish between responses that are missing completely at random, MCAR, and those that are missing at random, MAR (see Fitzmaurice et al. 1996; Paik 1997; Rubin 1976). When the responses are MCAR, the indicator variable $R_{it}$ reflecting the presence or absence of $Y_{it}$ does not depend on the previous responses $Y_{i1}, \cdots, Y_{i,t-1}$. In this instance, if we define $R_i = diag(R_{i1}, \cdots, R_{iT})$ and incorporate this matrix into the estimating equation given by (6) to yield

$$\sum_{i=1}^{K} \frac{\partial \mu_i^{'}}{\partial \beta} [\Sigma_i(\rho)]^{-1} R_i (y_i - \mu_i) = 0, \tag{7}$$

it is still possible to obtain an unbiased estimator for $\beta$ that will be consistent and efficient. Note that $\Sigma_i(\rho)$ is a $T \times T$ matrix with appropriate variance and covariance entries in the first $T_i$ rows and $T_i$ columns and zeroes in the last $T - T_i$ rows and columns. On the other hand, when the missing data mechanism for the responses is assumed to be MAR (implying that $R_{it}$ does depend on the previous responses $Y_{i1}, \cdots, Y_{i,t-1}$), it can be shown that $E[R_{it}(Y_{it} - \mu_{it})] \neq 0$. In this situation, the estimator for $\beta$ based on (7) will be biased and inconsistent. Upon realizing this to be the case, many studies have attempted to correct for this problem by using a modified inverse probability-weighted distance function

$$w_{it}^{-1} \{H_{i,t-1}(y); \alpha\} [R_{it}(Y_{it} - \mu_{it})], \tag{8}$$

where $H_{i,t-1}(y) \equiv H_{i,t-1} = (Y_{i1}, \cdots, Y_{i,t-1})$, so that the expectation of (8) is zero. Following Robins et al. (1995), for data that are MAR, we can write the probability weight $w_{it}\{H_{i,t-1}(y); \alpha\} = w_{it}$ as a function of past responses as follows. Specifically, imagine that the probability that the $i$-th individual responds at the $j$-th time point depends on the past lag $q$ responses, where $q \leq j - 1$. Letting $g_{ij}(y_{i,j-1}, \cdots, y_{i,j-q}; \alpha)$ represent this probability, we can write $g_{ij}(y_{i,j-1}, \cdots, y_{i,j-q}; \alpha) = P(R_{ij} = 1 \mid R_{i1} = 1, \cdots, R_{i,j-1} = 1; y_{i,j-1}, \cdots, y_{i,j-q})$, which can be modeled as

$$g_{ij}(y_{i,j-1},\cdots,y_{i,j-q};\alpha)=\frac{exp(1+\sum_{l=1}^{q}\alpha_l y_{i,j-l})}{1+exp(1+\sum_{l=1}^{q}\alpha_l y_{i,j-l})}, \tag{9}$$

where $\alpha_l$ is a parameter that reflects the dependence of $R_{ij}$ on $y_{i,j-l}$ for all $l=1,\cdots,q$. Robins et al. (1995) set

$$\begin{aligned} w_{it} &= P(R_{it}=1,R_{i,t-1}=1,\cdots,R_{i1}=1\mid H_{i,t-1}) \\ &= P(R_{it}=1\mid R_{i,t-1}=\cdots=R_{i1}=1;H_{i,t-1})\times \\ &\quad P(R_{i,t-1}=1\mid R_{i,t-2}=\cdots=R_{i1}=1;H_{i,t-2})\times \\ &\quad \cdots\times P(R_{i2}=1\mid R_{i1}=1;H_{i1})P(R_{i1}=1) \\ &= \prod_{j=1}^{t} g_{ij}(y_{i,j-1},\cdots,y_{i,j-q};\alpha). \end{aligned} \tag{10}$$

Since monotonic missingness is assumed

$$E\left[R_{it}Y_{it}\mid H_{i,t-1}\right]=P\left[R_{i1}=1,R_{i2}=1,\cdots,R_{it}=1;Y_{it}=1\mid H_{i,t-1}\right], \tag{11}$$

or, alternatively

$$\begin{aligned} E\left[R_{it}Y_{it}\mid H_{i,t-1}\right] &= P(R_{i1}=1)P\left[R_{i2}=1\mid R_{i1}=1;H_{i1}\right]\cdots \\ &\quad P\left[R_{it}=1\mid R_{i1}=1,R_{i2}=1,\cdots,R_{i,t-1}=1;H_{i,t-1}\right] \\ &\quad P\left[Y_{it}=1\mid H_{i,t-1}\right]. \end{aligned} \tag{12}$$

Using model (1) and $g_{ij}(y_{i,j-1},\cdots,y_{i,j-q};\alpha)$ given in (9), (12) becomes

$$\begin{aligned} E\left[R_{it}Y_{it}\mid H_{i,t-1}\right] &= \prod_{j=1}^{t} g_{ij}(y_{i,j-1},\cdots,y_{i,j-q};\alpha)\lambda_{it} \\ &= w_{it}\lambda_{it}, \end{aligned} \tag{13}$$

which implies that

$$E\left[\frac{R_{it}Y_{it}}{w_{it}}\mid H_{i,t-1}\right]=\lambda_{it}, \tag{14}$$

thus giving

$$E_{H_{i,t-1}}E\left[\frac{R_{it}Y_{it}}{w_{it}}\mid H_{i,t-1}\right]=E_{H_{i,t-1}}[\lambda_{it}]=\mu_{it}. \tag{15}$$

Similarly

$$E_{H_{i,t-1}} E\left[\frac{R_{it}\mu_{it}}{w_{it}} \mid H_{i,t-1}\right] = \mu_{it}, \tag{16}$$

suggesting that combining (15) and (16) yields

$$E_{H_{i,t-1}} E\left[\frac{R_{it}(Y_{it}-\mu_{it})}{w_{it}} \mid H_{i,t-1}\right] = 0. \tag{17}$$

This unconditional unbiasedness property of the weighted distance or estimating function $\left[\frac{R_{it}(Y_{it}-\mu_{it})}{w_{it}}\right]$ motivated many researchers to write a weighted generalized estimating equation (WGEE) and solve it for the $\beta$ involved in those $\mu_{it}$. The WGEE, first developed by Robins et al. (1995), is reproduced in brief, in Sect. 2.1. Note that to construct the WGEE, Robins et al. (1995) suggested the specification of a user-selected covariance matrix of $\{(Y_{it}-\mu_{it}), t = 1, \ldots, T_i\}$ by pretending that as though the data were complete. Recently, Sutradhar and Mallick (2010) have found that this widely used WGEE approach produces highly biased regression estimates, indicating consistency break down. In this paper, specifically in Sect. 3, we carry out an extensive simulation study considering various degrees of missingness and examine further the inconsistency problem encountered by the WGEE approach.

In Sect. 2.2, we consider a simpler version of a fully standardized GQL (FSGQL) approach discussed by Sutradhar (2013, Sect. 3.2.4) by constructing the weight matrix, that is, unconditional covariance matrix of $\{\left[\frac{R_{it}(Y_{it}-\mu_{it})}{w_{it}}\right], t = 1, \ldots, T_i\}$ using longitudinal independence $(i.e., \rho = 0)$. We will refer to this as the FSGQL(I) approach. In the simulation study in Sect. 3, we examine the relative performance of this FSGQL(I) approach with the existing WGEE as well as WGEE(I) (independence assumption-based WGEE) approach.

Further note that if the correlation model for the complete data were known through $\lambda_{it}$ in (14), one could exploit the conditional distance function $\left[\frac{R_{it}(Y_{it}-\lambda_{it})}{w_{it}} \mid H_{i,t-1}\right]$ to construct a conditional-weighted GQL (CWGQL) estimating equation and solve such an equation to obtain consistent regression estimates. We discuss this approach in Sect. 2.3 and include it in the simulation study in Sect. 3 to examine its performance as compared to the aforementioned approaches.

## 2 Estimation

### *2.1 WGEE Approach*

Robins et al. (1995, Eq. (10), p. 109) used the result in (17) to propose the WGEE

$$\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} [V_i(\alpha^*)]^{-1} \Delta_i (y_i - \mu_i) = 0, \tag{18}$$

(see also Paik 1997, Eq. (1), p. 1321) where $\Delta_i = diag(\delta_{i1}, \delta_{i2}, \cdots, \delta_{iT})$ with $\delta_{it} = R_{it}/w_{it}\{H_{i,t-1}(y); \alpha\} = R_{it}/w_{it}$. The quantity $V_i(\alpha^*)$ is a working covariance matrix of $Y_i$ (see Liang and Zeger 1986) that is used in an effort to increase the efficiency of the estimates. Of note is the fact while Robins et al. (1995) suggested a WGEE, they did not account for the missingness in the data when specifying $V_i(\alpha^*)$; they simply based their working covariance matrix on the complete data formulae. For this reason, this WGEE approach may be referred to as a partially standardized GEE (PSGEE) approach. See the previous article by Sutradhar (2013) in this chapter for details on the use of PSGEE. Note that a user-selected covariance matrix based on complete data that ignores the missing mechanism leads the WGEE to be unstable, in particular, when the proportion of missing data is high, causing breakdown in estimation, i.e., breakdown in consistency. However, this inconsistency issue has not been adequately addressed in the literature including the studies by Robins et al. (1995), Paik (1997), Rotnitzky et al. (1998) and Birmingham et al. (2003). One of the main reasons is that none of the studies used any stochastic correlation structure in conjunction with the missing mechanism to model the binary data in the incomplete longitudinal setup.

In this paper, in order to investigate the effect on the estimates of the regression parameter vector, we propose to replace the working covariance matrix $V_i(\alpha^*)$ in (18) with a proper unconditional covariance matrix that accommodates the missingness in the data. The proposed approach is presented in the next section.

## 2.2 FSGQL Approach

The unconditional unbiasedness property in (17), that is,

$$E_{H_{i,t-1}} E\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}} \mid H_{i,t-1}\right] = E_{H_{i,t-1}} E\left[\delta_{it}(Y_{it} - \mu_{it}) \mid H_{i,t-1}\right] = 0$$

motivates one to develop a FSGQL estimating equation for $\beta$, which requires the computation of the unconditional variance of $\delta_{it}(Y_{it} - \mu_{it})$. Thus, for all $t = 1, \ldots, T_i$, we now compute the unconditional covariance matrix, namely

$$\text{cov}[\Delta_i(y_i - \mu_i)] = \Sigma_i^*(\beta, \rho, \alpha), \text{ (say)},$$

by using the formula

$$\begin{aligned}\Sigma_i^*(\beta, \rho, \gamma) &= E_{H_i(y)}[\text{cov}\{\Delta_i(Y_i - \mu_i(\beta))\} | H_i(y)] \\ &\quad + \text{cov}_{H_i(y)}[E\{\Delta_i(Y_i - \mu_i(\beta))\} | H_i(y)],\end{aligned}$$

where $H_i(y)$ denotes the history of responses. For computational details under any specified correlation model, we refer to the previous article by Sutradhar

(2013, Sect. 3.2.4). For the binary AR(1) model in (1), the elements of the $T_i \times T_i$ unconditional covariance matrix $\Sigma_i^*(\beta, \rho, \alpha)$ are given by

$$\begin{aligned}
&\operatorname{cov}[\delta_{iu}(y_{iu} - \mu_{iu}), \delta_{it}(y_{it} - \mu_{it})] \\
&\equiv \begin{cases}
\sigma^*_{i,11} = \mu_{i1}[1 - \mu_{i1}] \\
\sigma^*_{i,tt} = E_{H_i(y)}[w_{it}^{-1}\{\mu_{it}(1 - \mu_{it}) + \rho(1 - 2\mu_{it})(y_{i,t-1} - \mu_{i,t-1})\}, (\text{for } t = 2, \ldots, T_i) \\
\sigma *_{i,ut} = \rho\rho^{t-1-u}\mu_{iu}(1 - \mu_{iu}), \ (\text{for } u = 1 < t) \\
\sigma^*_{i,ut} = \rho^2\rho^{t-u}\mu_{i(u-1)}(1 - \mu_{i(u-1)}), \ (\text{for } 1 < u < t).
\end{cases}
\end{aligned} \tag{19}$$

Note that the formulas in (19) under the present AR(1) binary model may be verified directly. For example, we compute the $t$-th diagonal element of the $\Sigma_i^*(\beta, \rho, \alpha)$ matrix as follows. Since $\delta_{it} = R_{it}/w_{it}\{H_{i,t-1}(y); \alpha\} = R_{it}/w_{it}$, we can write

$$\begin{aligned}
Var\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}}\right] &= Var_{H_{i,t-1}} E\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}} \mid H_{i,t-1}\right] \\
&\quad + E_{H_{i,t-1}} Var\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}} \mid H_{i,t-1}\right],
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
Var_{H_{i,t-1}} E\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}} \mid H_{i,t-1}\right] &= Var_{H_{i,t-1}}\left[\frac{1}{w_{it}} w_{it}(\lambda_{it} - \mu_{it})\right] \\
&= E_{H_{i,t-1}}\left[(\lambda_{it} - \mu_{it})^2\right]
\end{aligned} \tag{21}$$

since $\left[E_{H_{i,t-1}}(\lambda_{it} - \mu_{it})\right]^2 = 0$, and

$$\begin{aligned}
&E_{H_{i,t-1}} Var\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}} \mid H_{i,t-1}\right] \\
&= E_{H_{i,t-1}}\left[\frac{1}{w_{it}}(1 - w_{it})\left\{\lambda_{it}(1 - \lambda_{it}) + (\lambda_{it} - \mu_{it})^2\right\} + \lambda_{it}(1 - \lambda_{it})\right] \\
&= E_{H_{i,t-1}}\left[\frac{1}{w_{it}}\lambda_i(1 - \lambda_{it}) + \frac{1}{w_{it}}(\lambda_{it} - \mu_{it})^2 - (\lambda_{it} - \mu_{it})^2\right]
\end{aligned} \tag{22}$$

Substituting (21) and (22) into (20) gives

$$\begin{aligned}
Var\left[\frac{R_{it}(Y_{it} - \mu_{it})}{w_{it}}\right] &= \mu_{it}(1 - \mu_{it}) E_{H_{i,t-1}}\left(\frac{1}{w_{it}}\right) + \rho(1 - 2\mu_{it}) \times \\
&\quad \left[E_{H_{i,t-1}}\left(\frac{Y_{i,t-1}}{w_{it}}\right) - \mu_{i,t-1} E_{H_{i,t-1}}\left(\frac{1}{w_{it}}\right)\right]
\end{aligned} \tag{23}$$

since $\lambda_{it} = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1})$ by (1). The conditional expectations given the response history, $H_{i,t-1}$, in (23) are evaluated as follows:

$$E_{H_{i,t-1}}\left(\frac{1}{w_{it}}\right) = \sum_{y_{i1},y_{i2},\cdots,y_{i,t-1}} \frac{1}{w_{it}} \mu_{i1}^{y_{i1}}(1-\mu_{i1})^{1-y_{i1}} \prod_{j=2}^{t-1} (\lambda_{ij})^{y_{ij}}(1-\lambda_{ij})^{1-y_{ij}} \quad (24)$$

and

$$E_{H_{i,t-1}}\left(\frac{Y_{i,t-1}}{w_{it}}\right) = \sum_{y_{i1},y_{i2},\cdots,y_{i,t-1}} \left(\frac{y_{i,t-1}}{w_{it}}\right) \mu_{i1}^{y_{i1}}(1-\mu_{i1})^{1-y_{i1}} \prod_{j=2}^{t-1} (\lambda_{ij})^{y_{ij}}(1-\lambda_{ij})^{1-y_{ij}} \quad (25)$$

### 2.2.1 FSGQL(I) Approach

Note that in a complete longitudinal setup, one may obtain consistent regression estimates even if longitudinal correlations are ignored in developing the estimating equation but such estimates may not be efficient (Sutradhar 2011, Chap. 7). By this token, to obtain consistent regression estimates in the incomplete longitudinal setup, we may still use the independence assumption (i.e., use $\rho = 0$) but the missing mechanism must be accommodated to formulate the covariance matrix for the construction of the estimating equation. Thus, for simplicity, we now consider a specialized version of the FSGQL approach, namely FSGQL(I) approach, where the GQL estimating equation is developed by using the independence assumption ($\rho = 0$)-based covariance matrix. More specifically, under this approach, the covariance matrix $\Sigma_i^*(\beta, \rho = 0, \alpha)$ has the form

$$\Sigma_i^*(\beta, \rho = 0, \alpha) \equiv \begin{cases} \sigma_{i,11}^* = \mu_{i1}[1-\mu_{i1}] \\ \sigma_{i,tt}^* = E_{H_i(y)}[w_{it}^{-1}\{\mu_{it}(1-\mu_{it})\}], (\text{for } t = 2,\ldots,T_i) \\ \sigma*_{i,ut} = 0, \ (\text{for } u \neq t), \end{cases} \quad (26)$$

where $E_{H_i(y)}[w_{it}^{-1}]$ is computed by (24).

Now by replacing the "working" covariance matrix $V_i(\alpha^*)$ in the WGEE given in (18) with $\Sigma_i^*(\beta, \rho = 0, \alpha)$, one may obtain the FSGQL(I) estimate for $\beta$ by solving the estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i^{'}}{\partial beta} [\Sigma_i^*(\beta, \rho = 0, \alpha)]^{-1} \Delta_i(y_i - \mu_i) = 0. \quad (27)$$

## 2.3 CWGQL Approach

Note that instead of using the distance function with unconditional zero mean, one may like to exploit the distance function with zero mean conditionally. This is possible only when the expectation of the binary response conditional on the past history is known. In this case, by replacing $\mu_{it}$ with $\lambda_{it}$ in (17), one may construct the distance function which has mean zero conditional on the past history, that is,

$$E\left[\frac{R_{it}(Y_{it}-\lambda_{it}(\beta,\rho))}{w_{it}} \mid H_{i,t-1}\right]=0, \tag{28}$$

where for binary AR(1) model, for example, the conditional mean has the form

$$\lambda_{it}(\beta,\rho)=\mu_{it}+\rho(y_{i,t-1}-\mu_{i,t-1}), \tag{29}$$

for $t=2,\ldots,T$.

Suppose that

$$\lambda_i(\beta,\rho)=[\lambda_{i1}(\beta),\lambda_{i2}(H_{i,1}(y),\beta,\rho),\ldots,\lambda_{iT_i}(H_{i,T_i-1}(y),\beta,\rho)]'$$

with $\lambda_{i1}(\beta)=\mu_{i1}(\beta)$. To develop a GQL-type estimating equation in the conditional approach, one minimizes the distance function

$$\sum_{i=1}^{K}[\{\Delta_i(y_i-\lambda_i(\beta,\rho))\}'\{\mathrm{cov}(\Delta_i(y_i-\lambda_i(\beta,\rho)))|H_i(y)\}^{-1}\{\Delta_i(y_i-\lambda_i(\beta,\rho))\}'] \tag{30}$$

with respect to $\beta$, the parameter of interest. Given the history, let the conditional covariance matrix $\{\mathrm{cov}(\Delta_i(y_i-\lambda_i(\beta,\rho)))|H_i(y)\}$ be denoted by $\Sigma_{ich}(\beta,\rho)$. Then assuming that $\beta$ and $\rho$ in $\Sigma_{ich}(\beta,\rho)$ are known, minimizing the quadratic distance function (30) with respect to $\beta$ is equivalent to solving the equation

$$\begin{aligned}&\sum_{i=1}^{K}\frac{\partial[E\{\Delta_i\lambda_i(\beta,\rho)|H_i(y)\}']}{\partial\beta}\Sigma_{ich}^{-1}(\beta,\rho)\{\Delta_i(y_i-\lambda_i(\beta,\rho))\}\\ &=\sum_{i=1}^{K}\frac{\partial\lambda_i'(\beta,\rho)}{\partial\beta}\Sigma_{ich}^{-1}(\beta,\rho)\{\Delta_i(y_i-\lambda_i(\beta,\rho))\}=0.\end{aligned} \tag{31}$$

*Computational formula for* $\Sigma_{ich}(\beta,\rho,\gamma)$

For convenience, we first write

$$\begin{aligned}&\Delta_i=W_i^{-1}R_i, \text{ with}\\ &W_i=\mathrm{diag}[w_{i1},w_{i2},\ldots,w_{iT_i}], \text{ and } R_i=\mathrm{diag}[R_{i1},\ldots,R_{iT_i}].\end{aligned}$$

It then follows that

$$\begin{aligned}\Sigma_{ich}(\beta,\rho) &= \text{cov}[\{\Delta_i(y_i-\lambda_i(\beta,\rho))\}|H_i(y)] \\ &= W_i^{-1}\text{cov}[\{R_i(y_i-\lambda_i(\beta,\rho))\}|H_i(y)]W_i^{-1}. \end{aligned} \tag{32}$$

Now to compute the covariance matrix in (32), we write

$$R_i(y_i-\lambda_i(\beta,\rho)) = [R_{i1}(y_{i1}-\lambda_{i1}),\ldots,R_{it}(y_{it}-\lambda_{it}),\ldots,R_{iT_i}(y_{iT_i}-\lambda_{iT_i})]'.$$

It then follows that for $u < t$, for example,

$$\text{cov}[\{R_{iu}(y_{iu}-\lambda_{iu}),R_{it}(y_{it}-\lambda_{it})\}|H_{i,t-1}(y)] = 0, \tag{33}$$

and for $t = 1,\ldots,T_i$,

$$\begin{aligned}&\text{var}[R_{it}(y_{it}-\lambda_{it})|H_{i,t-1}(y)] = \text{var}[R_{it}|H_{i,t-1}(y)]\text{var}[y_{it}|H_{i,t-1}(y)] \\ &+ E^2[R_{it}|H_{i,t-1}(y)]\text{var}[y_{it}|H_{i,t-1}(y)] + \text{var}[R_{it}|H_{i,t-1}(y)]E^2[(y_{it}-\lambda_{it})|H_{i,t-1}(y)] \\ &= w_{it}(1-w_{it})\sigma_{ic,tt} + w_{it}^2\sigma_{ic,tt} \\ &= w_{it}\sigma_{ic,tt}, \end{aligned} \tag{34}$$

where $\sigma_{ic,tt}$ is the conditional variance of $y_{it}$ given the history. For example, in the binary case, $\sigma_{ic,tt} = \lambda_{it}(1-\lambda_{it})$.

### 2.3.1 CWGQL Estimating Equation

Now by substituting (34) and (33) into (32), one obtains

$$\Sigma_{ich}(\beta,\rho) = W_i^{-1}W_i\Sigma_{ic}W_i^{-1} = W_i^{-1}\Sigma_{ic}, \tag{35}$$

where $\Sigma_{ic} = \text{diag}[\sigma_{ic,11},\ldots,\sigma_{ic,T_iT_i}]$. Consequently, when this formula for $\Sigma_{ich}$ from (35) is applied to the conditional GQL (CGQL) estimating equation in (31), one obtains

$$\sum_{i=1}^{K}\frac{\partial\lambda_i'(\beta,\rho)}{\partial\beta}\Sigma_{ic}^{-1}(\beta,\rho)W_i\{\Delta_i(y_i-\lambda_i(\beta,\rho))\} = 0, \tag{36}$$

which is unaffected by the missing MAR mechanism. This is not surprising, as conditional on the history, $R_{it}$ and $y_{it}$ are independent. However, this fully conditional approach requires the modeling of the conditional means of the responses, which is equivalent to modeling the correlation structure.

### 2.3.2 Conditional Likelihood Estimation

In fact when conditional inference is used, one can obtain likelihood estimates for $\beta$ and $\rho$ by maximizing the exact likelihood function under the condition that $Y_{it}$ and $R_{it}$ are independent given the history. This is easier for the analysis of longitudinal binary data as compared to the longitudinal analysis for count data subject to MAR.

Since the $R_{it}$'s satisfy the monotonic restriction given in Sect. 1, and because $R_{it}$ and $Y_{it}$ are independent conditional on the history under the MAR mechanism, the likelihood function for the $i$th individual may be expressed as

$$\begin{aligned} L_i(\beta,\rho,\alpha) &= f_{i1}(y_{i1}) f_{i2|1}\{(y_{i2}, r_{i2}=1)|r_{i1}=1, y_{i1}\}\ldots \\ &\quad \times f_{iT_i|T_{i-1}}\{(y_{iT_i}, r_{iT_i}=1)|r_{i1}=1, r_{i2}=1,\ldots,r_{i(T_i-1)}=1, H_{i,t-1}(y)\} \\ &= \mu_{i1}^{y_{i1}}[1-\mu_{i1}]^{1-y_{i1}} \Pi_{t=1}^{T_i}[\{g_{it}\}\{\lambda_{it}^{y_{it}}(1-\lambda_{it})^{1-y_{it}}\}], \end{aligned} \tag{37}$$

where, by (9),

$$g_{it}(\alpha) = P[(R_{it}=1)|r_{i1}=1,\ldots,r_{i,t-1}=1, H_{i,t-1}(y)] = \frac{\exp(1+\alpha y_{i,t-1})}{1+\exp(1+\alpha y_{i,t-1})}.$$

## 3 Simulation Study

### *3.1 Comparison Between WGEE (AR(1)), WGEE(I) and FSGQL(I) Approaches: Multinomial Distribution Based Joint Generation of R and y*

In this section, we describe and report the results of a simulation study that centers on a comparison of the WGEE approach of Robins et al. (1995) for estimating the regression parameter vector with the proposed FSGQL approach. Recall that the WGEE in (18) was constructed by using a "working" covariance matrix $V_i(\alpha^*) = A_i^{\frac{1}{2}} R_i^*(\alpha^*) A_i^{\frac{1}{2}}$, of the response vector $y_i$. Note that this weight matrix was chosen ignoring the missing mechanism. Furthermore, there is no guideline to choose the "working" correlation matrix $R_i^*(\alpha^*)$. In the simulation study, we will consider a non-stationary longitudinal binary AR(1) model with true correlation structure $C_i(\rho)$ given by (4), for the responses subject to MAR. To examine the performance of the WGEE approach (18), we choose the best possible stationary AR(1) correlation form, namely,

$$R_i^*(\alpha^*) = (r_{ut}^*(\alpha^*)) = (\alpha^{*|t-u|}),$$

as compared to using MA(1) and EQC-based "working" correlation matrices. We will refer to this WGEE as the WGEE(AR(1)). Also we will consider the simplest version of the WGEE approach, namely WGEE(I), which is obtained based on the independence assumption by using $\alpha^* = 0$ in the "working" correlation matrix $R_i^*(\alpha^*)$. These two versions of the WGEE approach will be compared with the FSGQL(I) approach in (27) which was constructed by accommodating missing mechanism but by using longitudinal independence assumption, i.e., $\rho = 0$ or $C_i(\rho) = I_{T_i}$. For simplicity, in the present simulation study, we do not consider the true complete covariance matrix $\Sigma_i^*(\beta, \rho, \alpha)$-based FSGQL approach in (19).

### 3.1.1 Joint Generation of (R and y) Incomplete Binary Data: Multinomial Distribution Based

In order to generate an incomplete longitudinal binary data set subject to MAR, we follow the approach of Sutradhar and Mallick (2010). Specifically, the procedure initially assumes that every individual provides a response at time $t = 1$. Thus, since $R_{i1} = 1$ for all $i = 1, \cdots, K$, a binary response $y_{i1}$ is generated with marginal probability $\mu_{i1}$. Subsequently, $y_{it}$ is only observed for the $i$-th individual $(i = 1, \cdots, K)$ at time $t$ $(t = 2, \cdots, T)$ when $R_{it} = 1$ conditional on having observed the previous $t-1$ responses for that individual; in other words, conditional on $R_{i1} = 1, \cdots, R_{i,t-1} = 1$. Therefore, at time $t$ $(t = 2, \cdots, T)$, both $Y_{it}$ and $R_{it}$ are random variables conditional on the observed history up to time $t-1$, and, as such, one of the following three events occurs:

$$E_1 : [R_{it} = 1, Y_{it} = 1 \mid R_{i1} = \cdots = R_{i,t-1} = 1, H_{i,t-1}(y)],$$
$$E_2 : [R_{it} = 1, Y_{it} = 0 \mid R_{i1} = \cdots = R_{i,t-1} = 1, H_{i,t-1}(y)],$$
$$\text{or } E_3 : [R_{it} = 0 \mid R_{i1} = \cdots = R_{i,t-1} = 1, H_{i,t-1}(y)], \text{ which implies that } y_{it}$$

is not observed.

Let $z_{its} = 1$ for any $s = 1, 2, 3$ indicate that $E_s$ has occurred. Then, for $l \neq s$, $z_{itl} = 0$, and it must be the case that $\sum_{s=1}^{3} z_{its} = 1$. Let $p_{its} = P(z_{its} = 1)$ for $s = 1, 2, 3$. If we set $q = 1$ in (9), and use the resulting equation in conjunction with model (1), the $p_{its}$ may be expressed as

$$p_{it1} = P(z_{it1} = 1) = P[R_{it} = 1, Y_{it} = 1 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)],$$
$$p_{it2} = P(z_{it2} = 1) = P[R_{it} = 1, Y_{it} = 0 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)],$$

and

$$p_{it3} = P(z_{it3} = 1) = P[R_{it} = 0 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)],$$

which can be written as

$$p_{it1} = P[R_{it} = 1 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)]\, P[Y_{it} = 1 \mid H_{i,t-1}(y)]$$
$$= g_{it}(y_{i,t-1}; \alpha)\lambda_{it}, \tag{38}$$
$$p_{it2} = P[R_{it} = 1 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)]\, P[Y_{it} = 0 \mid H_{i,t-1}(y)]$$
$$= g_{it}(y_{i,t-1}; \alpha)(1 - \lambda_{it}), \tag{39}$$

and

$$p_{it3} = P[R_{it} = 0 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)] = 1 - g_{it}(y_{i,t-1}; \alpha), \tag{40}$$

where

$$g_{it}(y_{i,t-1}; \alpha) = exp(1 + \alpha y_{i,t-1})/[1 + exp(1 + \alpha y_{i,t-1})]. \tag{41}$$

Thus, Sutradhar and Mallick (2010) summarize the data generation routine for the $i$-th individual, $i = 1, \cdots, K$, as follows:

1. Generate $y_{i1}$ from a Bernoulli distribution with parameter $\mu_{i1}$.
2. For any $t > 1$, the values of $z_{its}$ for $s = 1, 2, 3$ are realized according to the multinomial probability distribution

$$P(z_{it1}, z_{it2}, z_{it3}) = \frac{1!}{z_{it1}!z_{it2}!z_{it3}!} p_{it1}^{z_{it1}} p_{it2}^{z_{it2}} p_{it3}^{z_{it3}}$$

   with $\sum_{s=1}^{3} z_{its} = 1$. For $z_{its} = 1$, allocate the response $y_{it}$ following $E_s$.
3. If $z_{its} = 1$, stop generating $y_{it}$ for this individual; otherwise repeat steps (1) and (2) for $t \leq T$.

### 3.1.2 Comparison Under Various Designs

Regarding the simulation study, for each of four designs, we set $K = 100$ and $T = 4$ and performed 1,000 replications. We considered three different values of longitudinal correlation parameter, setting $\rho = 0.2$, 0.5, and 0.8 in turn. In order to investigate the effect of the degree of missingness on the estimates of the regression parameter vector, for $\Delta_i = diag(\delta_{i1}, \delta_{i2}, \cdots, \delta_{iT})$ in (18) with $\delta_{it} = R_{it}/w_{it}\{H_{i,t-1}(y); \alpha\} = R_{it}/w_{it}$, we set $w_{it} = \prod_{j=1}^{t} g_{ij}(y_{i,j-1}; \alpha)$ according to (10) with $q = 1$. We then studied two levels for $\alpha$, namely $\alpha = 1$, and $\alpha = -3$ (We assume throughout the simulation study that both $\rho$ and $\alpha$ are known; hence, we do not concern ourselves with estimating these quantities). Note that, according to (41), $P[R_{it} = 0 \mid y_{i,t-1} = 1] = 0.12$ and $P[R_{it} = 0 \mid y_{i,t-1} = 0] = 0.27$ for $\alpha = 1$, while $P[R_{it} = 0 \mid y_{i,t-1} = 1] = 0.88$ and $P[R_{it} = 0 \mid y_{i,t-1} = 0] = 0.27$ for $\alpha = -3$. Thus, when $\alpha = -3$, the rate of missingness is extremely high, as expected.

**Table 1** (Based on data using joint generation approach) Simulated means (SM) and standard errors (SSE) based on 1,000 simulations for $\beta = 0.5$ and selected values of $\rho$ and $\alpha$, selected design

| | | | | WGEE(AR(1)) | | WGEE(I) | | FSGQL(I) | |
|---|---|---|---|---|---|---|---|---|---|
| Design | $\alpha$ | $\rho$ | Statistic | $\hat{\beta}$ | # | $\hat{\beta}$ | # | $\hat{\beta}$ | # |
| No covariate | $-3$ | 0.2 | SM | 0.574 | 231 | 0.471 | 174 | 0.501 | 1,000 |
| | | | SSE | 0.284 | | 0.293 | | 0.196 | |
| | | 0.5 | SM | 0.551 | 361 | 0.250 | 233 | 0.502 | 1,000 |
| | | | SSE | 0.295 | | 0.311 | | 0.204 | |
| | | 0.8 | SM | 0.556 | 500 | $-0.052$ | 345 | 0.503 | 1,000 |
| | | | SSE | 0.270 | | 0.286 | | 0.219 | |
| | 1 | 0.2 | SM | 0.478 | 1,000 | 0.502 | 1,000 | 0.501 | 1,000 |
| | | | SSE | 0.144 | | 0.146 | | 0.141 | |
| | | 0.5 | SM | 0.434 | 1,000 | 0.503 | 1,000 | 0.504 | 1,000 |
| | | | SSE | 0.164 | | 0.172 | | 0.169 | |
| | | 0.8 | SM | 0.372 | 1,000 | 0.499 | 1,000 | 0.500 | 1,000 |
| | | | SSE | 0.190 | | 0.206 | | 0.200 | |
| $p = 1$ stationary | $-3$ | 0.2 | SM | 0.503 | 440 | 0.514 | 383 | 0.499 | 1,000 |
| | | | SSE | 0.348 | | 0.358 | | 0.292 | |
| | | 0.5 | SM | 0.504 | 593 | 0.508 | 464 | 0.521 | 999 |
| | | | SSE | 0.351 | | 0.380 | | 0.294 | |
| | | 0.8 | SM | 0.509 | 663 | 0.518 | 559 | 0.529 | 999 |
| | | | SSE | 0.361 | | 0.408 | | 0.325 | |
| | 1 | 0.2 | SM | 0.500 | 1,000 | 0.501 | 1,000 | 0.500 | 1,000 |
| | | | SSE | 0.190 | | 0.192 | | 0.185 | |
| | | 0.5 | SM | 0.496 | 1,000 | 0.499 | 1,000 | 0.498 | 1,000 |
| | | | SSE | 0.234 | | 0.241 | | 0.233 | |
| | | 0.8 | SM | 0.516 | 1,000 | 0.514 | 1,000 | 0.515 | 1,000 |
| | | | SSE | 0.271 | | 0.291 | | 0.282 | |

Note that # refers to the number of simulations where convergence is achieved

Initially, we compared the WGEE(I) and FSGQL(I) approaches using a stationary design that essentially contained no covariates. For this design, we simply had a single $\beta_1 = 0.5$, while the associated $x_{it1} = 1$ for all $i = 1, \cdots, 100$ and $t = 1, \cdots, 4$. Table 1 presents the means and standard errors of the WGEE and FSGQL(I) estimates over the 1,000 replications for each of the six combinations of $\rho$ and $\alpha$. The number of replications that converged is also reported. When the degree of missingness is not overly severe $(\alpha = 1)$, there is little difference in the WGEE(I) and FSGQL(I) estimates. Both approaches produce essentially unbiased estimates, and all replications converge. However, when the degree of missingness is more pronounced $(\alpha = -3)$, the WGEE(I) estimates are significantly biased. In addition, regardless of the value of $\rho$, more than half of the replications did not converge. On the other hand, the FSGQL(I) estimates are still unbiased, and all replications continue to converge. We investigated the WGEE approach further by considering an AR(1) type "working" correlation structure instead of an independence assumption-based "working" correlation matrix. This WGEE

**Table 2** (Based on data using joint generation approach) Simulated means (SM) and standard errors (SSE) based on 1,000 simulations for $\beta_1 = \beta_2 = 0.5$ and selected values of $\rho$ and $\alpha$, selected design (D) with $p = 2$ stationary (S) and non-stationary (NS) covariates

| | | | | WGEE(AR(1)) | | | WGEE(I) | | | FSGQL(I) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D | $\alpha$ | $\rho$ | Statistic | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # |
| S | −3 | 0.2 | SM | 0.598 | 0.598 | 122 | 0.506 | 0.593 | 85 | 0.512 | 0.520 | 999 |
| | | | SSE | 0.284 | 0.389 | | 0.301 | 0.380 | | 0.200 | 0.287 | |
| | | 0.5 | SM | 0.550 | 0.510 | 277 | 0.243 | 0.485 | 146 | 0.510 | 0.518 | 1,000 |
| | | | SSE | 0.276 | 0.376 | | 0.290 | 0.359 | | 0.213 | 0.299 | |
| | | 0.8 | SM | 0.544 | 0.527 | 403 | −0.046 | 0.515 | 239 | 0.504 | 0.516 | 999 |
| | | | SSE | 0.273 | 0.342 | | 0.301 | 0.383 | | 0.219 | 0.317 | |
| | 1 | 0.2 | SM | 0.477 | 0.508 | 1,000 | 0.501 | 0.509 | 1,000 | 0.503 | 0.509 | 1,000 |
| | | | SSE | 0.139 | 0.202 | | 0.141 | 0.203 | | 0.137 | 0.199 | |
| | | 0.5 | SM | 0.431 | 0.500 | 1,000 | 0.500 | 0.500 | 1,000 | 0.499 | 0.499 | 1,000 |
| | | | SSE | 0.166 | 0.246 | | 0.171 | 0.258 | | 0.166 | 0.248 | |
| | | 0.8 | SM | 0.386 | 0.516 | 1,000 | 0.516 | 0.520 | 1,000 | 0.515 | 0.519 | 1,000 |
| | | | SSE | 0.205 | 0.297 | | 0.222 | 0.322 | | 0.216 | 0.311 | |
| NS | −3 | 0.2 | SM | 0.497 | 0.284 | 19 | 0.449 | 0.045 | 20 | 0.514 | 0.428 | 955 |
| | | | SSE | 0.509 | 1.234 | | 0.503 | 1.082 | | 0.453 | 1.406 | |
| | | 0.5 | SM | 0.635 | −0.693 | 31 | 0.174 | −0.789 | 36 | 0.507 | 0.520 | 962 |
| | | | SSE | 0.512 | 0.886 | | 0.495 | 1.151 | | 0.459 | 1.388 | |
| | | 0.8 | SM | 0.844 | −1.122 | 120 | 0.021 | −3.764 | 51 | 0.491 | 0.477 | 948 |
| | | | SSE | 0.545 | 0.665 | | 0.623 | 1.387 | | 0.468 | 1.447 | |
| | 1 | 0.2 | SM | 0.481 | 0.453 | 999 | 0.503 | 0.511 | 999 | 0.504 | 0.506 | 1,000 |
| | | | SSE | 0.390 | 0.598 | | 0.391 | 0.607 | | 0.384 | 0.600 | |
| | | 0.5 | SM | 0.439 | 0.324 | 1,000 | 0.509 | 0.495 | 1,000 | 0.509 | 0.492 | 1,000 |
| | | | SSE | 0.374 | 0.562 | | 0.406 | 0.640 | | 0.402 | 0.631 | |
| | | 0.8 | SM | 0.386 | 0.201 | 999 | 0.502 | 0.537 | 996 | 0.499 | 0.532 | 1,000 |
| | | | SSE | 0.290 | 0.442 | | 0.408 | 0.693 | | 0.417 | 0.686 | |

**Table 3** (Based on data using conditional approach) Simulated means (SM) and standard errors (SSE) based on 1,000 simulations for $\beta = 0.5$ and selected values of $\rho$ and $\alpha$, selected design

| | | | | WGEE(AR(1)) | | WGEE(I) | | FSGQL(I) | |
|---|---|---|---|---|---|---|---|---|---|
| Design | $\alpha$ | $\rho$ | Statistic | $\hat{\beta}$ | # | $\hat{\beta}$ | # | $\hat{\beta}$ | # |
| No | $-3$ | 0.2 | SM | – | 0 | 0.565 | 912 | 0.495 | 1,000 |
| covariate | | | SSE | – | | 0.238 | | 0.200 | |
| | | 0.5 | SM | – | 0 | 0.550 | 910 | 0.495 | 1,000 |
| | | | SSE | – | | 0.250 | | 0.206 | |
| | | 0.8 | SM | – | 0 | 0.523 | 907 | 0.495 | 1,000 |
| | | | SSE | – | | 0.261 | | 0.218 | |
| | 1 | 0.2 | SM | 0.402 | 6 | 0.471 | 1,000 | 0.495 | 1,000 |
| | | | SSE | 0.293 | | 0.151 | | 0.139 | |
| | | 0.5 | SM | 0.323 | 4 | 0.478 | 1,000 | 0.505 | 1,000 |
| | | | SSE | 0.374 | | 0.170 | | 0.166 | |
| | | 0.8 | SM | 0.391 | 4 | 0.466 | 1,000 | 0.493 | 1,000 |
| | | | SSE | 0.201 | | 0.199 | | 0.205 | |
| $p = 1$ | $-3$ | 0.2 | SM | 0.472 | 415 | 0.479 | 375 | 0.497 | 998 |
| stationary | | | SSE | 0.585 | | 0.598 | | 0.275 | |
| | | 0.5 | SM | 0.511 | 571 | 0.528 | 458 | 0.505 | 999 |
| | | | SSE | 0.621 | | 0.647 | | 0.287 | |
| | | 0.8 | SM | 0.489 | 696 | 0.503 | 574 | 0.515 | 1,000 |
| | | | SSE | 0.606 | | 0.643 | | 0.330 | |
| | 1 | 0.2 | SM | 0.498 | 1,000 | 0.499 | 1,000 | 0.499 | 1,000 |
| | | | SSE | 0.537 | | 0.538 | | 0.198 | |
| | | 0.5 | SM | 0.513 | 1,000 | 0.512 | 1,000 | 0.509 | 1,000 |
| | | | SSE | 0.568 | | 0.571 | | 0.245 | |
| | | 0.8 | SM | 0.509 | 1,000 | 0.514 | 1,000 | 0.514 | 1,000 |
| | | | SSE | 0.579 | | 0.595 | | 0.290 | |

Note that # refers to the number of simulations where convergence is achieved

approach is referred to as the WGEE(AR(1)) approach. Specifically, we set $V_i(\alpha^*) = A_i^{1/2} R_i^*(\alpha^*) A_i^{1/2}$, where $R_i^*(\alpha^*)$ is a $T \times T$ correlation matrix with $corr(Y_{it}, Y_{i,t+l}) = \alpha^{*l}$ and $A_i = diag(\sigma_{i,11}, \cdots, \sigma_{i,tt}, \cdots, \sigma_{i,T_iT_i}, 0, \cdots, 0)$ with $\sigma_{i,tt} = \mu_{it}(1 - \mu_{it})$. To avoid estimation of $\alpha^*$ we have used $\alpha^* = \rho$. The results obtained for each of the six combinations of $\rho$ and $\alpha$ (the missing dependence parameter) are also presented in Table 1. For $\alpha = -3$, the WGEE(AR(1)) estimates based on an AR(1) type structure are significantly better than those based on independence, and the number of replications that converged is also notably higher. Nonetheless, the independent FSGQL(I) estimates are still noticeably better than either of the WGEE estimates. Also of note is the fact that the WGEE(AR(1)) estimates based on the AR(1) structure for $\alpha = 1$ are outperformed by their independent covariance structure counterparts.

We also considered a stationary design consisting of one covariate with associated parameter $\beta_1 = 0.5$. Specifically, for all $t = 1, \cdots, 4$, we set $x_{it1} = -1$ for $i = 1, \cdots, K/4$, $x_{it1} = 0$ for $i = (K/4) + 1, \cdots, 3K/4$, and $x_{it1} = 1$ for $i = (3K/4) +$

**Table 4** (Based on data using conditional approach) Simulated means (SM) and standard errors (SSE) based on 1,000 simulations for $\beta_1 = \beta_2 = 0.5$ and selected values of $\rho$ and $\alpha$, selected design (D) with $p = 2$ stationary (S) and non-stationary (NS) covariates

| | | | | WGEE(AR(1)) | | | WGEE(I) | | | FSGQL(I) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D | $\alpha$ | $\rho$ | Statistic | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # |
| S | −3 | 0.2 | SM | 0.550 | 0.483 | 126 | 0.453 | 0.512 | 101 | 0.505 | 0.511 | 1,000 |
| | | | SSE | 0.277 | 0.400 | | 0.269 | 0.395 | | 0.207 | 0.297 | |
| | | 0.5 | SM | 0.594 | 0.493 | 279 | 0.279 | 0.469 | 150 | 0.514 | 0.510 | 998 |
| | | | SSE | 0.292 | 0.400 | | 0.293 | 0.424 | | 0.208 | 0.309 | |
| | | 0.8 | SM | 0.561 | 0.575 | 385 | −0.039 | 0.557 | 231 | 0.514 | 0.551 | 1,000 |
| | | | SSE | 0.262 | 0.379 | | 0.307 | 0.415 | | 0.222 | 0.325 | |
| | 1 | 0.2 | SM | 0.478 | 0.501 | 1,000 | 0.503 | 0.500 | 1,000 | 0.502 | 0.502 | 1,000 |
| | | | SSE | 0.143 | 0.205 | | 0.144 | 0.207 | | 0.141 | 0.202 | |
| | | 0.5 | SM | 0.439 | 0.506 | 1,000 | 0.509 | 0.506 | 1,000 | 0.509 | 0.504 | 1,000 |
| | | | SSE | 0.172 | 0.240 | | 0.179 | 0.251 | | 0.174 | 0.242 | |
| | | 0.8 | SM | 0.386 | 0.522 | 1,000 | 0.520 | 0.526 | 1,000 | 0.517 | 0.526 | 1,000 |
| | | | SSE | 0.200 | 0.291 | | 0.218 | 0.317 | | 0.211 | 0.307 | |
| NS | −3 | 0.2 | SM | 0.678 | −0.095 | 27 | 0.504 | −0.185 | 25 | 0.497 | 0.538 | 952 |
| | | | SSE | 0.529 | 1.346 | | 0.572 | 1.268 | | 0.458 | 1.358 | |
| | | 0.5 | SM | 0.654 | −1.098 | 45 | 0.216 | −1.614 | 26 | 0.486 | 0.421 | 946 |
| | | | SSE | 0.521 | 0.914 | | 0.634 | 1.117 | | 0.469 | 1.411 | |
| | | 0.8 | SM | 0.841 | −1.261 | 134 | 0.086 | −3.630 | 59 | 0.470 | 0.389 | 949 |
| | | | SSE | 0.562 | 0.851 | | 0.565 | 1.374 | | 0.471 | 1.451 | |
| | 1 | 0.2 | SM | 0.488 | 0.436 | 1,000 | 0.507 | 0.502 | 1,000 | 0.506 | 0.498 | 1,000 |
| | | | SSE | 0.372 | 0.582 | | 0.376 | 0.595 | | 0.370 | 0.584 | |
| | | 0.5 | SM | 0.450 | 0.317 | 1,000 | 0.512 | 0.500 | 999 | 0.510 | 0.498 | 1,000 |
| | | | SSE | 0.354 | 0.557 | | 0.393 | 0.644 | | 0.397 | 0.634 | |
| | | 0.8 | SM | 0.372 | 0.200 | 1,000 | 0.506 | 0.539 | 998 | 0.505 | 0.536 | 1,000 |
| | | | SSE | 0.291 | 0.435 | | 0.413 | 0.720 | | 0.423 | 0.711 | |

$1, \cdots, K$. The simulation results associated with this design are presented in Table 1 for each combination of $\rho$ and $\alpha$. When $\alpha = 1$, the performance of WGEE(AR(1)), WGEE(I) and FSGQL(I) are very similar. It is also important to note that when $\alpha = -3$, despite the fact that the average estimates for the regression parameters are better for WGEE under relatively higher longitudinal correlations of $\rho = 0.5$ and 0.8, the estimated standard errors are significantly smaller for the proposed FSGQL(I) technique. In addition, WGEE experiences convergence problems on a significant number of simulation replications; when an independent covariance structure is assumed, convergence rates ranged between 40% and 60%, approximately, and were only slightly better when an AR(1) structure was specified.

Two designs consisting of two covariates with associated regression parameters $\beta_1 = \beta_2 = 0.5$ were also studied; one consisted of two stationary covariates, the other of nonstationary ones. For the design consisting of two stationary covariates, we set $x_{it1} = 1$ for all $i = 1, \cdots, 100$ and $t = 1, \cdots, 4$ as in the design with no covariate, and $x_{it2}$ according to the values specified for the single covariate design described above. The two covariates in the nonstationary design were set as follows:

$$x_{it1} = \begin{cases} \frac{1}{2}, & \text{for } i = 1, \cdots, \frac{K}{4}; t = 1, 2 \\ 0, & \text{for } i = 1, \cdots, \frac{K}{4}; t = 3, 4 \\ -\frac{1}{2}, & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4}; t = 1 \\ 0, & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4}; t = 2, 3 \\ \frac{1}{2}, & \text{for } i = \frac{K}{4} + 1, \cdots, \frac{3K}{4}; t = 4 \\ \frac{t}{2T}, & \text{for } i = \frac{3K}{4} + 1, \cdots, K; t = 1, \cdots, 4 \end{cases}$$

and

$$x_{it2} = \begin{cases} \frac{t-2.5}{2T}, & \text{for } i = 1, \cdots, \frac{K}{2}; t = 1, \cdots, 4 \\ 0, & \text{for } i = \frac{K}{2} + 1, \cdots, K; t = 1, 2 \\ \frac{1}{2}, & \text{for } i = \frac{K}{2} + 1, \cdots, K; t = 3, 4 \end{cases}$$

The results for both the stationary and non-stationary two-covariate designs are presented in Table 2. For both designs, when $\alpha = 1$, the performance of WGEE(I) under an independent covariance structure and FSGQL(I) is very similar. The estimates obtained using WGEE(AR(1)) with an AR(1) structure appear to be biased. When $\alpha = -3$, and there is a significantly higher degree of missingness, the estimates obtained under WGEE are biased regardless of the assumed covariance structure and the level of longitudinal correlation; this is particularly the case in the nonstationary design. Also of note is the fact that the convergence rates under the WGEE approach are very poor, with the majority under 5% for the nonstationary design.

### 3.2 Comparison of WGEE(AR(1)), WGEE(I), and FSGQL(I) Approaches: Generating $R$ and $y$ conditionally

Because $R_{it}$ and $y_{it}$ are independent conditional on the history $H_{i,t-1}(y)$, instead of generating them by using a multinomial distribution discussed in Sect. 3.1.1, one may generate them by using a conditional approach as follows:

1. Generate $y_{i1}$ from $bin(\mu_{i1})$ for all $i, i = 1, \ldots, K$.
2. For $i$-th individual, generate $R_{i2}$ from $bin(g_{i2})$, where $g_{it}$ is given by (9) for $t = 2, \ldots, T$.
3. If $R_{i2} = 0$, consider $R_{ij} = 0$ and stop generating $y_{ij}$ $(j = 2, \cdots, T)$.
4. If $R_{i2} = 1$, generate $y_{i2}$ from $bin(\lambda_{i2})$, where $\lambda_{it}$ is the mean of $Y_{it}$ conditional on $y_{i,t-1}$ for $t = 2, \ldots, T$, as given by (1).
5. Repeat from step 2 for $j = 3, \cdots, T$.

The estimates for the same designs are obtained as in Sect. 3.1.2, and the simulation results are reported in Tables 3 and 4. The results are similar to those of Tables 1 and 2, except that WGEE approaches appear to encounter more convergence problems especially when proportion of missing values is large.

### 3.3 Performance of CWGQL Approach: Multinomial Distribution-Based Joint Generation of R and y

As opposed to the marginal approach where the unconditional mean function $\mu_{it}(\beta)$ is modeled, in the longitudinal setup it is more appropriate to model the conditional regression (mean) function. When complete longitudinal binary data follow an AR(1)-type correlation model, as pointed out in (1), the conditional regression function may be modeled as

$$\lambda_{it}(\beta, \rho) = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}), \text{for } t = 2, \ldots, T.$$

Furthermore, as pointed out in (28), because in the incomplete longitudinal setup with MAR mechanism one finds

$$E\left[\frac{R_{it}(Y_{it} - \lambda_{it}(\beta, \rho))}{w_{it}} \mid H_{i,t-1}\right] = 0,$$

the regression parameter $\beta$ in $\mu_{it}(\beta)$ modeled through $\lambda_{it}(\beta, \rho)$ can be estimated by solving the CWGQL estimating (36). For the same design parameters used in Sect. 3.1.2 for Tables 1 and 2, by generating incomplete data using the multinomial distribution discussed in Sect. 3.1.1, we have obtained the CWGQL estimates for $\beta$ under different scenarios as for the results shown in Tables 1 and 2. The CWGQL estimates along with their standard errors are reported in Tables 5 and 6.

**Table 5** (Based on data using joint generation approach) Simulated means (SM) and standard errors (SSE) for *CWGQL* approach with $\beta = 0.5$ and selected values of $\rho$ and $\alpha$, based on 1,000 simulations

| | | | No covariate | | $p = 1$ (stationary) | |
|---|---|---|---|---|---|---|
| $\alpha$ | $\rho$ | Statistic | $\hat{\beta}$ | # | $\hat{\beta}$ | # |
| −3 | 0.2 | SM | 0.505 | 1,000 | 0.494 | 1,000 |
| | | SSE | 0.178 | | 0.247 | |
| | 0.5 | SM | 0.505 | 1,000 | 0.509 | 1,000 |
| | | SSE | 0.188 | | 0.256 | |
| | 0.8 | SM | 0.507 | 1,000 | 0.517 | 1,000 |
| | | SSE | 0.204 | | 0.291 | |
| 1 | 0.2 | SM | 0.500 | 1,000 | 0.499 | 1,000 |
| | | SSE | 0.138 | | 0.183 | |
| | 0.5 | SM | 0.501 | 1,000 | 0.493 | 1,000 |
| | | SSE | 0.158 | | 0.224 | |
| | 0.8 | SM | 0.497 | 1,000 | 0.514 | 1,000 |
| | | SSE | 0.183 | | 0.265 | |

Note that # refers to the number of simulations where convergence is achieved

**Table 6** (Based on data using joint generation approach) Simulated means (SM) and standard errors (SSE) for *CWGQL* approach with $\beta_1 = \beta_2 = 0.5$ and selected values of $\rho$ and $\alpha$, based on 1,000 simulations

| | | | Stationary covariates | | | Nonstationary covariates | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\rho$ | Statistic | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # | $\hat{\beta}_1$ | $\hat{\beta}_2$ | # |
| −3 | 0.2 | SM | 0.522 | 0.513 | 1,000 | 0.517 | 0.515 | 789 |
| | | SSE | 0.185 | 0.260 | | 0.429 | 0.992 | |
| | 0.5 | SM | 0.516 | 0.510 | 1,000 | 0.507 | 0.645 | 770 |
| | | SSE | 0.201 | 0.278 | | 0.410 | 0.903 | |
| | 0.8 | SM | 0.506 | 0.508 | 1,000 | 0.497 | 0.730 | 597 |
| | | SSE | 0.203 | 0.298 | | 0.370 | 0.837 | |
| 1 | 0.2 | SM | 0.501 | 0.507 | 1,000 | 0.503 | 0.503 | 1,000 |
| | | SSE | 0.133 | 0.193 | | 0.372 | 0.581 | |
| | 0.5 | SM | 0.498 | 0.500 | 1,000 | 0.502 | 0.471 | 901 |
| | | SSE | 0.160 | 0.232 | | 0.350 | 0.515 | |
| | 0.8 | SM | 0.506 | 0.509 | 1,000 | 0.449 | 0.423 | 302 |
| | | SSE | 0.195 | 0.281 | | 0.254 | 0.455 | |

Note that # refers to the number of simulations where convergence is achieved

In order to examine the relative performance of the CWGQL approach with those of WGEE(AR(1)), WGEE(I), and FSGQL(I), it is sufficient to compare the CWGQL approach with the FSGQL(I) approach only. This is because it was found from the results in Tables 1 and 2 that the WGEE approaches may encounter serious convergence problems (showing consistency breakdown) and also may produce highly biased estimates, where the FSGQL(I) approach, in general,

does not encounter such convergence problems and produces almost unbiased estimates even if a large proportion of values are missing. Now as compared to the FSGQL(I) approach, the CWGQL approach appears to produce slightly more efficient estimates than the FSGQL(I) approach. For example, when $\alpha = 1$ (moderate missing) and $\rho = 0.5$, in the no-covariate case, the FSGQL(I) approach (Table 1) produces an average estimate of $\beta = 0.5$ as 0.504 with standard error 0.169, whereas the CWGQL approach (Table 5) produces $\beta$ estimate as 0.501 with standard error 0.158. Similarly when $\rho = 0.8$ and $\alpha = -3$ (high missing), in the stationary one-covariate case, FSGQL(I) produces an estimate with standard error 0.325 as compared to 0.291 for CWGQL. Similar results are found for the stationary two-covariate case. Also in these stationary cases, the CWGQL approach does not encounter any convergence problems even if the proportion of missing is high. In the non-stationary cases however, the CWGQL approach encounters some convergence problems when the proportion of missing is high, but the problem is less serious than the WGEE and WGEE(I) approaches.

## 4 Conclusion and Discussion

It was found that the existing WGEE (Robins et al. 1995) and WGEE(I) approaches in general encounter convergence problems when the proportion of missing is high, and the WGEE approach may produce highly biased estimates even when the proportion of missing is moderate or low. These results agree with the recent study reported by Sutradhar and Mallick (2010). The WGEE(I) approach, however, produces almost unbiased estimates and consequently this approach produces consistent estimates when the proportion of missing is moderate or low. However, it can be inefficient. The proposed FSGQL(I) approach does not appear to encounter any serious convergence problems even when the proportion of missing is high and the covariates are non-stationary. Also, it produces unbiased estimates similar to the WGEE(I) approach but with smaller standard errors, showing that FSGQL(I) is more efficient as expected than the WGEE(I) approach. Thus even with high proportion of missing, one may reliably use the proposed FSGQL(I) approach for regression estimation whether the covariates are stationary or time dependent. The general FSGQL approach is supposed to increase the efficiency as compared to the FSGQL(I) approach when correlations are large, but this will be studied in the future.

We have also reported some results on the performance of a conditional estimating equation, namely CWGQL estimating equation approach. This approach was found to produce regression estimates with more efficiency than the FSGQL(I) approach. However as compared to the FSGQL(I) approach it encounters convergence problems when covariates are time dependent and the proportion of missing is high. However, it experiences less convergence problems than the WGEE(I) approach.

**Acknowledgment** The authors would like to thank the audience of the symposium for their comments and suggestions. The authors would also like to thank a referee for valuable comments on the original submission. The research program of Patrick J. Farrell is supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

Birmingham, J., Rotnitzky, A., Fitzmaurice, G.M.: Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. J. R. Stat. Soc. B **65**, 275–297 (2003)

Fitzmaurice, G.M., Laird, N.M., Zahner, G.E.P.: Multivariate logistic models for incomplete binary responses. J. Am. Stat. Assoc. **91**, 99–108 (1996)

Liang, K.-Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)

Paik, M.C.: The generalized estimating equation approach when data are not missing completely at random. J. Am. Stat. Assoc. **92**, 1320–1329 (1997)

Qaqish, B.F.: A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika **90**, 455–463 (2003)

Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Am. Stat. Assoc. **90**, 106–121 (1995)

Rotnitzky, A., Robins, M., Scharfstein, D.O.: Semi-parametric regression for repeated outcomes with nonignorable nonresponse. J. Am. Stat. Assoc. **93**, 1321–1339 (1998)

Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–592 (1976)

Sutradhar, B.C.: An overview on regression models for discrete longitudinal responses. Stat. Sci. **18**, 377–93 (2003)

Sutradhar, B.C.: Dynamic Mixed Models for Familial Longitudinal Data. Springer, New York (2011)

Sutradhar, B.C.: Inference progress in missing data analysis from independent to longitudinal setup. In: Sutradhar, B.C. (ed.) ISS-2012 Proceedings volume on longitudinal data analysis subject to measurement errors, missing values, and/or outliers, Springer Lecture Notes Series, pp. 101–123 (2013)

Sutradhar, B.C., Mallick, T.S.: Modified weights based generalized quasilikelihood inferences in incomplete longitudinal binary models. Can. J. Stat. **38**, 217–231 (2010)

Zeger, S.L, Liang, K.-Y., Self, S.G.: The analysis of binary longitudinal data with time-independent covariates. Biometrika **72**, 31–38 (1985)

# Innovative Applications of Shared Random Parameter Models for Analyzing Longitudinal Data Subject to Dropout

**Paul S. Albert, Rajeshwari Sundaram, and Alexander C. McLain**

**Abstract** Shared random parameter (SRP) models provide a framework for analyzing longitudinal data with missingness. We discuss the basic framework and review the most relevant literature for the case of a single outcome followed longitudinally. We discuss estimation approaches, including an approximate approach which is relatively simple to implement. We then discuss three applications of this framework in novel settings. First, we show how SRP models can be used to make inference about pooled or batched longitudinal data subject to non-ignorable dropout. Second, we show how one of the estimation approaches can be extended for estimating high dimensional longitudinal data subject to dropout. Third, we show how to use jointly model complex menstrual cycle length data and time to pregnancy in order to study the evolution of menstrual cycle length accounting for non-ignorable dropout due to becoming pregnant and to develop a predictor of time-to-pregnancy from repeated menstrual cycle length measurements. These three examples demonstrate the richness of this class of models in applications.

## 1 Introduction

Modeling longitudinal data subject to missingness has been an active area of research in the last few decades. The missing-data mechanism is said to be missing completely at random if the probability of missing is independent of both the observed and unobserved data. Further, the mechanism is not missing

P.S. Albert (✉) • R. Sundaram
*Eunice Kennedy Shriver*, National Institute of Child Health and Human Development, Bethesda, MD, USA
e-mail: albertp@mail.nih.gov; Sundaramr2@mail.nih.gov

A.C. McLain
University of South Carolina, Columbia, SC, USA
e-mail: mclaina@mailbox.sc.edu

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_7, 

at random (NMAR) if the probability of missingness depends on the unobserved data (Rubin 1976; Little and Rubin 1987). It is well known that naive methods that do not account for NMAR can lead to biased estimation. The use of shared (random) parameter models has been been one approach that accounts for non-random missing data. In this formulation, a model for the longitudinal response measurements is linked with a model for the missing-data mechanism through a set of random effects that are shared between the two processes. Wu and Carroll (1988) proposed a model whereby the response process, which was modeled with a linear mixed model with a random intercept and slope was linked with the censoring process by including an individual's random slope as a covariate in a probit model for the censoring process. When the probit regression coefficient for the random slope is not zero, there is a dependence between the response and missing-data processes. Failure to account for this dependence can lead to biased estimation of important model parameters. Shared-parameter models (Follmann and Wu 1995) induce a type of non-randomly missing-data mechanism that has been called "informative missingness" (Wu and Carroll 1988). For a review and comparison with other methods, see Little (1995), Hogan and Laird (1997), and Vonesh et al. (2006). More recently Molenberghs et al. (2012) have discussed a fundamental non-identifiability of shared random effects models. Specifically, these models make non-verifiable assumptions about data not seen and there are multiple models in a wide class that can equally explain the observed data. Thus, shared random parameter (SRP) models make implicit assumptions that need to be justified from on a scientific basis and cannot be completely verified empirically.

This article discusses some applications of SRPs to some interesting novel applications. In Sect. 2, we set up the general model formulation and show how this mechanism induces a special type of nonignorable missingness. We also discuss both a full maximum-likelihood approach and conditional approach for parameter estimation that is easier to implement. We discuss some examples where a single longitudinal measurement is subject to non-ignorable dropout. Section 3 shows an example of batched laboratory data and how a SRP model can be used to account for the apparent non-ignorable missingness. In Sect. 4 we provide an example of the joint modeling of multiple or high dimensional longitudinal biomarker and time-to-event data. Section 5 shows an example of jointly modeling complex menstrual cycle data and time-to-pregnancy using a SRP approach. Lastly, we present a discussion in Sect. 6.

## 2 Model Formulation and Estimation

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ be a vector of longitudinal outcomes for the $i$th subject ($i = 1, 2, \ldots, I$) observed on $J$ occasions $t_1, t_2, \ldots, t_J$, and let $R_i = (R_{i1}, R_{i2}, \ldots, R_{iJ})'$ be a vector of random variables reflecting the missing data status (e.g., $R_{ij} = 0$ denoting a missed visit). Further, let $b_i = (b_{i1}, b_{i2}, \ldots, b_{iL})'$ be an L-element vector of random effects for the $i$th subject which can be shared between the response and

missing data mechanism. We assume that $b_i$ is multivariate normal with mean vector 0 and covariance matrix $\Sigma_b$. Covariates $X_{ij}$ are also measured which can influence both $Y_{ij}$ and $R_{ij}$.

The joint distribution of $Y_i, R_i, b_i$ can be written as

$$f(y_i, r_i, b_i) = g(y_i|b_i, r_i)m(r_i|b_i)h(b_i).$$

We make the assumption that conditional on the random effects, the responses do not depend on the missing data status, thus $g(y_i|b_i, r_i) = g(y_i|b_i)$. Furthermore, the elements of $Y_i$ are conditionally independent given $b_i$. By conditional independence, the density for the response vector $Y_i$ conditional on $b_i$, $g(y_i|b_i)$ can be decomposed into the product of the densities for the observed and unobserved values of $Y_i$. Namely, $g(y_i|b_i) = g(y_i^o|b_i)g(y_i^m|b_i)$, where $y_i^o$ and $y_i^m$ are vectors of observed and missing data responses, respectively, for the $i$th subject. The density of the observed random variables can be expressed as

$$\begin{aligned} f(y_i^o, r_i) &= \int_{y^m} \int_b f(y_i^o, y_i^m, r_i) dbdy_i^m \\ &= \int_{y^m} \int_b g(y_i^o|b_i)g(y_i^m|b_i)m(r_i|b_i)h(b_i)dbdy_i^m \\ &= \int_b g(y_i^o|b_i)m(r_i|b_i)h(b_i)\Big\{\int_{y_i^m} g(y_i^m|b)dy_i^m\Big\}db_i \\ &= \int_b g(y_i^o|b_i)m(r_i|b_i)h(b_i)db_i. \end{aligned} \tag{1}$$

Although the conditional independence of $Y_i|b_i$ is easy to verify when there is no missing data, it is difficult to verify for SRP models. Serial correlation could be incorporated (conditional on the random effects) using autoregressive or lagged responses (see, Zeger and Qaqish 1988; Albert 2000; Sutradhar and Mallick 2010). These lag-response modeling components can be formulated with the addition of a shared random effect that links the response and missing data mechanism together. Alternatively, Albert et al. (2002) link together the response and missing data mechanism with a shared latent process where the subject-specific random effect $b$ is replaced by a random process $b_i = (b_{i1}, b_{i2}, \ldots, b_{iJ})'$. They consider a random process that follows a continuous-time exponential correlation structure since observations are not equally spaced. Although the shared latent processes model is an attractive approach, it requires computationally intensive techniques such as Monte-Carlo EM for parameter estimation. In the remainder of this article, we focus on the SRP rather than the shared latent process model.

Tsiatis and Davidian (2004) provide a concise discussion of how the joint density is obtained for the case where missingness is monotone (i.e., patients only drop out of the study) and measured in continuous time.

The choice of a density function $g$ depends on the type of longitudinal response data being analyzed. For Gaussian longitudinal data, $g$ can be specified as a Gaussian distribution, and the model formulation can be specified as a linear mixed model

(Laird and Ware 1982). A simple linear mixed model which can be used as an illustration is

$$Y_{ij}|X_i, b_i = \beta_0 + \beta_1 X_i + b_i + \varepsilon_{ij}, \tag{2}$$

where $X_i$ is a subject-specific covariate such as treatment group, $b_i$ is a random effect which is often assumed normally distributed, and $\varepsilon_{ij}$ is an error term which is assumed normally distributed. Alternatively, for discrete or dichotomous longitudinal responses, $g$ can be formulated as a generalized linear mixed model (Follmann and Wu 1995; Ten Have et al. 1998; Albert and Follmann 2000).

The choice of the density for the missing data indicators, $m$, depends on the type of missing data being incorporated. When missing data is a discrete time to dropout, a monotone missing data mechanism, then a geometric distribution is often used for $m$ (Mori et al. 1994). For example, the probability of dropping out is

$$\Phi^{-1}\{P(R_{ij} = 0|R_{ij} = 1)\} = \alpha_0 + \alpha X_i + \theta b_i. \tag{3}$$

Various authors have proposed shared random effects models for the case in which dropout is a continuous event time (Schluchter 1992; Schluchter et al. 2001; DeGruttola and Tu 1994; Tsiatis et al. 1995; Wulfson and Tsiatis 1997; Tsiatis and Davidian 2001; Vonesh et al. 2006). When missing data includes only intermittently missed observations without dropout, then the product of Bernoulli densities across each of the potential observations may be a suitable density function for $g$. Alternatively, when multiple types of missing data such as both intermittent missingness and dropout need to be incorporated, a multinomial density function for $g$ can be incorporated (Albert et al. 2002).

The shared random effects model accounts for a MNAR data mechanism, which can be seen with the following argument. Suppressing the index $i$ for notational simplicity, suppose that the random effect $b$ is a scalar with $R_j$ indicating whether $Y_j$ is observed. MAR implies that the conditional density of $R_j$ given the complete data $Y$ does not depend on $Y^m$, while a MNAR implies that this conditional density depends on $Y^m$. The conditional density of $R_j$ given $Y = (Y^o, Y^m)$ is

$$f(r_j|y^m, y^o) = \frac{\int g(r_j|b)g(y^m, y^o|b)h(b)db}{\int g(y^m, y^o|b)h(b)db}$$
$$\int g(r_j|b)h(b|y^o, y^m)db.$$

A MNAR data mechanism follows since the conditional density depends on $y^m$ and since $h(b|y^o, y^m)$ depends on $y^m$. It is interesting to note that for models (2) and (3) when the residual variance is very small, error is very small, $h(r_j|y^m, y^o) \approx h(r_j|y^o)$ since $h(b|y^m, y^o) \approx h(b|y^o)$. In this situation, the missing data mechanism will be close to MAR, so simply fitting a likelihood-based model for $y^o$ will result in valid inference.

Albert and Follmann (2007) discuss various SRP modeling formulation for analyzing binary longitudinal data with applications to an opiates clinical trial.

There are various approaches for parameter estimation. First, maximization of the likelihood $L = \Pi_{i=1}^{I} f(y_i^o, r_i)$, where $f$ is given by (1) can be used to obtain the maximum-likelihood estimates (MLEs). Maximizing the likelihood may be computationally intensive since it involves integrating over the random effects distribution. For a high dimensional random effects distribution, this involves the numerically difficult evaluation of a high dimensional integral. Approaches such as Monte-Carlo EM or Laplace approximations of the likelihood (Gao 2004) may be good alternatives to direct evaluation of the integral. Fortunately, many applications involve only one or two shared random effects where the integral can be evaluated more straightforwardly with Gaussian quadrature, adaptive Gaussian quadrature, or other numerical integration techniques. Various statistical software packages can be used to fit these models including procedures in SAS and specialized code in R.

An alternative approach for parameter estimation, which conditions on $R_i$, has been proposed (Wu and Carroll 1988; Follmann and Wu 1995; Albert and Follmann 2000). In developing this approach, first note that the joint distribution of $(Y_i^o, R_i, b_i)$ can be re-written as

$$\begin{aligned} f(y_i^o, r_i, b_i) &= f(y_i^o, b_i | r_i) m(r_i) \\ &= f(y_i^o | b_i, r_i) h(b_i | r_i) m(r_i) \\ &= g(y_i^o | b_i) h(b_i | r_i) m(r_i). \end{aligned}$$

Thus, the conditional likelihood for $y_i^o | r_i$ is given by $L = \Pi_{i=1}^{I} \int g(y_i^o | b_i) h(b_i | r_i) db_i$. Note that this approximate conditional model can be directly viewed as a pattern mixture model as

$$\begin{aligned} f(y^o, r) &= \int g(y^o | b) h(b | r) db \, m(r) \\ &= p(y | r) m(r), \end{aligned}$$

Little (1993).

For illustration, we can estimate the treatment effect $\beta_1$ in the non-random dropout model (2) and (3) by noting that $b_i | d_i$ can be approximated by a normal distribution with mean $\omega_0 + \omega_1 d_i$. The conditional model can then be characterized by a linear mixed model of the form $Y_{ij} = \beta_0^* + \omega_1 d_i + \beta_1^* + b_i + \varepsilon_{ij}$. An important point is that the parameters of this model are conditional on the dropout time $d_i$ and are easily interpretable. What is of interest are inferences on the marginal distribution of $Y_{ij}$. To estimate $\beta_1$ in model (2), we need to marginalize over the dropout time distribution. Specifically, $E(Y|x) = E(E(Y|d,x)) = \beta_0^* + \beta_1^* x + \omega_1 E(d|x)$, where $E(d|x)$ is the conditional distribution of dropout given the covariate $x$. We can estimate this conditional distribution in a two group comparison with $\hat{E}(d|x) = \overline{d}_x$. Thus, $\beta_1 = E(Y|x=1) - E(Y|x=0) = \beta_1^* + \omega_1(\overline{d}_1 - \overline{d}_0)$. Variance estimation for the MLE approach can be obtained through standard asymptotic techniques (i.e., inverting the observed Fisher information matrix). For the conditional modeling

approach, the simplest approach is to perform a bootstrap procedure (Efron and Tibshirani 1993) where all measurements on a chosen individual are sampled with replacement.

## 3 An Analysis of Longitudinal Batched Gaussian Data Subject to Non-random Dropout

Due to cost or feasibility, longitudinal data may be measured in pools or batches whereby samples are combined or averaged across individuals at a given time point. In these studies, interest may be on comparing the longitudinal measurements across groups. A complicating factor may be that subjects are subject to dropout from the study. An example of this type of data structure is a large mouse study examining the effect of an experimental antioxidant on the weight profiles over time in mice. It was suspected that animals receiving the treatment would reach a lower adult weight than control animals and that the decline in weight among treated animals would be less than that for control animals. Ninety-five genetically identical animals were enrolled into the treatment and control groups, respectively (190 total animals). Within a group, five animals were placed in each of the 19 cages at birth. Due to the difficulty in repeatedly weighting each animal separately, the average weight per cage was recorded at approximately bi-weekly intervals over the life span of the animals (2–3 years). At each follow-up time, average batch weight was measured as the total batch weight divided by the number of animals alive in that batch.

Albert and Shih (2011) proposed a SRP model for each of the two groups separately. Initially, we present the model when individual longitudinal data are observed and then develop the model for batched longitudinal data. Denote $Y_{ij}$ as the $j$th longitudinal observation at time $t_j$ for the $i$th subject. As described in Sect. 2, for a dropout process where an individual dies between the $(d-1)th$ and the $d$th time point, $R_{i1} = R_{i2} = \ldots = R_{id-1} = 1$ and $R_{id} = 0$. The dropout time for the $i$th subject is denoted as $d_i$.

We assume a linear mixed model in each group of the form

$$Y_{ij} = \beta_0 + \beta_1 t_j + b_{0i} + b_{1i} t_j + \varepsilon_{ij}, \tag{4}$$

where $b_i = (b_{0i}, b_{1i}) \sim N(0, \Sigma_b)$ and $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$ is independent of $b_i$. Further, we denote

$$\Sigma_b = \begin{pmatrix} \sigma^2_{b0} & \sigma_{b0b1} \\ \sigma_{b0b1} & \sigma^2_{b1} \end{pmatrix}.$$

Model (4) can be made more general by including a change point or additional polynomial functions of time to the fixed and random effects. Similar to (2) and (3), the dropout mechanism can be modeled with a geometric distribution in

which, conditional on the random intercept $b_{0i}$ and random slope $b_{1i}$, $R_{ij}|(R_{ij-1} = 1, b_{0i}, b_{1i})$ is Bernoulli with probability

$$P(R_{ij} = 0|R_{ij-1} = 1, b_{i0}, b_{1i}) = \Phi(\alpha(t_j) + \theta_1 b_{0i} + \theta_2 b_{1i}), \tag{5}$$

where $\alpha(t_j)$ is a function of follow-up time $t_j$. As discussed in Sect. 2, incorporating SRPs between the response and dropout process induces a non-ignorable dropout mechanism.

For the longitudinal animal study, interest is on estimating changes in the longitudinal process over time while accounting for potential informative dropout. For batched samples, we do not observe the actual $Y_{ij}$, but rather the average measurement in each batch. At the beginning of the study, subjects are placed into batches, and these batches are maintained throughout. Since subjects are dying over time and the batch structure is maintained, there may be very few subjects in a batch as the study draws to an end. Define $B_{lj}$ as the set of subjects who are alive in the $l$th batch at the $j$th time point. Define $n_{lj}$ as the number of subjects contained in $B_{lj}$. Further, define $X_{lj} = \frac{1}{n_{lj}} \sum_{i \in B_{lj}} y_{ij}$, where in each group, $l = 1, 2, \ldots, L$, and where $L$ is the number of batches in that group. Animals are grouped into batches of five animals that are repeatedly weighed in the same cage. In this study, there are 19 batches in each group ($L = 19$ in each of the two groups).

When individual longitudinal measurements $Y_{ij}$'s are observed, maximum-likelihood estimation is relatively simple as described earlier in Sect. 2. Estimation is much more difficult when longitudinal measurements are collected in batches. In principle, we can obtain MLEs of the parameters in model (4)–(5), denoted by $\eta$, by directly maximizing the joint likelihood, where the individual contribution of the likelihood for the $l$th batch is

$$L(X_l, d_l; \eta) = \int_{b_l} f(X_l|b_l) f(d_l|b_l) f(b_l) db_l, \tag{6}$$

where $X_l = (X_{l1}, X_{l2}, \ldots, X_{lJ_l})'$, $J_l$ is the last observed time-point immediately before the last subject in the $l$th batch dies ($n_{lj} = 0$ for $j > J_l$) and $b_l$ and $d_l$ are a vectors of all the random effects and dropout times, respectively, for individuals in the $l$th batch. In the application considered here, $b_l$ is a vector contain ten random effects and $d_l$ is a vector containing the dropout times for a batch size of five mice per cage. In (6), $f(X_l|b_l) = \Pi_{j=1}^{J_l} f(X_{lj}|b_l)$, where $f(X_{lj}|b_l)$ is a univariate normal density with mean given by $\beta_0 + \beta_1 t_j + \frac{1}{n_{lj}} \sum_{i \in B_{lj}} (b_{0i} + b_{1i} t_j)$ (see (4)) and variance $\sigma_\varepsilon^2 / n_{lj}$. Further, $f(b_i)$ is a multivariate normal with block diagonal matrix (under an ordering where random effects on the same subjects are grouped together) and $f(d_l|b_l)$ is the product of geometric probabilities.

One approach to maximize the likelihood is to use the E–M algorithm. In the E-step we compute the expected value of the complete-data log likelihood (the

log-likelihood we would have if we observed $b_l$) given the observed data $X_l$ and $d_l$, and in the M-step, we maximize the resulting expectation. Specifically,

$$max_\eta \sum_{l=1}^{L} E[logL(X_l, b_i, d_l; \eta)|X_l, d_l]), \tag{7}$$

where $\eta$ is a vector of all parameters of the shared parameter model and $logL(X_l, b_l, d_l; \eta)$ is the log of the complete data likelihood for the $l$th batch. The standard E–M algorithm is implemented by iterating between an E- and an M-step, whereby the expectation in (7) is evaluated in the E-step and the parameters are updated through the maximization of (7) in the M-step. Unfortunately, the E-step is difficult to implement in closed form. As an alternative Albert and Shih (2011) proposed a Monte-Carlo (MC) EM algorithm where the E-step is evaluated using the Metropolis–Hastings algorithm; the details are included in this paper.

Although the shared parameter modeling approach is feasible, it can be computational intensive due to the Monte-Carlo Sampling. An alternative approach that is simpler to implement for the practitioner is the conditional model discussed in Sect. 2. The conditional approach can easily be adapted for approximate parameter estimation for the shared random effects model with batched longitudinal data.

The approximate conditional model approach discussed for *unbatched* longitudinal data can be applied to the batched data (i.e., observing $X_{lj}$'s rather than $Y_{ij}$'s). Since $Y_i|d_i$ in (4) is multivariate normal, $X_l|d_l$ is also multivariate normal. Denote

$$\Sigma_c = \begin{pmatrix} \sigma_{c0}^2 & \sigma_{c0c1} \\ \sigma_{c0c1} & \sigma_{c1}^2 \end{pmatrix}.$$

The conditional distribution of $X_l|d_l$ is multivariate normal with means and covariance matrix given by

$$E(X_{lj}|d_l) = \omega_0 + \omega_1 t_j + \omega_2 \sum_{i \in B_{lj}} d_i/n_{lj} + \omega_3 \sum_{i \in B_{lj}} d_i t_j/n_{lj}, \tag{8}$$

$$Cov(X_{lj}, X_{lj'}|d_l) = \frac{min(n_{lj}, n_{lj'})}{n_{lj} n_{lj'}} (\sigma_{c0}^2 + t_j t_{j'} \sigma_{c1}^2 + (t_j + t_{j'})\sigma_{c0c1}), \tag{9}$$

for $j \neq j'$, and

$$Var(X_{lj}|d_l) = \frac{1}{n_{lj}} (\sigma_{c0}^2 + t_j^2 \sigma_{c1}^2 + 2t_j \sigma_{c0c1} + \sigma_\varepsilon^2). \tag{10}$$

The multivariate normal likelihood with mean and variance given by (8)–(10) can be maximized using a quasi-Newton Raphson algorithm. This has been implemented in R using the optimum function. Once the MLEs of the conditional

model are computed, inference about the average intercept and slope can be performed by marginalizing over the dropout times. Similar to what was described for un-batched longitudinal data in Sect. 2, the average intercept and slope can be estimated by $\hat{\omega}_o + \hat{\omega}_2\overline{d}_1$ and $\hat{\omega}_1 + \hat{\omega}_3\overline{d}_2$, respectively. Similar to variance estimation for the SRP model, standard errors for the estimated mean intercept and slope can be estimated using the bootstrap by re-sampling cage-specific data.

We examined the statistical properties of the maximum-likelihood and the approximate conditional approach using simulations. First, we simulated data according to the SRP model (4) and (5) and fit the correctly specified SRP model, the approximate conditional model (APM), and an ignorable model (IM) that simply fit (4) without regard to the dropout process. Data are simulated under model (4) and (5) with $\sigma_{b0} = \sigma_{b1} = \sigma_\varepsilon = 1$, $\sigma_{b0b1} = 0$, $\alpha(t_j) = -1$, $\theta_1 = \theta_2 = 0.25$, and an intercept and slope of 0 and 1, respectively. The average estimated slopes under the SRP model, ACM, and the IM were 0.99 (SD = 0.12), 1.07 (0.11), and 0.83 (0.08), respectively. Not surprisingly, the SRP model is unbiased under the correct specification and the IM is highly biased. The ACM is approximately unbiased which is consistent with our previous theoretical discussions. Second, we simulated data according to the ACM. In this case, the ACM is unbiased, but both the SRP model and IM are severely biased. These simulations suggest that the ACM model may be more robust (under different model formulations) than the SRP model.

A detailed analysis of these data is presented in Albert and Shih (2011). We will summarize the analysis here. Scientific interest was on estimating and comparing the weight in animals at full growth (15 months) and subsequently the decline in weight in older age (slope) between the treatment group (an agent called Tempol) and a control group. Table 1 shows estimates and standard errors for the IM (simply fit the longitudinal model and discard the relationship between the two processes), SRP, and conditional approximation approaches. All methods show that Tempol treated animals have a statistically significant lower early adult weight (intercept) and a slower decline in weight into later adulthood (slope) as compared with genetically identical control animals.

## 4 Jointly Modeling Multivariate Longitudinal Measurements and Discrete Time-to-Event Data

An exciting area in biomedical research is investigating the relationship between biomarker measurements and time-to-event. For example, developing a predictor of the risk of pre-term birth from biomarker data is an important goal in obstetrical medicine. SRP models that link the two processes provides a nice way to do this. Unfortunately, this is problematic in relatively high dimensions.

Denote $Y_{1i}=(Y_{1i1}, Y_{1i2}, \ldots, Y_{1iJ_i})', Y_{2i}=(Y_{2i1}, Y_{2i2}, \ldots, Y_{2iJ_i})', \ldots Y_{Pi}=(Y_{Pi1}, Y_{Pi2}, \ldots, Y_{PiJ_i})'$ as the $P$ biomarkers measured repeatedly at $j = 1,2,..,J_i$ time points. Further, define $Y^*_{pi} = (Y^*_{pi1}, Y^*_{pi2}, \ldots, Y^*_{piJ_i})'$ as the longitudinal measurements without

**Table 1** Batched mouse data example

| Method | Parameters | Est. tempol | Est. control |
|---|---|---|---|
| Ignorable | Intercept | 28.1 | 38.4 |
| | | (0.24) | (0.76) |
| | Slope | −1.81 | −2.47 |
| | | (0.19) | (0.28) |
| Shared random parameter | Intercept | 27.9 | 36.9 |
| | | (0.24) | (0.42) |
| | Slope | −1.87 | −2.20 |
| | | (0.17) | (0.20) |
| Conditional | Intercept | 27.9 | 38.3 |
| | | (0.23) | 0.72) |
| | Slope | −1.48 | −2.36 |
| | | (0.19) | (0.31) |

The shared random parameter model and conditional model estimates of intercept and slope in the Tempol and control groups. Estimates are presented with standard errors in ( ). Standard errors were estimated using 250 samples of a non-parametric bootstrap

measurement error for the $p$th biomarker and $Y_i^* = (Y_{1i}^*, Y_{2i}^*, \ldots, Y_{Pi}^*)$. We consider a joint model for multivariate longitudinal and discrete time-to-event data in which the discrete event time distribution is modeled as a linear function of previous true values of the biomarkers without measurement error on the probit scale. Specifically,

$$P(R_{ij} = 0 | R_{ij-1} = 1; X_i^*) = \Phi(\alpha_{0j} + \sum_{p=1}^{P} \alpha_p Y_{pi(j-1)}^*), \tag{11}$$

where $i = 1,2,\ldots,I$, $j = 1,2,\ldots,J_i$, $R_{i0}$ is taken as 1, $\alpha_{0j}$ governs the baseline discrete event time distribution and $\alpha_p$ measures the effect of the $p$th biomarker ($p = 1,2,\ldots,P$) at time $t_{j-1}$ on survival at time $t_j$.

The longitudinal data is modeled assuming that the fixed and random effect trajectories are linear. Specifically, the multivariate longitudinal biomarkers can be modeled as

$$Y_{pij} = Y_{pij}^* + \varepsilon_{pij}, \tag{12}$$

where

$$Y_{pij}^* = \beta_{p0} + \beta_{p1} t_j + \gamma_{pi0} + \gamma_{pi1} t_j, \tag{13}$$

where $\beta_{p0}$ and $\beta_{p1}$ are the fixed effect intercept and slope for the $p$th biomarker, and $\gamma_{pi0}$ and $\gamma_{pi1}$ are the random effect intercept and slope for the $p$th biomarker on the $i$th individual. Denote

$$\beta = (\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \ldots, \beta_{P0}, \beta_{P1})'$$

and

$$b_i = (b_{1i0}, b_{1i1}, b_{2i0}, b_{2i1}, \ldots b_{Pi0}, b_{Pi1})'.$$

We assume that $b_i$ is normally distributed with mean 0 and variance $\Sigma_b$, where $\Sigma_b$ is a $2P \times 2P$ dimensional variance matrix and $\varepsilon_{pij}$ are independent error terms which are assumed to be normally distributed with mean 0 and variance $\sigma^2_{p\varepsilon}$ $(p = 1, 2, \ldots, P)$.

Albert and Shih (2010b) proposed a two-stage regression calibration approach for estimation, which can be described as follows. In the first stage, multivariate linear mixed models can be used to model the longitudinal data. In the second stage, the time-to-event model is estimated by replacing the random effects with corresponding empirical Bayes estimates. There are three problems with directly applying this approach. First, estimation in the first stage is complicated by the fact that simply fitting multivariate linear mixed models results in bias due to informative dropout; this is demonstrated by Albert and Shih (2010a) for the case of $P = 1$. Second, parameter estimation for multivariate linear mixed models can be computationally difficult when the number of longitudinal measurements ($P$) is even moderately large. Third, calibration error in the empirical Bayes estimation needs to be accounted for in the time-to-event model. The proposed approach will deal with all three of these problems.

The bias from informative dropout is a result of differential follow-up whereby the longitudinal process is related to the length of follow-up. That is, in (13), patients with large values of $Y^*_{pij}$ are more likely to have an early event when $\alpha_p > 0$ for $p = 1, 2, \ldots, P$. There would be no bias if all $J$ follow-up measurements were observed on all patients. As proposed by Albert and Shih (2010a) for univariate longitudinal data, we can avoid this bias by generating complete data from the conditional distribution of $Y_i = (Y_{1i}, Y_{2i}, \ldots, Y_{Pi})$ given $d_i$, denoted as $Y_i|d_i$. Since $Y_i|d_i$ under model (11–12) does not have a tractable form, we propose a simple approximation for this conditional distribution. The distribution of $Y_i|d_i$ can be expressed as

$$P(Y_i|d_i) = \int h(Y_i|b_i, d_i) g(b_i|d_i) db_i. \tag{14}$$

Since $d_i$ and the values of $Y_i$ are conditional independent given $b_i$, $h(Y_i|b_i, d_i) = h(Y_i|b_i)$, where $h(Y_i|b_i) = \Pi^P_{p=1} h(Y_{pi}|b_{pi0}, b_{pi1})$. The distribution of $Y_i|d_i$ can be expressed as a multivariate linear mixed model if we approximate $g(b_i|d_i)$ by a normal distribution. Under the assumption that $g(b_i|d_i)$ is normally distributed with mean $\mu_{d_i} = (\mu_{01d_i}, \mu_{11d_i}, \mu_{02d_i}, \mu_{12d_i}, \ldots, \mu_{0Pd_i}, \mu_{1Pd_i})'$ and variance $\Sigma^*_{bd_i}$, and by re-arranging mean structure parameters in the integrand of (14) so that the random effects have mean zero, $Y_i|d_i$ corresponds to the following multivariate linear mixed model

$$Y_{pij}|(d_i, b^*_{ip0d_i}, b^*_{ip1d_i}) = \beta^*_{p0d_i} + \beta^*_{p1d_i} t_j + b^*_{ip0d_i} + b^*_{ip1d_i} t_j + \varepsilon^*_{pij}, \qquad (15)$$

where $i = 1,2,\ldots,I$, $j = 1,2,\ldots,J_i$, and $p = 1,2,\ldots,P$. The parameters $\beta^*_{p0d_i}$ and $\beta^*_{p1d_i}$ are intercept and slope parameters for the $p$th longitudinal measurement and for patients who have an event time at time $d_i$ or who are censored at time $t_J$. In addition, the associated random effects $b^*_{id_i} = (b^*_{i10d_i}, b^*_{i11d_i}, b^*_{i20d_i}, b^*_{i21d_i}, \ldots, b^*_{iP0d_i}, b^*_{iP1d_i})'$ are multivariate normal with mean 0 and variance $\Sigma^*_{bd_i}$, and the residuals $\varepsilon^*_{pij}$ are assumed to have an independent normal distribution with mean zero and variance $\sigma^{*2}_{\varepsilon p}$. Thus, this conditional model involves estimating separate fixed effect intercept and slope parameters for each potential event-time and for subjects who are censored at time $t_J$. Likewise, separate random effects distributions are estimated for each of these discrete time points. For example, the intercept and slope fixed-effect parameters for the $p$th biomarker for those patients who have an event at time $d_i = t_3$ is $\beta^*_{p0t_3}$ and $\beta^*_{p1t_3}$, respectively. Further, the intercept and slope random effects for all $P$ biomarkers on those patients who have an event at time $d_i = t_3$, $b^*_{it_3}$, is multivariate normal with mean 0 and variance $\Sigma^*_{bt_3}$. A similar approximation has been proposed by Albert and Shih (2010a) for univariate longitudinal data $(P = 1)$.

Recall that by generating complete data from (15) we are able to avoid the bias due to informative dropout. However, when $P$ is large, direct estimation of model (15) is difficult since the number of elements in $b^*_{id_i}$ grows exponentially with $P$. For example, the dimension of the variance matrix $\Sigma^*_{bd_i}$ is $2P$ by $2P$ for $P$ longitudinal biomarkers. Fieuws and Verbeke (2005) proposed estimating the parameters of multivariate linear mixed models by formulating bivariate linear mixed models on all possible pairwise combinations of longitudinal measurements. In the simplest approach, they proposed fitting bivariate linear mixed models on all $\binom{P}{2}$ combinations of longitudinal biomarkers and averaging "overlapping" or duplicate parameter estimates. Thus, we estimate the parameters in the fully specified model (15) by fitting $\binom{P}{2}$ bivariate longitudinal models that only include pairs of longitudinal markers. Fitting these bivariate models is computationally feasible since only four correlated random effects are contained in each model. (i.e., $\Sigma^*_{bd_i}$ is a four-by-four dimensional matrix for each discrete event-time $d_i$.) Duplicate estimates of fixed effects and random-effect variances from all pairwise bivariate models are averaged to obtain final parameter estimates of the fully specified model (15). For example, when $P = 4$ there are $(P-1) = 3$ estimates of $\beta^*_{p0d_i}$, $\beta^*_{p1d_i}$, $\sigma^{*2}_{\varepsilon p}$ for the $p$th longitudinal biomarker that need to be averaged.

Model (15) is then used to construct complete longitudinal pseudo data sets which in turn are used to estimate the mean of the posterior distribution of an individual's random effects given the data. Specifically, multiple complete longitudinal data sets can be constructed by simulating $Y_{pij}$ values from the approximation to the distribution of $Y_i|d_i$ given by (15) where the parameters are replaced by their estimated values. Since the simulated data sets have complete follow-up on each individual, the bias in estimating the posterior mean of $b_i$ caused by informative dropout will be much reduced.

The posterior mean of distribution $b_i$ given the data can be estimated by fitting (11)–(13) to the generated complete longitudinal pseudo data. However, similar to fitting the conditional model (15), fitting model (11)–(13) is difficult due to the high dimension of $\Sigma_b$. Thus, we again use the pairwise estimation approach of Fieuws and Verbeke (2005), whereby we estimate the parameters of (2)–(3) by fitting all pairwise bivariate models and averaging duplicate parameter estimates to obtain final parameter estimates. For each generated complete longitudinal pseudo data set, the estimate of the posterior mean, denoted as $\hat{b}_i = (\hat{b}_{1i0}, \hat{b}_{1i1}, \ldots, \hat{b}_{Pi0}, \hat{b}_{Pi1})'$ can be calculated as

$$\hat{b}_i = \Sigma_b Z_i' V_i^{-1} (X_i - Z_i \hat{\beta}), \tag{16}$$

where $Z_i$ is a $PJ \times 2P$ design matrix corresponding to the fixed and random effects in (11)–(13), where $Z_i = diag(\underbrace{A', A', \ldots, A'}_{P \text{ Times}})$,

$$A = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_J \end{pmatrix},$$

and $V_i$ is the variance of $X_i$. Estimates of $Y^*_{pij}$, denoted as $\hat{Y}^*_{pij}$, are obtained by substituting $(\hat{\beta}_{p0}, \hat{\beta}_{p1}, \hat{b}_{pi0}, \hat{b}_{pi1})$ for $(\beta_{p0}, \beta_{p1}, b_{pi0}, b_{pi1})$ in (13).

To account for the measurement error in using $\hat{b}_i$ as compared with using $b_i$ in (11), we note that

$$P(R_{ij} = 0 | R_{i(j-1)} = 1; \hat{Y}^*_i) = \Phi\Bigg( \frac{\alpha_{0j} + \sum_{p=1}^{P} \alpha_p \hat{Y}^*_{pij}}{\sqrt{1 + Var\Big\{ \sum_{p=1}^{P} \alpha_p (\hat{Y}^*_{pij} - Y^*_{pij}) \Big\}}} \Bigg), \tag{17}$$

where $Var\Big\{ \sum_{p=1}^{P} \omega_p (\hat{Y}^*_{pi(j-1)} - Y^*_{pi(j-1)}) \Big\} = C_{ij}' Var(\hat{b}_i - b_i) C_{ij}$, $C_{ij} = (\omega_1, \omega_1 t_{j-1}, \omega_2, \omega_2 t_{j-1}, \ldots, \omega_p, \omega_p t_{j-1})$, $Var(\hat{b}_i - b_i) = \Sigma_b - \Sigma_b Z_i' \{V_i^{-1} - V_i^{-1} Z_i Q Z_i' V_i^{-1}\} Z_i \Sigma_b$, and where $Q = \sum_{i=1}^{I} (Z_i' V_i^{-1} Z_i)^{-1}$ (Laird and Ware 1982). Expression (17) follows from the fact that $E[\Phi(a + V)] = \Phi\big[(a + \mu)/\sqrt{1 + \tau^2}\big]$, where $V \sim N(\mu, \tau^2)$.

In the second stage, $\alpha_{0j}$ $(j = 1, 2, \ldots, J)$ and $\alpha_p$ $(p = 1, 2, \ldots, P)$ can be estimated by maximizing the likelihood

$$L = \prod_{i=1}^{I} \Big[ \prod_{j=1}^{J_i} \{1 - P(R_{ij} = 0 | R_{i(j-1)} = 1; \hat{Y}^*_i)\} \Big] P(R_{i(J_i+1)} = 0 | R_{iJ_i} = 1; \hat{Y}^*_i)^{J_i < J}, \tag{18}$$

where $P(R_{ij} = 0 | R_{i(j-1)} = 1, \hat{Y}^*_i)$ is given by (17). Thus, we propose the following algorithm for estimating $\alpha_{0j}$ and $\alpha_p$ $(p = 1, 2, \ldots, P)$.

1. Estimate the parameters of model (15) by fitting $\binom{P}{2}$ bivariate models to each of the pairwise combinations of longitudinal measurements and averaging duplicate parameter estimates. The bivariate models can be fit in R using code presented in Doran and Lockwood (2006).
2. Simulate complete longitudinal pseudo measurements (i.e., $Y_{pij}$ for $p = 1,2,\ldots,P$, $i = 1,2,\ldots,I$, $j = 1,2,\ldots,J$) from model (15) with model parameters estimated from step 1.
3. Estimate the parameters in model (12)–(13) without regard to the event time distribution from complete longitudinal pseudo measurements (simulated in step 2) by fitting all possible $\binom{P}{2}$ bivariate longitudinal models and averaging duplicate model parameter estimates.
4. Calculate $\hat{b}_i$ using (7) and $\hat{Y}^*_{pij}$ using (13) with $b_i$ replaced by $\hat{b}_i$ and $\beta$ being replaced by $\hat{\beta}$ estimated in step 3.
5. Estimate $\alpha_{0j}$ $(j = 1,2,..,J)$ and $\alpha_p$ $(p = 1,2,\ldots,P)$ using (17) and (18).
6. Repeat steps 2 to 5 $M$ times and average $\hat{\alpha}_{0j}$ and $\hat{\alpha}_p$ to get final estimates.

We choose $M = 10$ in the simulations and data analysis since this was shown to be sufficiently large for univariate longitudinal modeling discussed in Albert and Shih (2010a). Asymptotic standard errors of $\hat{\alpha}_{0j}$ and $\hat{\alpha}_p$ cannot be used for inference since they fail to account for the missing data uncertainty in our procedure. The bootstrap (Efron and Tibshirani 1993) can be used for valid standard error estimation.

This approach is most useful in situations where the number of longitudinal measurements is very large (e.g., panels of cytokine measurements followed longitudinally). For computational simplicity, we focus on a simulated example where three biomarkers are measured longitudinally at five time points on 300 individuals. Table 2 shows the results of these simulations. The results show that simply using the observed longitudinal data will result in severely biased estimation. The proposed approach results in unbiased estimation for the parameters of the joint model. The table also includes estimates for the situation in which we observe the biomarkers without measurement error (only possible to do in simulations). This *strawman* case simply shows us that we could do better in terms of efficiency if the biomarkers could be assessed with less measurement error.

**Table 2** Simulation shows estimates for a method that uses the true values of the markers without measurement error (NoME), the proposed method (Proposed), and an approach that uses the observed biomarkers (Observed)

| Parameters | True values | NoME | Proposed | Observed |
|---|---|---|---|---|
| $\alpha_0$ | $-1.75$ | $-1.77$ (0.115) | $-1.76$ (0.180) | $-1.37$ (0.089) |
| $\alpha_1$ | 0.40 | 0.408 (0.060) | 0.405(0.089) | 0.221(0.042) |
| $\alpha_2$ | 0 | 0.00 (0.058) | 0.00(0.0.077) | 0.001(0.042) |
| $\alpha_3$ | 0.40 | 0.405 (0.062) | 0.400(0.0.092) | 0.229(0.043) |

We simulate according to (11)–(13) with $\alpha_{0j} = \alpha_0$ for all $j$ and with stated values of $\alpha_1$, $\alpha_2$, and $\alpha_3$. Remaining values are given in Table 1 of Albert and Shih (2010b)

## 5 Jointly Modeling of Menstrual Cycle Length and Time-to-Pregnancy

An important scientific problem in reproductive epidemiology is both to characterize the menstrual cycle patterns in women who are attempting pregnancy and to develop predictive models for time-to-pregnancy. A SRP model that links together the complex menstrual cycle pattern with time-to-pregnancy is important for valid statistical analysis in both problems. Specifically, when interest focuses on characterizing longitudinal changes in the menstrual cycle, it is important to account for the dependence between the two processes since failure to do so results in informative dropout in the longitudinal process. Further, incorporating dependence between the two processes is important for developing a flexible class of prediction models of time-to-pregnancy.

The menstrual cycle pattern is complex since it is well known to be long tailed with a proportion of cycles being unusually long while a majority appear within normal ranges. Various authors (e.g., Guo et al. 2006 and references within) have proposed two component mixture models with one component reflecting a distribution with a long right tail and the other reflecting a normal distribution. McLain et al. (2012) propose a class a mixture model with a normal distribution for "normal" cycles and a long tailed distribution (reflecting the possibility of both extremely long and short irregular cycles). Further, McLain et al. introduce random effects which are shared between the two models as well as with a discrete time survival model characterizing time-to-pregnancy.

For illustration, we present a simplified version of McLain et al.'s modeling approach without external covariates. The menstrual cycle is modeled as a mixture of two components. First, for normal cycles

$$\begin{aligned} Y_{N,ij} &= \mu_i + \varepsilon_{N,ij} \\ &= \beta_0 + b_{N,i} + \varepsilon_{N,ij},\ \varepsilon_{N,ij} \sim N(0, \sigma_i^2), \end{aligned} \tag{19}$$

where $\sigma_i^2 = \sigma_0^2 exp(b + b_{S,i})$. For abnormal cycles, the following representation is assumed

$$Y_{A,ij} = \beta_0 + b_{A,i} + \varepsilon_{A,ij},\ \ \varepsilon_{A,ij} \sim EVD(0, \eta), \tag{20}$$

where $EVD(0, \eta)$ denotes an extreme-value type I distribution with location 0 and scale $\eta$. Further, for implementation it was assumed that $b_{A,ij} = \varepsilon b_{N,ij}$, that the between subject heterogeneity is a scalar shift between the normal and abnormal cycles. The distribution of the menstrual cycle length $Y_{ij}$ is completed through the specification of the mixture, $Y_{ij} = g_{ij} Y_{N,ij} + (1 - g_{ij}) Y_{A,ij}$, where $g_{ij}$ is an indicator that the $i$th women's $j$th cycle is normal, and $p = P(g_{ij} = 1)$. Finally, time-to-pregnancy is specified with a discrete-time survival model as in (5) and (11) with a different link function,

$$P(R_{ij} = 0 | R_{ij-1} = 1, b_{N,i}, b_{S,i}) = 1 - \exp(-V_{ij} \exp(\omega + \phi_\mu b_{A,i} + \phi_\sigma b_{S,i})), \quad (21)$$

where $V_{ij}$ is an indicator of whether the $i$th subject had sexual intercourse during the fertile window (period in which they can conceive) during the $j$th cycle. The parameters $\phi_\mu$ and $\phi_\sigma$ characterize the dependence between time-to-pregnancy and the mean structure and variance structure (among normal cycles), respectively. Negative values of these parameters characterize positive associations between the two processes in terms of mean and variance of the longitudinal process.

Parameter estimation was conducted by maximizing the joint likelihood as described in their manuscript. The major complication was evaluating the bi-variate integral in the joint likelihood (integrating over the bi-variate random effects). McLain et al. used Gaussian quadrature (Abramowitz and Stegun 1972) for this numerical integration.

McLain et al. (2012) fit this model to interesting time-to-pregnancy cohort data. Of interest is that they found estimates of $\phi_\mu$ which were positive and estimates of $\phi_\sigma$ which were negative ($\phi_\sigma$ estimates were significantly different from zero), reflecting that women with shorter cycles and more normal cycle variability had a longer time to pregnancy.

## 6 Discussion

This paper presents a summary of methodology and approaches for using SRP models to analyze longitudinal data subject to missingness. The basic approach which started with Wu and Carroll (1988) has been expanded in many directions. In this paper we reviewed some of these expansions focusing on making inference in batched or pooled longitudinal data subject to missingness, joint modeling of high dimensional longitudinal data and time-to-event data, and in the joint modeling of time-to-pregnancy and complex menstrual cycle patterns.

We discuss both a direct maximum-likelihood approaches and approximating approaches for fitting SRP models. The direct modeling approaches, although feasible for univariate longitudinal responses, can be very computationally intensive for either batched or high-dimensional longitudinal data. In fact, direct maximum-likelihood approach would be infeasible for joint modeling of high-dimensional and time-to-event data with all known approaches. The approximate conditional model, although not ideal in certain cases, provides a feasible solution to this difficult problem.

Most SRP models assume that the random effects are normally distributed. Various authors have investigated the robustness of inferences for different settings. Davidian et al. have shown that for a joint model of longitudinal and survival inferences on joint model parameters are relatively insensitive to random effects misspecification, particularly when the number of longitudinal measurements is relatively large.

Selection models where the non-ignorable missing data mechanism is incorporated by modeling the probability of missing as a direct function of the missed observation had it been observed are alternatives to SRP models. Pattern mixture modeling, another commonly used technique for analyzing longitudinal data with missingness, is one where we condition on the missing data pattern and make inference marginalized over the missing data pattern. In spirt, the pattern mixture model is the similar to the conditional model which is proposed as an approximation to the SRP model.

We recommend that model adequacy be examined in traditional ways such as the examination of residuals and fitted values as well as through goodness of fit tests. However, as pointed out by Molenberghs et al. (2012), SRP models (as well as other models for missing data) require assumptions about underlying mechanism that are impossible to fully verify empirically. Knowledge about the subject at hand needs to be incorporated in model development for proper inference.

**Acknowledgments** We thank the referee and editor for their thoughtful and constructive comments which lead to an improved manuscript. We also thank the audience of the International Symposium in Statistics (ISS) on Longitudinal Data Analysis Subject to Outliers, Measurement Errors, and/or Missing Values. This research was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

## References

Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1972)

Albert, P.S.: A transition model for longitudinal binary data subject to nonignorable missing data. Biometrics **56**, 602–608 (2000)

Albert, P.S., Follmann, D.A.: Modeling repeated count data subject to informative dropout. Biometrics **56**, 667–677 (2000)

Albert, P.S., Follmann, D.A.: Random effects and latent processes approaches for analyzing binary longitudinal data with missingness: a comparison of approaches using opiate clinical trial data. Stat. Methods Med. Res. **16**, 417–439 (2007)

Albert, P.S, Follmann, D.A., Wang, S.A., Suh, E.B.: A latent autoregressive model for longitudinal binary data subject to informative missingness. Biometrics **58**, 631–642 (2002)

Albert, P.S., Shih, J.H.: On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. Biometrics **66**, 983–991 (2010a)

Albert, P.S., Shih, J.H.: An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. Ann. Appl. Stat. **4**, 1517–1532 (2010b)

Albert, P.S., Shih, J.H.: Modelling batched Gaussian longitudinal weight data in mice subject to informative dropout. Published on-line before print. Stat. Methods Med. Res. (2011)

DeGruttola, V.T., Tu, X.M.: Modeling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics **50**, 1003–1014 (1994)

Doran, H.C., Lockwood, J.R.: Fitting value-added models in R. J. Educ. Behav. Stat. **31**, 205–230 (2006)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)

Fieuws, S., Verbeke, G.: Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. Biometrics **62**, 424–431 (2005)

Follmann, D., Wu, M.: An approximate generalized linear model with random effects for informative missing data. Biometrics **51**, 151–168 (1995)
Gao, S.: A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. Stat. Med. **23**, 211–219 (2004)
Guo, Y., Manatunga, A.K., Chen, S., Marcus, M.: Modeling menstrual cycle length using a mixture distribution. Biostatistics **7**, 151–168 (2006)
Hogan, J.W., Laird, N.M.: Model-based approaches to analyzing incomplete longitudinal and failure time data. Stat. Med. **16**, 259–272 (1997)
Laird, N.M., Ware, J.H.: Random -effects models for longitudinal data. Biometrics **38**, 963–978 (1982)
Little, R.J.A.: Pattern-Mixture models for multivariate incomplete data. J. Am. Stat. Assoc. **88**, 125–134 (1993)
Little, R.J.A.: Modeling the drop-out mechanism in repeated-measures studies. J. Am. Stat. Assoc. **90**, 1112–1121 (1995)
Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)
McLain, A., Lum, K.J., Sundaram, R.: A joint mixed effects dispersion model for menstrual cycle length and time-to-pregnancy. Biometrics **68**, 648–656 (2012)
Molenberghs, G., Njeru, E., Kenward, M.G., Verbeke, G.: Enriched-data problems and essential non-identifiability. Int. J. Stat. Med. Res. **1**, 16–44 (2012)
Mori, M., Woolson, R.F., Woodworth, G.G.: Slope estimation in the presence of informative censoring: modeling the number of observations as a geometric random variable. Biometrics **50**, 39–50 (1994)
Rubin, D.B.: Inference and missing data. Biometrika **63**, 581–592 (1976)
Schluchter, M.D.: Methods for the analysis of informative censored longitudinal data. Stat. Med. **11**, 1861–1870 (1992)
Schluchter, M.D., Greene, T., Beck, G.J.: Analysis of change in the presence of informative censoring: application to a longitudinal clinical trial of progressive renal disease. Stat. Med. **20**, 989–1007 (2001)
Sutradhar, B.C., Mallick, T.S.: Modified weights based on generalized inferences in incomplete longitudinal binary models. Can. J. Stat. **38**, 217–231 (2010)
Ten Have, T.R., Pulkstenis, E., Kunselman, A., Landis, J.R.: Mixed effects logistic regression models for longitudinal binary response data with informative dropout. Biometrics **54**, 367–383 (1998)
Tsiaits, A.A., Davidian, M.: Joint modeling of longitudinal and time-to-event data: an overview. Stat. Sin. **14**, 809–834 (2004)
Tsiatis, A.A., Davidian, M.: A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika **88**, 447–459 (2001)
Tsiatis, A.A., DeGruttola, V., Wulfsohn, M.S.: Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. J. Am. Stat. Assoc. **90**, 27–37 (1995)
Vonesh, E.F., Greene, T., Schluchter, M.D.: Shared parameter models for the joint analysis of longitudinal data with event times. Stat. Med. **25**, 143–163 (2006)
Wu, M.C., Carroll, R.J.: Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics **44**, 175–188 (1988)
Wulfson, M.S., Tsiatis, A.A.: A joint model for survival and longitudinal data measured with error. Biometrics **53**, 330–339 (1997)
Zeger, S.L., Qaqish, B.: Markov regression models for time series: a quasilikelihood approach. Biometrics **44**, 1019–1031 (1988)

# Response-Dependent Sampling with Clustered and Longitudinal Data

**Michael A. McIsaac and Richard J. Cook**

**Abstract** Prospective cohort studies typically involve repeated assessment of individuals to determine whether they have a particular health condition. The usual goal in such studies is to relate the presence of the condition to disease markers or exposure variables. Disease markers are often too difficult or costly to measure for all individuals in a sample. In such settings, two- and multi-phase sampling designs are routinely adopted to enable researchers to select individuals on whom these expensive markers are to be assessed. In this article we review the rationale and format of two-phase sampling designs in retrospective and cross-sectional studies. We then develop frameworks for multi-phase designs in the context of studies with clustered or longitudinal responses. Model-based and semi-parametric methods are discussed for estimation and inference.

## 1 Introduction

Two-phase sampling designs have proven useful in epidemiology for ensuring efficient use of resources when estimating the effect of expensive or otherwise difficult to measure exposure variables on a response. Under such designs, a regression model is often specified with a binary response indicating disease status and a covariate vector recording the exposure variable of interest along with possible auxiliary covariates. The first phase of sampling generates data on the response and auxiliary covariates. A sub-sample of these individuals is chosen at a second phase of sampling, and the expensive exposure variable is measured for these individuals. Viewed as a whole, the full sample features missing exposure data in individuals

M.A. McIsaac (✉) • R.J. Cook
Department of Statistics and Actuarial Science, University of Waterloo,
Waterloo, Ontario, Canada, N2L 3G1
e-mail: mamcisaa@uwaterloo.ca; rjcook@uwaterloo.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_8, 

selected in phase I but not selected in the phase II sub-sample, with the missing data mechanism determined by the nature of the phase-II sampling probabilities.

There is a wide range of statistical approaches for regression with incomplete covariate data, including methods based on maximum likelihood (Lawless et al. 1999), mean score equations (Reilly and Pepe 1995; Reilly 1996), inverse probability weighted estimating functions, and augmented inverse probability weighted estimating functions (Robins et al. 1994; Tsiatis 2006). These approaches differ in the nature of the assumptions required and the extent to which data from individuals with incomplete exposure data are utilized. Maximum likelihood, while potentially optimally efficient, requires one to model the distribution of the exposure variable given any auxiliary variables, and misspecification of this model can lead to an inconsistent estimator (Horton and Laird 2001). The mean score method involves specification of unbiased estimating functions by nonparametrically estimating the conditional distribution of the exposure variable given the response and auxiliary variables based on the phase-II sample (Reilly and Pepe 1995). In their simplest form, inverse probability weighted estimating equations restrict attention to individuals in the phase-II sample and hence do not require modelling of the covariate distribution. The resulting estimates are consistent provided the weights are correctly specified, but they are typically less efficient than maximum likelihood estimates (Lawless et al. 1999). Augmented inverse probability weighted estimating equations aim to improve efficiency by exploiting information in the individuals who only provide information in the phase-I sample (Robins et al. 1994; Tsiatis 2006).

When planning studies, the challenge is to specify the phase-II selection model which will lead to the most efficient estimators of the parameters of interest; this is typically the coefficient of the exposure variable. To do this one must adopt a response model and a framework for inference which accommodate the incomplete exposure data. Factors influencing the choice of the framework for inference include the kinds of assumptions one is willing to make, the degree of importance placed on robustness, and efficiency. Given any particular framework, the asymptotic distribution of the resulting estimators is then required to inform the design (i.e. specification of the phase-II sampling probabilities).

Much of the work to date on two-phase designs involves univariate outcomes reflecting disease status. The purpose of this article is to consider statistical issues in two-phase designs with more complex disease outcomes, motivated by our involvement in the following two studies.

*Example 1 (A Study of Genetic Risk in Psoriatic Arthritis).* The Centre for Prognosis Studies in the Rheumatic Diseases maintains a clinical registry of patients at the Toronto Western Hospital with psoriatic arthritis. Patients have been recruited and followed since its inception in 1976 and it is now the largest cohort of patients with PsA in the world. Upon entry to the clinic patients undergo a detailed clinical and radiological examination and provide serum samples which are subsequently stored. Follow-up clinical and radiological assessments are scheduled annually and biannually, respectively, in order to track changes in joint damage. Disease

progression can be modelled in a number of ways including the development of newly damaged joints (Sutradhar and Cook, 2009), the involvement of new types of joints (Tolusso and Cook 2009; Chandran et al. 2010), and the onset of a particular condition. These approaches, however, involve composite outcomes because they aggregate information over multiple joints. We consider analyses based on models for the onset of damage in the sacroiliac (SI) joints, which signals the onset of spondyloarthritis. Damage of the SI joints is determined by radiological examination with the extent of damage in each joint graded using a standardized scale (Rahman et al. 1998). Serum biomarkers and genetic factors can play important roles in identifying patients at high risk for developing psoriatic spondyloarthritis (Rahman et al. 1998), and as a consequence, biomarker studies are of considerable importance.

We consider data from patients from the first assessment at which serum samples are taken which can be used for genetic testing. We restrict attention to individuals who have not experienced damage in their sacroiliac joints as of this assessment and a clustered (paired) response is based on the onset of damage in the left and right sacroiliac joints between the baseline and a follow-up assessment. The candidate genetic risk factor in this setting is the human leukocyte antigen B27, a factor known to be associated with progression of other diseases involving connective tissue and joints, and the auxiliary variable is a marker of inflammation called C-reactive protein (CRP) (del Rincon et al. 2003). Genetic typing is costly and it is desirable to carry this out for a subset of individuals in the cohort.

*Example 2 (The Canadian Longitudinal Study of Aging).* The Canadian Longitudinal Study on Aging (CLSA) involves the establishment of a pan-Canadian cohort to enable estimation of the incidence rates of several chronic diseases and to study associated risk factors. It involves 50,000 individuals aged 45 to 85 years old who are to be followed for 20 years or until the time of death. All participants in the CLSA will provide some information to the study, while a subset of 30,000 will be chosen for additional, in-depth examination. This sub-cohort will undergo a more intensive clinical examination, provide imaging data, and give biological specimens every three years; specimens will be stored in biobanks in a controlled environment to facilitate subsequent testing. Thus the biobank will serve as a valuable resource for affiliated investigators to study risk factors predictive of disease onset and progression. Samples will be too expensive to process for all 30,000 individuals in the cohort undergoing intensive follow-up, so it will be of central importance to determine how individuals should be selected for testing of stored specimens (Raina et al. 2009). We therefore explore the extension of the two-phase sampling problem to longitudinal data. Since interest lies in the onset of disease, we focus on transitional models and formulate the exposure effects on transition probabilities. We study various designs for sampling and analysis to investigate how optimal selection procedures can be derived at a particular time point given the available partial histories. Specifically, we examine the improved precision in estimation that can result when more information is used in deriving optimal selection probabilities.

The remainder of this article is organized as follows. Notation is defined and the format of two-phase response-dependent sampling schemes is described in Sect. 2. In Sect. 3 we consider the setting of clustered responses with cluster-level exposure and auxiliary variables. Marginal models (Liang and Zeger 1986) are adopted in this setting with analysis frameworks based on maximum likelihood, mean score estimating functions, and inverse probability weighted pseudolikelihood. In Sect. 4 we give a framework for two-phase designs in longitudinal studies where interest lies in modelling the effect of an exposure variable on the onset of disease under a first-order Markov model. Asymptotic theory and optimal designs are provided for each setting. Concluding remarks and topics for further research are given in Sect. 5.

## 2 Response-dependent Sampling with Correlated Data

### *2.1 Notation and Study Design*

Two-phase sampling has been widely used to enhance precision of estimators of key parameters with resource constraints (Chatterjee et al. 2003; Pickles et al. 1995; Whittemore and Halpern 1997). This sampling framework is particularly appealing whenever the measurement of a covariate of central importance incurs considerable cost relative to the cost of associated auxiliary variables. Two-phase sampling involves the collection of outcome and inexpensive auxiliary data in a large phase-I sample, which is exploited to determine how individuals should be selected into a phase-II subsample for measurement of the expensive covariate (Reilly and Pepe 1995; Zhao et al. 2009). The efficiency gain that comes from such a two-phase sampling framework depends on the parameter of interest, the method of analysis, and the way in which the phase-I data are exploited in the design of the phase-II selection probabilities (Reilly 1996).

We begin with a discussion of likelihood-based inference which requires full model specification but enables optimal efficiency. To cover the case of clustered and longitudinal data simultaneously, we adopt a general formulation whereby $Y_i = (Y_{i1}, \ldots, Y_{iK})'$ denotes a $K \times 1$ response vector for individual $i$; we let $X_i$ and $V_i$ denote the expensive exposure variable and the auxiliary variable, respectively. Let $f(Y_i|X_i, V_i; \beta)$ denote the conditional joint density or mass function for $Y_i$ given $(X_i, V_i)$ indexed by a $p \times 1$ parameter $\beta$. Let $g(X_i|V_i; \alpha)$ denote the conditional distribution of $X_i|V_i$ indexed by a $q \times 1$ parameter $\alpha$ and let the $r \times 1$ parameter $\gamma$ index the marginal distribution of $V$. The random variables are governed by the joint model $f(Y, X, V; \beta, \alpha, \gamma) = f(Y|X, V; \beta)g(X|V; \alpha)h(V; \gamma)$, but $(\alpha, \gamma)$ are nuisance parameters which are routinely eliminated by conditioning on $(X, V)$ when data are complete.

In a two-phase study, $\{(Y_i, V_i), i = 1, 2, \ldots, N\}$ are observed for all $N$ individuals selected in the phase-I sample and $X_i$ is observed in the $n$ individuals selected for inclusion in the phase-II sample. If $R_i = I(X_i \text{ is observed})$, then selection into the

phase-II sample is governed by the phase-II selection probabilities $\pi(Y,V;\delta) = P(R=1|Y,V;\delta)$, where $\delta$ indexes this distribution. Note that within this two-phase sampling framework, we consider missingness by design, so we can be confident that data are *missing at random* (MAR)—i.e. $P(R=1|Y,X,V;\delta) = P(R=1|Y,V;\delta)$ (Little and Rubin 2002). If the phase-II selection probabilities do not exploit the phase-I data—i.e. $P(R=1|Y,V;\delta) = P(R=1;\delta)$—then individuals are selected for the phase-II sample by simple random sampling and the expensive exposure variable will be *missing completely at random* (MCAR). Phase-two selection probabilities which exploit phase-I data can result in more efficient estimators.

## 2.2 Methods of Analysis

A variety of frameworks are available for the analyses of clustered data (Neuhaus 1992). Mixed-effect models (Laird and Ware 1982; Stiratelli et al. 1984) are effective when one wishes to assess the effects of within-cluster covariates. These models account for the dependence of responses within clusters by introducing unobservable, cluster-specific latent variables. When one wishes to explore the effects of cluster-level covariates on marginal means, analyses are often more naturally carried out via population-average approaches which may involve full model specification (Heagerty and Zeger 2000; Heagerty 2002); first order generalized estimating equations can also be adopted (Liang and Zeger 1986) or second order generalized estimating equations could be used, the latter being most often considered for clustered binary responses (Prentice 1988; Zhao and Prentice 1990). Autoregressive models are appropriate when response data arise serially and it is of interest to determine how changes occur over time (Zeng and Cook 2007; Sutradhar 2008). These methods of analyses can be extended in different ways to account for data which are incomplete (Lawless et al. 1999; Robins et al. 1995; Troxel et al. 1997).

We consider three likelihood-based methods for estimation of regression coefficients in marginal mean models and conditional means when covariate data are incomplete due to a MAR mechanism.

### 2.2.1 Maximum Likelihood

The full likelihood for these data is

$$L_F(\beta,\alpha,\gamma,\delta) = \prod_{i=1}^{N} \left[ f(Y_i,X_i,V_i;\beta,\alpha,\gamma)\, P(R_i=1|Y_i,V_i;\delta) \right]^{R_i} \times \left[ f(Y_i,V_i;\beta,\alpha,\gamma)\, P(R_i=0|Y_i,V_i;\delta) \right]^{1-R_i}.$$

One may restrict attention to the partial likelihood

$$L(\beta,\alpha,\gamma)=\prod_{i=1}^{N}\left[f(Y_i,X_i,V_i;\beta,\alpha,\gamma)\right]^{R_i}\left[f(Y_i,V_i;\beta,\alpha,\gamma)\right]^{1-R_i}$$

provided $\delta$ is functionally independent of $(\beta',\alpha',\gamma')'$. In the special case that data are complete, orthogonality of the parameters enables focus on the partial likelihood

$$L(\beta)=\prod_{i=1}^{N}f(Y_i|X_i,V_i;\beta) \tag{1}$$

(Breslow and Chatterjee 1999; Lawless et al. 1999). More generally however, if $X_i$ is not observed for some clusters and the missing data mechanism is MAR, then the observed data partial likelihood is

$$L(\theta)=\prod_{i=1}^{N}\left[f(Y_i|X_i,V_i;\beta)g(X_i|V_i;\alpha)\right]^{R_i}\left[E_{X|V}\left\{f(Y_i|\mathrm{X},V_i;\beta)\right\}\right]^{1-R_i}, \tag{2}$$

where it can be seen that estimation of the parameters of interest, $\beta$, must occur jointly with the estimation of the nuisance parameter $\alpha$ in $\theta=(\beta',\alpha')'$.

Parameter estimates can be found by maximizing the likelihood in (2) directly, or by implementing an EM algorithm (Dempster et al. 1977) and iteratively maximizing the complete-data likelihood

$$L_c(\theta)=\prod_{i=1}^{N}\left[f(Y_i|X_i,V_i;\beta)g(X_i|V_i;\alpha)\right]^{R_i}\left[f(Y_i|X_i,V_i;\beta)g(X_i|V_i;\alpha)\right]^{1-R_i} \tag{3}$$

(Little and Rubin 2002). The expectation step involves computing $Q(\theta;\theta^k)=E_{X|Y,V}[\log L_c(\theta);\theta^k]$, where $\theta^k$ is the estimate of $\theta$ at the $k$th iterations and $E_{X|Y,V}[\log L_c(\theta);\theta^k]$ is

$$\sum_{i=1}^{N}\Big\{R_i\left[\log\, f(Y_i|X_i,V_i;\beta)+\log\, g(X_i|V_i;\alpha)\right]$$
$$+\,(1-R_i)\left[E_{X|Y,V}\{\log\, f(Y_i|X,V_i;\beta);\theta^k\}+E_{X|Y,V}\{\log\, g(X|V_i;\alpha);\theta^k\}\right]\Big\}.$$

The maximization step yields updated estimates $\theta^{(k+1)}$ obtained by solving

$$\frac{\partial Q(\theta;\theta^k)}{\partial\beta}=\sum_{i=1}^{N}\left\{R_iU_\beta(Y_i|X_i,V_i)+(1-R_i)E_{X|Y,V}[U_\beta(Y_i|X,V_i);\theta^k]\right\}=0 \tag{4}$$

and

$$\frac{\partial Q(\theta;\theta^k)}{\partial\alpha}=\sum_{i=1}^{N}\left\{R_iU_\alpha(X_i|V_i)+(1-R_i)E_{X|Y,V}[U_\alpha(X|V_i);\theta^k]\right\}=0,$$

where $U_\alpha(X_i|V_i)=\partial\log g(X_i|V_i;\alpha)/\partial\alpha$, and $U_\beta(Y_i|X_i,V_i)=\partial\log f(Y_i|X_i,V_i;\beta)/\partial\beta$.

Note that this method simultaneously estimates $\beta$ and $\alpha$ and hence if $X|V$ is not correctly modelled, estimates of $\beta$ will be inconsistent. We therefore consider alternative methods of analysis which, while motivated by the likelihood approach, do not require specification of the model for $X|V$. These pseudolikelihood approaches are potentially less efficient, but can provide consistent estimators of $\beta$ without making any model assumptions about the covariate distributions.

### 2.2.2 The Mean Score Method

Each step in the iterative EM procedure involves using (4) to update $\beta$ by estimating the conditional expectation of the pseudoscore function $U_\beta(Y|X,V)$ for individuals with incomplete data. This expectation can alternatively be estimated empirically in a single step (Lawless et al. 1999) rendering the so-called mean score equation of Reilly and Pepe (1995):

$$\overline{U}(\beta) = \sum_{i=1}^{N} \left\{ R_i U_\beta(Y_i|X_i,V_i) + (1-R_i)\hat{E}_{X|Y,V}[U_\beta(Y_i|X,V_i)] \right\} = 0. \tag{5}$$

The problem then reduces to obtaining a robust nonparametric estimate of $g(X|Y,V)$ in order to compute $\hat{E}_{X|Y,V}(\cdot)$. When data are MAR and $(Y,V)$ is discrete, the conditional distribution can be consistently estimated nonparametrically using the phase-II sample since $g(X|Y,V,R=1) = g(X|Y,V)$.

### 2.2.3 Weighted Pseudolikelihood

Recall that with complete data on all individuals we would want to maximize the likelihood function (1) or, equivalently, solve the score equations

$$U(\beta) = \sum_{i=1}^{N} U_\beta(Y_i|X_i,V_i) = \sum_{i=1}^{N} \partial \log f(Y_i|X_i,V_i;\beta)/\partial\beta = 0. \tag{6}$$

When the data are incomplete, rather than making auxiliary distributional assumptions, we may wish to restrict attention to individuals who provide complete information. Such complete-case estimators often induce bias when data are not MCAR (Little and Rubin 2002), but if contributions to (6) are weighted by the inverse of the probability $X_i$ is observed, the resultant estimators will be consistent (Lawless et al. 1999; Robins et al. 1994). That is, we can maximize the weighted log-pseudolikelihood or, equivalently, solve the weighted pseudoscore equations

$$\overline{\overline{U}}(\beta) = \sum_{i=1}^{N} \overline{\overline{U}}_i(\beta) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i,V_i;\delta)} U_\beta(Y_i|X_i,V_i) = 0. \tag{7}$$

Solving $\overline{\overline{U}}(\beta) = 0$ yields a consistent estimator of $\beta$ since (7) is an unbiased estimating function. To see this, we take the expectation of a single term in the sum and drop the subscript $i$ for convenience to get

$$E_{Y|X,V}\left[\frac{E_{R|Y,X}\{R\}}{\pi(Y,V;\delta)}U_\beta(Y|X,V)\right] = E_{Y|X,V}\left[U_\beta(Y|X,V)\right] = 0\,,$$

since $R \perp X|(Y,V)$ if $X$ is MAR and $U_\beta(Y|X,V)$ is an unbiased estimating function.

We now turn our attention to the particular problems of optimal two-phase design with clustered and longitudinal data and restrict attention to the case of binary responses.

## 3 Response-dependent Sampling with Clustered Binary Data

### *3.1 The Response Model for Clustered Data*

Let $Y_i = (Y_{i1}, Y_{i2})'$ denote the bivariate binary response for cluster $i$, and let $X_i$ and $V_i$ be the univariate expensive and auxiliary covariates, respectively, defined at the cluster level (i.e. all subjects in a given cluster have the same values of these covariates). In the context of the study from the University of Toronto Psoriatic Arthritis Clinic, the responses correspond to the status of the left and right sacroiliac joints. The expensive covariate represents the human leukocyte antigen (HLA) marker B27 and the auxiliary variable is the inexpensive marker of inflammation, CRP, measured at the baseline visit. We consider a regression model for the marginal mean and let $\mu_{ij} = E[Y_{ij}|X_i,V_i] = P(Y_{ij} = 1|X_i,V_i)$. Specifically we adopt the logistic model

$$\text{logit}\ \ \mu_{ij} = \beta_0 + \beta_x X_i + \beta_v V_i, \tag{8}$$

where the covariates are assumed to have a common affect on both responses. We adopt the model of Lipsitz et al. (1991) and so account for the association between $Y_{i1}$ and $Y_{i2}$ given $(X_i,V_i)$ via a common conditional odds ratio. That is, we let $\mu_{ikl} = P(Y_{i1} = k, Y_{i2} = l|X_i,V_i;\beta)$, where $\beta = (\beta_0,\beta_x,\beta_v,\psi)'$, with

$$\psi = \frac{P(Y_{i1}=1,Y_{i2}=1|X_i,V_i)/P(Y_{i1}=0,Y_{i2}=1|X_i,V_i)}{P(Y_{i1}=1,Y_{i2}=0|X_i,V_i)/P(Y_{i1}=0,Y_{i2}=0|X_i,V_i)} = \frac{\mu_{i11}/\mu_{i01}}{\mu_{i10}/\mu_{i00}}$$

the odds of subunit 1 in cluster $i$ responding given $X_i$ and $V_i$ when subunit 2 responds, versus the respective odds when subunit 2 doesn't respond, Then

$$P(Y_{i1}=1,Y_{i2}=1|X_i,V_i;\beta) = \begin{cases} \frac{c_i - [c_i^2 - 4\psi(\psi-1)\mu_{i1}\mu_{i2}]^{1/2}}{2(\psi-1)} & \text{if } \psi \neq 1 \\ \mu_{i1}\mu_{i2} & \text{if } \psi = 1, \end{cases}$$

where $c_i = 1-(1-\psi)(\mu_{i1}+\mu_{i2})$. The marginal means and the odds ratio completely specify the bivariate distribution of the clustered binary responses. We consider binary covariates $X$ and $V$ which arise so that

$$\text{logit } P(X_i = 1|V_i;\alpha) = \alpha_0 + \alpha_v V_i$$

and

$$\text{logit } P(V_i = 1;\gamma) = \gamma.$$

The discrete nature of the covariates $(X,V)$ means there is no issue of misspecification in this part of the model.

## 3.2 The Selection Model

We specify the second-phase sampling design for these bivariate data through the choice of selection parameters $\delta$ in the probabilities $\pi(Y_i,V_i;\delta) = P(R_i = 1|Y_i,V_i;\delta)$, where we consider the selection model

$$\text{logit } \pi(Y_i,V_i;\delta) = \delta_0+\delta_1 Y_{i1}+\delta_2 Y_{i2}+\delta_3 V_i+\delta_4 Y_{i1}Y_{i2}+\delta_5 Y_{i1}V_i+\delta_6 Y_{i2}V_i+\delta_7 Y_{i1}Y_{i2}V_i.$$

Note that since the covariate $V$ and the responses $Y_1$ and $Y_2$ are binary, the use of this saturated selection model is equivalent to specifying stratum-specific sampling probabilities which indicate the selection probabilities that should be used within each of the eight strata defined by the phase-I data $(Y_1,Y_2,V)$.

## 3.3 Mean Score Method with Discrete Phase-One Data

When both $Y$ and $V$ are discrete variables and $g(X|Y,V) = g(X|Y,V,R=1)$, a natural estimate of the conditional distribution $g(X|Y,V)$ is

$$\hat{g}(X|Y,V) = \frac{n^{(1)}_{X,Y,V}}{n^{(1)}_{Y,V}},$$

where $n^{(1)}_{Y,X,V} = \sum_{i:R_i=1} I(Y_i = Y, X_i = X, V_i = V)$ and $n^{(1)}_{Y,V} = \sum_{i:R_i=1} I(Y_i = Y, V_i = V)$. The conditional expectation of the pseudoscore is then estimated as

$$\hat{E}_{X|Y,V}[U_\beta(Y_j|X,V_j)] = \sum_x U_\beta(Y_j|X,V_j)\frac{n^{(1)}_{Y_j,X,V_j}}{n^{(1)}_{Y_j,V_j}}$$

so

$$\sum_{j:R_j=0} \hat{E}_{X|Y,V}[U_\beta(Y_j|X,V_j)] = \sum_{i:R_i=1} U_\beta(Y_j|X_i,V_j)\frac{n^{(0)}_{Y_j,V_j}}{n^{(1)}_{Y_j,V_j}},$$

where $n^{(0)}_{Y,V} = \sum_{i:R_i=0} I(Y_i = Y, V_i = V)$. Therefore the mean score estimating equations (5) reduce to

$$\overline{U}(\beta) = \sum_{i=1}^{N} R_i \left(1 + \frac{n^{(0)}_{Y_i,V_i}}{n^{(1)}_{Y_i,V_i}}\right) U_\beta(Y_i|X_i,V_i) = 0,$$

which can be seen to be a weighted pseudolikelihood approach (7) where the selection probabilities are estimated empirically using

$$\pi(Y,V;\hat{\delta}) = \left(1 + \frac{n^{(0)}_{Y,V}}{n^{(1)}_{Y,V}}\right)^{-1} = \frac{\sum_i I(R_i = 1, Y_i = Y, V_i = V)}{\sum_i I(Y_i = Y, V_i = V)}$$

(Lawless et al. 1999; Zhao 2005). The weighted pseudolikelihood approach will remain consistent if known weights are replaced with consistently estimated weights, as is done here with the mean score method. In fact, it is often advantageous to utilize estimated weights even when the true weights are known since the estimation of weights in (5) incorporates information from all individuals available at the first phase of sampling, while (7) only considers the completely observed individuals selected at phase two; therefore, this mean score approach will generally be more efficient than the weighted pseudolikelihood approach that incorporates the known selection probabilities (Lawless et al. 1999; Robins et al. 1994).

### *3.4 Frameworks for Analysis and Design Criteria*

Different designs can exploit phase-I data in different ways. The different second-phase sampling designs will result in different levels of efficiency of the resultant estimators, and the optimal designs will depend on the chosen method of analysis. We consider five sampling designs: simple random sampling, balanced sampling, optimal maximum likelihood sampling, optimal weighted pseudolikelihood sampling, and optimal mean score sampling. These designs (which are described in more depth below) require different amounts of information at phase-I. Simple random sampling ignores all phase-I data. Balanced sampling designs require only the size of the phase-I strata. The optimal designs are derived to minimize the asymptotic variance of the estimator of $\beta_x$ and they require knowledge of the parameter values at the design stage. In practice, these parameter values would be

unknown; however, it would be possible to base these optimal design derivations on initial parameter estimates found using a small pilot study. This process has been shown to work well in several settings (Reilly and Pepe 1995; Reilly 1996; Pepe et al. 1994; Whittemore and Halpern 1997).

We consider the problem where $N$, the size of the phase-I sample is fixed and budgetary constraints require that the expected number of individuals selected at phase-II, $P(R=1)*N$, is also fixed. Optimal designs aim to minimize the variance of the estimator of $\beta_x$ subject to this budgetary constraint. We consider Bernoulli sampling (Lawless et al. 1999) wherein all $N$ individuals are observed at phase-I and selection decisions for inclusion in phase-II are made independently and according to pre-specified selection probabilities $\pi(Y,V;\delta)$.

Truly optimal designs are not always feasible as they may sometimes result in selection probabilities that exceed one (Reilly and Pepe 1995) and may degenerate and result in selection probabilities that are near zero for some strata (Breslow and Cain 1988). In general, small selection probabilities are problematic as they may preclude testing of certain interactions, and both the mean score method and the weighted pseudolikelihood require selection probabilities be bounded away from zero. We, therefore, constrain all of our selection probabilities to be in the range $(0.05,1)$. As in Reilly and Pepe (1995), when optimal selection probabilities fall outside of this range, we fix the offending selection probability at the boundary and optimize the remaining selection probabilities. The balanced design can suffer from a similar problem in that a truly balanced design can often require selection probabilities that are larger than 1 in smaller strata. In this situation, we fix the offending selection probabilities at 1 and select the remaining individuals in a balanced way from the other strata.

### 3.4.1 Simple Random Sampling

Simple random sampling uses phase-II selection probabilities that are the same for all individuals irrespective of their phase-I data: i.e. $\pi(Y,V;\delta)=P_R$ for some constant $P_R$. The data that arise from this design are MCAR. This naive sampling scheme does not exploit information available in the phase-I data and so it will be used as a baseline to assess the efficiency gains of more sophisticated designs.

### 3.4.2 Balanced Sampling

Breslow and Cain (1988) and Breslow and Chatterjee (1999) advocate a balanced sampling design. Phase-I data are used to stratify the sample, and the phase-II sample is chosen to contain the same number of individuals from each stratum. This design is not optimally efficient but is thought to offer a "reasonable compromise between the competing demands of efficiency and the need to check model assumptions" (Breslow and Chatterjee 1999).

It is not always clear how the phase-I data should be used to stratify the sample. For the clustered data problem, we will consider two balanced sampling designs. In the first balanced sampling design, the phase-I sample will be divided into the eight classes defined by all possible values of $(Y_1, Y_2, V)$. However, since we are defining efficiency in terms of the variance of the estimator of $\beta_x$, and (8) assumes a common effect of $X$ on either response, it may be more in the spirit of the balanced design to sample equally from the six strata defined by $(Y_1 + Y_2, V)$; therefore, we also consider this second balanced design when analysing the clustered data. Note that in our asymptotic calculations, these designs are based on expected phase-I stratum sizes, which come from having knowledge of the true parameters at the design stage.

### 3.4.3 Optimal Likelihood Sampling

If $\hat{\theta}$ is the estimator of $\theta = (\alpha', \beta')'$ which maximizes (2) and is estimated from data obtained with phase-II selection probabilities defined by $\delta$, then asymptotically

$$\sqrt{N}(\hat{\theta} - \theta) \sim N(0, \mathscr{I}_{\theta;\delta}^{-1} \Gamma_{\theta;\delta} \mathscr{I}_{\theta;\delta}^{-1}),$$

where $\mathscr{I}_{\theta;\delta} = E\left[ -\partial S_i(\theta)/\partial \theta' \right]$, $\Gamma_{\theta;\delta} = E\left[S_i(\theta) S_i'(\theta)\right]$ and $S_i(\theta)$ is the score function corresponding to the observed-data likelihood in (2). We say that this estimator has asymptotic variance $\mathscr{I}_{\theta;\delta}^{-1}$, since $\mathscr{I}_{\theta;\delta} = \Gamma_{\theta;\delta}$ (Cox and Hinkley 1974).

The expected information is affected by the choice of the phase-II selection parameter, $\delta$, so optimal maximum likelihood designs, $\pi(Y, V; \delta^{opt})$, can be found for any specified $\theta$. This is done here by numerically identifying the phase-II selection probabilities that minimize the asymptotic variance of the maximum likelihood estimator of $\beta_x$ subject to the budgetary constraints. The budget limits how many individuals can be sampled in the second phase; we set

$$P(R = 1) = \sum_{Y,V} \pi(Y, V; \delta) P(Y, V) = P_R \tag{9}$$

so that given the size of the phase-I sample, $N$, the expected phase-II sample size is fixed at $N \cdot P_R$, for some prespecified sampling fraction $P_R$. This sampling design will be optimally efficient for maximum likelihood estimation of $\beta_x$ whenever the covariate model and the parameters used in the design are correctly specified, as they are in the asymptotic calculations in the next section.

### 3.4.4 Optimal Mean Score Sampling

Reilly and Pepe (1995) show that the mean score estimator is asymptotically normal with an asymptotic variance that can be written as

$$\mathscr{I}_{\beta}^{-1} + \mathscr{I}_{\beta}^{-1} \Omega_{\beta;\delta} \mathscr{I}_{\beta}^{-1},$$

where

$$\mathscr{I}_\beta = E\big[-\partial U_\beta(Y|X,V)/\partial\beta'\big],$$

and

$$\Omega_{\beta;\delta} = \sum_{Y,V} P(Y,V)[\pi(Y,V;\delta)^{-1} - 1]\cdot var_{X|Y,V}[U_\beta(Y|X,V)],$$

with $var_{X|Y,V}[U_\beta(Y|X,V)]$ given by

$$\big\{E_{X|Y,V}\big[U_\beta(Y|X,V)U'_\beta(Y|X,V)\big] - E_{X|Y,V}\big[U_\beta(Y|X,V)\big]E_{X|Y,V}\big[U'_\beta(Y|X,V)\big]\big\}.$$

Therefore, the optimal second-phase selection probabilities, which give the greatest precision in estimating $\beta_x$ subject to the budgetary constraint (9), can be written as

$$\pi(Y,V;\delta^{opt}) = \frac{P_R\big\{\mathscr{I}_\beta^{-1} var_{X|Y,V}[U_\beta(Y|X,V)]\mathscr{I}_\beta^{-1}\big\}^{1/2}_{[k,k]}}{\sum_{Y,V} P(Y,V)\big\{\mathscr{I}_\beta^{-1} var_{X|Y,V}[U_\beta(Y|X,V)]\mathscr{I}_\beta^{-1}\big\}^{1/2}_{[k,k]}},$$

where $\{A\}_{[k,k]}$ refers to the entry of the asymptotic variance matrix corresponding to $\beta_x$ (Pepe et al. 1994; Reilly and Pepe 1995; Reilly 1996).

### 3.4.5 Optimal Weighted Pseudolikelihood Sampling

Asymptotically, the weighted pseudolikelihood estimator, $\tilde{\beta}$, is distributed as

$$\sqrt{N}(\tilde{\beta} - \beta) \sim N(0, \mathscr{I}_\beta^{-1}\Gamma_{\beta;\delta}\,\mathscr{I}_\beta^{-1}),$$

where

$$\Gamma_{\beta;\delta} = E\left[\frac{R_i}{\pi(Y,V;\delta)^2} U_\beta(Y|X,V)U'_\beta(Y|X,V)\right]$$

(Lawless et al. 1999; Robins et al. 1994). The asymptotic variance of the weighted pseudolikelihood estimator, $\mathscr{I}_\beta^{-1}\Gamma_{\beta;\delta}\,\mathscr{I}_\beta^{-1}$, can be written explicitly as a function of the selection probabilities by noting that $\mathscr{I}_\beta$ is functionally independent of $\delta$, and

$$\Gamma_{\beta;\delta} = \sum_{Y,V} P(Y,V)\pi(Y,V;\delta)^{-1} E_{X|Y,V}\big[U_\beta(Y|X,V)U'_\beta(Y|X,V)\big].$$

Therefore, as in Reilly and Pepe (1995), a Lagrange multiplier approach can be taken to minimize the asymptotic variance matrix entry corresponding to the

estimator of $\beta_x$ subject the budgetary constraint in (9). These optimal second-phase selection probabilities are

$$\pi(Y,V;\delta^{opt}) = \frac{P_R\left\{\mathscr{I}_\beta^{-1}E_{X|Y,V}\left[U_\beta(Y|X,V)U'_\beta(Y|X,V)\right]\mathscr{I}_\beta^{-1}\right\}^{1/2}_{[k,k]}}{\sum_{Y,V}P(Y,V)\left\{\mathscr{I}_\beta^{-1}E_{X|Y,V}\left[U_\beta(Y|X,V)U'_\beta(Y|X,V)\right]\mathscr{I}_\beta^{-1}\right\}^{1/2}_{[k,k]}},$$

where again $\{A\}_{[k,k]}$ refers to the entry of the asymptotic variance matrix corresponding to $\beta_x$.

### 3.5 *Asymptotic Relative Efficiencies*

In order to assess the efficiency gain that can result from exploiting available auxiliary data in the selection of individuals for measurement of expensive covariate information, balanced and optimal phase-II sampling designs were derived for a range of parameter values. The asymptotic efficiencies of the estimators resulting from these designs were calculated relative to the asymptotic efficiency of a simple random sampling design. We considered the three methods of analysis: maximizing the observed data likelihood (ML), the mean score method (MS), and maximizing the weighted pseudolikelihood (WP). For each of these methods of analysis we considered four designs: simple random sampling (SRS), balanced sampling over all eight strata defined by $(Y_1,Y_2,V)$ (BAL 8), balanced sampling over the six strata defined by $(Y_1+Y_2,V)$ (BAL 6), and the sampling design which is asymptotically optimal for precise estimation of $\beta_x$ with the given method of analysis (OPT).

The efficiency of each design $D$ was calculated relative to simple random sampling through

$$RE_x(D,A) = \frac{asvar_D(\hat{\beta}_x^A)}{asvar_{SRS}(\hat{\beta}_x^A)}, \tag{10}$$

where, for example, $asvar_{BAL8}(\hat{\beta}_x^{ML})$ represents the asymptotic variance of the estimator of $\beta_x$ that comes from using ML analysis with the BAL 8 design. We also consider the relative efficiency of the designs for estimating the effect of the auxiliary variable $\beta_v$ and $\psi$. Note that the "optimal" design will not necessarily be efficient for estimation of parameters other than $\beta_x$, although Reilly (1996) reported that in their examples optimal designs for one parameter "achieved an improvement in the precision of almost all parameters".

The asymptotic relative efficiencies of the different sampling designs is presented in Figs. 1, 2, and 3, for estimation of $\beta_x$, $\beta_v$, and $\psi$, respectively. The relative efficiencies are presented for a range of values of the association parameter $\psi$ while the other parameters were chosen so that $E[Y_1]=E[Y_2]=0.2; E[X]=0.25; E[R]=0.25; \beta_x=\log(1.25); \beta_v=\log(1.25); \alpha_v=\log(1.25)$. The fourth panel in each of these figures presents the asymptotic variance of the estimators that result from

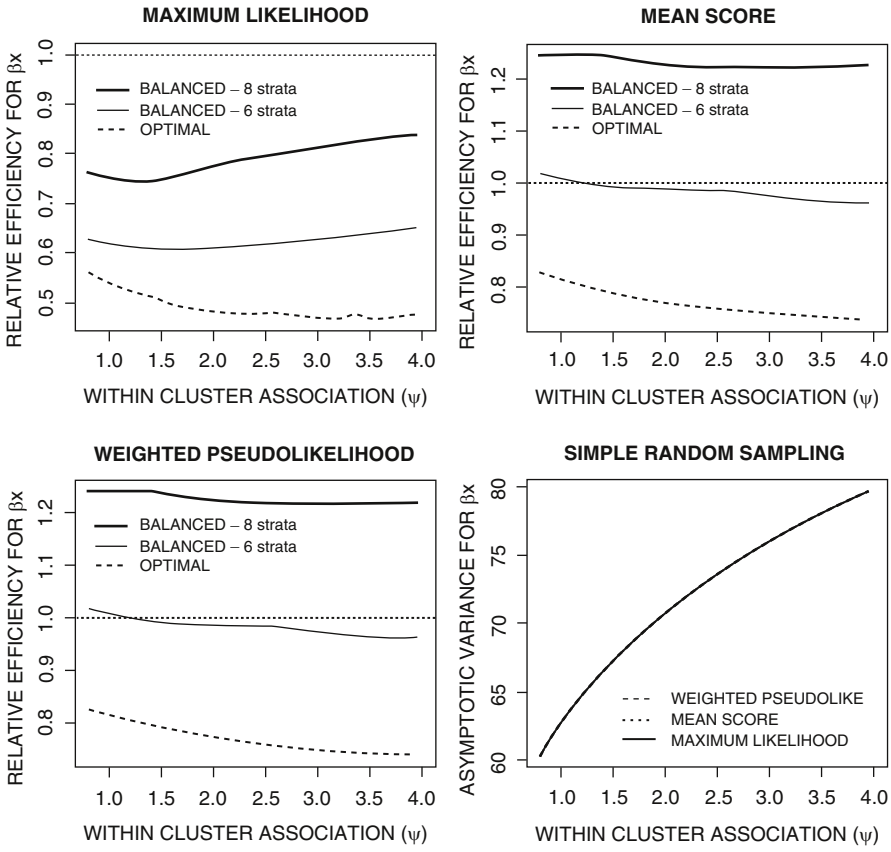


**Fig. 1** Asymptotic efficiency of estimators of $\beta_x$ under balanced and optimal designs relative to simple random sampling when using ML, MS, and WP. The asymptotic efficiencies are shown relative to the asymptotic variance of the SRS estimators which are shown in the fourth panel. $E[Y_1] = E[Y_2] = 0.2; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$

using SRS. Therefore, each of the first three panels within Figs. 1–3 leads to a comparison of efficiency amongst phase-II sampling designs for the specified method of analysis, while the fourth panel allows for a comparison of efficiency between methods of analysis.

It can be seen that the optimal design allows for a great increase in the efficiency of estimation of $\beta_x$; implementing an optimal phase-II sampling strategy can result in efficiency gains of 30–50 % over SRS, depending on the method of analysis (Fig. 1). In fact, for all methods of analysis, the optimal design results in more efficient estimators than SRS for all parameters, not just $\beta_x$ (Figs. 1–3). This is similar to that which was reported by Reilly (1996), where optimizing for efficient estimation of one parameter led to efficiency gains everywhere.

The balanced designs sometimes result in efficiency gains and sometimes result in a loss of efficiency. In the estimation of $\psi$ using WP analysis (Fig. 3, panel 3),

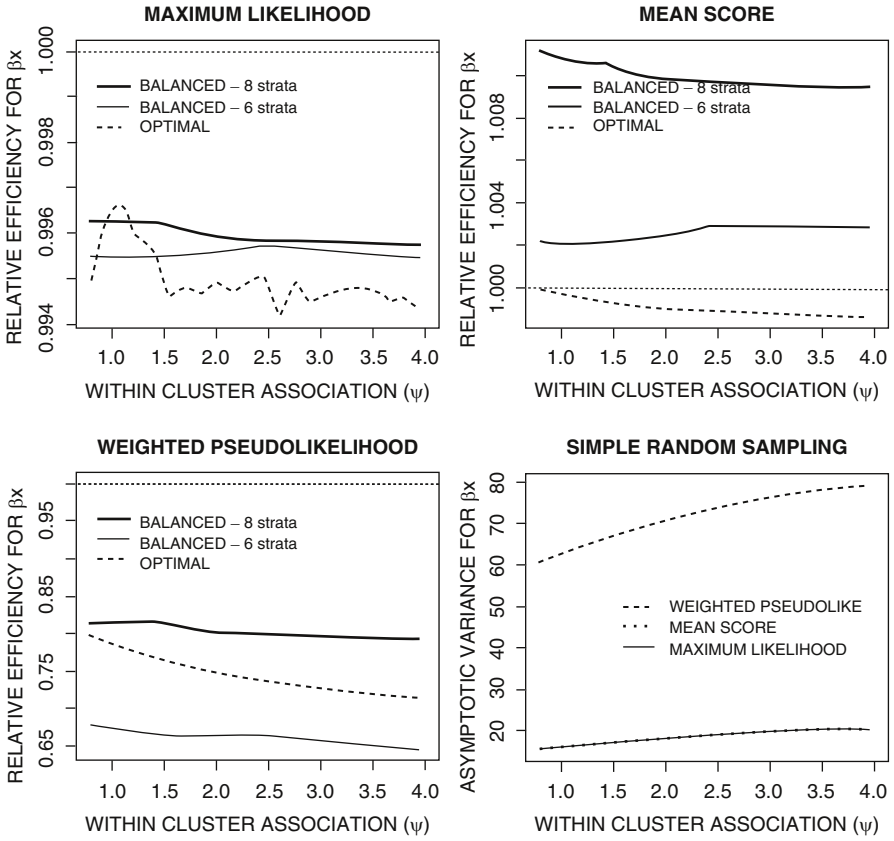


**Fig. 2** Asymptotic efficiency of estimators of $\beta_v$ under balanced and optimal designs relative to simple random sampling when using ML, MS, and WP. The asymptotic efficiencies are shown relative to the asymptotic variance of the SRS estimators which are shown in the fourth panel. $E[Y_1] = E[Y_2] = 0.2; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$

the balanced designs are both more efficient than the optimal design. However, for estimation of $\beta_x$ (Fig. 1), a balanced design can be seen to be much less efficient than the naive SRS for both MS and WP analysis. The BAL 6 design is generally more efficient than the BAL 8 design, but neither design is consistently more efficient than SRS.

Estimators of $\beta_x$ from SRS designs are very similar for ML, MS, and WP analysis (Fig. 1, panel 4), but use of WP is very inefficient for estimation of the other parameters (Figs. 2 and 3, panel 4).

It can also be seen that there is little difference amongst the designs for estimating $\psi$ or $\beta_v$ when using ML or MS (Figs. 2 and 3, panels 1 and 2). However, SRS is severely inefficient for estimating these parameters with WP analysis (Figs. 2 and 3, panel 3). So, the efficiency of estimators of $\psi$ and $\beta_v$ using WP analysis is greatly affected by the choice of sampling design, but even with the most efficient phase-II sampling design, WP estimators will still be less efficient than ML or MS estimators.

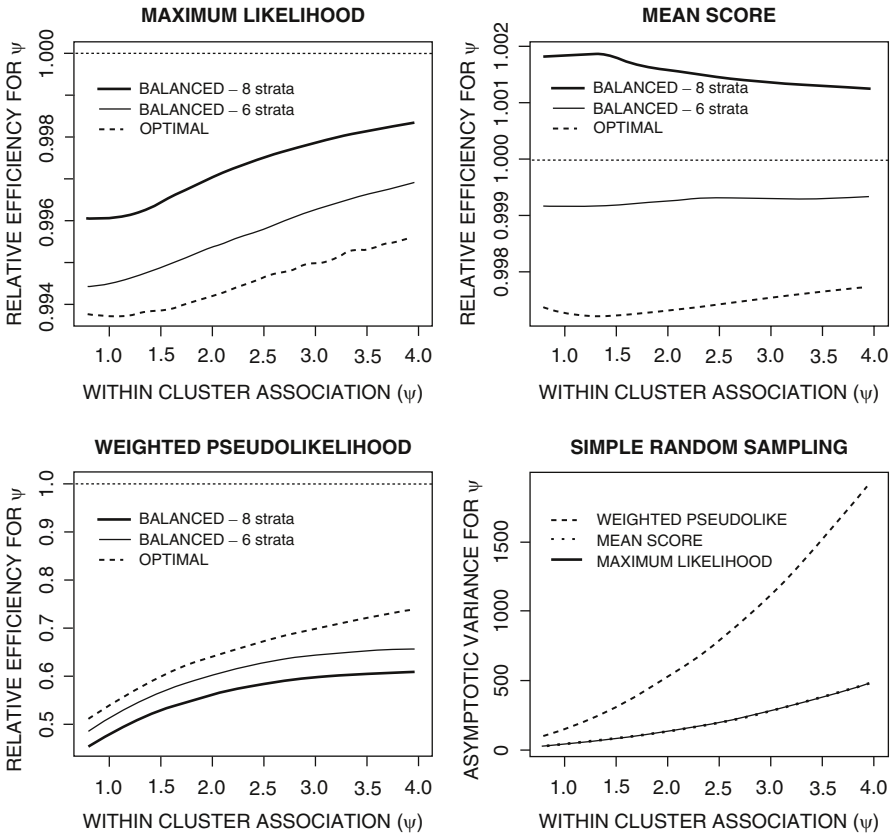


**Fig. 3** Asymptotic efficiency of estimators of $\psi$ under balanced and optimal designs relative to simple random sampling when using ML, MS, and WP. The asymptotic efficiencies are shown relative to the asymptotic variance of the SRS estimators which are shown in the fourth panel. $E[Y_1] = E[Y_2] = 0.2; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$

# 4 Response-dependent Sampling with Longitudinal Binary Data

## *4.1 The Response Model for Longitudinal Data*

Here we consider the analysis of binary data arising from a longitudinal study where the binary response variable is measured at baseline and at each of two prespecified follow-up timepoints. We assume that $Y_{i0} = 0$ and denote the response vector for individual $i$ as $Y_i = (Y_{i0}, Y_{i1}, Y_{i2})'$. We again consider binary covariates $X_i$ and $V_i$, where $V_i$ is known for all individuals at time 0, but $X_i$ will only be collected for individuals selected into a phase-II sample. This setting is a simplified version

of the kind of data collected in the CLSA when interest lies in estimating the effect of a risk factor for the onset of a disease. Here $Y_{ik}$ indicates the presence of disease at the assessment $k$, $k = 1,2$, and interest lies in the effect of covariates on $P(Y_{ik} = 1|Y_{i,k-1} = 0)$, the probability of disease developing between the $(k-1)$st and $k$th assessments, where we assume an irreversible disease process with $P(Y_{ik} = 0|Y_{i,k-1} = 1) = 0$. Specifically, here it is of interest to examine how the change in disease status (e.g. onset of diabetes) is affected by a time-invariant and expensive binary covariate $X_i$ (e.g. a genetic factor), after accounting for an available baseline auxiliary covariate $V_i$.

We again consider analyses through maximization of the observed data likelihood (ML), the mean score method (MS), and weighted pseudolikelihoods (WP). For these data, we are not interested in estimating marginal parameters as in (8), rather we are primarily interested in the transitional effect of the covariate $X$ in the response model

$$\text{logit}\, P(Y_{ik} = 1|Y_{i,k-1} = 0, X_i, V_i; \beta) = \beta_0 + \beta_1 I(k = 2) + \beta_x X_i + \beta_v V_i, \qquad k = 1,2.$$

Due to the irreversible nature of the disease process, the joint response model on which the likelihood methods are based is

$$\begin{aligned} P(Y_i|X_i, V_i; \beta) &= I(Y_{i,1} = 1)P(Y_{i,1} = 1|Y_{i,0} = 0, X_i, V_i; \beta)I(Y_{i,1} = 0)I(Y_{i,2} = 0) \\ &\quad \times [1 - P(Y_{i,1} = 1|Y_{i,0} = 0, X_i, V_i; \beta)][1 - P(Y_{i,2} = 1|Y_{i,1} = 0, X_i, V_i; \beta)] \\ &\quad \times I(Y_{i,1} = 0)I(Y_{i,2} = 1)[1 - P(Y_{i,1} = 1|Y_{i,0} = 0, X_i, V_i; \beta)] \\ &\quad \times P(Y_{i,2} = 1|Y_{i,1} = 0, X_i, V_i; \beta). \end{aligned}$$

## *4.2 The Selection Model*

Here we consider balanced and optimal designs for the selection of a phase-II sample at each of the three timepoints. This allows us to examine how the efficiency of designs is affected by the amount of auxiliary information available at phase-I for choosing the phase-II sample. Note that simple random sampling is not affected by the time at which the phase-II sample is chosen as this design does not exploit the data available at phase-I.

The selection model at time $t$ can be expressed as $P(R_i = 1|Y_{i1}, \ldots, Y_{it}, V_i; \delta^{(t)})$. At each progressive timepoint, more phase-I information is available for exploitation in deriving efficient phase-II selection probabilities. At timepoint 0, the phase-I sample can be divided into two strata based on the available information on $V$, so $\pi(Y, V; \delta^{(0)}) = \pi(V; \delta^{(0)})$; at timepoint 1, the phase-I sample can be stratified into four classes based on the available information on $V$ and $Y_1$, so $\pi(Y, V; \delta^{(0)}) = \pi(Y_1, V; \delta^{(0)})$; at timpoint 2, the phase-I sample can be stratified into six classes based on the available information on $V$, $Y_1$, and $Y_2$, where it is known that $P(Y_2 = 0|Y_1 = 1) = 0$.

Simple random sampling is the same at each timepoint, but the efficiency of the balanced and optimal designs will be affected by the amount of information available at phase-I. Therefore, for this study of transitional effects, we consider 7 designs for each method of analysis: simple random sampling (SRS), balanced sampling using the phase-I data available at each timepoint (call these BAL 0, BAL 1, and BAL 2 at timepoints 0, 1, and 2, respectively) and the sampling designs which are optimal for estimating $\beta_x$ given the specified method of analysis and the data that are available at the time of selection (call these OPT 0, OPT 1, and OPT 2). We will again present the efficiencies of the designs relative to simple random sampling, as calculated in (10).

The asymptotic variances and optimal designs can be found as in the previous section; however, summations are no longer over strata defined by $(Y_1, Y_2, V)$, but rather over strata defined by the data that are available at the time of selection. This decrease in phase-I data essentially places added constraints on the optimal sampling designs derived in the previous section; for example, at timepoint 0, when only $V$ is available for phase-II sampling decisions, then $\pi(Y, V; \delta) = \pi(V; \delta)$ for all $Y = (Y_1, Y_2) \in \{(0,0), (0,1), (1,1)\}$.

### *4.3 Asymptotic Relative Efficiencies*

We derived optimal designs for a range of values of $P_R$, which defines the budgetary constraint in (9). Other parameters were chosen so that $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$. The relative efficiencies of the different sampling designs is presented in Figs. 4, 5, and 6, for ML analysis, MS analysis, and WP analysis, respectively. We consider the relative efficiency of each of the considered designs for estimating $\beta_0$, $\beta_1$, $\beta_x$, and $\beta_v$.

As expected, the optimal sampling design offered large efficiency gains over simple random and balanced designs when estimating $\beta_x$. As before, these optimal designs also added efficiency to the estimation of other parameters (Figs. 4, 5, and 6). Having more information at the time of sampling increased the efficiency of the optimal design for the estimation of all parameters. However BAL 2, the balanced design at timepoint 2, was generally less efficient than BAL 1, the balanced design which was based only on the auxiliary information available at timepoint 1. This indicates that, as was seen in the comparison of BAL 6 and BAL 8 in the previous section, having more phase-I information does not necessarily improve the efficiency of balanced designs.

The asymptotic variance of the ML and MS estimators under SRS was very similar; however, the optimal design offered a greater increase in efficiency for the ML estimator of $\beta_x$ than for the MS estimator (Figs. 4 and 5, panel 3). The balanced designs were often less efficient than the naive simple random sampling approach to gathering data for estimation of $\beta_x$ (Figs. 4, 5, and 6). The use of a balanced design appears to be particularly inefficient when analysis is to be carried out through the

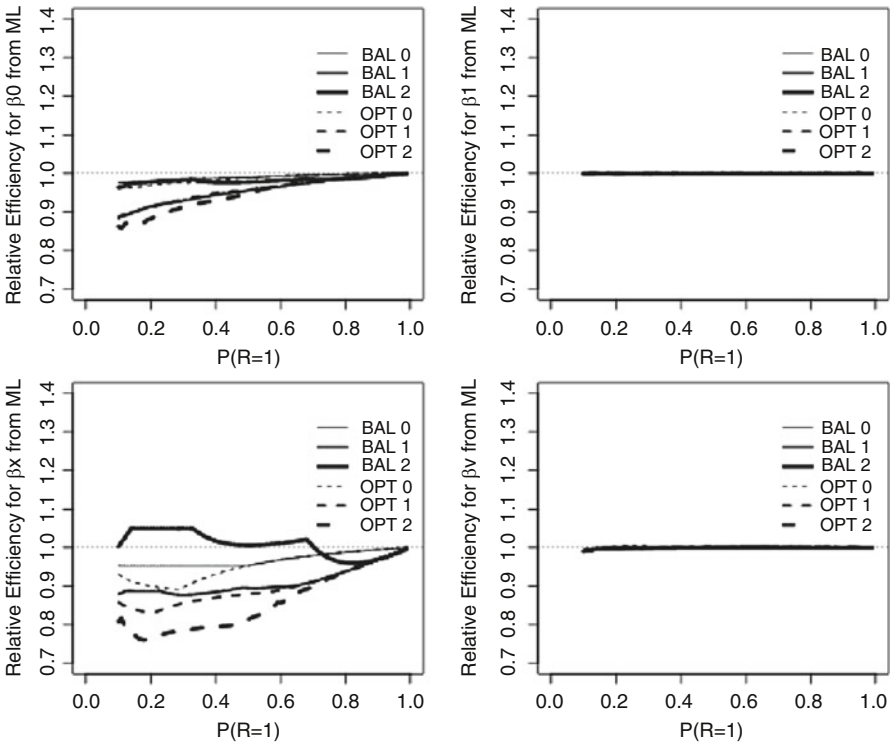


**Fig. 4** Asymptotic efficiency of estimators under balanced and optimal designs relative to simple random sampling when using maximum likelihood analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[V] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$ Note: the asymptotic variance of the ML estimators of $\beta_0$, $\beta_1$, $\beta_x$, and $\beta_v$ under SRS with $P(R = 1) = 0.5$ are, respectively, 9.86, 12.99, 33.00, and 16.58

mean score method or the weighted pseudolikelihood (Figs. 5 and 6). Note that as the sampling fraction increases, smaller strata are selected in their entirety by the balanced designs (the selection probabilities must be capped at 1, as discussed previously); this accounts of the lack of smoothness in the change in asymptotic efficiency of the balanced designs. Some lack of smoothness can also be seen in the plot of the optimal ML designs; this occurs because these optimal ML designs are found numerically.

## 5 Discussion

To our knowledge this article was among the first to study the two-phase sampling designs involving clustered or longitudinal data. Given the increased interest in studies involving cross-sectionally clustered data and the recent trend towards the

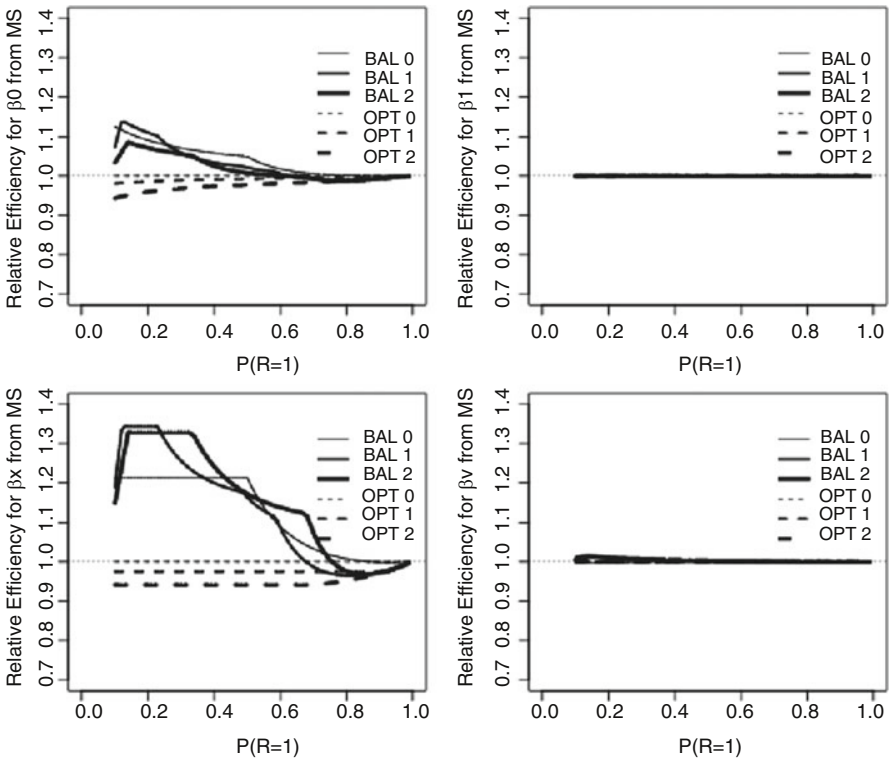


**Fig. 5** Asymptotic efficiency of estimators under balanced and optimal designs relative to simple random sampling when using the mean score method for analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[V] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$ Note: the asymptotic variance of the MS estimators of $\beta_0$, $\beta_1$, $\beta_x$, and $\beta_v$ under SRS with $P(R = 1) = 0.5$ are, respectively, $9.86, 12.99, 33.00$, and $16.58$

design of massive cohort studies of health and disease, the insights that result from this work are important.

For the setting of clustered data, the first decision to make is typically on the method of analysis and there are a variety of frameworks one can adopt. We restricted attention to bivariate response data and marginal models for characterizing the effects of exposure. In this setting, maximum likelihood and the mean-score methods can be more efficient than weighted pseudolikelihood for estimation of the exposure effect (with maximum likelihood generally being the superior of the two) but this comes at the expense of making assumptions and modelling the covariate distribution. Interestingly, the three analysis methods have approximately the same efficiency when using simple random sampling. When covariates change within clusters an alternative model formulation could be based on random effects models. To our knowledge, there has been no work on two-phase designs within the framework of random effect models for clustered data and this is an area of current interest.

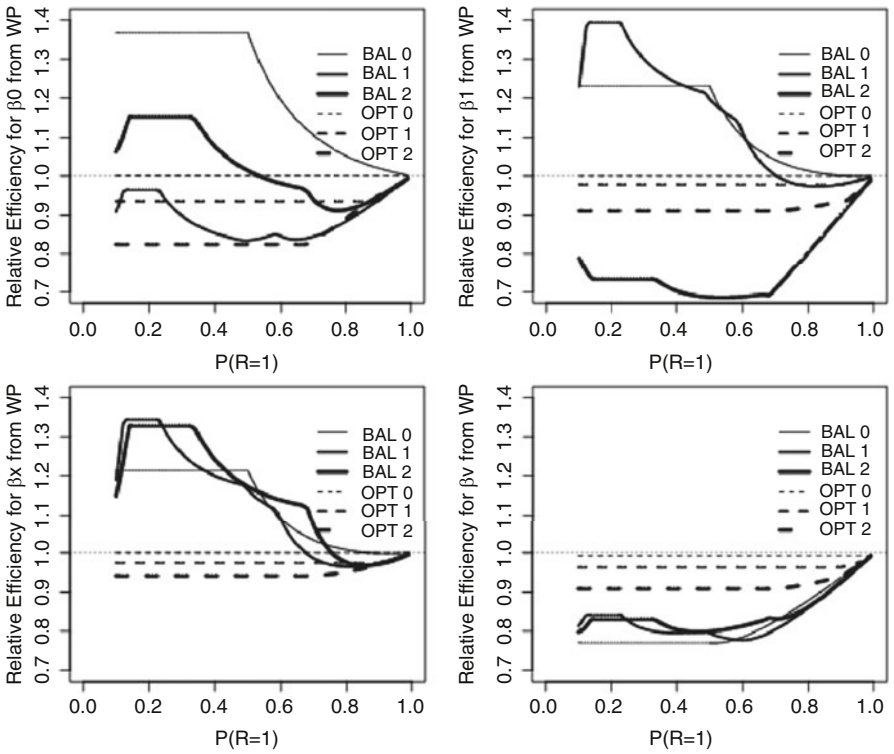


**Fig. 6** Asymptotic efficiency of estimators under balanced and optimal designs relative to simple random sampling when using a weighted pseudolikelihood analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[V] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$ Note: the asymptotic variance of the WP estimators of $\beta_0$, $\beta_1$, $\beta_x$, and $\beta_v$ under SRS with $P(R = 1) = 0.5$ are, respectively, 17.37, 25.97, 33.00, and 33.00

We have adopted a very simple response model with a binary $X$ and binary auxiliary variable. When the exposure variable is continuous, a robust implementation of the mean score method may be more appealing, and weighted pseudolikelihood would also have more appeal since no modelling of exposure is required. In ongoing work (not reported here) we found that optimal designs based on maximum likelihood analyses may be more sensitive to small changes in the parameters used at the design stage than optimal mean score designs. So, if models for exposure variable are difficult to formulate with confidence, the robustness of the mean score and weighted pseudolikelihood approaches may be more appealing. When the auxiliary variable is continuous, discretizing seems the most practical approach to addressing the curse of dimensionality and this has been recommended by several authors (Lawless et al. 1999).

When comparing the effect of different frameworks for analysis and design, it is interesting to note that the conclusions about optimality bear only on the criteria

adopted for the optimal design. The intercept, effect of the auxiliary variable and association parameters do not necessarily behave in the same way.

The pragmatic approach of using balanced sampling designs as a compromise between robustness and efficiency does not yield clear and consistent recommendations; the resulting estimators sometimes perform well and sometimes perform poorly. It is therefore unclear what auxiliary information should be considered when implementing a balanced design in the more complex settings we consider here.

There are several directions of future research that are natural to consider. We focus on clusters of size two because of interest in the two sacroilliac joints among patients with psoriatic arthritis. However, clusters can naturally be much larger as would be the case if all joints were to be modelled. Dealing with larger cluster sizes is in principle straightforward but may suggest the use of second-order generalized estimating functions rather than likelihood analyses. One may elect to retain the robustness of a first order analysis by refraining from higher order assumptions, or invoke fourth moment assumptions to try to optimize efficiency at the expense of robustness in the estimating equation framework.

We have also restricted attention to a first order Markov model in the longitudinal context with only three assessments. Longer term follow-up, as is planned for the Canadian Longitudinal Study in Aging (Raina et al. 2009), raises questions about the need for more elaborate response models, the need for greater collapsing of strata, and issues surrounding time-varying covariates. These and other issues are subject to further research.

**Acknowledgments** Michael McIsaac's research was supported by an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Discovery Grants to Richard Cook from NSERC (RGPIN 155849) and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Canada Research Chair in Statistical Methods for Health Research. The authors thank Dr. Dafna Gladman and Dr. Vinod Chandran for collaboration and helpful discussions regarding the research at the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto. The authors gratefully acknowledge the careful review and comments from a referee and Dr. Brajendra Sutradhar.

## References

Breslow, N.E., Cain, K.C.: Logistic regression for two-stage case-control data. Biometrika. **75**(1), 11–20 (1988)

Breslow, N.E., Chatterjee, N.: Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. Appl. Stat. **48**(4), 457–468 (1999)

Chandran, V., Tolusso, D.C., Cook, R.J., Gladman, D.D.: Risk factors for axial inflammatory arthritis in patients with psoriatic arthritis. J. Rheumatol. **37**(4), 809–815 (2010)

Chatterjee, N., Chen, Y., Breslow, N.E.: A pseudoscore estimator for regression problems with two-phase sampling. J. Am. Stat. Assoc. **98**(461), 158–168 (2003)

Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman & Hall, London (1974)

del Rincon, I., Williams, K., Stern, M.P., Freeman, G.L., O'Leary, D.H., Escalante, A.: Association between carotid atherosclerosis and markers of inflammation in rheumatoid arthritis patients and healthy subjects. Arthritis Rheum. **48**(7), 1833–1840 (2003)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**(1), 1–38 (1977)
Heagerty, P.J., Zeger, S.L.: Marginalized multilevel models and likelihood inference. Stat. Sci. **15**, 1–26 (2000)
Heagerty, P.J.: Marginalized transition models and likeliood inference for longitudinal categorical data. Biometrics **58**(2), 342–351 (2002).
Horton, N.J., Laird, N.M.: Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. Biometrics **57**, 34–42 (2001).
Lawless, J.F., Kalbfleisch, J.D., Wild, C.J.: Semiparametric methods for response-selective and missing data problems in eegression. J. Roy. Stat. Soc. B **61**(2), 413–438 (1999)
Laird, N., Ware, J.H.: Random-effects models for longitudinal data. Biometrics **38**(4), 963–974 (1982)
Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**(1), 13–22 (1986)
Lipsitz, S.R., Laird, N.M., Harrington, D.P.: Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika **78**(1), 153–160 (1991)
Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data, 2nd edn. Wiley, New York (2002)
Neuhaus, J.M.: Statistical methods for longitudinal and clustered designs with binary responses. Stat. Meth. Med. Res. **1**, 249–273 (1992)
Pepe, M.S., Reilly, M., Fleming, T.R.: Auxiliary outcome data and the mean-score method. J. Stat. Plann. Infer. **42**, 137–160 (1994)
Pickles, A., Dunn, G., Vazquez-Barquero, J.L.: Screening for stratification in two-phase ("two-stage") epidemiological surveys. Stat. Meth. Med. Res. **4**, 73–89 (1995)
Prentice, R.L.: Correlated binary regression with covariates specific to each binary observation. Biometrics **44**(4), 1033–1048 (1988)
Rahman, P., Gladman, D.D., Cook, R.J., Zhou, Y., Young, G., Salonen, D.: Radiological assessment in psoriatic arthritis. Rheumatology **37**(7), 760–765 (1998)
Raina, P.S, Wolfson, C., Kirkland, S.A., Griffith, L.E., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C.M., Hogan, D., Wister, A., Payette, H., Shannon, H., Brazil, K.: The Canadian longitudinal study on aging (CLSA). Can. J. Aging **28**(3), 221–229 (2009)
Reilly, M.: Optimal sampling strategies for two phase studies. Am. J. Epidemiol. **143**, 92–100 (1996)
Reilly, M., Pepe, M.S.: A mean score method for missing and auxiliary covariate data in regression models. Biometrika **82**(2), 299–314 (1995)
Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**(427), 846–866 (1994)
Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of Missing Data. J. Am. Stat. Assoc. **90**(429), 106–121 (1995)
Stiratelli, R., Laird, N., Ware, J.H.: Random-effects models for serial observations with binary response. Biometrics **40**(4), 961–971 (1984)
Sutradhar, B.C.: On auto-regression type dynamic mixed models for binary panel data. Metron **66**(2), 209–221 (2008)
Sutradhar, R., Cook, R.J.: A bivariate mover-stayer model for interval-censored recurrent event data: application to joint damage in rheumatology. Comm. Stat. Theor. Meth. **18**, 3389–3405 (2009)
Tolusso, D.C., Cook, R.J.: Robust estimation of state occupancy probabilities for interval-censored multistate data: an application involving spondylitis in psoriatic arthritis. Comm. Stat. Theor. Meth. **38**(18), 3307–3325 (2009)
Troxel, A.B., Lipsitz, S.R., Brennan, T.A.: Weighted estimating equations with nonignorable nonresponse data. Biometrics **53**(3), 857–869 (1997)
Tsiatis, A.A.: Semiparametric Theory and Missing Data. Springer, New York (2006)
Whittemore, A.S., Halpern, J.: Multi-stage sampling in genetic epidemiology. Stat. Med. **16**, 153–167 (1997)

Zeng, L., Cook, R.J.: Transition models for multivariate longitudinal binary data. J. Am. Stat. Assoc. **102**, 211–223 (2007)

Zhao, L.P., Prentice, R.L.: Correlated binary regression using a quadratic exponential model. Biometrika **77**(3), 642–648 (1990)

Zhao, Y.: Design and efficient estimation in regression analysis with missing data in two-phase studies. PhD thesis, University of Waterloo (2005)

Zhao, Y., Lawless, J.F., McLeish, D.L.: Likelihood methods for pegression models with expensive variables missing by design. Biom. J. **51**(1), 123–136 (2009)

# Part III
# Longitudinal Data Analysis Subject to Outliers

# Robust Inference Progress from Independent to Longitudinal Setup

**Brajendra C. Sutradhar**

**Abstract** In practice one may not have always smooth data. When bulk of the data are smooth but the complete data set apparently contains a few contaminated observations or outliers, one encounters difficulties to choose an inference technique because of the fact that the traditional inference techniques developed for smooth data analysis may no longer provide unbiased and consistent estimates for the desired parameters such as regression parameters in linear or generalized linear models (GLMs) setup. In this paper, we first briefly review some of the widely used bias corrected techniques in linear model setup. But, as opposed to the linear models in normal or other continuous exponential family based variables, the robust inference for discrete data in the GLMs setup, such as for count and binary data, is, however, not adequately discussed in the literature. The advantages and drawbacks of an existing outliers resistant Mallow's type quasi-likelihood (MQL) estimation approach in GLMs setup are reviewed in brief. We then discuss a recently proposed fully standardized MQL (FSMQL) approach that provides almost unbiased estimates ensuring its higher consistency performance. One encounters further challenges when the data in GLMs setup are repeatedly collected over a period of time. This is mainly because one then requires to modify the FSMQL type estimation approaches such that the modified approach also accommodates the correlation structure of the repeated data. A recently proposed robust generalized QL (RGQL) approach is reviewed for the purpose.

## 1 Introduction

In a regression setup, the responses whether linear, count, or binary, are generated as a function of certain suitable covariates. If bulk of the responses appear to be close to the mean function of the responses with a few remaining responses appearing

B.C. Sutradhar (✉)
Memorial University, St. John's, NF, Canada, A1C5S7
e-mail: bsutradh@mun.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_9, 

at a significant distance from the mean function, then these latter few responses are considered to be potential outliers. In general these outliers occur because of the corresponding covariates which may be contaminated in some ways, and they are referred to as the mean shifted outliers. In some situations, a response may be considered as an outlier because of its inflated variance as compared to the bulk of the responses. It is of main interest to understand the regression model appropriate for bulk of the good responses. But the use of few outlying responses may distort the inference for the bulk of the responses. There are at least two ways this inference problem has been tackled in the literature.

First, it is attempted to detect the outliers and exclude them for the overall inference. For some justifications on this, one may be referred to Hampel et al. (1986, Sect. 1.4) among others. For the purpose, many researchers have discussed the so-called maximum studentized residual (MSR) and maximum normed residual (MNR) tests for detection of outliers in a linear regression setup for independent data. For example, one may refer to the work of Srikantan (1961), Stefansky (1971, 1972), Tietjen et al. (1973), Prescott (1975), Lund (1975), Bailey (1977), Johnson and Prescott (1975), Ellenberg (1973, 1976), Cook and Prescott (1981), Doornbos (1981), and Beckman and Cook (1983, Sect. 4), among others. The powers of these two statistics in detecting outliers may also be affected by the ways the parameters of the regression models are estimated. For a discussion on this, see, for example, a relatively recent work by Sutradhar et al. (2007). In second approach, a robust weighted distance function is constructed such that the suspected outliers get smaller weights. Next the distance function is minimized for the estimation of the regression effects. Some of the existing widely used robust procedures are: Minimax estimation, M-estimation, L-estimation, and R-estimation. For details on these procedures, see, for example, Hampel et al. (1986), Rousseeuw and Leroy (1987), and Huber (2004), and the references therein.

In the independent setup, some authors such as Cantoni and Ronchetti (2001), among others, have suggested a Mallow's type quasi-likelihood (MQL) robust estimation approach to obtain a consistent estimate for the regression effects involved in the model. For the MQL construction, they use the Huber's robust function but did not use the inverse of the variance of such a function to make the MQL standardized. Recently, Bari and Sutradhar (2010a) have improved this estimating equation and introduced a fully standardized MQL (FSMQL) estimating equation that provides regression estimates with smaller bias. In this paper, we review these MQL and FSMQL approaches for the estimation of the regression effects involved in generalized linear models (GLMs), for example for binary and count data.

Also, there have been some studies using QL or generalized estimating equations (GEE) approaches for robust regression estimation in the longitudinal setup. For example, Preisser and Qaqish (1999) have used a resistant GEE (REGEE) approach, which was improved by Cantoni (2004) (see also Sinha 2006 for a random effects approach) by using a semi-standardized MQL (SSMQL; see also Bari and Sutradhar 2010b) approach. In the second part of the paper, we review these approaches including the robust GQL (RGQL) approach discussed by Bari and Sutradhar

(2010b) and point out their advantages and drawbacks. Both count and binary longitudinal models are considered.

## 2 Robust Inference in Regression Models in Independent Setup

### *2.1 Inference for Linear Models*

There exists a vast literature for robust inference in linear models for independent data in the presence of one or more outliers. See, for example, Rousseeuw and Leroy (1987), Huber (2004, Chap. 7), and a relatively recent paper by Sutradhar et al. (2007). These studies mainly deal with outliers in normal responses. For simplicity consider a simple linear regression model

$$y = X\beta + \varepsilon, \tag{1}$$

where $y = (y_1, \ldots, y_i, \ldots, y_K)'$ is a $K \times 1$ response vector, $X$ is known design matrix of order $K \times p$, $\beta$ is a $p \times 1$ vector of unknown parameters, and $\varepsilon$ is an $K \times 1$ error variable distributed as $\varepsilon \sim N(0, \sigma^2 I_K)$, $I_K$ being the $K \times n$ identity matrix. Usually, each observation in a realization $(y, X)$ contributes to the evaluation of the regression coefficient $\beta$. The contribution of one observation, however, may be discordant to the point of sensibly determining the value of a regression parameter. Such an observation is said to be an outlier. To see how an outlier can perturb the linear model (1), two types of outliers are generally considered. They are (a) mean shifted outliers, also referred to as the additive outliers, and (b) variance inflated outliers, also referred to as the innovative or multiplicative outliers.

To construct an additive outlier model, one can perturb the linear model (1) and write

$$y = X\beta + \tilde{\varepsilon}, \tag{2}$$

where $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_i, \ldots, \tilde{\varepsilon}_K)'$ is related to $\varepsilon$ in (1) as

$$\tilde{\varepsilon}_j = \begin{cases} \varepsilon_j + \delta_1, & \text{for } j = i \\ \varepsilon_j, & \text{for } j \neq i, \end{cases} \tag{3}$$

where for $|\delta_1| > 0$, $y_i = x_i'\beta + \tilde{\varepsilon}_i$ is certainly a discordant observation when compared to the other $K - 1$ observations. It is clear from (1) and (3) that

$$\tilde{\varepsilon} \sim N(\delta, \sigma^2 I_K),$$

where

$$\delta = [01'_{i-1}, \delta_1, 01'_{K-i}]'.$$

To construct a variance inflated outlier model, one can perturb the model (1) as

$$y = X\beta + \varepsilon^*, \tag{4}$$

where $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_i^*, \ldots, \varepsilon_K^*)'$ is related to $\varepsilon$ in (1) as

$$\varepsilon_j^* = \begin{cases} \varepsilon_j/\sqrt{\omega}, & \text{for } j = i \\ \varepsilon_j, & \text{for } j \neq i, \end{cases} \tag{5}$$

where for $\omega \to 0$, the $i$th observation $y_i$ will have large variance leading this observation to be an outlier. It is clear from (1) and (5) that

$$\varepsilon^* \sim N(0, V_\omega = \sigma^2 \text{diag}[1'_{i-1}, 1/\omega, 1'_{K-i}]).$$

Thus, under model (2), bulk $(K-1)$ of the error variables follow $N(0, \sigma^2)$ distribution and 1 follows $N(\delta_1, \sigma^2)$. This is equivalent to say that the $\tilde{\varepsilon}_i$ in model (2) are independent, identically distributed with the common underlying distribution

$$F(\tilde{\varepsilon}) = (1 - \frac{1}{K})\Phi\left(\frac{\tilde{\varepsilon} - 0}{\sigma}\right) + \frac{1}{K}\Phi\left(\frac{\tilde{\varepsilon} - \delta_1}{\sigma}\right),$$

(Huber 2004, Example 1.1) where $\Phi(\cdot)$ is the standard normal cumulative. Similarly, one may say that $\varepsilon_i^*$ under model (4) are independent, identically distributed with common underlying distribution

$$F(\varepsilon^*) = (1 - \frac{1}{K})\Phi\left(\frac{\varepsilon^* - 0}{\sigma}\right) + \frac{1}{K}\Phi\left(\frac{\varepsilon^* - 0}{\sigma/\sqrt{\omega}}\right).$$

### 2.1.1 Robust Estimation of Regression Effects

It is understandable that the ordinary least square (LS) estimator

$$\hat{\beta}_{LS} = [X'X]^{-1}X'y \tag{6}$$

is biased for $\beta$ under model (2)–(3) and will be unbiased but inefficient under model (4)–(5). There exist various robust approaches for the consistent estimation of $\beta$ irrespective of the underlying model whether it is (2)–(3) or (4)–(5). Here we briefly describe two of the approaches, for example.

Huber's Robust Weights Based Iterative Re-weighted Least Square Approach

This estimate is obtained via an iterative re-weighted least squares (RWLS) method (Street et al. 1988). For $p$ components of $\beta$, in this approach one solves the robust weights based estimating equation

$$\sum_{j=1}^{K} \xi_j x_{ju}(y_j - x_j'\beta) = 0,\ u = 1\ldots, p, \tag{7}$$

where $x_{ju}$ is the $u$th component of the $x_j$ vector, and

$$\xi_j = \frac{\psi(r_j)}{r_j}, \tag{8}$$

with $\psi(r_j)$ as the Huber's bounded function of $r_j$ given by

$$\psi(z) = \max\left[-a, \min(z, a)\right],\ \text{with}\ a = 1.25,$$

where $r_j = (y_j - x_j'\beta^*_{r(0)})/\tilde{s}$ for $j = 1,\ldots,n$, with $\beta^*_{r(0)}$ as an initial robust estimate of $\beta$ which may be obtained by minimizing the $L_1$ distance $\sum_{j=1}^{K} |y_j - x_j'\beta|$, and $\tilde{s}$ as a robust estimate of $\sigma$ given by

$$\tilde{s} = \text{Median}\left\{\text{largest K-p+1 of the } \frac{|y_j - x_j'\beta^*_{r(0)}|}{0.6745}\right\}.$$

Note that if $r_j = 0$, one uses $\xi_j = 1$. The solution to (7) may then be obtained as

$$\beta^*_{r(1)} = (X'\Omega X)^{-1}X'\Omega y, \tag{9}$$

where $\Omega = \text{diag}[\xi_1,\ldots,\xi_K]$. This $\beta^*_{r(1)}$ replaces $\beta^*_{r(0)}$ and provides us with a new start and new weights for an improved estimate of $\beta$ to be obtained by (9). This cycle of iterations continues until convergence. Let the final solution be denoted by $\hat{\beta}_{r(1)}$.

An Alternative Weights Based Iterative RWLS Approach

Rousseeuw and Leroy (1987, Chap. 5) suggest a least median of squares (LSM) approach where the scale parameter to compute the residual is estimated using robust weights different than Huber's weights used in the last section. In fact one can use the iterative least square approach discussed in the last section by replacing Huber's weights with these new weights suggested by Rousseeuw and Leroy (1987, p. 202). See, for example, Sutradhar et al. (2007) for a comparison

between RWLS approaches using Huber's and Rousseeuw and Leroy weights. To be specific, Rousseeuw and Leroy robust weights are defined as

$$\tilde{w}_j = \begin{cases} 1, \text{ if } |d_{j(\beta^*_{r(0)})}/\tilde{s}_0| \leq 2.5 \\ 0, \text{ otherwise,} \end{cases} \tag{10}$$

where $d_{j(\beta^*_{r(0)})} = y_j - x_j'\beta^*_{r(0)}$ and $\tilde{s}_0$ is given by

$$\tilde{s}_0 = 1.4826(1+5/(K-p))\sqrt{\text{Median } d^2_{j(\beta^*_{r(0)})}}\,.$$

These robust weights in (10) are then used to compute an $\tilde{\Omega}$ matrix as

$$\tilde{\Omega} = diag[\tilde{w}_1, \ldots, \tilde{w}_j, \ldots, \tilde{w}_K],$$

which is then used to obtain a first step improved robust estimate for $\beta$ as

$$\beta^{**}_{r(1)} = (X'\tilde{\Omega}X)^{-1}X'\tilde{\Omega}y. \tag{11}$$

The cycle of iterations continues until convergence. Let this final RWLS estimate be denoted by $\hat{\beta}_{r(2)}$.

### 2.1.2 Robust Estimation of Variance Component

Note that in the linear model setup, the LS estimate of $\sigma^2$ is obtained by computing the residual sum of squares based on the least square estimate of $\beta$. That is, $\hat{\sigma}^2_{ls} = \sum_{j=1}^K (y_j - x_j'\hat{\beta}_{ls})^2/(K-p)$. Under the linear model in the presence of outliers, one may obtain LS estimate of $\sigma^2$ simply by replacing $\hat{\beta}_{ls}$ with $\hat{\beta}_{r(1)}$ or $\hat{\beta}_{r(2)}$ obtained in the last section. Thus the LS estimator for $\sigma^2$ has the formula

$$\tilde{\sigma}^2_{ls(1)} = \sum_{j=1}^K (y_j - x_j'\hat{\beta}_{r(1)})^2/(K-p), \tag{12}$$

or

$$\tilde{\sigma}^2_{ls(2)} = \sum_{j=1}^K (y_j - x_j'\hat{\beta}_{r(2)})^2/(K-p). \tag{13}$$

Huber's Robust Weights Based Iterative RWLS Estimator for $\sigma^2$

Following Street et al. (1988), one obtains this estimator as

$$\hat{\sigma}^2_{r(1)} = \sum_{j=1}^{K} \xi_j \left[ y_j - x_j' \hat{\beta}_{ls} \right]^2 \Big/ \Big( \sum_{j=1}^{K} \xi_j - tr\{(X'\Omega^2 X)(X'\Omega X)^{-1}\} \Big), \qquad (14)$$

where $\xi_j$ $(j = 1, \ldots, K)$ is the $j$th robust weight to protect the estimate against possible outliers, and $\Omega = diag(\xi_1, \ldots, \xi_j, \ldots, \xi_K)$. To be specific, $\xi_j$ is defined as $\xi_j = \psi(r_j)/r_j$ with $r_j = (y_j - x'\hat{\beta}_{ls})/s^*$, where

$$s^* = \text{Median} \left\{ \text{largest K-p+1 of the } \frac{|y_j - x_j'\hat{\beta}_{ls}|}{0.6745} \right\}.$$

Note that the $\psi$ function involved in $\xi_j$ in (14) is the same Huber's robust function used in (8).

Rosseeuw and Leroy Weights Based Robust Estimator for $\sigma^2$

This robust estimator is computed following Rousseeuw and Leroy (1987, p. 202, Eq. (1.5)). More specifically, in this approach, robust weights are defined as

$$w_j = \begin{cases} 1, \text{ if } |d_{j(\hat{\beta}_{ls})}/s_0| \leq 2.5 \\ 0, \text{ otherwise,} \end{cases}$$

where $d_{j(\hat{\beta}_{ls})} = y_j - x_j'\hat{\beta}_{ls}$ and $s_0$ is given by

$$s_0 = 1.4826(1 + 5/(K-p)) \sqrt{\text{Median } d^2_{j(\hat{\beta}_{ls})}}.$$

Next, these weights are exploited to compute the estimator, say $\hat{\sigma}^2_{r(2)}$, as

$$\hat{\sigma}^2_{r(2)} = \Big( \sum_{j=1}^{K} w_j d^2_{j(\hat{\beta}_{ls})} \Big) \Big/ \Big( \sum_{j=1}^{K} w_j - p \Big). \qquad (15)$$

### 2.1.3 Finite Sample Performance of the Robust Estimators: An Illustration

Sutradhar et al. (2007) conducted a simulation study to examine the performance of the robust methods as compared to the LS method in estimating the parameters in a linear model when the data contain a few variance inflated outliers. Here, we refer to some of the results of this study, for example. Consider a linear model with $p = 2$ covariates so that $\beta = (\beta_1, \beta_2)'$. For the associated $K \times 2$ design matrix $X$, consider their design configuration:

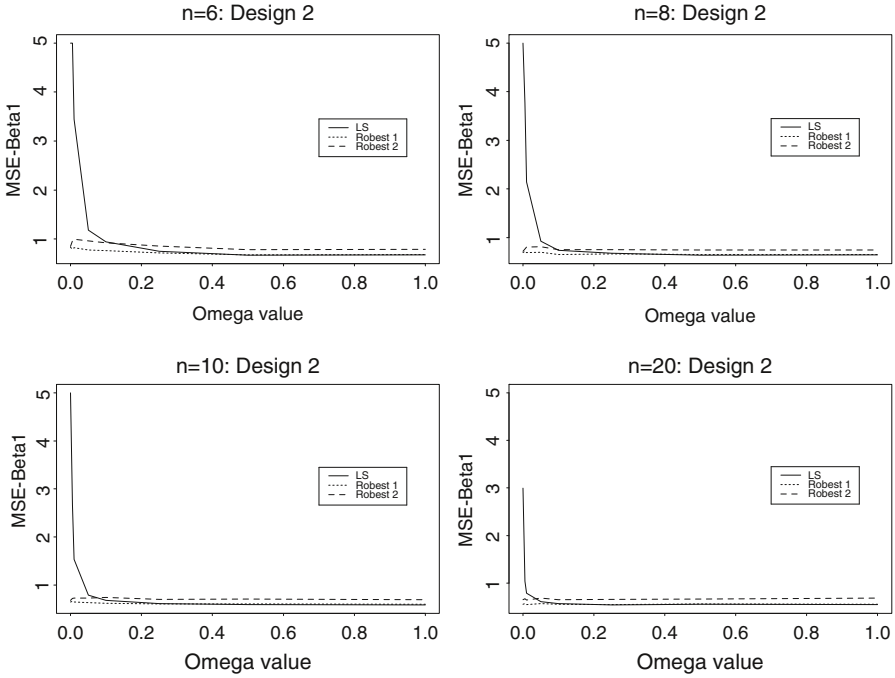


**Fig. 1** Mean squared error (MSE) of $\hat{\beta}_{ls,1}$ (LS estimator of $\beta_1$), $\hat{\beta}_{r(1),1}$ (first robust estimator of $\beta_2$), and $\hat{\beta}_{r(2),1}$ (second robust estimator of $\beta_1$)

$$D_2 : x_1 = 1, x_2 = 0, \text{all other } x(s) \text{ at } 0.5.$$

With regard to the sample size, consider $K(\equiv n) = 6, 8, 10,$ and 20 to examine the effect of small as well as moderately large samples on the estimation. Furthermore, select two locations for the possible outlier, namely locations at $i = 2$ and 3 for $K = 6$; $i = 2$ and 4 for $K = 8$; $i = 2$ and 6 for $K = 10$; and $i = 2$ and 11 for $K = 20$. Also, without any loss of generality, choose $\sigma^2 = 1$, $\beta_1 = 1$, and $\beta_2 = 0.5$. For variance inflation, eight values of $\omega_i$, namely $\omega_i = 0.001, 0.005, 0.01, 0.05, 0.10, 0.25, 0.50,$ and 1.0, were considered. Note that $\omega_i = 1.0$ represents the case where the data do not contain any outliers, whereas a small value of $\omega_i$ indicates that $y_i$ is generated with a large variance implying that $y_i$ can be an influential outlier. The data were simulated 10,000 times. Under each simulation, the LS estimate of $\beta$ and $\sigma^2$ were obtained, which are denoted by $\hat{\beta}_{ls} = (\hat{\beta}_{ls,1}, \hat{\beta}_{ls,2})'$ and $\hat{\sigma}^2_{ls}$, respectively. As far as the robust estimation of $\beta$ and $\sigma^2$ is concerned, these parameters were estimated by using two robust approaches. More specifically, $\hat{\beta}_{r(1)} = (\hat{\beta}_{r(1),1}, \hat{\beta}_{r(1),2})'$ is obtained by using (9), $\hat{\beta}_{r(2)} = (\hat{\beta}_{r(2),1}, \hat{\beta}_{r(2),2})'$ is obtained by using (11), and similarly $\tilde{\sigma}^2_{r(1)}$ and $\tilde{\sigma}^2_{r(2)}$ are obtained from (14) and (15), respectively. The mean squared errors (MSEs) of these estimators based on 10,000 simulations are displayed in Figs. 1–3, for the estimates of $\beta_1, \beta_2$, and $\sigma^2$, respectively.

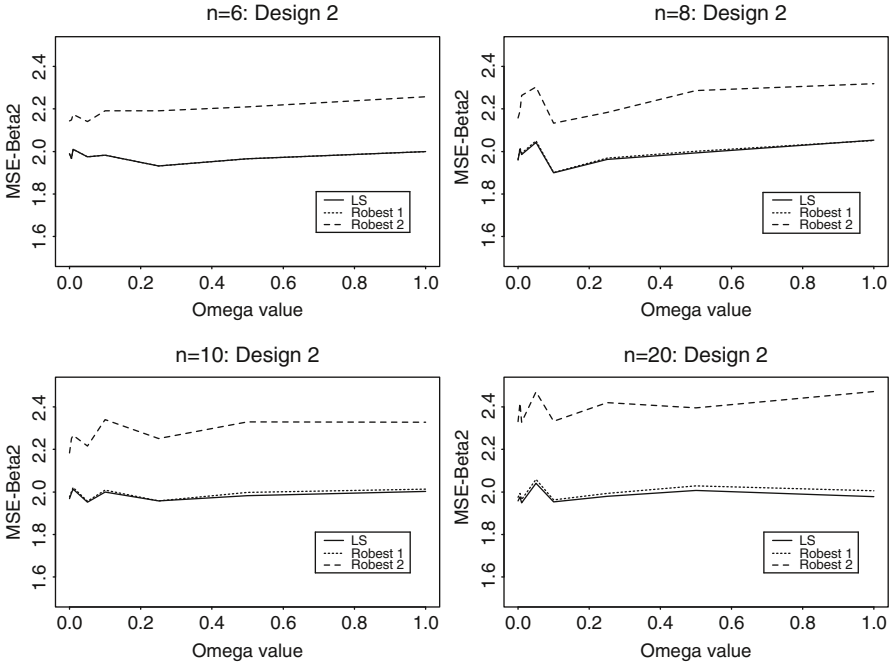


**Fig. 2** Mean squared error (MSE) of $\hat{\beta}_{ls,2}$ (LS estimator of $\beta_2$), $\hat{\beta}_{r(1),2}$ (first robust estimator of $\beta_2$), and $\hat{\beta}_{r(2),2}$ (second robust estimator of $\beta_2$)

In summary, the results of this simulation study indicate that in the presence of a variance inflated outlier, the second robust approach performs worse as compared to the first robust and LS methods in estimating $\beta_1$ and $\beta_2$. In estimating $\sigma^2$, the LS method performs very poorly when compared with the robust methods.

## 2.2 *Robust Estimation in GLM Setup For Independent Discrete Data*

As opposed to the linear models in normal or other continuous exponential family based variables, the robust inference for discrete data in the GLMs setup, such as for count and binary data, is, however, not adequately discussed in the literature. For $i = 1, \ldots, K$, let $y_i$ be a discrete response, such as count or binary, collected from the $i$th individual, and $x_i = (x_{i1}, \ldots, x_{iu}, \ldots, x_{ip})'$ be the corresponding $p$-dimensional observed covariate vector. Note that when the data contain a single outlier, any of the $K$ responses $y_1, \ldots, y_i, \ldots, y_K$ can be that outlier. Now, in the spirit of the mean shifted linear outlier model (2)–(3), suppose that we consider $y_j,\ j \neq i,\ i = 1, \ldots, K$, for example, to be the outlier because of the covariate for the $j$th individual, namely $x_j$ is contaminated. Note that if $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iu}, \ldots, \tilde{x}_{ip})'$ denotes the $p$-dimensional

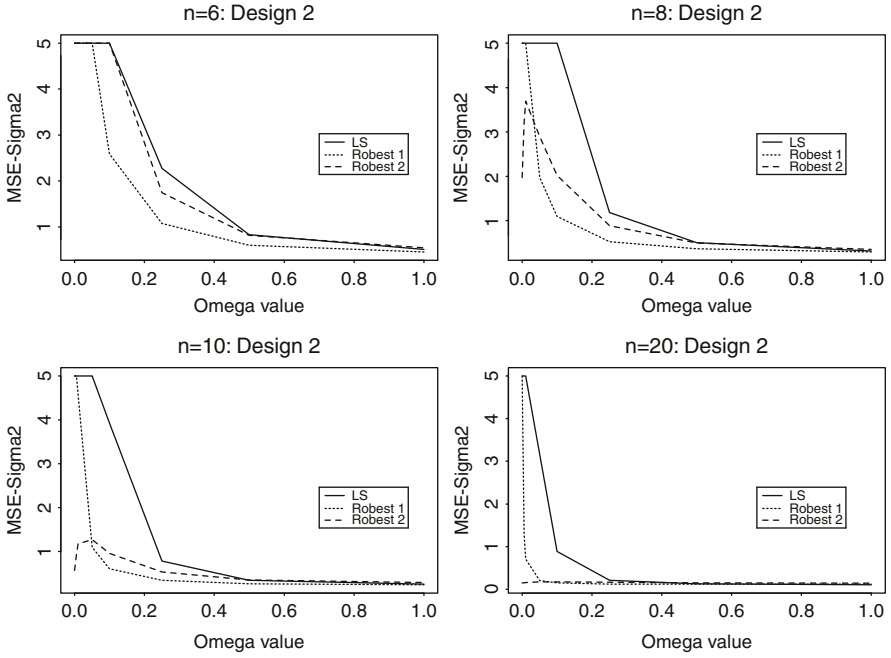


**Fig. 3** Mean squared error (MSE) of $\hat{\sigma}^2_{ls}$ (LS estimator of $\sigma^2$), $\tilde{\sigma}^2_{r(1)}$ (first robust estimator of $\sigma^2$), and $\breve{\sigma}^2_{r(2)}$ (second robust estimator of $\sigma^2$)

uncontaminated covariate vector corresponding to $y_i$ for all $i = 1, \ldots, K$, then for a positive vector $\delta$, the observed covariates $\{x_i\}$ may be related to the uncontaminated covariates $\{\tilde{x}_i\}$ as

$$x_j = \tilde{x}_j + \delta,$$

$$\text{but } x_i = \tilde{x}_i, \text{ for } i \neq j,\ i = 1, \ldots, K. \tag{16}$$

It is of primary interest to estimate $\beta = (\beta_1, \ldots, \beta_u, \ldots, \beta_p)'$, the effects of uncontaminated covariates $\tilde{x}_i$ on the response $y_i$. But, as not all the $\tilde{x}_i$'s are observed, one cannot use them to estimate $\beta$, instead the observed contaminated $x_i$'s are used, which causes bias and hence inconsistency in the estimators.

### 2.2.1 Understanding Outliers in Count and Binary Data

*K* Count Observations with a Single Outlier

First assume that in the absence of outliers, $y_1, \ldots, y_i, \ldots, y_K$ are generated following the Poisson density $P(Y_i = y_i) = [\exp(-\mu_i)\mu_i^{y_i}]/y_i!$, with $\mu_i = \exp(\tilde{x}_i'\beta)$ with $\tilde{x}_i = (\tilde{x}_{i1}, \tilde{x}_{i2})'$. Suppose that the values of these two covariates arise from

$$\tilde{x}_{i1} \overset{iid}{\sim} N(0.5, 0.25) \text{ and } \tilde{x}_{i2} \overset{iid}{\sim} N(0.5, 0.5),$$

respectively, for all $i = 1, \ldots, K$. Suppose that $j$ is the index for the outlying observation that takes a value between 1 and $K$.

Now, to consider $y_j$ as an outlying value, that is, to have a data set of size $K$ with one outlier, one may then shift the values of $\tilde{x}_{j1}$ and $\tilde{x}_{j2}$ as

$$x_{j1} = \tilde{x}_{j1} + \delta \text{ and } x_{j2} = \tilde{x}_{j2} + \delta,\ \delta > 0,$$

respectively, but retain $x_{i1} = \tilde{x}_{i1}$ and $x_{i2} = \tilde{x}_{i2}$, for all $i \neq j$. As far as the shifting is concerned, suppose that $\delta = 2.0$. Thus, $y_1, \ldots, y_K$ refer to a sample of $K$ count observations with $y_j$ as the single outlier.

### $K$ Binary Observations with a Single Outlier

Note that the existing literature (Copas 1988, p. 226; Carroll and Pederson 1993; Sinha 2004) does not provide a clear definition for the outliers in binary data. Remark that Cantoni and Ronchetti (2001) have suggested a practically useful MQL robust inference technique for independent data subject to outliers in GLM setup. However even though GLMs include count and binary models, since the concordant counts (bulk of the observations of similar nature) in the Poisson case and the concordant success numbers in the binomial case can be exploited in a similar way to recognize any possible outliers in the respective data sets, Cantoni and Ronchetti's (2001) definitions of outliers are appropriate only for the Poisson and binomial cases. Thus, even though binary is a special case of the binomial setup, Cantoni and Rochetti's (2001) robust inference development does not appear to be appropriate for the binary data. In view of these difficulties with regard to the robust inferences for the binary case, Bari and Sutradhar (2010a) have provided a new definition for the outliers for the binary data. More specifically, they dealt with one and two sided outliers in the binary data. For convenience these definitions are summarized as follows.

*One sided outlier* For

$$Pr[Y_i = 1] = E[Y_i] = \mu_i = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)},$$

and

$$p_{sb} = max\{\mu_i\},\ p_{lb} = min\{\mu_i\},$$

suppose that the bulk $(K-1)$ of the binary observations occur with small probabilities such that

$$Pr[Y_i = 1] = \begin{cases} \leq p_{sb} \text{ for } i \neq j, i = 1, \ldots, K, \\ > p_{sb} \text{ for } i = j, \end{cases} \tag{17}$$

or, with large probabilities such that

$$Pr[Y_i = 1] = \begin{cases} \geq p_{lb} \text{ for } i \neq j, i = 1, \ldots, K, \\ < p_{lb} \text{ for } i = j, \end{cases} \tag{18}$$

Here the binary $y_j$, whether 1 or 0, satisfying (17) is referred to as an upper sided outlier or satisfying (18) is referred to as a lower sided outlier, whereas the remaining $K-1$ responses denoted by $y_i$ for $i \neq j$ constitute a group of "concordant" observations.

*Two sided outlier* It may happen in practice that probabilities for the bulk of the observations lie in the range $p_{sb} \leq P(Y_i = 1) \leq p_{lb}$, leading to a situation where one may encounter a two sided outlier. To be specific, $y_j = 0$ or 1 will be an outlier if either $P(Y_j = 1) > p_{lb}$ or $P(Y_j = 1) < p_{sb}$.

*Generation of K binary observations with an outlier* We now illustrate the generation of $K$ binary observations including one outlier. For the purpose one may first generate $K$ binary responses $y_1, \ldots, y_i, \ldots, y_K$ assuming that they do not contain any outliers. To be specific, generate these $K$ "good" responses following the binary logistic model $P(Y_i = 1) = [\exp(\tilde{x}_i'\beta)]/[1 + \exp(\tilde{x}_i'\beta)]$, with two covariates so that $\tilde{x}_i = (\tilde{x}_{i1}, \tilde{x}_{i2})'$ and $\beta = (\beta_1, \beta_2)'$. As far as the covariate values are concerned, similar to the Poisson case, consider two covariates $\tilde{x}_{i1}$ and $\tilde{x}_{i2}$ as

$$\tilde{x}_{i1} \overset{iid}{\sim} N(-1.0, 0.25) \text{ and } \tilde{x}_{i2} \overset{iid}{\sim} N(-1.0, 0.5),$$

respectively, for $i = 1, \ldots, K$.

Next, to create an outlier $y_j$ where $j$ can take any value between 1 and $K$, change the corresponding covariate values $\tilde{x}_{j1}$ and $\tilde{x}_{j2}$ as

$$x_{j1} = \tilde{x}_{j1} + \delta_1 \text{ and } x_{j2} = \tilde{x}_{j2} + \delta_2, \ \delta_1, \delta_2 > 0,$$

respectively. Note that for large positive $\delta_1$ and $\delta_2$, these modified covariates will be increased in magnitude yielding larger probability for $y_j = 1$. One may then treat $y_j$ as an outlier. For convenience, suppose that one uses $\delta_1 = 3.0$ and $\delta_2 = 4.0$. As far as the remaining covariates are concerned, they are kept unchanged. That is, for $i \neq j$ $(i = 1, \ldots, K)$, consider $x_{i1} = \tilde{x}_{i1}$ and $x_{i2} = \tilde{x}_{i2}$.

### 2.2.2 Naive and Existing Robust QL Estimation Approaches

Naive QL (NQL) Estimation of $\beta$

Had there been no outliers, one could have obtained the consistent estimate of $\beta$ by solving the well-known QL (quasi-likelihood) estimating equation

$$\sum_{i=1}^{K}\left[\frac{\partial\tilde{\mu}_i}{\partial\beta}V^{-1}(\tilde{\mu}_i)(y_i-\tilde{\mu}_i)\right]=0, \tag{19}$$

(see Wedderburn 1974; McCullagh and Nelder 1989; Heyde 1997) where, for example, $\tilde{\mu}_i = E[Y_i] = \exp(\tilde{x}_i'\beta)$ and $V(\tilde{\mu}_i) = var[Y_i] = \tilde{\mu}_i$ for Poisson count data; and $\tilde{\mu}_i = E[Y_i] = \exp(\tilde{x}_i'\beta)/[1+\exp(\tilde{x}_i'\beta)]$ and $V(\tilde{\mu}_i) = var[Y_i] = \tilde{\mu}_i(1-\tilde{\mu}_i)$ for binary data. But, as the uncontaminated $\tilde{x}_i$'s are unobserved, it is not possible to use (19) for the estimation of $\beta$. Now suppose that following (19) but by using the observed covariates $\{x_i\}$, one writes the naive quasi-likelihood (NQL) estimating equation for $\beta$ given by

$$\sum_{i=1}^{K}\left[\frac{\partial\mu_i}{\partial\beta}V^{-1}(\mu_i)(y_i-\mu_i)\right]=0, \tag{20}$$

where, for example, $\mu_i = \exp(x_i'\beta)$ and $V(\mu_i) = \mu_i$ for Poisson count data; and $\mu_i = \exp(x_i'\beta)/[1+\exp(x_i'\beta)]$ and $V(\mu_i) = \mu_i(1-\mu_i)$ for binary data. Since $\beta$ is the effect of $\tilde{x}_i$ on $y_i$ for all $i = 1,\ldots,K$, it then follows that the quasi-likelihood estimator obtained from (20) will be biased and hence inconsistent for $\beta$.

### Partly Standardized Mallows Type QL (PSMQL) Estimation of $\beta$

As a remedy to the inconsistency of the quasi-likelihood estimator obtained from (20), Cantoni and Ronchetti (2001) (see also references therein), among others, have suggested a Mallow's type quasi-likelihood (MQL) robust estimation approach to obtain a consistent estimate for the regression effects $\beta$. For the purpose, for $r_i = \frac{y_i-\mu_i}{\sqrt{V(\mu_i)}}$, they first define the Huber robust function as

$$\psi_c(r_i) = \begin{cases} r_i, & |r_i| \le c, \\ c\ \text{sign}(r_i), & |r_i| > c, \end{cases} \tag{21}$$

where $c$ is referred to as the so-called tuning constant. This robust function is then used to construct the MQL estimating equation given by

$$\sum_{i=1}^{K}\left[w(x_i)\frac{\partial\mu_i}{\partial\beta}V^{-\frac{1}{2}}(\mu_i)\psi_c(r_i)-a(\beta)\right]=0, \tag{22}$$

where $a(\beta) = \frac{1}{K}\sum_{i=1}^{K} w(x_i)\frac{\partial\mu_i}{\partial\beta}V^{-\frac{1}{2}}(\mu_i)E[\psi_c(r_i)]$, with $\mu_i = E(Y_i)$, $V(\mu_i) = var(Y_i)$, and $w(x_i) = 1$ for the binomial data as in Huber's linear regression case, but $w(x_i) = \sqrt{(1-h_i)}$ for the Poisson data, where $h_i$ is the $i$th diagonal element of the hat matrix $H = X(X'X)^{-1}X'$, with $X = (x_1,\ldots,x_i,\ldots,x_K)'$ being the $K\times p$ covariate matrix.

Note that in order to minimize the robust distance function $\psi_c(r_i)$, the MQL estimating (22) was constructed by using the variance $V(\mu_i) = var(Y_i)$ as a weight function and $\frac{\partial \mu_i}{\partial \beta}$ as a gradient function, whereas a proper estimating equation should use $var(\psi_c(r_i))$ and $\frac{\partial \psi_c(r_i)}{\partial \beta}$ as the weight and gradient functions, respectively. One may therefore refer to the estimating (22) as a partly standardized MQL (PSMQL) estimating equation. This PSMQL estimating (22) provides regression estimates with smaller bias than the traditional maximum likelihood or NQL estimating (20). But, as discussed in Bari and Sutradhar (2010a), this improvement does not appear to be significant enough to recommend the use of the PSMQL estimation approach. Moreover, this PSMQL approach is not suitable for inferences in binary regression models.

### FSMQL Estimation of $\beta$

As an improvement over the PSMQL estimation, Bari and Sutradhar (2010a) have proposed a FSMQL estimation approach where the regression effects $\beta$ is obtained by solving the FSMQL estimating equation

$$\sum_{i=1}^{K}\left[w(x_i)\frac{\partial}{\partial \beta}\left\{\psi_c(r_i)-\frac{1}{K}\sum_{i=1}^{K}E\left(\psi_c(r_i)\right)\right\}\left\{var\left(\psi_c(r_i)\right)\right\}^{-1}\right.$$
$$\left.\times\left\{\psi_c(r_i)-\frac{1}{K}\sum_{i=1}^{K}E\left(\psi_c(r_i)\right)\right\}\right]=0. \tag{23}$$

Note that this FSMQL estimating (23) is constructed by replacing the "working" variance and gradient functions $V(\mu_i)$ and $\frac{\partial \mu_i}{\partial \beta}$ in (22), with the true variance and gradient functions $var(\psi_c(r_i))$ and $\frac{\partial \psi_c(r_i)}{\partial \beta}$, respectively. Also, $w(x_i) = \sqrt{(1-h_i)}$ is used in both binary and Poisson cases. Furthermore, the specific formulas for the true weight function $var(\psi_c(r_i))$ and the gradient function $\frac{\partial \psi_c(r_i)}{\partial \beta}$ for the count and binary cases are available from Bari and Sutradhar (2010a, Sects. 2.1 and 2.2).

Bari and Sutradhar (2010a) also considered another version of the FSMQL estimating (23), which was developed by using the deviance $\psi_c(r_i) - E(\psi_c(r_i))$ instead of $\psi_c(r_i) - \frac{1}{K}\sum_{i=1}^{K}E(\psi_c(r_i))$. This alternative FSMQL estimating equation has the form

$$\sum_{i=1}^{K}\left[w(x_i)\frac{\partial}{\partial \beta}\left\{\psi_c(r_i)-E\left(\psi_c(r_i)\right)\right\}\left\{var\left(\psi_c(r_i)\right)\right\}^{-1}\left\{\psi_c(r_i)-E\left(\psi_c(r_i)\right)\right\}\right]=0. \tag{24}$$

For convenience, one may refer to (23) and (24) as the $\text{FSMQL}_1$ and $\text{FSMQL}_2$ estimating equations, respectively.

*Robust Function and Properties for Count Data*

For the count data, consider the Huber robust function $\psi_c(r_i)$ as in (21). The expectation and variance of this function are available from Cantoni and Ronchetti (2001, Appendix A, p. 1028). The gradient of the robust function and its expectation may then be computed as follows (see also Bari and Sutradhar 2010a, Appendix):

$$\frac{\partial \psi_c(r_i)}{\partial \beta} = \begin{cases} -\frac{\mu_i}{V^{\frac{1}{2}}(\mu_i)} x_i, & |r_i| \le c, \\ 0, & |r_i| > c, \end{cases} \tag{25}$$

and

$$\frac{\partial E(\psi_c(r_i))}{\partial \beta} = -c\left[\frac{\partial}{\partial \beta}F_{Y_i}(i_2) + \frac{\partial}{\partial \beta}F_{Y_i}(i_1)\right] + \frac{\mu_i}{V^{\frac{1}{2}}(\mu_i)}\left[\left\{x_i P(Y_i = i_1) + \frac{\partial}{\partial \beta}P(Y_i = i_1)\right\}\right.$$

$$\left. - \left\{x_i P(Y_i = i_2) + \frac{\partial}{\partial \beta}P(Y_i = i_2)\right\}\right], \tag{26}$$

where

$$\frac{\partial}{\partial \beta}P(Y_i = i_1) = P(Y_i = i_1)(i_1 - \mu_i)x_i, \ \frac{\partial}{\partial \beta}P(Y_i = i_2) = P(Y_i = i_2)(i_2 - \mu_i)x_i,$$

$$\frac{\partial}{\partial \beta}F_{Y_i}(i_1) = \sum_{j=0}^{i_1}\frac{\partial}{\partial \beta}P(Y_i = j), \text{ and } \frac{\partial}{\partial \beta}F_{Y_i}(i_2) = \sum_{j=0}^{i_2}\frac{\partial}{\partial \beta}P(Y_i = j).$$

*Robust Function and Properties for Binary Data*

*(a) Robust function in the presence of one sided outlier*

Suppose that the bulk of the binary observations occur with small probabilities. In this case, the robust function $\psi_c(r_i)$ $(i = 1, \ldots, n)$ may be defined as

$$\psi_c(r_i) = \begin{cases} \frac{y_i - \mu_i}{V^{\frac{1}{2}}(\mu_i)}, & P(Y_i = 1) \le p_{sb}, i \ne j, i = 1, \ldots, K, \\ \frac{y_i - \mu_i^{(c_1)}}{V^{(c_1)\frac{1}{2}}(\mu_i^{(c_1)})}, & P(Y_i = 1) > p_{sb}, i = j, \end{cases} \tag{27}$$

where $\mu_i = \frac{exp(x_i'\beta)}{1+exp(x_i'\beta)}$, $V(\mu_i) = \mu_i(1-\mu_i)$ for all $i = 1, \ldots, K$, and $p_{sb} = max\{\mu_i\}$, $i \ne j$, is a bound for all $K-1$ small probabilities.

Note that as opposed to the case given in (27), if the bulk of the binary observations occur with large probabilities, then the robust function $\psi_c(r_i)$ $(i = 1, \ldots, K)$ is defined as

$$\psi_c(r_i) = \begin{cases} \frac{y_i - \mu_i}{V^{\frac{1}{2}}(\mu_i)}, & P(Y_i = 1) \geq p_{lb}, i \neq j, i = 1, \ldots, K, \\ \frac{y_i - \mu_i^{(c_2)}}{V^{(c_2)\frac{1}{2}}(\mu_i^{(c_2)})}, & P(Y_i = 1) < p_{lb}, i = j, \end{cases} \tag{28}$$

where $p_{lb} = min\{\mu_i\}$, $i \neq j$, is a bound for all $K-1$ large probabilities.

*(b) Robust function in the presence of two sided outlier*

In this case, the robust function $\psi_c(r_i)$ $(i = 1, \ldots, K)$ may be defined as

$$\psi_c(r_i) = \begin{cases} \frac{y_i - \mu_i^{(c_1)}}{V^{(c_1)\frac{1}{2}}(\mu_i^{(c_1)})}, & P(Y_i = 1) > p_{lb}, i = j, \\ \frac{y_i - \mu_i}{V^{\frac{1}{2}}(\mu_i)}, & p_{sb} \leq P(Y_i = 1) \leq p_{lb}, i \neq j, i = 1, \ldots, K, \\ \frac{y_i - \mu_i^{(c_2)}}{V^{(c_2)\frac{1}{2}}(\mu_i^{(c_2)})}, & P(Y_i = 1) < p_{sb}, i = j, \end{cases} \tag{29}$$

where $\mu_j^{(c_1)}$ and $V^{(c_1)}(\mu_j^{(c_1)})$ are defined as in (27), whereas $\mu_j^{(c_2)}$ and $V^{(c_2)}(\mu_j^{(c_2)})$ are defined as in (28).

*(b(i)) Basic properties of the robust function $\psi_c(r_i)$: Binary case*

It is convenient to write these properties for the two sided outlier case. The results for the one sided outlier may be obtained as a special case. The expectation, variance, and gradient of the robust function in the presence of a two sided outlier are available from Bari and Sutradhar (2010a, Appendix). For convenience, these properties are summarized as follows.

Let $\psi_c(r_i)$ denote the robust function defined as in (29). The expectation and variance of $\psi_c(r_i)$ are given by

$$E(\psi_c(r_i)) = \frac{\mu_i - \mu_i^{(c_1)}}{V^{(c_1)\frac{1}{2}}(\mu_i^{(c_1)})} P_1 + \frac{\mu_i - \mu_i^{(c_2)}}{V^{(c_2)\frac{1}{2}}(\mu_i^{(c_2)})} P_3, \tag{30}$$

and

$$var(\psi_c(r_i)) = \frac{(1 - 2\mu_i^{(c_1)})\mu_i + \mu_i^{(c_1)^2}}{V^{(c_1)}(\mu_i^{(c_1)})} P_1 + P_2 + \frac{(1 - 2\mu_i^{(c_2)})\mu_i + \mu_i^{(c_2)^2}}{V^{(c_2)}(\mu_i^{(c_2)})} P_3 - [E(\psi_c(r_i))]^2, \tag{31}$$

where $P_1$, $P_2$, and $P_3$ are the probabilities for a binary observation to satisfy the conditions $P(Y_i = 1) > p_{lb}$, $p_{sb} \leq P(Y_i = 1) \leq p_{lb}$, and $P(Y_i = 1) < p_{sb}$, respectively.

In practice, the probabilities $P_1$, $P_2$, and $P_3$ may be computed from the data by using the sample proportions given by, for example,

$$P_1 = \frac{\text{Number of observations satisfying } P(Y_i = 1) > p_{lb}}{\text{Total observation } (K)}.$$

The gradient of the robust function $\psi_c(r_i)$ [defined in (29)] and its expectation are given by

$$\frac{\partial \psi_c(r_i)}{\partial \beta} = \begin{cases} 0, & P(Y_i = 1) > p_{lb}, i = j, \\ \frac{-\mu_i(1-\mu_i)x_i}{V^{\frac{1}{2}}(\mu_i)}, & p_{sb} \le P(Y_i = 1) \le p_{lb}, i \ne j, i = 1, \ldots, K, \\ 0, & P(Y_i = 1) < p_{sb}, i = j, \end{cases} \tag{32}$$

and

$$\frac{\partial E(\psi_c(r_i))}{\partial \beta} = \frac{(1-\mu_i)\mu_i x_i}{V^{(c_1)\frac{1}{2}}(\mu_i^{(c_1)})} P_1 + \frac{(1-\mu_i)\mu_i x_i}{V^{(c_2)\frac{1}{2}}(\mu_i^{(c_2)})} P_3. \tag{33}$$

To illustrate the finite sample based relative performance of the competitive robust approaches, namely PSMQL (22), FSMQL$_1$ (23), and FSMQL$_2$ (24) approaches, we refer to some of the simulation results from Bari and Sutradhar (2010a). In the presence of a single outlier, the count and binary data were generated as in Sect. 2.2.1. With $K = 60$ observations including an outlier, the relative bias (RB) of an estimator, for example, for $\beta_k$ $(k = 1, \ldots, p)$ given by

$$\text{RB}\,(\hat{\beta}_k) = \frac{|\hat{\beta}_k - \beta_k|}{\text{s.e.}\,(\hat{\beta}_k)} \times 100, \tag{34}$$

were computed based 1,000 simulations. The results are shown in Table 1.

**Table 1** (*For count and binary data with one outlier*) Simulated means (SM), simulated standard errors (SSE), and relative biases (RB) of the PSMQL, FSMQL$_1$, and FSMQL$_2$ estimates of the regression parameters $\beta_1 = 1.0$ and $\beta_2 = 0.5$, for sample size 60 and selected values of the tuning constant $c = 1.4$ under the Poisson model, and tuning constant $\mu^{c_1} = 0.9$ under the binary model, in the presence of one outlier

| | | | | Estimation method | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PSMQL | | FSMQL$_1$ | | FSMQL$_2$ | |
| Model | $K$ | Tuning constant | Statistic | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ |
| Count | 60 | $c = 1.4$ | SM | 0.507 | 0.600 | 0.899 | 0.517 | 0.893 | 0.488 |
| | | | SSE | 0.206 | 0.188 | 0.307 | 0.239 | 0.279 | 0.210 |
| | | | RB | 240 | 53 | 33 | 7 | 38 | 6 |
| Binary | 60 | $\mu^{c_1} = 0.9$ | SM | 1.161 | 0.194 | 0.994 | 0.503 | 1.003 | 0.486 |
| | | | SSE | 0.777 | 0.760 | 0.782 | 0.777 | 0.779 | 0.764 |
| | | | RB | 21 | 40 | 1 | 0 | 0 | 2 |

The results of the table show that both fully standardized robust procedures $\text{FSMQL}_1$ and $\text{FSMQL}_2$ perform much better in estimating $\beta$ as compared to the existing PSMQL robust approach.

## 3 Robust Inference in Longitudinal Setup

### *3.1 Existing GEE Approaches for Robust Inferences*

Let $\mu_i(x_i) = E(Y_i) = (\mu_{i1}, \ldots, \mu_{it}, \ldots, \mu_{iT})'$ denote the mean, and $\Sigma_i(x_i, \rho) : T \times T$ be the true covariance matrix of the response vector $y_i$ where $x_i$ represents all true covariates, i.e., $x_i \equiv x_{i1}, \ldots, x_{it}, \ldots, x_{iT}$. For convenience, the covariance matrix $\Sigma_i(x_i, \rho)$ is often expressed as $\Sigma_i(x_i, \rho) = A_i^{\frac{1}{2}} C_i(\rho) A_i^{\frac{1}{2}}$, where $A_i = \text{diag}[\sigma_{i11}, \ldots, \sigma_{itt}, \ldots, \sigma_{iTT}]$ and $C_i(\rho)$ is the correlation matrix for repeated binary or count data. Note that if the longitudinal data do not contain any outliers, then one may obtain consistent and highly efficient estimate of $\beta$ by solving the GQL estimating equation

$$\sum_{i=1}^{K} \left[ \frac{\partial \mu_i'(x_i)}{\partial \beta} \Sigma_i^{-1}(x_i, \hat{\rho})(y_i - \mu_i(x_i)) \right] = 0, \tag{35}$$

(see Sutradhar 2003) where $\hat{\rho}$ is a suitable consistent, for example, a moment estimate of $\rho$.

Note that in practice it may, however, happen that a small percentage such as 1% of longitudinal observations are suspected to be outliers. Suppose that $m$ of the $KT$ responses are referred to as the outliers when their corresponding covariates are shifted by an amount $\delta$, $\delta$ being a real valued vector. For convenience, we denote the new set of covariates as

$$\tilde{x}_{it} = \begin{cases} x_{it} & \text{for } (i,t) \not\equiv (i',t') \\ x_{it} + \delta & \text{for } (i,t) \equiv (i',t') \end{cases},$$

and use these observed covariates $\tilde{x}_{it}$ for the estimation of $\beta$. It is, therefore, clear that since $\beta$ is the effect of the true covariate $x_{it}$ on $y_{it}$, the solution of the observed covariates $\tilde{x}_i$ based naive GQL (NGQL) estimating equation

$$\sum_{i=1}^{K} \left[ \frac{\partial \mu_i'(\tilde{x}_i)}{\partial \beta} \Sigma_i^{-1}(\tilde{x}_i, \hat{\rho})(y_i - \mu_i(\tilde{x}_i)) \right] = 0, \tag{36}$$

will produce biased and hence inconsistent estimate for $\beta$. To overcome this inconsistency problem, Preisser and Qaqish (1999), among others, have proposed to solve a resistant generalized quasi-likelihood estimating equation (REGEE) given by

$$\sum_{i=1}^{K}\left[\frac{\partial\mu_i'(\tilde{x}_i)}{\partial\beta}V_i^{-1}(\tilde{x}_i,\alpha)(\psi_i^*-c_i)\right]=0, \tag{37}$$

where $\psi_i^*$ is a down-weighting function, $c_i=E(\psi_i^*)$, and $V_i(\tilde{x}_i,\alpha)$ is a "working" covariance matrix (Liang and Zeger 1986). Note that the REGEE in (37) does not appear to be a proper weighted estimating equation. This is because, first, $V_i(\tilde{x}_i,\alpha)$ is only a substitute of $\Sigma_i(\tilde{x}_i,\rho)$ matrix, whereas in the presence of outliers, one needs to use $\Omega_i^*=var(\psi_i^*)$ in order to obtain efficient regression estimates. Secondly, the REGEE (37) uses $\frac{\partial\mu_i'(\tilde{x}_i)}{\partial\beta}$ as the gradient function, whereas the consistency of the estimates may depend on the proper gradient function constructed by taking the derivative of the $\psi_i^*-c_i$ function with respect to $\beta$.

Cantoni (2004) has provided an improvement over the REGEE by introducing the proper gradient function in the estimating equation. To be specific, as compared to Preisser and Qaqish (1999) (see also Eq. (36)), Cantoni (2004) constructed an improved resistant generalized estimating equation (IREGEE) given by

$$\sum_{i=1}^{K}\left[E\left\{\frac{\partial(\psi_i^*-c_i)}{\partial\beta'}\right\}'V_i^{-1}(\tilde{x}_i,\alpha)(\psi_i^*-c_i)\right]=0, \tag{38}$$

where $E\left[\frac{\partial(\psi_i^*-c_i)}{\partial\beta'}\right]$ is a proper gradient of the robust function $\psi_i^*-c_i$, with

$$E\left[\frac{\partial(\psi_i^*-c_i)}{\partial\beta'}\right]=E\left[\frac{\partial(\psi_i^*-c_i)}{\partial\mu_i'(\tilde{x}_i)}\right]\frac{\partial\mu_i}{\partial\beta'}=\Gamma_i\frac{\partial\mu_i}{\partial\beta'}.$$

Note that the estimating (38) still uses a "working" covariance matrix $V_i(\tilde{x}_i,\alpha)$, whereas an efficient estimating equation (Sutradhar and Das 1999) should use the proper covariance matrix of the robust function, namely $\Omega_i^*=var(\psi_i^*)$. Further, similar to Cantoni (2004), Sinha (2006) has attempted to develop certain robust inferences to deal with outliers in the longitudinal data. But, Sinha (2006) has modeled the longitudinal correlations through random effects, which, therefore addresses a different problem than longitudinal data problems.

Recently, Bari and Sutradhar (2010b) has proposed an auto-correlation class based robust GQL (RGQL) approach for inferences in binary and count panel data models in the presence of outliers. This RGQL approach produces consistent and highly efficient regression estimates, and it is a generalization of the FSMQL approach for independent data to the longitudinal setup. The RGQL approach is summarized in the next section.

## 3.2 *RGQL Approach for Robust Inferences in Longitudinal Setup*

Note that when the covariates are stationary, that is, time independent, one may develop a general auto-correlation class based robust GQL estimation approach.

Bari and Sutradhar (2010b) have considered non-stationary covariates and exploited the most likely AR(1) type correlation structures for both count and binary data. These correlation structures are discussed in detail in Sutradhar (2010), see also Sutradhar (2011). For convenience we summarize these correlation structures as follows.

Recall that $x_{it} = (x_{it1}, \ldots, x_{itu}, \ldots, x_{itp})'$ is the $p \times 1$ vector of covariates corresponding to $y_{it}$ when the data do not contain any outliers, and $\beta$ denote the effects of the covariate $x_{it}$ on $y_{it}$. The AR(1) correlation models for repeated responses $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$ based on the uncontaminated covariates $x_{i1}, \ldots, x_{it}, \ldots, x_{iT}$, for binary and count data are given below.

*AR(1) model for repeated binary data*

For $\mu_{it} = \frac{exp(x_{it}'\beta)}{1+exp(x_{it}'\beta)}$, for all $t = 1, \ldots, T$, the AR(1) model for the binary data may be written as

$$\begin{aligned} &y_{i1} \sim bin(\mu_{i1}) \text{ and} \\ &y_{it}|y_{i,t-1} \sim bin[\mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1})], \end{aligned} \tag{39}$$

(Zeger et al. 1985; Qaqish 2003) where $\rho$ is a correlation index parameter. The binary AR(1) model (39) has the auto-correlation structure given by

$$corr(Y_{iu}, Y_{it}) = \begin{cases} \rho^{t-u} \left[\frac{\sigma_{iuu}}{\sigma_{itt}}\right]^{1/2}, \text{ for } u < t \\ \rho^{u-t} \left[\frac{\sigma_{itt}}{\sigma_{iuu}}\right]^{1/2}, \text{ for } u > t \end{cases}, \tag{40}$$

where $\sigma_{iuu} = \mu_{iu}(1 - \mu_{iu})$, for example, is the variance of $y_{iu}$. Note that $\rho$ parameter in (39)–(40) must satisfy the range restriction

$$max\left[-\frac{\mu_{it}}{1-\mu_{i,t-1}}, -\frac{1-\mu_{it}}{\mu_{i,t-1}}\right] \leq \rho \leq min\left[\frac{1-\mu_{it}}{1-\mu_{i,t-1}}, \frac{\mu_{it}}{\mu_{i,t-1}}\right]. \tag{41}$$

*AR(1) model for repeated count data*

As opposed to the binary AR(1) model (39), the AR(1) model for the count data is defined as

$$\begin{aligned} &y_{i1} \sim Poisson(\mu_{i1}) \\ &y_{it} = \rho * y_{i,t-1} + d_{it}, \; t = 2, \ldots, T, \end{aligned} \tag{42}$$

(see McKenzie 1988; Sutradhar 2003), where $y_{i,t-1} \sim Poisson(\mu_{i,t-1})$ and $d_{it} \sim Poisson(\mu_{it} - \rho\mu_{i,t-1})$, with $\mu_{it} = E(Y_{it}) = \exp(x_{it}'\beta)$. In (42), $d_{it}$ and $y_{i,t-1}$ are assumed to be independent. Also, for given count $y_{i,t-1}$,

$$\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho),$$

where $b_j(\rho)$ stands for a binary variable with $P[b_j(\rho) = 1] = \rho$ and $P[b_j(\rho) = 0] = 1 - \rho$. The AR(1) model (42) for count data has the auto-correlation structure given by

$$corr(Y_{iu}, Y_{it}) = \rho^{t-u} \sqrt{\frac{\mu_{iu}}{\mu_{it}}}, \tag{43}$$

with $\rho$ satisfying the range restriction

$$0 < \rho < min\left[1, \frac{\mu_{it}}{\mu_{i,t-1}}\right], t = 2, \cdots, T. \tag{44}$$

### 3.2.1 RGQL Estimating Equation

For $\xi_i = [\psi_c(r_{i1}), \ldots, \psi_c(r_{it}), \ldots, \psi_c(r_{iT})]'$, its expectation $\lambda_i$ is available from Cantoni and Ronchetti (2001) for the count data, and from Sect. 2.2.2 for the binary case. Recall from (38) that based on "working" covariance of the responses (Liang and Zeger 1986), Cantoni (2004) has suggested an IREGEE approach for estimating $\beta$ in the presence of outliers. One may obtain consistent $\beta$ estimate by solving a slightly different equation than (38) given by

$$\sum_{i=1}^{K} \left[ W_i \frac{\partial}{\partial \beta} \left\{ \xi_i - K^{-1} \sum_{i=1}^{K} \lambda_i \right\}' V_i^{-1}(\alpha) \left\{ \xi_i - K^{-1} \sum_{i=1}^{K} \lambda_i \right\} \right] = 0, \tag{45}$$

where $W_i = diag[w_{i1}, \ldots, w_{it}, \ldots, w_{iT}]$ is the $T \times T$ covariate dependent diagonal weight matrix so that covariates corresponding to the outlying response yield less weight for the corresponding robust function. To be specific, the $t$-th diagonal element of the $W_i$ matrix is computed as $w_{it} = \sqrt{1 - h_{itt}}$, $h_{itt}$ being the $t$-th diagonal element of the hat matrix $H_i = \tilde{X}_i(\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'$ with $\tilde{X}_i = [\tilde{x}_{i1}, \ldots, \tilde{x}_{it}, \ldots, \tilde{x}_{iT}]'$. See, for example, Cantoni and Ronchetti (2001). Also in (45), $V_i(\alpha) = \text{cov}(Y_i) = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}$ is a "working" covariance matrix of $y_i$, with $R(\alpha)$ as the associated "working" correlation matrix. Note that there are twofold problems with this estimating equation. First, for efficiency increase, it would have been appropriate to use $\text{cov}(\xi_i) = cov[\psi_c(r_{i1}), \ldots, \psi_c(r_{it}), \ldots, \psi_c(r_{iT})]$ as the weight matrix instead of the true covariance matrix $\Sigma_i(\alpha) = \text{cov}(Y_i)$. Secondly, Cantoni (2004) did not even use $\Sigma_i$, rather has used a "working" covariance matrix $V_i(\alpha) = \text{cov}(Y_i)$.

To overcome this inefficiency problem encountered by Cantoni's approach, Bari and Sutradhar (2010b) have suggested a robust function based GQL (RGQL) estimating equation for $\beta$ as

$$\sum_{i=1}^{K}\left[W_i\frac{\partial}{\partial\beta}\left\{\xi_i-K^{-1}\sum_{i=1}^{K}\lambda_i\right\}'\Omega_i^{-1}\left\{\xi_i-K^{-1}\sum_{i=1}^{K}\lambda_i\right\}\right]=0, \tag{46}$$

where

$$\Omega_i=cov(\xi_i)=(\omega_{iut}), \tag{47}$$

with

$$\omega_{iut}=E\left[\psi_c(r_{iu})\psi_c(r_{it})\right]-\left\{E(\psi_c(r_{iu}))E(\psi_c(r_{it}))\right\}, \tag{48}$$

where, as mentioned above, the formulas for $E[\psi_c(r_{it})]$ are available for both count and binary data.

Computation of $\Omega_i$ for the Binary Data

Note that the computation of the product moment $E[\psi_c(r_{iu})\psi_c(r_{it})]$ in (48) is manageable for the binary case, but it is extremely difficult for the count data. For example, suppose that $y_{it}$, $t=1,\ldots,T$, used in the robust functions $\psi_c(r_{it})$, follow an AR(1) type correlation structure given by (40), where $\mu_{it}=\frac{exp(x_{it}'\beta)}{1+exp(x_{it}'\beta)}$ and $\rho$ is a correlation index parameter. Next, suppose that the binary data contain two sided outliers. One may then follow (29) and compute all nine combinations for the product term $\psi_c(r_{iu})\psi_c(r_{it})$ and compute the expectations of all these nine terms, and derive the formulas as

$$E\left[\psi_c(r_{iu})\psi_c(r_{it})\right]=\rho^{t-u}\sigma_{iuu}a_{iut}+\left[E(\psi_c(r_{iu}))E(\psi_c(r_{it}))\right], \tag{49}$$

where

$$\begin{aligned}a_{iut}=&\frac{P_1^2}{\sqrt{\sigma_{itt}^{(c_1)}\sigma_{iuu}^{(c_1)}}}+P_1P_2\left[\frac{1}{\sqrt{\sigma_{itt}\sigma_{iuu}^{(c_1)}}}+\frac{1}{\sqrt{\sigma_{itt}^{(c_1)}\sigma_{iuu}}}\right]\\&+P_1P_3\left[\frac{1}{\sqrt{\sigma_{itt}^{(c_2)}\sigma_{iuu}^{(c_1)}}}+\frac{1}{\sqrt{\sigma_{itt}^{(c_1)}\sigma_{iuu}^{(c_2)}}}\right]+P_2P_3\left[\frac{1}{\sqrt{\sigma_{itt}^{(c_2)}\sigma_{iuu}}}+\frac{1}{\sqrt{\sigma_{itt}\sigma_{iuu}^{(c_2)}}}\right]\\&+\frac{P_2^2}{\sqrt{\sigma_{itt}\sigma_{iuu}}}+\frac{P_3^2}{\sqrt{\sigma_{itt}^{(c_2)}\sigma_{iuu}^{(c_2)}}},\end{aligned}$$

for $u<t$. We may then easily compute $\omega_{iut}$ by using (49) and (48).

Further note that for the one sided outlier case, the $E[\psi_c(r_{iu})\psi_c(r_{it})]$ can be obtained from (49) as follows. For the one sided down-weighting function $\psi_c(r_{it})$ given in (28), one may compute the expectation of $\psi_c(r_{iu})\psi_c(r_{it})$ from (49) by

changing the limits obtained by replacing $p_{lb}$ with 0. Similarly, the product moment based on the down-weighting function $\psi_c(r_{it})$ given in (27), can be obtained from (49) by changing the limits obtained by replacing $p_{sb}$ with 1.

Under the AR(1) binary correlation structure (40), the outlier based moment estimation formula for $\rho$ derived from (49), is given by

$$\hat{\rho}_M = \frac{\frac{\sum_{i=1}^{K}\sum_{u=1}^{T-1}[\psi_c(r_{iu})-E(\psi_c(r_{iu}))][\psi_c(r_{i,u+1})-E(\psi_c(r_{i,u+1}))]w_{iu}w_{i,u+1}}{K(T-1)}}{\frac{\sum_{i=1}^{K}\sum_{u=1}^{T}[\psi_c(r_{iu})-E(\psi_c(r_{iu}))]^2/var[\psi_c(r_{iu})]}{KT}} \Big/ \frac{\sum_{i=1}^{K}\sum_{u=1}^{T-1}\sigma_{iuu}a_{iut}w_{iu}w_{i,u+1}}{K(T-1)}. \tag{50}$$

Alternatively, for any lag 1 dependent [irrespective of the correlation structure such as AR(1) or MA(1)] binary or count data with possible outliers, the lag 1 correlation index parameter $\rho$ may be estimated as

$$\hat{\rho}_M = \frac{\frac{\sum_{i=1}^{K}\sum_{u=1}^{T-1}[\psi_c(r_{iu})w_{iu}-\bar{\xi}_{u,w}][\psi_c(r_{i,u+1})w_{i,u+1}-\bar{\xi}_{u+1,w}]}{K(T-1)}}{\frac{\sum_{i=1}^{K}\sum_{u=1}^{T}[\psi_c(r_{iu})w_{iu}-\bar{\xi}_{u,w}]^2}{KT}}, \tag{51}$$

where $\bar{\xi}_{t,w} = \frac{1}{K}\sum_{i=1}^{K}\psi_c(r_{it})w_{it}$.

Computation of $\Omega_i$ for Count Data

Note that as opposed to the binary case, the construction of the $\Omega_i$ matrix is difficult for the count data case. One may, however, alternatively compute this $\Omega_i$ matrix by using the general formula

$$cov(\xi_i) = \Omega_i = A_{i\xi}^{\frac{1}{2}} C_{i\xi} A_{i\xi}^{\frac{1}{2}}, \tag{52}$$

where $A_{i\xi} = [var(\psi_c(r_{i1})), \ldots, var(\psi_c(r_{it})), \ldots, var(\psi_c(r_{iT}))]$ and $C_{i\xi} = (c_{i\xi,ut})$, with $c_{i\xi,ut} = corr[\psi_c(r_{iu}), \psi_c(r_{it})]$ for $u,t = 1,\ldots,T$. For (52), the formulas for $var[\psi_c(r_{it})]$ for the binary data are given in Sect. 2.2.2, and for the count data they are available from Cantoni and Ronchetti (2001, Appendix). As far as the computation of the $C_{i\xi}$ matrix is concerned, one may approximate this matrix by a constant matrix $C_{\xi}^{*}$, say, by pretending that the covariates are stationary even though they are non-stationary (i.e., time dependent). Under this assumption, the $(u,t)th$ component of the constant matrix $C_{\xi}^{*}$ may be computed as

$$C_{\xi}^{*} = (c_{\xi,ut}^{*}),$$

where

$$c^*_{\xi,ut} = \frac{\frac{1}{K}\sum_{i=1}^{K}[\psi_c(r_{iu}) - \bar{\xi}_u][\psi_c(r_{it}) - \bar{\xi}_t]}{\sqrt{\frac{1}{K}\sum_{i=1}^{K}[\psi_c(r_{iu}) - \bar{\xi}_u]^2 \frac{1}{K}\sum_{i=1}^{K}[\psi_c(r_{it}) - \bar{\xi}_t]^2}}, \tag{53}$$

with $\bar{\xi}_t = \frac{1}{K}\sum_{i=1}^{K}\psi_c(r_{it})$, for all $t = 1, \ldots, T$.

Note that the REGEE approach encounters convergence problems and also this approach produces regression estimates with much larger relative biases than the RGQL approach. See, for example, the finite sample relative performance of the RGQL and REGEE approaches shown through intensive simulation studies reported in Bari and Sutradhar (2010b).

**Acknowledgments** The author fondly acknowledges the stimulating discussion by the audience of the symposium and wishes to thank for their comments and suggestions.

## References

Bailey, B.: Tables of the Bonferroni *t*-test. J. Am. Stat. Assoc. **72**, 469–478 (1977)
Bari, W., Sutradhar, B.C.: On bias reduction in robust inference for generalized linear models. Scand. J. Stat. **37**, 109–125 (2010a)
Bari, W., Sutradhar, B.C.: Robust inferences in longitudinal models for binary and count panel data in the presence of outliers. Sankhya B **72**, 11–37 (2010b)
Beckman, R.J., Cook, R.D.: Outlier....s. Technometrics **25**, 119–163 (1983)
Cantoni, E.: A robust approach to longitudinal data analysis. Can. J. Stat. **32**, 169–180 (2004)
Cantoni, E., Ronchetti, E.: Robust inference for generalized linear models. J. Am. Stat. Assoc. **96**, 1022–1030 (2001)
Carroll, R.J., Pederson, S.: On robustness in the logistic regression model. J. R. Stat. Soc. B **55**, 693–706 (1993)
Cook, R.D., Prescott, P.: On the accuracy of Bonferroni significance levels for detecting outliers in linear models. Technometrics, **23**, 59–63 (1981)
Copas, J.B.: Binary regression models for contaminated data (with discussion). J. R. Stat. Soc. B **50**, 225–265 (1988)
Doornbos, R.: Testing for a single outlier in a linear model. Biometrics **37**, 705–711 (1981)
Ellenberg, J.H.: The joint distribution of the standardized least squares residuals from a general linear regression. J. Am. Stat. Assoc. **68**, 941–943 (1973)
Ellenberg, J.H.: Testing for a single outlier from a general linear model. Biometrics **32**, 637–645 (1976)
Hampel, F.R., Rousseeuw, P.J., Ronchetti, E.M., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions. Wiley, New York (1986)
Huber, P.J.: Robust Statistics. Wiley, New York (2004)
Heyde, C.C.: Quasi-likelihood and its applications. Springer-Verlag, New York (1997)
Johnson, B.A., Prescott, P.: Critical values of a test to detect outliers in factorial experiments. Appl. Stat. **24**, 56–59 (1975)
Liang, K.-Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
Lund, R.E.: Tables for an approximate test for outliers in linear regression. Technometrics **17**, 473–476 (1975)

McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, London (1989)
McKenzie, E.: Some ARMA models for dependent sequences of Poisson counts. Adv. Appl. Probab. **20**, 822–835 (1988)
Preisser, J.S., Qaqish, B.F.: Robust regression for clustered data with applications to binary regression. Biometrics **55**, 574–579 (1999)
Prescott, P.: An approximate test for outliers in linear models. Technometrics **17**, 129–132 (1975)
Qaqish, B.F.: A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika **90,** 455–463 (2003)
Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (1987)
Sinha, S.K.: Robust analysis of generalized linear mixed models. J. Am. Stat. Assoc. **99**, 451–460 (2004)
Sinha, S.K.: Robust inference in generalized linear model for longitudinal data. Can. J. Stat. **34**, 1–18 (2006)
Srikantan, K.S.: Testing for a single outlier in a regression model. Sankhya A **23**, 251–260 (1961)
Stefansky, W.: Rejecting outliers by maximum normed residual. Ann. Math. Stat. **42**, 35–45 (1971)
Stefansky, W.: Rejecting outliers in factorial designs. Technometrics **14**, 469–479 (1972)
Street, J.O., Carroll, R.J., Ruppert, D.: A note on computing robust regression estimates via iteratively reweighted least squares. Am. Stat. **42**, 152–154 (1988)
Sutradhar, B.C.: An overview on regression models for discrete longitudinal responses. Stat. Sci. **18**, 377–393 (2003)
Sutradhar, B.C.: Inferences in generalized linear longitudinal mixed models. Can. J. Stat. **38**, 174–196 (2010)
Sutradhar, B.C.: Dynamic Mixed Models for Familial Longitudinal Data. Springer, New York (2011)
Sutradhar, B.C., Chu, D.P.T., Bari, W.: Estimation effects on powers of two simple test statistics in identifying an outlier in linear models. J. Stat. Comput. Simul. **77**, 305–328 (2007)
Sutradhar, B.C., Das, K.: On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika **86**, 459–465 (1999)
Tietjen, G.L., Moore, R.H., Beckman, R.J.: Testing for a single outlier in simple linear regression. Technometrics **15**, 717–721 (1973)
Wedderburn, R.W.M. : Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika **61**, 439–447 (1974)
Zeger, S.L., Liang, K.Y., Self, S.G.: The analysis of binary longitudinal data with time independent covariates. Biometrika **72**, 31–38 (1985)