Analysis of Correlated Data with Measurement Error in Responses or Covariates

by

Zhijian Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Correlated data frequently arise from epidemiological studies, especially familial and longitudinal studies. Longitudinal design has been used by researchers to investigate the changes of certain characteristics over time at the individual level as well as how potential factors influence the changes. Familial studies are often designed to investigate the dependence of health conditions among family members. Various models have been developed for this type of multivariate data, and a wide variety of estimation techniques have been proposed. However, data collected from observational studies are often far from perfect, as measurement error may arise from different sources such as defective measuring systems, diagnostic tests without gold references, and self-reports. Under such scenarios only rough surrogate variables are measured. Measurement error in covariates in various regression models has been discussed extensively in the literature. It is well known that naive approaches ignoring covariate error often lead to inconsistent estimators for model parameters.

In this thesis, we develop inferential procedures for analyzing correlated data with response measurement error. We consider three scenarios: (i) likelihood-based inferences for generalized linear mixed models when the continuous response is subject to nonlinear measurement errors; (ii) estimating equations methods for binary responses with misclassifications; and (iii) estimating equations methods for ordinal responses when the response variable and categorical/ordinal covariates are subject to misclassifications.

The first problem arises when the continuous response variable is difficult to measure. When the true response is defined as the long-term average of measurements, a single measurement is considered as an error-contaminated surrogate. We focus on generalized linear mixed models with nonlinear response error and study the induced bias in naive estimates. We propose likelihood-based methods that can yield consistent and efficient estimators for both fixed-effects and variance parameters. Results of simulation studies and analysis of a data set from the Framingham Heart Study are presented.

Marginal models have been widely used for correlated binary, categorical, and ordinal data. The regression parameters characterize the marginal mean of a single outcome, without conditioning on other outcomes or unobserved random effects. The generalized estimating equations (GEE) approach, introduced by Liang and Zeger (1986), only models the first two moments of the responses with associations being treated as nuisance characteristics. For some clustered studies especially familial studies, however, the association structure may be of scientific interest. With binary data Prentice (1988) proposed additional estimating equations that allow one to model pairwise correlations. We consider marginal models for correlated binary data with misclassified responses. We develop "corrected" estimating equations approaches that can yield consistent estimators for both mean and association parameters. The idea is related to Nakamura (1990) that is originally developed for correcting bias induced by additive covariate measurement error under generalized linear models. Our approaches can also handle correlated misclassifications rather than a simple misclassification process as considered by Neuhaus (2002) for clustered binary data under generalized linear mixed models. We extend our methods and further develop marginal approaches for analysis of longitudinal ordinal data with misclassification in both responses and categorical covariates. Simulation studies show that our proposed methods perform very well under a variety of scenarios. Results from application of the proposed methods to real data are presented.

Measurement error can be coupled with many other features in the data, e.g., complex survey designs, that can complicate inferential procedures. We explore combining survey weights and misclassification in ordinal covariates in logistic regression analyses. We propose an approach that incorporates survey weights into estimating equations to yield design-based unbiased estimators.

In the final part of the thesis we outline some directions for future work, such as transition models and semiparametric models for longitudinal data with both incomplete observations and measurement error. Missing data is another common feature in applications. Developing novel statistical techniques for dealing with both missing data and measurement error can be beneficial.

Acknowledgements

I wish to acknowledge the support of various individuals and organizations. Particularly, I would like to express my deep gratitude to my supervisors, Dr. Grace Y. Yi and Dr. Changbao Wu for their mentorship, guidance, encouragement and support during my PhD studies. Their dedications to students and academic profession are to be highly appreciated. I would also like to thank my committee members Dr. Richard Cook and Dr. Mary Thompson for reviewing my thesis and their constructive suggestions.

I am indebted to the Department of Statistics and Actuarial Science at the University of Waterloo for providing such an excellent doctoral program and academic environment. Many thanks to the faculty members and staff. Many thanks to my friends, colleagues and fellow graduate students for continued encouragement and support.

Finally, I would like to thank my parents and my sister for their love and support.

Contents

List of Tables

List of Figures

Guide to Notation

In this section, we provide brief explanation and representative examples of the notation used in this thesis. Matrices and column vectors are typically denoted using bold letters, e.g., $\mathbf{M}$, and transpose is denoted using $\mathbf{M}^{\mathrm{T}}$. For precise definitions, see the text.

symbol	**description**
i	index for independent observational unit
j	index for repeated measurements within an independent unit
n	sample size
m	number of repeated measurements in a cluster
Y, $\mathbf{Y}$	true response
S, $\mathbf{S}$	surrogate response
X, $\mathbf{X}$	error-prone covariate
W, $\mathbf{W}$	surrogate for the error-prone covariate
Z, $\mathbf{Z}$	precisely measured covariates
H, $\mathbf{H}$	Misclassification indicator
C, $\mathbf{C}$	pairwise product of two binary variables
F, $\mathbf{F}$	product of two misclassification indicators (Chapter 3)
b, $\mathbf{b}$	random component (Chapters 1 and 2); intermediate quantity (Chapters 3 and 4); index for bootstrap samples (Chapter 5)
a	intermediate quantity
μ	expectation of the response
$\boldsymbol{\beta}$	regression parameters in a mean model
ξ	expectation of the pairwise product of binary variables
$\boldsymbol{\alpha}$	second-order association parameters (Chapters 3 and 4); parameters in marginal distribution of an ordinal covariate (Chapter 5)
$\boldsymbol{\theta}$	response parameters

τ	probability associated with response misclassification process
$\mathbf{L}$	covariates involved in a misclassification process
$\boldsymbol{\gamma}$	regression coefficients in response measurement error or misclassification process
ζ	superpopulation in survey context (Chapters 1 and 5); expectation of the product of two misclassification indicators (Chapter 3)
$\boldsymbol{\varphi}$	regression coefficients in covariate misclassification process
$\boldsymbol{\nu}$	second-order association parameters for response misclassification process
$\boldsymbol{\eta}$	vector of all nuisance parameters
ϵ	random error in linear and linear mixed models (Chapters 1 and 2)
$\boldsymbol{\epsilon}$	residual vector in estimating functions (Chapters 1, 3, and 4)
e	measurement error
$h(\cdot)$	link function in a measurement error model
σ^2	variance of a continuous random variable
ρ	correlation
ψ	odds ratio for binary responses (Chapter 3); global odds ratio for ordinal responses (Chapter 4)
λ	odds ratio for misclassifications (Chapter 3); cumulative probability of an ordinal response (Chapter 4)
$\mathbf{u}$	a set of covariates involved in the second-order association model
ς	bivariate cumulative probability of two ordinal responses
ϕ	parameters involved in a dependence model
δ	indicator for observations in the validation subsample
$\mathcal{L}$	likelihood function
ℓ	log-likelihood function
$\mathrm{I}(\cdot)$	indicator function
$\mathbf{U}$	estimating function of $\boldsymbol{\theta}$
$\mathbf{Q}$	estimating function of $\boldsymbol{\eta}$
$\mathbf{D}$	derivatives of marginal means
$\mathbf{V}$	covariance matrix
$\mathbf{R}$	correlation matrix

$\mathbf{B}$	diagonal matrix with entries given by marginal variances
$\mathcal{I}$	Fisher information for $\boldsymbol{\theta}$
$\mathcal{J}$	Fisher information for $\boldsymbol{\eta}$
$\boldsymbol{\Gamma}$	$\mathrm{E}\left[\partial \mathbf{U}^{\mathrm{T}} / \partial \boldsymbol{\theta}\right]$
$\boldsymbol{\Sigma}$	$\mathrm{E}\left[\mathbf{U}\mathbf{U}^{\mathrm{T}}\right]$
$\mathbf{M}$	empirical version for $\mathrm{E}\left[\partial \mathbf{U}^{\mathrm{T}} / \partial \boldsymbol{\theta}\right]$
$\mathbf{J}$	empirical version for $\mathrm{E}\left[\partial \mathbf{Q}^{\mathrm{T}} / \partial \boldsymbol{\eta}\right]$
$\boldsymbol{\Omega}$	a variant of $\mathbf{U}$ that accounts for the uncertainty in estimated $\boldsymbol{\eta}$ (Chapters 3 and 4)
$\mathbf{P}$	classification probability matrix for a categorical response variable
$\mathbf{G}$	classification probability matrix for a categorical covariate (Chapter 4)
A	intermediate quantity in an approximate likelihood (Chapter 2); intermediate quantity in a replication study for misclassified responses (Chapter 3)
π	the ratio of the circumference of a circle to its diameter (Chapter 2); probability in the misclassification process for a covariate (Chapters 1, 4 and 5)
t	value of a Gaussian quadrature point (Chapter 2); index for iterations of an algorithm (Chapters 3-5)
w	weight of a Gaussian quadrature point
N	size of the finite population
s	sample from a complex survey
p	subscript indicating pseudo likelihood (Chapter 2); sampling scheme (Chapters 1 and 5)
d	number of replicates (Chapters 2 and 3); survey weight (Chapters 1 and 5)

Chapter 1

Introduction

1.1 Overview

The fundamental task for many epidemiological studies is to investigate the relationship between a set of predictor variables (covariates) and a particular outcome variable (response), which can be either continuous or discrete. Statistical models are often used to characterize the effects of the covariates on the response. These models involve parameters that are of scientific interest, and inference about the parameters is often the main goal for statistical analysts. To do so, observations are often assumed independent, for which regression models such as linear models or generalized linear models (GLMs) can be employed.

Correlated data arise from many epidemiological studies, especially clustered studies, in which data are collected on members within a cluster, and longitudinal studies, in which measurements are collected on the same subject repeatedly over time. For example, members from a familial pedigree are genetically related, and their health conditions are typically correlated. Some longitudinal studies are designed to investigate how a characteristic changes over time. Rigorously controlled experiments such as prospective randomized single-center and multi-center clinical trials are often involved (Hedeker and Gibbons, 2006). In medical studies, the measurement might be blood pressure, cholesterol level, lung volume, or serum glucose (Laird and Ware, 1982). Multiple measurements may be obtained from each individual at regularly or irregularly spaced measurement occasions and possibly under changing experimental conditions. For technical convenience, longitudinal data may be thought of as a special kind of clustered data by treating a subject as a cluster so that available statistical tools for analysis of clustered data can also be applied to longitudinal data (Song, 2007). Unlike the univariate case, the correlation among repeated measurements must be accounted for when analyzing data from these studies in order to make valid inferences. Many models have been developed to take into account the correlation, and various estimation methods have been proposed. These models can be roughly divided into two broad classes: conditional models (e.g., random effects or mixed models, transition models), and marginal models.

Variables are often assumed to be perfectly measured when we apply standard statistical tools. In reality, however, data collected from observational studies and surveys are often far from perfect, as *measurement error* may arise from many sources. For example, ambiguous words in a badly designed survey questionnaire may lead to incorrect interpretations of the respondents. When a diagnostic test for a particular disease is not gold standard, we may obtain a false positive or false negative result for the infection status. In some studies, variables cannot be precisely measured, although rough surrogate variables may be obtained.

In this thesis, we develop inferential procedures for analyzing correlated data with response measurement error. We consider three scenarios: (i) likelihood-based inferences for generalized linear mixed models when the continuous response is subject to nonlinear measurement errors; (ii) estimating equations methods for binary responses with misclassifications; and (iii) estimating equations methods for ordinal responses when the response variable and categorical/ordinal covariates are subject to misclassifications.

1.2 Methods for Analysis of Longitudinal and Clustered Data

Suppose data contain n independent clusters. Let Y_{ij} denote the response for the jth observation in cluster i, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Let $\mathbf{X}_{ij}$ denote a vector of covariates whose effects are of interest. If observations are independent of each other, the data can be fitted using a GLM given by

$$g(\mu_{ij}) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1.1}$$

where $\mu_{ij} = \mathrm{E}[Y_{ij}|\mathbf{X}_{ij}]$, $g(\cdot)$ is a link function that relates μ_{ij} to the linear predictor, and $\boldsymbol{\beta}$ is a vector of regression parameters quantifying the covariate effects. The link function $g(\cdot)$ is monotone and differentiable. For continuous responses, the link function is usually the identity function $g(u) = u$. For binary responses, commonly used link functions include the logit link $g(u) = \log\{u/(1-u)\}$, the complementary log-log link $g(u) = \log\{-\log(1-u)\}$, and the probit link $g(u) = \Phi^{-1}(u)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable. In the presence of within-cluster associations, however, GLMs are no longer good solutions. In this section, we provide an overview of three general approaches to the analysis of clustered/longitudinal data.

1.2.1 Mixed models

A flexible class of mixed models can be applied to normally distributed continuous outcomes, categorical outcomes, and other non-normally distributed outcomes such as counts. They are often used in studies where we cannot fully control the circumstances under which measurements are taken. Because of the considerable variation among clusters, data from these studies can be analyzed using some variant of a two-stage model. The joint probability distribution of the repeated measurements has the same form for each cluster, but a portion of the parameters may vary across clusters. These parameters, or “random effects”, have a certain distribution in the population that constitutes the second stage of the model.

For convenience, we use the term "cluster" to represent the independent unit in both clustered studies and longitudinal studies. As a result, a cluster may refer to a family, in which observations on all members are collected, or a subject, on which repeated measures are collected over time. A generalized linear mixed model (GLMM) has the form

$$g(\mu_{ij}^b) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i, \tag{1.2}$$

where $\boldsymbol{\beta}$ is a vector of fixed-effects parameters, $\mathbf{b}_i$ is a vector of random effects associated with covariates $\mathbf{Z}_{ij}$ (usually part of $\mathbf{X}_{ij}$), and $\mu_{ij}^b = \mathrm{E}[Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i]$ is the conditional mean of the response. Here an implicit assumption $\mathrm{E}[Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i] = \mathrm{E}[Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i]$ is often made (Pepe and Anderson, 1994). The vector of random effects $\mathbf{b}_i$ follows a certain distribution, say, $f(\mathbf{b}_i)$ with variance $\boldsymbol{\sigma}_{\mathbf{b}}$. The link function $g(\cdot)$ relates μ_{ij}^b to the linear predictor. The main task of statistical inference is to estimate the response parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\sigma}_b^2)^{\mathrm{T}}$, with primary interest in $\boldsymbol{\beta}$ (McCulloch and Searle, 2001). When Y_{ij} is continuous and $g(\cdot)$ is the identity function, a linear mixed model (LMM) is given by

$$Y_{ij} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i + \epsilon_{ij}, \tag{1.3}$$

where the random error ϵ_{ij} is often assumed to follow the normal distribution with mean 0 and variance σ_ϵ^2. For example, in a study of changes in lung volume during childhood (Laird and Ware, 1982), conditions related to the growth of the children change over time, contributing to variation of lung volume among individuals. It is then reasonable to assume that the relationship between lung volume and the cube of height is linear but the regression parameters may vary among children.

A key feature distinguishing mixed models from usual regression models is that the subject-specific random effects are unobserved components. A straightforward strategy for the estimation of $\boldsymbol{\beta}$ and the parameters specifying the distribution of the random effects is to use maximum likelihood (ML) method based on the marginal distribution of the observations. However, the likelihood function involves integration over the random components and is not in a closed form for most cases of GLMMs.

Some authors proposed iterative algorithms for computing the ML estimates or restricted maximum likelihood (REML) estimates in LMMs with normal variance components (e.g., Harville, 1977; Fellner, 1986). Schall (1991) adapted the algorithm of Harville (1977) to yield approximate ML or REML estimates in GLMMs. These models assume that the random effects are independent of the covariates in standard applications, e.g., the example of analyzing the effect of air pollutants on pulmonary function development in children considered by Laird and Ware (1982). However, Neuhaus and McCulloch (2006) showed that when the random effects are correlated with one of the covariates, naively fitting a GLMM ignoring this correlation leads to inconsistent estimators. The authors proposed conditional ML method that partitions the covariate into between- and within-cluster components to reduce bias. Mixed models are full likelihood-based and can easily handle both time-invariant and time-varying covariates (Hedeker and Gibbons, 2006). Therefore, they are among the most widely used methods for analysis of clustered or longitudinal data.

1.2.2 Marginal models

Marginal approaches have been widely used in longitudinal and familial studies focusing on the population-averaged dependence of the responses on the covariates. A link function is specified to connect the marginal expectation of a response to the linear predictor without conditioning on the other outcomes or unobserved random components, as opposed to conditional models (e.g., transition models, and mixed models). Marginal models generally do not impose a full parametric assumption for the joint distribution of the multivariate responses. Instead, least assumptions on the first and second moments of the responses are made. In a landmark paper, Liang and Zeger (1986) introduced the generalized estimating equations (GEE) approach for analyzing longitudinal data, in which the mean parameters are of primary interest while the association between outcomes is considered as a nuisance characteristic. As a result, the GEE approach models the marginal mean of the responses assuming a common correlation structure across all clusters. The parameters associated with the "working" correlation structure can be estimated from Pearson residuals via the method of moments.

Let $\mu_{ij} = \mathrm{E}[Y_{ij}|\mathbf{X}_i]$ $(j = 1, \ldots, m_i;\ i = 1, \ldots, n)$ be the marginal mean of the response given the covariates. Marginal models specify the relationship between μ_{ij} and the covariate effects in the form of a GLM given by (1.1). The mean parameters now have different interpretations than those in mixed models. Again, $\mathrm{E}[Y_{ij}|\mathbf{X}_i] = \mathrm{E}[Y_{ij}|\mathbf{X}_{ij}]$ is often assumed (Pepe and Anderson, 1994). Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$. Define

$$\mathbf{U}_{1i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\boldsymbol{\epsilon}_{1i},$$

where $\boldsymbol{\epsilon}_{1i} = \mathbf{Y}_i - \boldsymbol{\mu}_i$, $\mathbf{D}_{1i} = \partial\boldsymbol{\mu}_i^{\mathrm{T}}/\partial\boldsymbol{\beta}$, $\mathbf{V}_{1i} = \mathbf{B}_{1i}^{1/2}\mathbf{R}_{1i}(\boldsymbol{\alpha})\mathbf{B}_{1i}^{1/2}$, $\mathbf{B}_{1i} = \mathrm{diag}(v_{i1}, \ldots, v_{im_i})$, $v_{ij} = \mathrm{var}(Y_{ij}|\mathbf{X}_{ij})$ is the marginal variance of Y_{ij}, and $\mathbf{R}_{1i}(\boldsymbol{\alpha})$ is a working correlation matrix for $\mathbf{Y}_i$ parameterized by $\boldsymbol{\alpha}$. The GEE approach estimates $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^{n} \mathbf{U}_{1i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0}. \tag{1.4}$$

Here the correlation parameters $\boldsymbol{\alpha}$ are treated as nuisance parameters. By assuming a common correlation structure (e.g., independent, exchangeable, AR(1), or unspecified), $\boldsymbol{\alpha}$ can be estimated from Pearson residuals $(Y_{ij} - \mu_{ij})/\sqrt{\mu_{ij}(1 - \mu_{ij})}$ via the method of moments given $\boldsymbol{\beta}$ (Liang and Zeger, 1986). The GEE estimate of $\boldsymbol{\beta}$ is essentially a multivariate analog of the quasi-score function estimate based on quasi-likelihood method. Estimation can be carried out using the iterative Fisher scoring algorithm. An advantage of the GEE approach is that inference of $\boldsymbol{\beta}$ is robust against misspecification of $\mathbf{R}_{1i}(\boldsymbol{\alpha})$ for large sample size n. If $\mathbf{R}_{1i}(\boldsymbol{\alpha})$ is approximately correct, i.e., $\mathbf{R}_{1i}(\boldsymbol{\alpha}) \approx \mathrm{corr}(\mathbf{Y}_i|\mathbf{X}_i)$, solving equation (1.4) yields efficient estimate of $\boldsymbol{\beta}$. Even if the correlation structure is misspecified, the GEE method still yields a consistent estimator for $\boldsymbol{\beta}$ with some loss of efficiency (Crowder, 1995, 2001).

Many authors have studied the estimation of the correlation matrix (e.g., Prentice, 1988; Liang et al., 1992; Chaganty, 1997). Prentice (1988) suggested that the correlation among clustered binary responses may also be of scientific interest and proposed additional second-order estimating equations for the association parameters. This approach allows one to model the pairwise correlations and can improve the efficiency of the estimation of response probability regression parameters. Let $C_{ijj'} = Y_{ij}Y_{ij'}$ for

$j < j'$ and $\mathbf{C}_i = (C_{ijj'}, j < j')^{\mathrm{T}}$. Let $\mu_{ijj'} = \mathrm{E}[C_{ijj'}|\mathbf{X}_i]$ and $\boldsymbol{\xi}_i = (\mu_{ijj'}, j < j')^{\mathrm{T}}$. The first and second order estimating equations used by Prentice (1988) for binary data to simultaneously model $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are

$$\sum_{i=1}^{n} \mathbf{U}_{1i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\boldsymbol{\epsilon}_{1i} \quad = \quad \mathbf{0}, \tag{1.5}$$

$$\sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\boldsymbol{\epsilon}_{2i} \quad = \quad \mathbf{0}, \tag{1.6}$$

where $\boldsymbol{\epsilon}_{2i} = \mathbf{C}_i - \boldsymbol{\xi}_i$, $\mathbf{D}_{2i} = \partial \boldsymbol{\xi}_i^{\mathrm{T}} / \partial \boldsymbol{\alpha}$, and $\mathbf{V}_{2i}$ is a working covariance matrix for $\mathbf{C}_i$. Often, $\mathbf{V}_{2i} = \mathrm{diag}\{\mu_{ijj'}(1 - \mu_{ijj'}), j < j'\}$ is assumed in order to avoid modeling third and higher moments of the responses. Here $\mathbf{V}_{1i}$ is the covariance matrix rather than just a working covariance matrix for $\mathbf{Y}_i$, which if different from that in the GEE approach.

To allow higher-order associations, Zhao and Prentice (1990) considered reparametrization of a quadratic exponential model for correlated binary data in terms of marginal mean parameters and correlations and proposed pseudo-ML estimation procedures for these parameters. Because of desirable properties and easier interpretation, odds ratio is commonly used by investigators as a measure of association between paired binary responses. For instance, Lipsitz et al. (1991) modified the moment-based estimating equations of Prentice (1988) by modeling the pairwise association with the odds ratio. They showed through simulations that the marginal parameter estimates for the logistic regression model appear slightly more efficient when using the odds ratio parametrization. Similarly, Fitzmaurice and Laird (1993) discussed likelihood-based methods for analyzing longitudinal binary data using odds-ratio representation, extending the approach of Zhao and Prentice (1990) under quadratic exponential family. The procedure of Prentice (1988), which uses cross-products for association presentation, can become computationally infeasible as the cluster size gets large. Carey et al. (1993) proposed the alternating logistic regressions (ALR) approach for simultaneously regressing the response on explanatory variables as well as modeling second-order associations in terms of pairwise odds ratios. For $j < j'$, let $\xi_{ijj'} = \mathrm{E}[Y_{ij}|Y_{ij'} = y_{ij'}]$ be the conditional expectation given by (Diggle, 1992)

$$\xi_{ijj'} \quad = \quad \mathrm{logit}^{-1}\left\{(\log \psi_{ijj'})y_{ij'} + \log\left(\frac{\mu_{ij} - \mu_{ijj'}}{1 - \mu_{ij} - \mu_{ij'} + \mu_{ijj'}}\right)\right\}.$$

Let $\dot{\boldsymbol{\xi}}_i = (\xi_{ijj'}, j < j')^{\mathrm{T}}$. The set of first-order estimating equations from the ALR model has the same form of (1.5), but the set of second-order estimating equations is given by

$$\sum_{i=1}^{n} \dot{\mathbf{U}}_{2i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \dot{\mathbf{D}}_{2i} \dot{\mathbf{V}}_{2i}^{-1} \dot{\boldsymbol{\epsilon}}_{2i}, \tag{1.7}$$

where $\dot{\boldsymbol{\epsilon}}_{2i}$ is a residual vector with components given by $\dot{\epsilon}_{ijj'} = Y_{ij} - \xi_{ijj'}$, $\dot{\mathbf{D}}_{2i} = \partial \dot{\boldsymbol{\xi}}_i^{\mathrm{T}} / \partial \boldsymbol{\alpha}$, and $\dot{\mathbf{V}}_{2i} = \mathrm{diag}\{\xi_{ijj'}(1 - \xi_{ijj'}), j < j'\}$ is a working covariance matrix.

The employment of additional estimating equations for association parameters can improve efficiency of the estimators for the mean parameters, provided that the second-order association structure is modeled correctly. However, Sutradhar and Das (1999) indicated that estimates of mean parameters obtained under a working independence assumption are sometimes more efficient than those with a misspecified non-diagonal working correlation structure.

1.2.3 Transition models

Transition models focus on conditional regression parameters rather than marginal mean parameters. They are typically used for analysis of longitudinal binary and categorical data by incorporating both the covariates effects and the dependence on previous outcomes. A stochastic model for analysis of serial binary data was introduced by Azzalini (1994), which models the influence of covariates on current response by a marginal regression but separately characterizes the serial dependence by a first-order Markov association. A first-order Markov model assumes that the current response variable is dependent on the history only through the immediate previous response. Heagerty and Zeger (2000) described a class of marginalized models, which specifies a conditional model for the underlying process of data generation but permits estimation of marginal mean parameters. A likelihood-based method for analysis of binary serial data was proposed by Heagerty (2002), who generalized the model of Azzalini (1994) to a broad class of marginalized transition models (MTM) that permits marginal regression analysis and allows a general pth-order dependence structure (Chen et al., 2009).

Suppose we have binary response data $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})$ observed on subject i at occasions $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. A marginal generalized linear model specifies

$$g(\mu_{ij}^M) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},$$

where $\mu_{ij}^M = \mathrm{E}[Y_{ij}|\mathbf{X}_i]$ is the marginal mean of the response, and $\boldsymbol{\beta}$ is a vector of coefficients quantifying the effects of the covariates $\mathbf{X}_{ij}$ on average response. A first-order Markov model describes the dependence of the current outcome on the immediate previous outcome through transition probabilities $p_{ij,1} = \Pr(Y_{ij} = 1|Y_{i,j-1} = 1)$ and $p_{ij,0} = \Pr(Y_{ij} = 1|Y_{i,j-1} = 0)$. Therefore, it can be seen that the first-order Markov model of Azzalini (1994) is a two-stage model. First, a marginal mean regression model can be structured as

$$\mu_{ij}^M = p_{ij,1}\mu_{i,j-1}^M + p_{ij,0}\left(1 - \mu_{i,j-1}^M\right).$$

Second, the transition probabilities are modeled using odds ratio

$$\psi_{ij} = \frac{p_{ij,1}/(1 - p_{ij,1})}{p_{ij,0}/(1 - p_{ij,0})},$$

which measures the strength of the serial dependence (Azzalini, 1994). Heagerty and Zeger (2000) described the dependence using the conditional expectation $\mu_{ij}^C = \mathrm{E}[Y_{ij}|Y_{i,j-1}, \mathbf{X}_i]$ under a logit model

$$\mathrm{logit}(\mu_{ij}^C) = \Delta_{ij} + \phi_{ij,1}Y_{i,j-1},$$

where regression coefficient $\phi_{ij,1} = \log\psi_{ij}$ is the log odds ratio and is dependent on both $\mathbf{X}_{ij}$ and $Y_{i,j-1}$. The intercept Δ_{ij} in the model can be shown to be equal to $\mathrm{logit}(p_{ij,0})$ and is determined by $\boldsymbol{\beta}$ and $\phi_{ij,1}$. Furthermore, a linear regression model can be specified,

$$\phi_{ij,1} = \mathbf{u}_{ij,1}^{\mathrm{T}} \boldsymbol{\alpha}_1,$$

where the parameter $\boldsymbol{\alpha}_1$ determines how the dependence of $\phi_{ij,1}$ on $Y_{i,j-1}$ varies as a function of a set of covariates $\mathbf{u}_{ij,1}$. For a pth-order dependence model, MTM(p), the logit-linear model for conditional expectation $\mu_{ij}^C = \mathrm{E}[Y_{ij}|\mathbf{X}_i, Y_{i,j-1}, \ldots, Y_{i,j-p}]$ is given by

$$\begin{aligned} \mathrm{logit}(\mu_{ij}^C) &= \Delta_{ij} + \sum_{k=1}^{p} \phi_{ij,k} Y_{i,j-k}, \\ \phi_{ij,k} &= \mathbf{u}_{ij,k}^{\mathrm{T}} \boldsymbol{\alpha}_k, \quad j = 1, \ldots, p, \end{aligned}$$

where the serial dependence is modeled in an additive form (Heagerty, 2002).

1.3 Measurement Error/Misclassification

Measurement error has been a longstanding concern in epidemiological studies. When referring to a categorical variable, it is termed *misclassification*. Variables obtained from self-report questionnaires are known to contain error, e.g., dietary intake, and nutrition consumption, among others. Self-report bias is one of the major sources of measurement error in data from surveys. Other examples of measurement errors include many variables of medical interests, such as exposures to indoor or outdoor pollutants, nutrition or drug intakes.

When covariates in the statistical models are subject to error, naive estimators for model parameters are often inconsistent; see, for instance, Fuller (1987), Cook and Stefanski (1994), and Prentice (1982), among others. On the other hand, measurement errors may also exist in responses. One typical example is the long-term average of systolic blood pressure, as it cannot be precisely measured with a single reading. When a diagnostic test for a particular disease is not gold standard or the measuring device is defective, the binary outcome may also contain misclassification. Much of the research interest in this area has been focused on measurement error in covariates, particularly in continuous covariates. A large body of literature on methodology can be found to be related to this problem, e.g., Cook and Stefanski (1994), Wang et al. (1998), Suh and Schafer (2002) and Yi and Cook (2005). Error in response, however, has received relatively less attention. Some contributions include Neuhaus (1999, 2002), who studied estimation bias and inefficiency due to misclassification in binary responses, and Buonaccorsi (1996), who discussed nonlinear measurement error in a continuous response variable.

In this section, we give a short introduction to bias analysis for independent data with measurement error in a covariate. We also outline some statistical approaches to correcting the bias induced by covariate measurement error. A brief review of the literature on response measurement error is also given.

1.3.1 Measurement error in a continuous covariate

Measurement error in continuous covariates have been discussed extensively under GLMs, see, e.g., Carroll et al. (1984), Stefanski and Buzas (1995), among others. To develop methods for eliminating or reducing bias induced by measurement error, we must make some basic assumptions for the measurement error process. Different measurement error mechanisms lead to different approaches to bias correction. The literature distinguishes between functional modeling, which does not impose any distributional assumption on the true error-prone covariates, and structural modeling, which hypothesizes a distributional structure for those covariates (e.g., Wang et al., 1998; Gustafson, 2004).

Let Y_i be the response for subject i, $i = 1, \ldots, n$. Let X_i be a continuous covariate subject to measurement error and $\mathbf{Z}_i$ be a vector of precisely measured covariates. The expectation $\mu_i = \mathrm{E}[Y_i | X_i, \mathbf{Z}_i]$ is related to the covariates in a GLM

$$g(\mu_i) = X_i \beta_x + \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta}_z,$$

where β_x and $\boldsymbol{\beta}_z$ are regression parameters associated with the effects of X_i and $\mathbf{Z}_i$, respectively.

Instead of observing the true value of X_i, we observe an error-contaminated surrogate version W_i. There are two ways of characterizing the relationship between X_i and W_i: one models the dependence of X_i on W_i, and the other models the dependence of W_i on X_i, given other variables. Much of the research focuses on a *classical* additive measurement error model: given Y_i and $\mathbf{Z}_i$,

$$W_i = X_i + e_i, \tag{1.8}$$

where e_i follows a distribution with mean 0 and variance σ_e^2, e.g., a normal distribution, and is often assumed to be independent of X_i. In some other cases, it is more reasonable to assume that the measurement error process follows

$$X_i = W_i + e_i. \tag{1.9}$$

This is called the *Berkson* measurement error model (Berkson, 1950), in which the realization of the surrogate W_i comes before that of X_i. Berkson error may predominate over classical error in exposure assessment in some epidemiological studies. For example, a person's actual exposure to indoor air pollutant may be unobserved, but the air pollutant in that person's neighborhood is measured. Therefore, Berkson error model fits this kind of error structure, as the indoor pollutant level depends on the outdoor pollutants.

Here we demonstrate the impact of a mismeasured continuous covariate on the estimates of regression coefficients through an example used by Yi (2007). Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_x X_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $X_i \sim \text{Normal}(\mu_x, \sigma_x^2)$ and $\epsilon_i \sim \text{Normal}(0, \sigma_\epsilon^2)$. Let the measurement error process for X_i follow the classical additive model (1.8) with $e_i \sim \text{Normal}(0, \sigma_e^2)$. Naively fitting a linear model to the observed data $\{(Y_i, W_i);\ i = 1, \ldots, n\}$ leads to a misspecified model

$$Y_i = \beta_0^* + \beta_x^* W_i + \epsilon_i^*,$$

where β_0^* and β_x^* are regression coefficients under the false model, and ϵ_i^* is assumed to follow a normal distribution with mean 0 and variance σ_ϵ^{*2}. Let $\bar{Y} = \sum_{i=1}^n Y_i/n$, $\bar{X} = \sum_{i=1}^n X_i/n$, $\bar{W} = \sum_{i=1}^n W_i/n$, $\bar{\epsilon} = \sum_{i=1}^n \epsilon_i/n$, and $\bar{e} = \sum_{i=1}^n e_i/n$. The naive least squares estimator for β_x is given by

$$\hat{\beta}_x^* = \frac{\sum_{i=1}^n (W_i - \bar{W})(Y_i - \bar{Y})}{\sum_{i=1}^n (W_i - \bar{W})^2}.$$

With some algebra, we have

$$\begin{aligned}
\hat{\beta}_x^* &= \frac{\sum_{i=1}^n (W_i - \bar{W})\{\beta_x(X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon})\}}{\sum_{i=1}^n (W_i - \bar{W})^2} \\
&= \beta_x \frac{\sum_{i=1}^n (W_i - \bar{W})(X_i - \bar{X})}{\sum_{i=1}^n (W_i - \bar{W})^2} + \frac{\sum_{i=1}^n (W_i - \bar{W})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (W_i - \bar{W})^2} \\
&= \beta_x \frac{\sum_{i=1}^n (X_i - \bar{X} + e_i - \bar{e})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X} + e_i - \bar{e})^2} + \frac{\sum_{i=1}^n (W_i - \bar{W})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (W_i - \bar{W})^2} \\
&= \beta_x \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (X_i - \bar{X})(e_i - \bar{e})}{\sum_{i=1}^n (X_i - \bar{X})^2 + 2\sum_{i=1}^n (X_i - \bar{X})(e_i - \bar{e}) + \sum_{i=1}^n (e_i - \bar{e})^2} \\
&\quad + \frac{\sum_{i=1}^n (W_i - \bar{W})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (W_i - \bar{W})^2} \\
&\xrightarrow{p} \beta_x \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} \right), \qquad \text{as } n \to \infty,
\end{aligned}$$

where the convergence in probability is based on the assumptions of independence between X_i and e_i and independence between W_i and ϵ_i. Therefore, the naive analysis leads to attenuated estimate of the regression coefficient associated with the mismeasured covariate, and the attenuation increases as the variance of the measurement error increases.

Unlike classical additive error, Berkson error causes little or no bias in the estimates of regression coefficients, as the measurement error e_i is simply absorbed into ϵ_i in the response model. That is,

$$\begin{aligned}
Y_i &= \beta_0 + \beta_x W_i + (\beta_x e_i + \epsilon_i) \\
&= \beta_0 + \beta_x W_i + \epsilon_i^*,
\end{aligned}$$

where ϵ_i^* has variance $(\beta_x^2\sigma_e^2 + \sigma_\epsilon^2)$. Because of the inflation of the random error variance, Berkson error reduces the power of a study.

1.3.2 Misclassification in a categorical covariate

Categorical variables are frequently included in covariates in regression analyses. A categorical variable is said to be subject to misclassification if the recorded category may differ from the true category. Unlike the continuous case, the surrogate categorical variable now cannot be expressed as a sum of the true plus a noise variable. Similar to the difference between classical model and Berkson model for continuous measurement error, Spiegelman et al. (2000) distinguished a reclassification model from a misclassification model for categorical variables. The reclassification model specifies the distribution of the true category given the observed category, the form of which needs to be identified and empirically verified using validation data. An example considered by the authors is the estimation of effect of high saturated fat intake on the risk of breast cancer using data from the Women's Health Initiative (Prentice et al., 1988), where average daily saturated fat intake is dichotomized with cutoff $\leq$ 30g/day. The binary variable for high saturated fat intake contains misclassification, because it is difficult to measure individual long-term average diet, components of which are the exposures in the regression analysis.

In this thesis, we only consider the classical-type misclassification. A misclassification process is modeled in terms of (mis)classification probabilities. These probabilities describe that given the true category and the precisely measured covariates, how likely we observe the recorded category. Let X_i be a categorical covariate with $(K + 1)$ levels taking values $0, \ldots, K$, and let W_i be a surrogate for X_i. Let $\pi_{iqr} = \Pr(W_i = r|X_i = q, \mathbf{Z}_i)$ be the probability that the recorded category is r when the true category is q, $q, r = 0, \ldots, K$. Regression models such as generalized logit models can be employed to characterize the dependence of the misclassification process on $\mathbf{Z}_i$.

Misclassification is known to induce bias in the effect estimates in regression models (Gustafson, 2004). Some papers dealing with misclassified covariates in epidemiologic studies are available in the literature (see, e.g., Greenland, 1980, 1982, 1988, 2008; Rosner, 1996). Christopher and Kupper (1995), Veierød and Laake (2001) and Paulino et al. (2003) assessed the bias results in linear regression, Poisson regression and binomial regression with misclassification. Buonaccorsi et al. (2005) investigated the impact of misclassification of a categorical covariate X on the estimates of the coefficients associated with the precisely measured covariates $\mathbf{Z}$. They showed that as long as the error-prone X is correlated with $\mathbf{Z}$, the naive estimates of the coefficients of $\mathbf{Z}$ are biased even when the misclassification process is independent of $\mathbf{Z}$.

1.3.3 Approaches for handling covariate error

There have been numerous methods for correcting bias induced by covariate measurement error. These approaches sometimes are referred to as functional methods and structural methods based on whether distributional assumptions are made for the mismeasured covariates. Functional modeling, which does not specify the structures of the error-prone covariates, is appealing in situations where we do not have much knowledge about the behaviors of the covariates. Some popular functional methods include the corrected scores approach of Nakamura (1990, 1992), regression calibration, and the simulation-extrapolation (SIMEX) approach originally proposed by Cook and Stefanski (1994). Structural methods come into play when it is necessary to specify a marginal distribution for the error-prone covariates, or it is of interest to study their marginal behaviors. However, concerns may arise that the resulting estimates and inferences may depend upon the parametric models chosen.

Likelihood-based methods

To perform likelihood-based analysis, full modeling assumption is usually required for every key component of the data. Suppose the probability density function of response Y_i is given by $f_{Y|X,Z}(Y_i|X_i, \mathbf{Z}_i)$ conditional on covariates $(X_i, \mathbf{Z}_i)$, and the density function of X_i conditional on $\mathbf{Z}_i$ is given by $f_{X|Z}(X_i|\mathbf{Z}_i)$. Often, the marginal distribution of the precisely measured $\mathbf{Z}_i$ is left unspecified. We also assume that the measurement error process is fully parameterized with probability density function

$f_{W|X,Z}(W_i|X_i, \mathbf{Z}_i)$. The observed-data likelihood is then given by

$$\begin{aligned}
&\prod_{i=1}^{n} f_{Y,W|Z}(Y_i, W_i|\mathbf{Z}_i) \\
&\quad \propto \prod_{i=1}^{n} \int f_{Y,W,X|Z}(Y_i, W_i, X_i|\mathbf{Z}_i) dX_i \\
&\quad = \prod_{i=1}^{n} \int f_{Y|W,X,Z}(Y_i|W_i, X_i, \mathbf{Z}_i) f_{W|X,Z}(W_i|X_i, \mathbf{Z}_i) f_{X|Z}(X_i|\mathbf{Z}_i) dX_i \\
&\quad = \prod_{i=1}^{n} \int f_{Y|X,Z}(Y_i|X_i, \mathbf{Z}_i) f_{W|X,Z}(W_i|X_i, \mathbf{Z}_i) f_{X|Z}(X_i|\mathbf{Z}_i) dX_i,
\end{aligned}$$

in which nondifferential measurement error mechanism is used. Here, nondifferential measurement error mechanism means that Y_i depends only on the true covariates $(X_i, \mathbf{Z}_i)$ but not on the observed surrogate W_i, given $(X_i, \mathbf{Z}_i)$.

Robustness to model assumptions is a concern for likelihood-based methods. In situations where the assumptions are proper, maximum likelihood estimators are generally more efficient compared to simpler methods (Carroll et al., 2006, p. 181). A major challenge for likelihood-based approaches is that they are usually computationally demanding.

Estimating equation method

We now describe estimating equation methods, in which only the mean and variance structures of the response are specified. An estimating function $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i)$ is called an unbiased estimating function of $\boldsymbol{\beta}$ if it satisfies

$$\mathrm{E}[\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i)] = \mathbf{0}, \quad i = 1, \ldots, n.$$

An unbiased estimating function leads to a consistent estimator for $\boldsymbol{\beta}$ under certain regularity conditions. That is, as $n \to \infty$, the solution $\hat{\boldsymbol{\beta}}$ to

$$\sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, Z_i) = \mathbf{0}$$

converges in probability to the true $\boldsymbol{\beta}$. When W_i is observed instead of X_i, the naive estimating function $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, W_i, Z_i)$ is no longer unbiased. However, if a modified version $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$ is unbiased under expectations conditional on true data $(Y_i, X_i, \mathbf{Z}_i)$, i.e.,

$$\mathrm{E}[\mathbf{U}^*(\boldsymbol{\beta}; Y_i, W_i, Z_i)] \quad = \quad \mathrm{E}_{Y,X,Z}[\mathrm{E}_{W|Y,X,Z}\left\{\mathbf{U}^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)\right\}] = \mathbf{0},$$

then solving

$$\sum_{i=1}^{n} \mathbf{U}^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i) = \mathbf{0}$$

still gives consistent estimator for $\boldsymbol{\beta}$ (Nakamura, 1990, 1992). It suffices to construct $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$ such that

$$\mathrm{E}_{W|Y,X,Z}[\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)] = \mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i). \tag{1.10}$$

That is, $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, W_i, \mathbf{Z}_i)$ is an unbiased estimator for $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, X_i, \mathbf{Z}_i)$ under conditional expectations given true data $(Y_i, X_i, \mathbf{Z}_i)$ and hence is called "corrected" estimating functions (or "corrected" score functions). "Corrected" score functions exist for some regression models in the GLM family such as Gaussian, Poisson, Gamma, inverse Gaussian and Wald regression model. For logistic regression model, however, a corrected score does not exist, although simulation based methods such as Monte Carlo averaging method can be used for constructing approximate versions (Novick and Stefanski, 2002).

The two types of methods described above take different approaches to modeling the measurement error process. Likelihood-based methods are representative examples of structural approaches, as they specify a full probability model for the underlying true covariate. They are widely used due to the consistency and high efficiency of the maximum likelihood estimators, as well as their good asymptotic properties. The estimating equation approach only models the measurement error structure but leaves the probability distribution of the error-prone variable completely unspecified. The SIMEX approach of Cook and Stefanski (1994) is another popular functional approach, which requires an additive measurement error model. It uses a re-sampling method to establish the relationship between the bias in the estimates of regression coefficients and the measurement error variance, and then extrapolate to the case where there is no measurement error. The implementation of the SIMEX approach is easy. However, it is computationally intensive (e.g., Stefanski and Cook, 1995; Wang et al., 1998).

Another relatively straightforward approach is the regression calibration, which imputes the underlying true values of the covariates using a calibration function and applies standard analysis tools to the imputed data. The calibration function and associated parameters, however, are often unknown and need to be estimated from validation data or replicates. Therefore, adjusting standard errors is required in order to account for the uncertainty in the estimated calibration function parameters, using either the bootstrap variance estimation or the sandwich method. Regression calibration is a very convenient way to reduce the bias induced by measurement error. However, the regression calibration model is only an approximate, working model for the observed data. It is typically used in ad hoc ways, simply as a modeling device and not based on any fundamental considerations such as classical or Berkson error model. When the model is highly non-linear, this method may not work well (Carroll et al., 2006).

Covariate measurement error in data from clustered and longitudinal studies has been considered by some authors (e.g., Prentice, 1986; Wang et al., 1998; Lin and Carroll, 1999). Wang and Davidian (1996) considered the influence of measurement error on variance component estimators in nonlinear mixed models. Wang et al. (1998) investigated the bias induced by classical additive error in a generalized linear mixed measurement error model (GLMMeM) and proposed to use SIMEX with the quadratic extrapolation function for estimation of the mixed model parameters. Buonaccorsi et al. (2000) considered the estimation of both regression coefficients and variance parameters for a class of linear mixed models with measurement error in a time-varying covariate. They found that regression calibration suitable and highly efficient for fixed-effects, because the fixed-effects and the variance components are orthogonal in the context of linear mixed models. The authors also showed that a "corrected regression calibration" method, which is equivalent to the pseudo-maximum likelihood approach, can be used to correct the bias of the estimates of the variance components. Xiao et al. (2010) considered measurement error in multiple covariates and obtained consistent estimators by extending the generalized method of moments (e.g., Griliches and Hausman, 1986; Wansbeek, 2001).

1.3.4 Response measurement error

Compared to the rich literature on covariate measurement error, response measurement error has received relatively less attention. In linear regression, classical measurement error in responses increases the variability of the estimated coefficients without causing bias (Carroll et al., 2006). Therefore, classical measurement error in responses is often ignored in linear regression analysis, as in part, it can be absorbed into the noise term of the response model. For nonlinear response measurement error, however, this does not apply. Buonaccorsi (1996) considered nonlinear response error in linear regression models and proposed the pseudo-maximum likelihood approach with an illustration of a four-parameter logistic measurement error structure. Yanez et al. (1998) presented a method of adjusting for response error in the modeling of association of a set of explanatory variables with the change of the outcome variable such as blood pressure. Moore et al. (2000) reviewed the sources of measurement error in income surveys.

Much of the research on response measurement error has been focused on binary and categorical cases, i.e., misclassifications. Some early works include Tenebein (1970, 1972) and Hochberg (1977) on studies of association in contingency tables with element misclassification using doubly sampled data. Here a doubly sampling scheme consists of two mechanisms: the observations in a larger sample are classified into a contingency table by an inexpensive but fallible method, while the units of a subsample are classified jointly by the fallible method and by some expensive but reliable method. Ekholm and Palmgren (1987) employed the GLM for analysis of doubly sampled data by considering the problem as misclassification in both the explanatory factor and the binary response. Chua and Fuller (1987) considered response error associated with self-reported categorical data from surveys. Bollinger and David (1997) used pseudo-maximum likelihood estimation methods for Food Stamp participation by incorporating demographic and economic covariates in models for underreporting and overreporting. Neuhaus (1999) examined the magnitude of bias and efficiency loss due to misclassification in binary regression with a single covariate and obtained some approximate bias-correction factor for regression parameter. Roy et al. (2005) developed likelihood-based analysis for the probit regression model with measurement error in covariates and classification error in binary responses.

Neuhaus (2002) studied the influence of response misclassification in generalized linear mixed models for analysis of data from clustered and longitudinal studies. The author showed that the class of GLMMs enjoy a closure property under misclassified responses analogous to single-response GLMs (that is, the resulting model still belongs to GLMMs but with a different link function), and the asymptotic relative efficiency of the naive estimates to error-free estimates can be obtained. Roy et al. (2009) considered multivariate probit models for correlated binary data with covariate and response errors. They proposed likelihood-based methods for reducing the induced bias in the marginal effects as well as the correlation parameters.

1.3.5 Identifiability

A general concern with measurement error problems is model identifiability. That is, whether it is possible or not to know the exact parameters if one actually had an infinite number of observations. When a problem is not identifiable, it means that a key piece of information is unavailable.

Identifiability generally depends on the form of the model and the assumptions made for the components in the model. Carroll et al. (2006) addressed this issue for likelihood-based approaches. In some nonlinear measurement error models, parameters associated with both the response model and the measurement error model may be identified without extra information, e.g., validation data or replication data. It is the nonlinearity in the model that makes identifiability possible. However, estimation without additional data is generally not practical for linear models with variables and measurement error that are normally or close to normally distributed (Carroll et al., 2006, p. 184).

Identifiability is also the major practical issue for misclassification problems, as misclassification probabilities are very weakly identified. That means a very large sample is often required in order to obtain stable estimates or achieve convergence of an algorithm. The difficulty to estimate with any precision carries over to estimation of the underlying risk function (Carroll et al., 2006, p. 347). If extra information is not available, misclassification parameters may be identified theoretically but not in a practical sense. Copas (1988) and Neuhaus (2002) stated that without additional data the best one can do is to conduct sensitivity analysis for possible values of the misclassification probabilities.

To get around the identifiability issue, it is often assumed that extra information is available in the form of validation data, multiple measurements, or instrument variables. Carroll and Wand (1991) described semiparametric estimation and inference in a logistic regression model with measurement error in the predictors, where a smaller validation data set is available in addition to the primary data set. Similarly, Lee and Sepanski (1995) introduced consistent methods for the estimation of linear and nonlinear regression models with measurement errors in variables in the presence of validation data. The methods allowed the measurement errors be correlated with the true explanatory variables in the model. Hu (2008) considered nonlinear models with a misclassified discrete explanatory variable that is also allowed to be correlated with other explanatory variables. The author provided a nonparametric approach to the problem of identification and estimation using instrumental variables, for which certain monotonicity restrictions may be required on the latent model.

In this thesis we do not focus on addressing the identifiability issues in measurement error problems. Instead, in each chapter we first treat the error parameters as known and develop methods to correct the induced bias in the estimates of response parameters. Estimation of error parameters using possible additional information is then discussed.

1.4 Analysis of Survey Data

Surveys are an important and popular tool for collecting data. Analytical use of survey data especially health survey data has become more and more common, with focus on the association of particular outcome variables with explanatory variables at the population level. Estimating equation methods have been widely used, and their statistical properties have been studied by some authors, see, e.g., Godambe and Thompson (1986) and Binder and Patak (1994), among others.

Let N be the size of a finite population. Let Y_i and $\mathbf{X}_i$ be the response variable and a vector of auxiliary variables for individual i, $i = 1, \ldots, N$. We assume that the finite population $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ is generated from a superpopulation model ζ, which involves a vector of parameters $\boldsymbol{\theta}$. The finite population parameter, noted as $\boldsymbol{\theta}_N$, can be regarded as the solution to the population (or "census") estimating equations

$$\sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i) = \mathbf{0}, \tag{1.11}$$

where $\mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i)$ are unbiased estimating functions of $\boldsymbol{\theta}$. Here, unbiasedness means

$$\mathrm{E}_\zeta[\mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i] = \mathbf{0}, \tag{1.12}$$

with E_ζ denoting expectation under the superpopulation model ζ (Godambe and Thompson, 1986). Let $\mu_i = \mathrm{E}_\zeta[Y_i|\mathbf{X}_i]$. Different choices of $\mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i)$ in (1.11) lead to different population characteristics. For example, $\mathbf{U}_i(\boldsymbol{\theta}; Y_i) = Y_i - \boldsymbol{\theta}$ gives the population mean $\boldsymbol{\theta}_N = (1/N)\sum_{i=1}^{N} Y_i$, $\mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i) = \mathbf{X}_i(Y_i - \mu_i)$ with $\mu_i = \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta})\{1 + \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta})\}^{-1}$ gives the logistic regression vector $\boldsymbol{\theta}_N$, and $\mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i) = \mathbf{X}_i(Y_i - \mu_i)$ with $\mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}$ gives the population regression vector $\boldsymbol{\theta}_N = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^{\mathrm{T}}$ (Rao et al., 2002). Under the superpopulation model, $\boldsymbol{\theta}_N$ can be viewed as an estimate of the model parameter $\boldsymbol{\theta}$.

The data of the entire finite population are not available unless a census is conducted. Let s be a sample of n individuals obtained from the finite population using a complex survey design p. Solving the sample based estimating equations

$$\sum_{i \in s} \mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i) = 0, \tag{1.13}$$

yield estimates of both $\boldsymbol{\theta}_N$ and $\boldsymbol{\theta}$ simultaneously, provided that the superpopulation model is correctly specified. Another approach is to incorporate the feature of the complex survey design and construct estimating functions unbiased for the population estimating function under the survey design p. It is natural to solve

$$\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i) = 0, \tag{1.14}$$

where d_i is the design weight for individual i. Without non-response issues, $d_i = 1/\Pr(i \in s)$. Note that

$$\mathrm{E}_p \left[\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i) \right] = \sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\theta}; Y_i, \mathbf{X}_i),$$

where E_p denotes expectation under the design p, the weighted sample estimating functions in (1.14) is design unbiased for population estimating functions, and the solution is consistent for the population parameter $\boldsymbol{\theta}_N$ even if the superpopulation model is misspecified. In this case, finite population parameter $\boldsymbol{\theta}_N$ is of interest (Godambe and Thompson, 1986).

A major problem with the estimation of regression parameters is that data collected from surveys often contain measurement error. One source of measurement error is that questionnaires may not be well designed. Another source of error, particularly for large scale surveys, comes from mistakes during the course of data recording, coding, and editing. For example, the weight of the respondent is reported in pounds but may be recorded as in kilograms. When measurement error is coupled with complex survey design features, it adds another degree of difficulty and requires development of new tools and alternative approaches. Analysis of survey data in the presence of measurement error has been discussed by some authors, e.g., Fuller (1987, 1995), with a focus on the estimation of population mean, total, and quantiles.

Regression analysis of survey data with measurement error, however, has received relatively less attention.

Here we use a simple example to illustrate the possibility of extending existing bias correction methods to the survey context. Consider a finite population generated from a superpopulation model

$$Y_i = X_i\beta + \epsilon_i, \quad i = 1, \ldots, N, \tag{1.15}$$

where ϵ_i are independently normally distributed with mean 0 and variance σ_ϵ^2. The objective is to simultaneously estimate β, namely the slope of superpopulation model, and β_N, the finite population slope, from a sample of n subjects. The original estimating equation can be given by

$$\sum_{i \in s} U_i(\beta; Y_i, X_i) = \sum_{i \in s} X_i(Y_i - X_i\beta) = 0.$$

When the observed surrogate W_i for X_i follows the classical additive error model (1.8), an unbiased estimating function of β is given by $U_i^*(\beta; Y_i, W_i) = (Y_i - \beta W_i)W_i + \beta\sigma_e^2$ (Nakamura, 1990). It can be shown that by incorporating survey weights, the solution to estimating equation

$$\sum_{i \in s} d_i U_i^*(\beta; Y_i, W_i) = 0,$$

is both model and design unbiased for β_N.

1.5 Data Sets

In this section we describe two data sets that are used in the following chapters of the thesis.

1.5.1 Framingham Heart Study

The Framingham Heart Study is a longitudinal investigation of the development of cardiovascular disease. The study began in 1948 and 5,209 subjects were initially enrolled in the original cohort. The cohort has been followed for morbidity and mortality, and participants have continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests. A total of 5124 second-generation (adult children of the original participants and the spouses of these adult children) were recruited into a second cohort in 1971 and participated in similar examinations. The objectives of the cohort study are to study the incidence and prevalence of cardiovascular disease and identify its constitutional and environmental risk factors, as well as to study the trend of the influence of the risk factors over time. Measurement error problems arising from the Framingham Heart Study have been discussed by many researchers, with a focus on error-in-covariate in statistical regression models. For example, Carroll et al. (1984) and Wang et al. (1998) considered relating the probability of developing coronary heart disease to some baseline risk factors including systolic blood pressure (SBP), a covariate treated as error-contaminated.

On the other hand, studying the risk factors for SBP measurements may also be of clinical interest. SBP and its discreet versions are used as outcome variables in this thesis, which are subject to measurement error.

1.5.2 Canadian Community Health Survey

The Canadian Community Health Survey (CCHS) is an ongoing large scale survey conducted by Statistics Canada. Cycle 3.1 in 2005 targets persons aged 12 years or older who live in private dwellings in the ten provinces and the three territories. Persons living on Indian Reserves or Crown lands, clientele of institutions, full-time members of the Canadian Armed Forces and residents of certain remote regions are excluded from the survey. The primary objectives of the survey are to provide estimates of health determinant, health status and health system utilization across Canada, and to gather data at the sub-provincial levels of geography (Statistics Canada, 2005).

For administrative purposes, each province is divided into health regions (HR) according to the types of regions: major urban centres, cities, and rural regions, and each territory is designated as a single HR. During Cycle 3.1 of the CCHS, data were collected in 122 HRs in the ten provinces, in addition to one HR per territory, totalling 125 HRs. Three sampling frames are used to select the sample of households: 49% of the sample of households came from an area frame, 50% came from a list frame of telephone numbers and the remaining 1% came from a Random Digit Dialling (RDD) sampling frame. The CCHS uses the area frame designed for the Canadian Labour Force Survey (LFS). The sampling plan of the LFS is a multistage stratified cluster design in which the dwelling is the final sampling unit. Geographic or socio-economic strata are created within each HR. Within the strata, between 150 and 250 dwellings are regrouped to create clusters. Some urban centres have separate strata for apartments or for census Enumeration Areas (EA) to pinpoint households with high income, immigrants and the native people. In each stratum, six clusters or residential buildings (sometimes 12 or 18 apartments) are chosen with probability proportional to size (PPS), with the number of households as the size variable. The list frame of telephone numbers was used in all but five HRs (the two RDD only HRs and the three territories) to complement the area frame. One list frame stratum was then created for each HR based on postal codes that were obtained from names, addresses and telephone numbers. Within each stratum the required number of telephone numbers was selected using simple random sampling from the list. As for the RDD frame, additional telephone numbers were selected to account for the numbers not in service or out-of-scope. The hit rate observed under the list frame approach varied from 75% to 88% depending on the province, which was much higher than that for the RDD frame. In four HRs, a Random Digit Dialling (RDD) sampling frame of telephone numbers was used to select the sample of households.

For all selected households, a single person aged 12 and older was randomly chosen from members of the household. After removing the out-of-scope units, 168,464 households were selected to participate in the CCHS Cycle 3.1. Data were obtained from 132947 respondents, yielding a response rate of 79%. Data were collected on general health, chronic health conditions, drinking or smoking status, including self-reported weight and height. A subsample of 7376 respondents aged 12 or older were also selected, who were asked later in the interview to directly measure weight and height. Among the 7376 individuals selected in the subsample, 4735 individuals responded. The main reason for non-response was refusal (Statistics Canada, 2005). Such validation subsample is useful in studies of risk factors for obesity as well as the effect of obesity on health conditions. It provides information on the relationship between a precise measurement and an error-contaminated measurement of weight or height that makes it possible to correct estimation bias induced by the self-reported data.

1.6 Outline of Thesis

The structure of the thesis is as follows. In Chapter 2, we consider the estimation of regression coefficients in a mixed model where the continuous response variable is subject to nonlinear measurement error. We first discuss the model formulation for the response process and the measurement error process. We then conduct bias analysis for a naive approach that completely ignores measurement error. We also investigate another naive approach, which fits mixed models to transformed data. Estimation and inference using likelihood-based methods are presented, and a two-stage pseudo likelihood approach is developed for cases where validation data are available. We conduct some simulation studies to investigate the performance of the proposed methods. Finally, a real data set from the Framingham Heart Study is analyzed.

In Chapter 3, we discuss the problem of misclassification in correlated binary responses arising from longitudinal studies or familial studies. We start with the model formulation for the mean response model and the misclassification process. A method for correcting the bias induced by misclassified binary responses is proposed, and generalized estimating equations analysis and the asymptotic properties are established. Misclassifications within the same cluster can be correlated when the observations are collected by the same person or using the similar defective measuring device. Some feasible ways to construct estimating equations for first and second-order model parameters while eliminating the bias induced by correlated misclassifications are explored.

Chapter 4 discusses the analysis of correlated ordinal data with misclassification in both the response variable and a categorical covariate. We consider marginal methods for estimating first and second order parameters associated with the cumulative probabilities of the ordinal responses. Estimating equations are constructed and asymptotic properties of the methods are discussed. We conduct simulations to show the good performances of the proposed methods. We then illustrate the use of the methods by a data analysis example.

Chapter 5 combines covariate measurement error problem and survey design features. We discuss the analytic use of survey data with binary responses and a misclassified ordinal covariate. Some issues about modeling the distribution of the ordinal covariate and the misclassification process are also addressed. We propose to use the expected score method for parametric estimation and use bootstrap method for variance calculation. A limited simulation study is conducted to investigate the performance of the expected score method. The proposed method is then applied to data from the CCHS cycle 3.1.

Finally, in Chapter 6 we summarize the overall findings and outline future work. Large scale longitudinal surveys have been widely used for studying labor force and population health in a country. Complex survey features can be incorporated in marginal models for categorical and ordinal data with misclassification. Incomplete observations arise frequently in both the outcome variable and the covariates, e.g., subjects may drop out of the studies. Some authors considered using an inverse probability weight matrix in the estimating equations approaches for dealing with incomplete longitudinal observations (see, e.g., Robins et al., 1995; Yi and Cook, 2002). In the presence of misclassification, modification to the weight matrix is needed. When the transition probability from one response category to another is also the focus in a longitudinal study, multi-state Markov transition models can be employed. Marginalized methods (e.g., Heagerty, 2002) can be extended to accommodate both misclassification and missing data.

Chapter 2

Correlated Data with Response Measurement Error under Generalized Linear Mixed Models

2.1 Introduction

The primary interest of epidemiological studies often focuses on investigating the association of a continuous or categorical outcome variable with covariates. For standard statistical analysis, we assume that all variables in the data are precisely observed. In some observational studies, however, measurements of variables may contain error due to imperfect measuring system and/or other reasons. Examples include the measurement of blood pressure using nonstandard device and the determination of disease infection status using poor diagnostic tests. There has been much interest in statistical inference for cases of error-in-covariates, and there exists a large body of references on this topic; see, for instance, Jiang et al. (1999), Wang et al. (1998), and Yi and Cook (2005). Measurement error in response, however, has received less attention, since it is believed that ignoring error in response would still lead to valid inferences. Unfortunately, this is only true for certain situations such as linear regression models with classical additive measurement error in responses. Buonaccorsi (1996) discussed some numerical assessment of bias in estimators from naive analysis ignoring nonlinear response measurement error under linear models. He proposed some solutions for correcting the bias. Neuhaus (1999, 2002) discussed binary responses and showed that naive analysis ignoring measurement error may lead to incorrect conclusions.

Generalized linear mixed models (GLMMs) are of practical importance and have been popular in analyzing correlated data, especially for clustered/familial data. These models are also widely used in statistical genetic analysis of animal breeding data, in which the sires of animals are considered random effects. GLMMs enable the accommodation of non-normally distributed responses and the specification of a possibly nonlinear link function between the mean of the response and the predictors. For example, some reproductive traits in animal breeding are scored as counts (e.g., litter size in pigs), and a mixed Poisson regression model is a possibility (Tempelman and Gianola, 1996). For longitudinal studies, in which repeated measurements are collected on the same subject over time, GLMMs are also widely employed in analyses to account for subject-specific variations (Diggle et al., 2002).

The Framingham Heart Study is a prospective study of the development of cardiovascular disease. This study has been the basis for a considerable amount of epidemiologic research. It is well known that some variables are measured with error. For example, Carroll et al. (1984) considered binary regression models with different link functions to relate the probability of developing heart disease to risk factors including systolic blood pressure (SBP), a variable that contains measurement error. Similarly, Yi (2008) and Yi et al. (2010) considered the effects of covariate measurement error on the estimation of response parameters for longitudinal studies with missing observations. Other research papers on covariate error using data from Framingham Heart Study include Hall and Ma (2007) and Zucker (2005), among others.

In this chapter, we study the impact of measurement error in response variables under GLMMs. We investigate asymptotic bias in the naive estimators for fixed effect parameters when the response measurement error is ignored. Some available approaches that can be used for handling nonlinear measurement errors are evaluated. We present the approximate likelihood method that can yield consistent and highly efficient estimators. In Section 2.5, we conduct a simulation study to compare the performances of various approaches. In Section 2.6, we illustrate the proposed method using a real data set from the Framingham Heart Study. Our primary interest is to study the relationship between long-time average SBP and risk factors such as age, smoking status, and serum cholesterol level (see, e.g., Jaquet et al., 1998; Primatesta et al., 2001; Ferrara et al., 2002). Some discussion and concluding remarks are given in Section 2.7.

2.2 Model Formulation

2.2.1 Response model

Suppose there are n independent clusters in the sample. Let Y_{ij} denote the response for the jth observation in cluster i, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. For longitudinal studies, Y_{ij} represents the response from the jth clinic visit for subject i. Let $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ be vectors of covariates associated with fixed effects and random effects for subject j and cluster i, respectively. Let $\mathbf{X}_i = (\mathbf{X}_{i1}^{\mathrm{T}}, \ldots, \mathbf{X}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{Z}_i = (\mathbf{Z}_{i1}^{\mathrm{T}}, \ldots, \mathbf{Z}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. A GLMM for the data is given by

$$g(\mu_{ij}^b) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i, \tag{2.1}$$

where $\mu_{ij}^b = \mathrm{E}[Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i]$ is the conditional expectation of Y_{ij} given $\mathbf{X}_i$, $\mathbf{Z}_i$ and random effects $\mathbf{b}_i$, and $\boldsymbol{\beta}$ is a vector of regression coefficients for the fixed effects. The random effects $\mathbf{b}_i$ follow a certain distribution, say, $f_b(\mathbf{b}_i; \boldsymbol{\sigma}_{\mathbf{b}})$, with unknown parameters $\boldsymbol{\sigma}_{\mathbf{b}}$. The link function $g(\cdot)$, which is monotone and differentiable, relates μ_{ij}^b to the subject-specific linear predictor $\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i$. When Y_{ij} is binary, common choices of $g(\cdot)$ can be the logit link, probit link, or complementary log-log link. The log link is usually employed when Y_{ij} is a Poisson or Gamma variable.

With continuous Y_{ij}, the choice of identity function $g(\cdot)$ in (2.1) leads to the linear mixed model (LMM)

$$Y_{ij} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i + \epsilon_{ij}, \tag{2.2}$$

which has been extensively discussed in the literature; see, for instance, Laird and

Ware (1982) and McCulloch and Searle (2001), among others. The error term ϵ_{ij} is often assumed to be normally distributed with mean 0 and unknown variance σ_ϵ^2. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\sigma}_b^{\mathrm{T}}, \sigma_\epsilon^2)^{\mathrm{T}}$ be the vector of response parameters. It is straightforward to formulate the marginal likelihood for cluster i as

$$\mathcal{L}_i = \int \prod_{j=1}^{m_i} f_Y(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i) f_b(\mathbf{b}_i) d\mathbf{b}_i, \tag{2.3}$$

where the integration is over the multi-dimensional random components $\mathbf{b}_i$. The likelihood of the data from all clusters is then given by $\mathcal{L} = \prod_{i=1}^n \mathcal{L}_i$.

Making inference about GLMMs often involves integrals that are intractable, because the random effects may enter the model nonlinearly. We will discuss this later in Section 2.4.4.

2.2.2 Measurement error models

In practice, Y_{ij} may not be measured precisely. Instead, we observe a surrogate S_{ij} that may be different from the true measurement. Parametric models for measurement error process are often employed in order to develop methods to eliminate or reduce estimation bias induced by measurement error. A common strategy is to specify the conditional distribution of S_{ij} given true data $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ of cluster i. It is often assumed that $\mathrm{E}[S_{ij}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i] = \mathrm{E}[S_{ij}|Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}]$ (Pepe and Anderson, 1994). If the measurement error process is independent of covariates, then the expectation of S_{ij} only involves the underlying true response, i.e.,

$$\mathrm{E}[S_{ij}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i] = h(Y_{ij}; \boldsymbol{\gamma}_{(i)}) \tag{2.4}$$

where $h(\cdot)$ is a function that involves a vector of error parameters $\boldsymbol{\gamma}_{(i)}$ for cluster i. The dependence of $\boldsymbol{\gamma}_{(i)}$ on i corresponds to situations where different measuring systems are used for different clusters. If the same measuring system is applied to all clusters, subscript i can be dropped from the parameters. Thus, the mean structure for the measurement error process involves a common parameter vector, say, $\boldsymbol{\gamma}$.

Here we introduce the formulations for two widely used classes of measurement error models.

Additive error models

Buonaccorsi (1996) discussed the formulation for nonlinear response measurement error model in an additive form. The relationship between the observed surrogate and the true response is given by

$$S_{ij} = h(Y_{ij}; \boldsymbol{\gamma}) + e_{ij}, \tag{2.5}$$

where e_{ij} has mean 0 and variance σ_e^2. It is common to assume that $e_{ij} \sim \text{Normal}(0, \sigma_e^2)$. Let $\boldsymbol{\eta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \sigma_e^2)^{\mathrm{T}}$ be the vector of parameters associated with the measurement error process. When $h(\cdot)$ is the identity function, (2.5) is the classical additive error model. In this case, naively fitting a linear mixed model (LMM) ignoring measurement error still leads to consistent estimator for $\boldsymbol{\beta}$, since e_{ij} is simply absorbed into the random error ϵ_{ij} of the response model. The naive estimator for the variance parameter σ_ϵ^2, however, will be incorrect due to extra variation induced by e_{ij}.

When $h(\cdot)$ is a nonlinear function, naive estimators for $\boldsymbol{\theta}$ from error-contaminated data are generally biased; see Buonaccorsi (1996) for an example on a four-parameter logistic model.

Multiplicative error models

Multiplicative covariate measurement errors arise as commonly as additive measurement errors, such as energy consumption, and air-borne exposures in occupational epidemiology (e.g., Lyles and Kupper, 1997; Carroll et al., 2006). Several authors have considered linear regression with multiplicative error in the covariates. For example, Hwang (1986) proposed a method-of-moments correction procedure to reduce the bias in regression parameters. In the context of measurement error in response, this type of model is expressed as

$$S_{ij} = h(Y_{ij}; \boldsymbol{\gamma}) \cdot e_{ij}, \tag{2.6}$$

where e_{ij} is independent of Y_{ij} and follows a distribution with mean 1 and variance σ_e^2, e.g., a log-normal distribution or Gamma distribution.

Although multiplicative errors are commonly seen, we show here that (2.6) can be transformed into an error model with an additive noise term. By taking logarithm on both sides of (2.6), we have

$$\log(S_{ij}) \quad = \quad \log\{h(Y_{ij};\boldsymbol{\gamma})\} + \log(e_{ij}).$$

Let $h^*(Y_{ij};\boldsymbol{\gamma},\sigma_e^2) = \{\log\{h(Y_{ij};\boldsymbol{\gamma})\} + \mathrm{E}[\log(e_{ij})]\}$ and $e_{ij}^* = \{\log(e_{ij}) - \mathrm{E}[\log(e_{ij})]\}$. Then we have $\log(S_{ij}) = h^*(Y_{ij};\boldsymbol{\gamma},\sigma_e^2) + e_{ij}^*$, which is of the same form as (2.5) with the noise term having mean 0. The modified function $h(\cdot)$, however, may involve both $\boldsymbol{\gamma}$ and σ_e^2. When e_{ij} follows log-normal distribution with mean 1 and variance σ_e^2, for instance, the log-transformed variable $\log(e_{ij})$ is normally distributed with mean $-\log(\sigma_e^2+1)/2$ and variance $\log(\sigma_e^2+1)$.

In following sections we focus the discussion on additive error, for which $h(\cdot)$ involves only $\boldsymbol{\gamma}$ but not σ_e^2.

2.3 Bias Analysis

In this section we assess the impact of measurement error on estimation of response parameters from two naive approaches that may be used in practice. The first approach ignores measurement error completely and fits a standard GLMM to the data treating S_{ij} as the response. The second approach constructs surrogate responses $\tilde{Y}_{ij} = h^{-1}(S_{ij};\boldsymbol{\gamma})$ ignoring measurement error e_{ij} and fits standard mixed models to the transformed data, provided that $h(\cdot)$ is known. For ease of exposition, we consider cases where the clusters are of equal size, i.e., $m_i = m$.

2.3.1 Naive analysis ignoring error

When the function $h(\cdot)$ is unknown, practitioners may naively fit a GLMM to the observed data. Ignoring measurement error in response amounts to fitting a misspecified model

$$g(\mu_{ij}^{b*}) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^* + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i^*, \qquad i = 1, \ldots, n,\ j = 1, \ldots, m,$$

where $\mathbf{b}_i^*$ is the random effects assuming the same distribution $f_b(\cdot)$ but with different covariance parameters $\boldsymbol{\sigma}_b^*$, and $\mu_{ij}^{b*} = \mathrm{E}[S_{ij}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i^*]$. For continuous Y_{ij} following linear mixed model (2.2), the misspecified model is given by

$$S_{ij} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^* + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i^* + \epsilon_{ij}^*, \qquad i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{2.7}$$

where ϵ_{ij}^* is assumed to have a distribution with mean 0 and variance σ_ϵ^{*2}, say, Normal$(0, \sigma_\epsilon^{*2})$. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\sigma}_b^{*\mathrm{T}}, \sigma_\epsilon^*)^{\mathrm{T}}$.

We now adapt the arguments in White (1982) to study the effects of mismeasured responses. The working likelihood contributed from cluster i is given by

$$\mathcal{L}_i^w(\boldsymbol{\theta}^*) = \int \prod_{j=1}^{m} f_{Y|X,Z,b}(S_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i^*) f_b(\mathbf{b}_i^*) d\mathbf{b}_i^*.$$

Let $\ell_i^w(\boldsymbol{\theta}^*) = \log \mathcal{L}_i^w(\boldsymbol{\theta}^*)$. Maximizing $\ell^w(\boldsymbol{\theta}^*) = \sum_{i=1}^n \ell_i^w(\boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\theta}^*$ gives a false ML estimator $\hat{\boldsymbol{\theta}}^*$. It can be shown that, as $n \to \infty$, $\hat{\boldsymbol{\theta}}^*$ converges in probability to a limit that is the solution to a set of estimating equations

$$\mathrm{E}_{true}\left[\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}^*} \ell_i^w\right] = \mathbf{0}, \tag{2.8}$$

where the expectation is taken with respect to the true distributions of all random variables $(\mathbf{S}_i, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$. The integrals involved in $\ell_i^w(\boldsymbol{\theta}^*)$, however, are often intractable. Thus, there is no simple closed form for the relationship between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$, though approximation can be obtained using numerical integrations.

To gain insights on the impact of ignoring error in response, we further consider a simple LMM involving a random slope for a single covariate X_{ij}, i.e.,

$$Y_{ij} \;=\; \beta_0 + (\beta_1 + b_i)X_{ij} + \epsilon_{ij}, \tag{2.9}$$

where $b_i \sim \text{Normal}(0, \sigma_b^2)$. The misspecified model (2.7) is now simplified to

$$S_{ij} = \beta_0^* + \beta_1^* X_{ij} + X_{ij} b_i^* + \epsilon_{ij}^*, \tag{2.10}$$

where b_i^* and ϵ_{ij}^* are assumed to be normally distributed with mean 0 and variance respectively given by σ_b^{*2} and σ_ϵ^{*2}. We have

$$\begin{aligned}
\mathcal{L}_i^w(\boldsymbol{\theta}^*) &= \int \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^{*2}}} \right)^m \frac{1}{\sqrt{2\pi\sigma_b^{*2}}} \\
&\quad \times \exp\left\{ -\frac{\sum_{j=1}^m (S_{ij} - \beta_0^* - X_{ij}\beta_1^* - X_{ij}b_i^*)^2}{2\sigma_\epsilon^{*2}} - \frac{b_i^{*2}}{2\sigma_b^{*2}} \right\} db_i^* \\
&= \left(\frac{1}{\sqrt{2\pi}} \right)^m \left(\frac{1}{\sqrt{\sigma_\epsilon^{*2}}} \right)^{m-1} \exp\left[-\frac{\sum_{j=1}^m (S_{ij} - \beta_0^* - X_{ij}\beta_1^*)^2}{2\sigma_\epsilon^{*2}} \right. \\
&\quad \left. + \frac{\sigma_b^{*2} \left\{ \sum_{j=1}^m (S_{ij} - \beta_0^* - X_{ij}\beta_1^*) X_{ij} \right\}^2}{2\sigma_\epsilon^{*2} \left(\sigma_b^{*2} \sum_{j=1}^m X_{ij}^2 + \sigma_\epsilon^{*2} \right)} \right] \times \frac{1}{\sqrt{\sigma_b^{*2} \sum_{j=1}^m X_{ij}^2 + \sigma_\epsilon^{*2}}}.
\end{aligned}$$

The corresponding working log-likelihood is then given by

$$\begin{aligned}
\ell_i^w(\boldsymbol{\theta}^*) &= -\frac{m}{2}\log(2\pi) - \frac{m-1}{2}\log\sigma_\epsilon^{*2} - \frac{1}{2}\log\left(\sigma_b^{*2} \sum_{j=1}^m X_{ij}^2 + \sigma_\epsilon^{*2} \right) \\
&\quad - \frac{\sum_{j=1}^m (S_{ij} - \beta_0^* - X_{ij}\beta_1^*)^2}{2\sigma_\epsilon^{*2}} + \frac{\sigma_b^{*2} \left\{ \sum_{j=1}^m (S_{ij} - \beta_0^* - X_{ij}\beta_1^*) X_{ij} \right\}^2}{2\sigma_\epsilon^{*2} \left(\sigma_b^{*2} \sum_{j=1}^m X_{ij}^2 + \sigma_\epsilon^{*2} \right)}.
\end{aligned}$$

The misspecified score function therefore can be obtained from taking derivatives of $\ell_i^w(\boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\theta}^*$. The components for β_0^* and β_1^*, for instance, are given by

$$\begin{aligned}
\begin{pmatrix} \frac{\partial}{\partial\beta_0^*}\ell_i^w(\boldsymbol{\theta}^*) \\ \frac{\partial}{\partial\beta_1^*}\ell_i^w(\boldsymbol{\theta}^*) \end{pmatrix} &= \sum_{j=1}^m \left[\frac{1}{\sigma_\epsilon^{*2}} \begin{pmatrix} S_{ij} - \beta_0^* - X_{ij}\beta_1^* \\ (S_{ij} - \beta_0^* - X_{ij}\beta_1^*) X_{ij} \end{pmatrix} \right. \\
&\quad \left. - \frac{\sigma_b^{*2} \left\{ \sum_{j'=1}^m (S_{ij'} - \beta_0^* - X_{ij'}\beta_1^*) X_{ij'} \right\}}{\sigma_\epsilon^{*2} \left(\sigma_b^{*2} \sum_{j'=1}^m X_{ij'}^2 + \sigma_\epsilon^{*2} \right)} \begin{pmatrix} X_{ij} \\ X_{ij}^2 \end{pmatrix} \right].
\end{aligned} \tag{2.11}$$

Based on (2.8), taking expectations on both sides of (2.11) gives

$$\begin{aligned}\mathbf{0} \;=\; & \sum_{j=1}^{m} \mathrm{E}_{X,Z}\mathrm{E}_{b|X,Z}\mathrm{E}_{Y|b,X,Z}\left[\frac{1}{\sigma_\epsilon^{*2}}\begin{pmatrix} h(Y_{ij};\boldsymbol{\gamma}) - \beta_0^* - X_{ij}\beta_1^* \\ \{h(Y_{ij};\boldsymbol{\gamma}) - \beta_0^* - X_{ij}\beta_1^*\}\, X_{ij}\end{pmatrix}\right. \\ & \left. - \;\frac{\sigma_b^{*2}\left(\sum_{j'=1}^{m}\{h(Y_{ij'};\boldsymbol{\gamma}) - \beta_0^* - X_{ij'}\beta_1^*\}\, X_{ij'}\right)}{\sigma_\epsilon^{*2}\left(\sigma_b^{*2}\sum_{j'=1}^{m} X_{ij'}^2 + \sigma_\epsilon^{*2}\right)}\begin{pmatrix} X_{ij} \\ X_{ij}^2\end{pmatrix}\right]. \qquad (2.12)\end{aligned}$$

We consider two special cases for the error structure: linear measurement error, and exponential measurement error. The first case, which is commonly seen in epidemiologic studies, specifies a linear relationship between S_{ij} and Y_{ij} as

$$S_{ij} = \gamma_0 + \gamma_1 Y_{ij} + e_{ij}, \qquad (2.13)$$

where γ_0 represents a bias of the measuring device at $Y_{ij} = 0$, and γ_1 is a scale factor. We can easily show that simple relationships between the true and false parameters are given by $\beta_0^* = \gamma_0 + \gamma_1\beta_0$, $\beta_1^* = \gamma_1\beta_1$, $\sigma_b^{*2} = \gamma_1^2\sigma_b^2$, and $\sigma_\epsilon^{*2} = \gamma_1^2\sigma_\epsilon^2 + \sigma_e^2$. The results hold for general LMM with multiple covariates, as the distribution of the observed surrogate response given covariates is still within the LMM framework but with scaled fixed effects and variance components.

The second special measurement error model we consider is an exponential error model given by

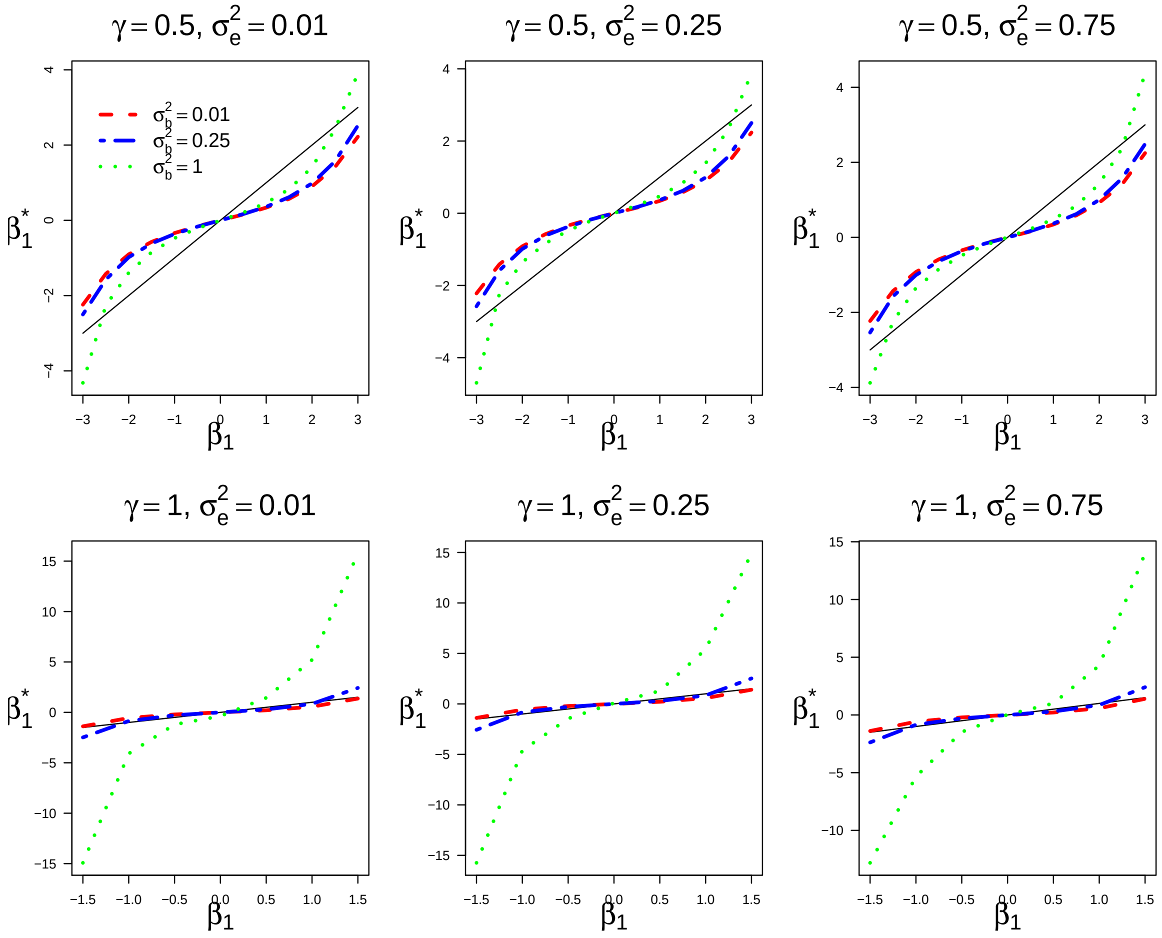
$$S_{ij} = \exp(\gamma Y_{ij}) + e_{ij}, \qquad (2.14)$$

where $e_{ij} \sim \mathrm{Normal}(0, \sigma_e^2)$ and is independent of Y_{ij}. This error model is of interest when Y_{ij} is the logarithm of an underlying variable that is impossible to obtain. As shown in Section 2.8.1, there is no closed form for the bias in the naive fixed-effect estimator due to the expectations over nonlinear functions.

Here we specifically undertake a numerical study to illustrate the bias induced by response error under model (2.14). We focus on the bias in β_1 given the values of β_0, σ_b^2, and σ_ϵ^2. The model parameters are specified by $\beta_0 = -1$, $\sigma_\epsilon^2 = 0.01$, and $\sigma_b^2 = 0.01$, 0.25, and 1. Various combinations of error parameters γ and σ_e^2 are considered. Figure 2.1 displays nonlinear curves of the naive β_1^* versus the true β_1. When $\gamma = 0.5$, for instance, the naive estimates of β_1 are attenuated for small values of β_1 but are inflated for large values. When $\gamma = 1$, however, β_1^* is much larger than β_1 in all settings. The shapes of the bias curves under the various settings are also different.

In general, the direction and magnitude of the bias induced by nonlinear response error depend on both $h(\cdot)$ and associated parameters $\boldsymbol{\eta}$ in the measurement error process. Variance parameters also play significant roles in the bias of the naive estimates.

Figure 2.1: Bias in β_1^* from the completely naive approach induced by an exponential error model. The dashed line (− − −), twodash line (− —), and dotted line (. . .) are for $\sigma_b^2 = 0.01, 0.25$, and 1, respectively.



2.3.2 Naive analysis of transformed data

If the specific form of $h(\cdot)$ is known, another straightforward naive approach is to construct surrogate response $\tilde{Y}_{ij} = h^{-1}(S_{ij}; \boldsymbol{\gamma})$ and perform standard statistical analysis treating $\tilde{Y}_{ij}$ as true.

It is easy to see that when $h(\cdot)$ is a linear function, the transformed surrogate $\tilde{Y}_{ij}$ is an unbiased surrogate for the true Y_{ij}. When $h(\cdot)$ is nonlinear, however, the unbiasedness of $\tilde{Y}_{ij}$ does not hold in general. Naively fitting a LMM to the transformed data leads to estimation of a vector of false parameters, say, $\tilde{\boldsymbol{\theta}}$, other than the true $\boldsymbol{\theta}$. The magnitude of bias mainly depend on measurement error variance σ_e^2 . When σ_e^2 is extremely small, the transformed surrogate $\tilde{Y}_{ij}$ approximates Y_{ij} very well. In this case, estimation bias induced by response measurement error is generally ignorable.

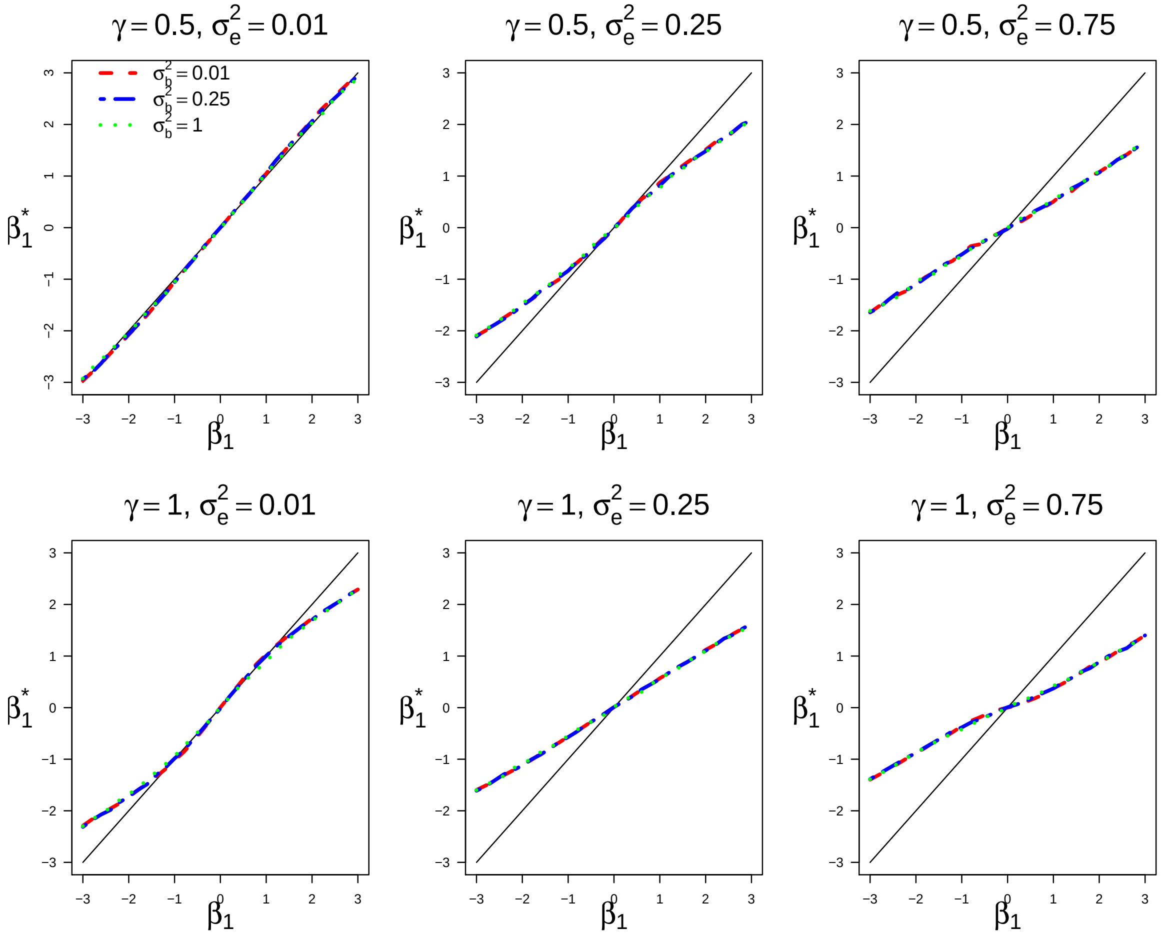
To investigate the asymptotic bias in this naive estimator using the transformed data, we again consider the simple LMM given by (2.9). A similar procedure can be employed to develop a set of equations to relate $\tilde{\boldsymbol{\theta}}$ to the true $\boldsymbol{\theta}$. The working log-likelihood from cluster i with the transformed surrogates is given by

$$\begin{aligned}\tilde{\ell}_i^w(\tilde{\boldsymbol{\theta}}) &= -\frac{m}{2}\log(2\pi) - \frac{m-1}{2}\log\tilde{\sigma}_\epsilon^2 - \frac{1}{2}\log\left(\tilde{\sigma}_b^2\sum_{j=1}^m X_{ij}^2 + \tilde{\sigma}_\epsilon^2\right) \\ &\quad - \frac{\sum_{j=1}^m(\tilde{Y}_{ij} - \tilde{\beta}_0 - X_{ij}\tilde{\beta}_1)^2}{2\tilde{\sigma}_\epsilon^2} + \frac{\tilde{\sigma}_b^2\left\{\sum_{j=1}^m(\tilde{Y}_{ij} - \tilde{\beta}_0 - X_{ij}\tilde{\beta}_1)X_{ij}\right\}^2}{2\tilde{\sigma}_\epsilon^2\left(\tilde{\sigma}_b^2\sum_{j=1}^m X_{ij}^2 + \tilde{\sigma}_\epsilon^2\right)}.\end{aligned}$$

The first derivatives of $\tilde{\ell}_i^w(\tilde{\boldsymbol{\theta}})$ with respect to $\tilde{\boldsymbol{\beta}}$ are of the same form as those in equations (2.11). Again the expectation of the working score function involves integrals that do not have simple closed forms for cases with nonlinear $h(\cdot)$.

Here again we conduct a numerical study under the scenarios described in previous section to investigate the relationship between $\tilde{\beta}_1$ and β_1. The bias curves are shown in Figure 2.2. One can see that the bias is dramatically reduced compared to that from the naive analysis ignoring error. Also, the size of the bias increases as the size of β_1 increases. Furthermore, the values of γ and σ_e^2 have significant impact on the bias. In general, the size of the bias increases as σ_e^2 increases.

Figure 2.2: Bias in $\tilde{\beta}_1$ from the naive analysis of the transformed data induced by an exponential error model. The dashed line (– – –), twodash line (– —), and dotted line (. . .) are for $\sigma_b^2 = 0.01, 0.25$, and 1, respectively.



2.4 Inference Methods

In this section we discuss likelihood inference methods for several practical cases: (i) $\boldsymbol{\eta}$ is known, (ii) $\boldsymbol{\eta}$ is unknown but a validation subsample is available, and (iii) replicates for the surrogates are available. We propose some strategies on the estimation of model parameters.

2.4.1 $\boldsymbol{\eta}$ is known

Conditional on fixed $\boldsymbol{\eta}$, the marginal likelihood of the observed data from cluster i can be written as

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\eta}) \;=\; & \int \left\{ \prod_{j=1}^{m} \int f_{S|Y,X,Z,b}(S_{ij}|Y_{ij},\mathbf{X}_{ij},\mathbf{Z}_{ij},\mathbf{b}_i;\boldsymbol{\eta}) \right. \\ & \left. \times\, f_{Y|X,Z,b}(Y_{ij}|\mathbf{X}_{ij},\mathbf{Z}_{ij},\mathbf{b}_i;\boldsymbol{\theta})dY_{ij} \right\} f_b(\mathbf{b}_i;\boldsymbol{\sigma}_b)d\mathbf{b}_i. \qquad (2.15)\end{aligned}$$

Let $\ell_i(\boldsymbol{\theta},\boldsymbol{\eta}) = \log\mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\mathbf{U}_i(\boldsymbol{\theta},\boldsymbol{\eta}) = \partial\ell_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}$. The ML estimator $\hat{\boldsymbol{\theta}}$ can be obtained from maximizing $\ell(\boldsymbol{\theta},\boldsymbol{\eta}) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\theta},\boldsymbol{\eta})$ provided that $\boldsymbol{\eta}$ is fixed at its true value, say, $\boldsymbol{\eta}_0$. This leads to solving a set of equations

$$\sum_{i=1}^{n}\mathbf{U}_i(\boldsymbol{\theta},\boldsymbol{\eta}_0) = \mathbf{0}.$$

From standard likelihood theory, the ML estimator $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$. As $n \to \infty$, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \text{MVN}(\mathbf{0},\boldsymbol{\mathcal{I}}^{-1})$, where $\boldsymbol{\mathcal{I}} = \text{E}\left[-\partial\mathbf{U}_i(\boldsymbol{\theta}),\boldsymbol{\eta}_0/\partial\boldsymbol{\theta}^{\text{T}}\right]$. Using Bartlett's identity and the law of large numbers, $\boldsymbol{\mathcal{I}}$ can be consistently estimated by, $n^{-1}\sum_{i=1}^{n}\mathbf{U}_i(\hat{\boldsymbol{\theta}},\boldsymbol{\eta}_0)\mathbf{U}_i(\hat{\boldsymbol{\theta}},\boldsymbol{\eta}_0)^{\text{T}}$.

2.4.2 $\boldsymbol{\eta}$ is estimated from validation data

In reality $\boldsymbol{\eta}$ is often unknown. In the absence of additional information such as a validation data set or replicates of the measurements, parameter identifiability may be a major issue (e.g., Carroll et al. 2006, p. 184). For nonlinear $h(\cdot)$, $\boldsymbol{\theta}$ may be theoretically identified in some particular situations without extra information. The estimators, however, are usually unstable. In this and next subsections, we discuss estimation and inference procedures when a validation data set and replicates of surrogates are respectively available.

Validation data arise commonly in the study when some observations are selected into a subsample and the true values of the responses are obtained. Let $\delta_{ij} = 1$ if Y_{ij} is available and $\delta_{ij} = 0$ otherwise. Let $N_v = \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij}$ be the size of the validation subsample. Here the selection is assumed to be a random process that is independent of the observed data. The full marginal likelihood of the main data and the validation data contributed from cluster i is given by

$$\begin{aligned}\mathcal{L}_{Fi}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \int \left[\prod_{j=1}^{m} \left\{ f_{S|X,Z,b}(S_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \right\}^{1-\delta_{ij}} \right. \\ &\qquad \left. \times \left\{ f_{S,Y|X,Z,b}(S_{ij}, Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \right\}^{\delta_{ij}} \right] f_b(\mathbf{b}_i; \boldsymbol{\sigma}_b) d\mathbf{b}_i, \quad (2.16)\end{aligned}$$

where

$$\begin{aligned}& f_{S|X,Z,b}(S_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \\ &\quad = \int f_{S|Y,X,Z,b}(S_{ij}|Y_{ij}; \boldsymbol{\eta}) f_{Y|X,Z,b}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}) dY_{ij},\end{aligned}$$

and

$$\begin{aligned}& f_{S,Y|X,Z,b}(S_{ij}, Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \\ &\quad = f_{S|Y,X,Z,b}(S_{ij}|Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\eta}) f_{Y|X,Z,b}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}).\end{aligned}$$

Under the assumption that S_{ij} is conditionally independent of $\mathbf{b}_i$ given Y_{ij}, we can rewrite (2.16) as

$$\begin{aligned}\mathcal{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta}) &= \left\{\int\left[\prod_{j=1}^{m}\left\{f_{S|X,Z,b}(S_{ij}|\mathbf{X}_{ij},\mathbf{Z}_{ij},\mathbf{b}_i;\boldsymbol{\theta},\boldsymbol{\eta})\right\}^{1-\delta_{ij}}\right.\right.\\ &\quad \left.\left.\times\ \left\{f_{Y|X,Z,b}(Y_{ij}|\mathbf{X}_{ij},\mathbf{Z}_{ij},\mathbf{b}_i;\boldsymbol{\theta})\right\}^{\delta_{ij}}\right]f_b(\mathbf{b}_i;\boldsymbol{\sigma}_b)d\mathbf{b}_i\right\}\\ &\quad \times\ \prod_{j=1}^{m}\left\{f_{S|Y}(S_{ij}|Y_{ij};\boldsymbol{\eta})\right\}^{\delta_{ij}}.\end{aligned}$$

Let

$$\begin{aligned}\mathcal{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta}) &= \left\{\int\left[\prod_{j=1}^{m}\left\{f_{S|X,Z,b}(S_{ij}|\mathbf{X}_{ij},\mathbf{Z}_{ij},\mathbf{b}_i;\boldsymbol{\theta},\boldsymbol{\eta})\right\}^{1-\delta_{ij}}\right.\right.\\ &\quad \left.\left.\times\ \left\{f_{Y|X,Z,b}(Y_{ij}|\mathbf{X}_{ij},\mathbf{Z}_{ij},\mathbf{b}_i;\boldsymbol{\theta})\right\}^{\delta_{ij}}\right]f_b(\mathbf{b}_i;\boldsymbol{\sigma}_b)d\mathbf{b}_i\right\},\end{aligned}$$

and $\mathcal{L}_{\eta i}(\boldsymbol{\eta})=\prod_{j=1}^{m}\left\{f_{S|Y}(S_{ij}|Y_{ij};\boldsymbol{\eta})\right\}^{\delta_{ij}}$. Therefore,

$$\mathcal{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})=\mathcal{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})\ \times\ \mathcal{L}_{\eta i}(\boldsymbol{\eta}).$$

Let $\ell_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})=\log\mathcal{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})$. When the dimension of $(\boldsymbol{\theta},\boldsymbol{\eta})$ is large, direct maximization of $\sum_{i=1}^{n}\ell_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})$ can be computationally demanding.

We propose to use a two-stage estimation procedure as an alternative to the joint estimation procedure. This approach employs stepwise maximization of the log-likelihood function. Let $\ell_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})=\log\mathcal{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})$ and $\ell_{\eta i}(\boldsymbol{\theta},\boldsymbol{\eta})=\log\mathcal{L}_{\eta i}(\boldsymbol{\theta},\boldsymbol{\eta})$. Let $\mathbf{U}_i^*(\boldsymbol{\theta},\boldsymbol{\eta})=\partial\ell_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}$ and $\mathbf{Q}_i^*(\boldsymbol{\eta})=\partial\ell_{\eta i}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$. In the first stage, estimator for $\boldsymbol{\eta}$ is obtained by solving

$$\sum_{i=1}^{n}\mathbf{Q}_i^*(\boldsymbol{\eta})=\mathbf{0}. \tag{2.17}$$

Let $\hat{\boldsymbol{\eta}}$ be the solution to (2.17). In the second stage, replace $\boldsymbol{\eta}$ with $\hat{\boldsymbol{\eta}}$ and solve

$$\sum_{i=1}^{n} \mathbf{U}_i^*(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \mathbf{0}. \tag{2.18}$$

Let $\hat{\boldsymbol{\theta}}_p$ denote the solution to (2.18), i.e., pseudo-ML estimator. As $N_v \to \infty$, $\hat{\boldsymbol{\eta}}$ converges to the true $\boldsymbol{\eta}$ in probability. Therefore, $\hat{\boldsymbol{\theta}}_p$ is consistent for $\boldsymbol{\theta}$ as $n \to \infty$ and $N_v/n \to \rho$, where $0 < \rho < 1$.

Since there is uncertainty associated with the estimated $\boldsymbol{\eta}$, we must account for the extra variation that transfers to $\hat{\boldsymbol{\theta}}_p$. Let $\boldsymbol{\mathcal{I}}_{11}^* = \sum_{i=1}^n \mathrm{E}\left[-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right]$, $\boldsymbol{\mathcal{I}}_{12}^* = \boldsymbol{\mathcal{I}}_{21}^{*\mathrm{T}} = \sum_{i=1}^n \mathrm{E}\left[-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right]$, and $\boldsymbol{\mathcal{J}}^*(\boldsymbol{\eta}) = \sum_{i=1}^n \mathrm{E}\left[-\partial \mathbf{Q}_i^*(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right]$. The asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_p$ can be obtained by following the arguments in Buonaccorsi (1996),

$$\boldsymbol{\Sigma}^* = \boldsymbol{\mathcal{I}}_{11}^{*-1}(\boldsymbol{\theta}, \boldsymbol{\eta}) + \boldsymbol{\mathcal{I}}_{11}^{*-1}(\boldsymbol{\theta}, \boldsymbol{\eta})\boldsymbol{\mathcal{I}}_{12}^*(\boldsymbol{\theta}, \boldsymbol{\eta})\boldsymbol{\mathcal{J}}^{*-1}(\boldsymbol{\eta})\boldsymbol{\mathcal{I}}_{21}^*(\boldsymbol{\theta}, \boldsymbol{\eta})\boldsymbol{\mathcal{I}}_{11}^{*-1}(\boldsymbol{\theta}, \boldsymbol{\eta}). \tag{2.19}$$

A sketch of the proof is outlined in Section 2.8.2. An approximate estimate of $\boldsymbol{\Sigma}^*$ can be obtained by replacing $\boldsymbol{\mathcal{I}}_{11}^*$, $\boldsymbol{\mathcal{I}}_{12}^*$, and $\boldsymbol{\mathcal{J}}^*(\boldsymbol{\eta})$ with their empirical counterparts $\mathbf{M}_{11}^* = \sum_{i=1}^n \left\{-\partial \mathbf{U}_i^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right\}$, $\mathbf{M}_{12}^* = \sum_{i=1}^n \left\{-\partial \mathbf{U}_i^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right\}$, and $\mathbf{J}^* = \sum_{i=1}^n \left\{-\partial \mathbf{Q}_i^*(\hat{\boldsymbol{\eta}})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right\}$, respectively.

2.4.3 Inference with replicates

In some situations we may have replicates for the surrogate measurements due to the design of the study. Such amount of additional information can be used for identifying the response model and the measurement error model when a validation subsample is not available (Carroll et al. 2006).

Let S_{ijr} be the rth surrogate replicate for subject j in cluster i, $r = 1, \ldots, d_{ij}$. For $r \neq r'$, we assume that S_{ijr} and $S_{ijr'}$ are conditionally independent given $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$. We consider two scenarios: (a) σ_e^2 is the only nuisance parameters to be estimated, (b) both σ_e^2 and $\boldsymbol{\gamma}$ are unknown.

A method for estimating σ_e^2

Here we discuss an approach for estimation of measurement error variance σ_e^2 that is the only unknown nuisance parameter. One can see that σ_e^2 can be estimated alone using the surrogate replicates when e_{ijr} are independent of each other conditional on $\mathbf{Y}_i$.

Let $\bar{S}_{ij.} = (1/d_{ij}) \sum_{r=1}^{d_{ij}} S_{ijr}$, $j = 1, \ldots, m$, $i = 1, \ldots, n$. Under the assumption of conditional independence between S_{ijr} and $S_{ijr'}$, an unbiased estimator for σ_e^2 is given by

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^n \sum_{j=1}^m \sum_{r=1}^{d_{ij}} (S_{ijr} - \bar{S}_{ij.})^2}{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - 1)}. \qquad (2.20)$$

Note that when e_{ijr}'s are normally distributed, $\left\{\sum_{r=1}^{d_{ij}} (S_{ijr} - \bar{S}_{ij.})^2\right\} / \sigma_e^2$ follows a chi squared distribution with $(d_{ij} - 1)$ degrees of freedom. Therefore, the variance of $\hat{\sigma}_e^2$ is given by

$$\text{var}(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{\sum_{i=1}^n \sum_{j=1}^m (d_{ij} - 1)}.$$

An approximate variance can be obtained by replacing σ_e^2 in the formula above with its estimate $\hat{\sigma}_e^2$.

This approach is appealing in situations where σ_e^2 is the only nuisance parameter to be estimated. With classical additive measurement error, estimation bias in the naive estimate of σ_ϵ^2 can be corrected by subtracting $\hat{\sigma_e}^2$. Another situation is that the nonlinear function $h(\cdot)$ and parameter $\boldsymbol{\gamma}$ are known, e.g., by the design of the study, from history data, or Box-Cox transformation (see, e.g., Hall and Ma 2007). Thus, the two-stage estimation and inference procedures for $\boldsymbol{\theta}$ can easily be used.

Joint estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$

The marginal likelihood of the replication data from cluster i is now given by

$$\begin{aligned}\mathcal{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= \int f_b(\mathbf{b}_i; \boldsymbol{\sigma}_b) \prod_{j=1}^{m} \Big\{ \int f_{Y|X,Z,b}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}) \\ &\qquad \times \prod_{r=1}^{d_{ij}} f_{S|Y,X,Z,b}(S_{ijr}|Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i; \boldsymbol{\eta}) dY_{ij} \Big\} d\mathbf{b}_i. \qquad (2.21)\end{aligned}$$

Unlike cases where validation data can be used for making inference about the measurement error process, here $\boldsymbol{\eta}$ generally can not be estimated by solving a set of equations that are free of $\boldsymbol{\theta}$. The underlying true responses are now completely unobserved. Therefore, the two-stage estimation procedure cannot be employed in a replication study except for some special situations where σ_e^2 is the only error parameter to be estimated. A joint estimation procedure for $(\boldsymbol{\theta}, \boldsymbol{\eta})$ by maximizing $\mathcal{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is required. Let $\mathcal{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \partial \mathcal{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}$ and $\mathcal{Q}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \partial \mathcal{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}$ be the score functions. ML estimators can be obtained by simultaneously solving

$$\sum_{i=1}^{n} \begin{pmatrix} \mathcal{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ \mathcal{Q}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \end{pmatrix} = \mathbf{0}.$$

Let $(\hat{\boldsymbol{\theta}}_R, \hat{\boldsymbol{\eta}}_R)$ be the solution. Under suitable regularity conditions, $n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}}_R - \boldsymbol{\theta} \\ \hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta} \end{pmatrix}$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix given by

$$\boldsymbol{\Sigma}_R = \mathrm{E}\left[\left\{ \mathcal{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta})^{\mathrm{T}}, \mathcal{Q}_i(\boldsymbol{\theta}, \boldsymbol{\eta})^{\mathrm{T}} \right\}^{\mathrm{T}} \left\{ \mathcal{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta})^{\mathrm{T}}, \mathcal{Q}_i(\boldsymbol{\theta}, \boldsymbol{\eta})^{\mathrm{T}} \right\} \right].$$

2.4.4 Numerical approximation

The likelihood functions discussed above involve integrations over unobserved random components and underlying responses. The random effects are assumed to follow known distributions such as normal distribution, gamma distribution, or t-distribution. There are computational challenges in implementing the above procedures, as the integrals typically do not have closed forms.

A common approach to dealing with the computation is to linearize the model with respect to the random effects, e.g., using a first-order population-averaged approximation to the marginal distribution by expanding about the average random effect (Vonesh and Carter 1992). Another approach is to use numerical approximation to integrals, such as Laplace's approximation (e.g., Wolfinger 1993, and Vonesh 1996) or Gaussian quadratures, to obtain an approximate likelihood function with a closed form. The basic form of linearization using Laplace's approximation is a second-order Taylor series expansion of the integrand $f(\mathbf{u})$ and is given by

$$\int_{\mathcal{R}^d} f(\mathbf{u})d\mathbf{u} \approx (2\pi)^{d/2} f(\mathbf{u}_0) \left| -\frac{\partial^2 \log f(\mathbf{u}_0)}{\partial \mathbf{u} \partial \mathbf{u}^{\mathrm{T}}} \right|^{-1/2},$$

where d is the dimension of $\mathbf{u}$, and $\mathbf{u}_0$ is the mode of $f(\mathbf{u})$, i.e., the solution to $\partial \log f(\mathbf{u})/\partial \mathbf{u} = \mathbf{0}$. To construct the Laplace approximation we need expressions for the first two derivatives of $\log f(\mathbf{u})$.

For one dimensional case, we use Gaussian-Hermite quadrature for approximating an integral where the integrand contains a weight function e^{-u^2}. Specifically, the integral is approximated by a sum

$$\int_{-\infty}^{\infty} e^{-u^2} f(u)du \approx \sum_{k=1}^{K} w_k f(t_k),$$

where K is the number of points, and t_k and w_k are the value and the weight of the kth designated point, respectively. As K increases, the accuracy of the approximation increases (McCulloch and Searle 2001).

As an example, we consider the likelihood function in (2.15), where the random effect is one-dimensional and follows a normal distribution Normal$(0, \sigma_b^2)$. It can be written as

$$\begin{aligned}\mathcal{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta}) &= \int\left\{\prod_{j=1}^{m}\int\frac{1}{2\pi\sigma_e^2}\exp\left[-\frac{\{S_{ij}-h(Y_{ij};\boldsymbol{\gamma})\}^2}{2\sigma_e^2}\right]\right.\\ &\quad\left.\times\frac{1}{2\pi\sigma_\epsilon^2}\exp\left[-\frac{\left\{Y_{ij}-(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+Z_{ij}b_i)\right\}^2}{2\sigma_\epsilon^2}\right]dY_{ij}\right\}\\ &\quad\times\frac{1}{2\pi\sigma_b^2}\exp\left(-\frac{b_i^2}{2\sigma_b^2}\right)db_i.\end{aligned}$$

Its approximate is given by

$$\tilde{\mathcal{L}}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta}) = \sum_{k=1}^{K}\left\{\frac{w_k}{\sqrt{\pi}}\prod_{j=1}^{m}\sum_{k'=1}^{K}\frac{w_{k'}}{\sqrt{\pi}}\frac{1}{\sqrt{2\pi\hat{\sigma}_e^2}}\exp\left(-\frac{e_{ij,kk'}^2}{2\hat{\sigma_e}^2}\right)\right\}, \qquad (2.22)$$

where $e_{ij,kk'} = S_{ij} - h(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + Z_{ij}t_k\sqrt{2\sigma_b^2} + t_{k'}\sqrt{2\sigma_\epsilon^2};\boldsymbol{\gamma})$. Here we use the same quadrature order for all integrations. Further let

$$A_{ij,k} = \sum_{k'=1}^{K}\frac{w_{k'}}{\sqrt{\pi}}\frac{1}{\sqrt{2\pi\hat{\sigma}_e^2}}\exp\left(-\frac{e_{ij,kk'}^2}{2\hat{\sigma_e}^2}\right).$$

The approximate score function is given by

$$\begin{aligned}\tilde{\mathbf{U}}_i^*(\boldsymbol{\theta},\boldsymbol{\eta}) &= \frac{\partial\log\tilde{\mathcal{L}}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})}{\partial\boldsymbol{\theta}}\\ &= \frac{\sum_{k=1}^{K}\left[\left\{w_k/\sqrt{\pi}\prod_{j=1}^{m}A_{ij,k}\right\}\left\{\sum_{j=1}^{m}\partial\log A_{ij,k}/\partial\boldsymbol{\theta}\right\}\right]}{\sum_{k=1}^{K}\left(w_k/\sqrt{\pi}\prod_{j=1}^{m}A_{ij,k}\right)},\end{aligned}$$

where $\partial\log A_{ij,k}/\partial\boldsymbol{\theta} = (1/A_{ij,k})\,\partial A_{ij,k}/\partial\boldsymbol{\theta}$, and

$$\frac{\partial A_{ij,k}}{\partial\boldsymbol{\theta}} = \sum_{k'=1}^{K}\left\{\frac{w_{k'}}{\sqrt{\pi}}\frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{e_{ij,kk'}^2}{2\sigma_e^2}\right)\left(-\frac{e_{ij,kk'}}{\sigma_e^2}\frac{\partial e_{ij,kk'}}{\partial\boldsymbol{\theta}}\right)\right\}.$$

Therefore, maximization of $\sum_{i=1}^{n}\tilde{\mathcal{L}}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})$ can be done using numerical optimization methods in available statistical software packages (e.g. optim() or nlminb() in R).

For cases with replication data, we further define

$$\begin{aligned} e_{ijr,kk'} &= S_{ijr} - h(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + Z_{ij} t_k \sqrt{2\sigma_b^2} + t_{k'}\sqrt{2\sigma_\epsilon^2}; \boldsymbol{\gamma}), \\ &\qquad\qquad r = 1, \ldots, d_{ij},\ j = 1, \ldots, m,\ i = 1, \ldots, n, \end{aligned}$$

and

$$\mathcal{A}_{ij,k} = \sum_{k'=1}^{K} \frac{w_{k'}}{\sqrt{\pi}} \left(\frac{1}{\sqrt{2\pi\sigma_e^2}} \right)^{d_{ij}} \exp\left(-\frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'}^2}{2\sigma_e^2} \right).$$

An approximate likelihood function is given by

$$\tilde{\mathcal{L}}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{k=1}^{K} \left\{ \frac{w_k}{\sqrt{\pi}} \prod_{j=1}^{m} \mathcal{A}_{ij,k} \right\}.$$

The score functions of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are then given by

$$\begin{aligned} &\begin{pmatrix} \tilde{\mathcal{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ \tilde{\mathcal{Q}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \end{pmatrix} \\ &= \frac{\sum_{k=1}^{K} \left[\left(w_k/\sqrt{\pi} \prod_{j=1}^{m} \mathcal{A}_{ij,k} \right) \left\{ \sum_{j=1}^{m} \left(1/\mathcal{A}_{ij,k}\right) \partial\mathcal{A}_{ij,k}/\partial(\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}} \right\} \right]}{\sum_{k=1}^{K} \left(w_k/\sqrt{\pi} \prod_{j=1}^{m} \mathcal{A}_{ij,k} \right)}, \end{aligned}$$

where

$$\frac{\partial\mathcal{A}_{ij,k}}{\partial\boldsymbol{\theta}} = -\sum_{k'=1}^{K} \left\{ \frac{w_{k'}}{\sqrt{\pi}} \left(\frac{1}{\sqrt{2\pi\sigma_e^2}} \right)^{d_{ij}} \exp\left(-\frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'}^2}{2\sigma_e^2} \right) \frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'} \partial e_{ijr,kk'}/\partial\boldsymbol{\theta}}{\sigma_e^2} \right\},$$

and

$$\begin{aligned} \frac{\partial\mathcal{A}_{ij,k}}{\partial\boldsymbol{\eta}} &= \begin{pmatrix} \partial\mathcal{A}_{ij,k}/\partial\boldsymbol{\gamma} \\ \partial\mathcal{A}_{ij,k}/\partial\sigma_e^2 \end{pmatrix} \\ &= \begin{pmatrix} \sum_{k'=1}^{K} \left\{ \frac{w_{k'}}{\sqrt{\pi}} \left(\frac{1}{\sqrt{2\pi\sigma_e^2}} \right)^{d_{ij}} \exp\left(-\frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'}^2}{2\sigma_e^2} \right) \left(-\frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'} \partial e_{ijr,kk'}/\partial\boldsymbol{\gamma}}{\sigma_e^2} \right) \right\} \\ \sum_{k'=1}^{K} \left\{ \frac{w_{k'}}{\sqrt{\pi}} \left(\frac{1}{\sqrt{2\pi\sigma_e^2}} \right)^{d_{ij}} \exp\left(-\frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'}^2}{2\sigma_e^2} \right) \left(-\frac{1}{2\sigma_e^2} + \frac{\sum_{r=1}^{d_{ij}} e_{ijr,kk'}^2}{2\sigma_e^2} \right) \right\} \end{pmatrix}. \end{aligned}$$

Similarly, the optimization tools in R can be used for this case.

As the number of random effects particularly nested random effects grows, quadrature quickly becomes computationally infeasible. The optimization may converge very slowly due to the high-dimensional integration. In some situations, a single quadrature node is sufficient, which is equivalent to Laplace approximation and is a computationally more expedient alternative.

2.5 Simulation Studies

2.5.1 Design of simulation

We conduct some simulation studies to assess the performance of the proposed methods. We generate clustered binary responses from a simple LMM with a random slope for 100 clusters of size 5. Specifically, the covariate X_{ij} and random component b_i are generated independently under Normal$(0, 1)$ and Normal$(0, \sigma_b^2)$, respectively. Given $(\mathbf{X}_i, b_i)$, the binary responses $Y_{i1}, Y_{i2}, \ldots, Y_{i5}$ are then generated conditionally independently from the LMM given by (2.9). Response parameters are specified by $\beta_0 = -1$, $\beta_1 = \log(0.5)$, $\sigma_b^2 = 0.04$, and $\sigma_\epsilon^2 = 0.04$.

We consider several measurement error models described in previous sections for generating the surrogate response S_{ij}:

(M1) $S_{ij} = \exp(\gamma Y_{ij}) + e_{ij}$, and

(M2) $S_{ij} = \gamma_0 + \gamma_1 Y_{ij} + e_{ij}$,

where e_{ij} is independent of $\mathbf{Y}_i$ and $\mathbf{X}_i$ and follows a normal distribution with mean 0 and variance σ_e^2. For error model M1, the error parameters are specified by $\gamma = 0.5$ and $\sigma_e^2 = 0.04$. For error model M2, the parameters are specified by $\gamma_0 = 0.5$, $\gamma_1 = 0.5$, and $\sigma_e^2 = 0.04$.

We consider two cases regarding the knowledge of $\boldsymbol{\eta}$: either treated as known or estimated from internal validation data. We obtain the validation subsample by randomly selecting one subject from each cluster.

For each parameter setting, we generate 2000 data sets. The internals in the likelihood-based approach are approximated by Gaussian quadrature of order 15. For comparison, we also include results from the two naive approaches.

2.5.2 Simulation results

We assess the performances of the estimators based on four measures: relative bias in percent (%RB), sample standard deviation of the estimates (SD), average of model-based standard errors (ASE), and coverage probability of the 95% confidence interval (CP).

Table 2.1 reports simulation results for the exponential measurement error model (M1) under scenarios where $\boldsymbol{\eta}$ is known or estimated from validation data. We first look at the quantities for the fixed-effect parameter β_1. As expected, the first naive approach (NAI1) ignoring response measurement error leads to very biased (attenuated) estimate of β_1, and the coverage rate of the 95% confidence interval is close to 0. The second naive approach (NAI2) using the transformed surrogate responses gives slightly better estimates of β_1. The magnitude of the relative bias (upward), although smaller than that from NAI1, is still substantial. The pseudo-ML (PML) estimator for β_1 from the likelihood-based approach is much more consistent under both scenarios than the two naive estimators, and the coverage rate of its 95% confidence interval is very close to the nominal value.

Table 2.2 reports the results for the linear measurement error model (M2) under scenarios where $\boldsymbol{\eta}$ is known or estimated from validation data. Again the estimator for β_1 from the naive approach ignoring error is biased. The value is scaled approximately by a factor of γ_1, which agrees with the analytical result shown in Section 2.3. The naive estimator using the transformed surrogates yields consistent estimators for β_0, β_1, and σ_b^2. The estimator for σ_ϵ^2, however, is very biased. As a result, the coverage rates of the confidence intervals for σ_b^2 and σ_ϵ^2 are far from the nominal value of 95%. In contrast, the likelihood-based approach gives consistent estimators for all the fixed-effect parameters and the variance parameters, and associated standard errors also approximate the empirical standard deviations very well. The coverage rates of the

95% confidence intervals are close to the nominal value.

2.6 Application

We illustrate our proposed method by analyzing data from the Framingham Heart Study. The data set includes exams #2 and #3 for $n = 1615$ male subjects aged 31-65 (Carroll et al., 2006, p. 112). Two SBP readings were taken during each exam. One of the clinical interests is to understand the relationship between SBP and potential risk factors such as baseline smoking status and age. (e.g., Jaquet et al., 1998; Primatesta et al., 2001; Ferrara et al., 2002). We let T_{ij} be the true SBP measurement defined as the long-term average of SBP for subject i at time j, where $j = 1$ for exam #2, and $j = 2$ for exam #3, and $i = 1, \ldots, n$. The risk factors, however, may not have linear effects on SBP directly. Some exploratory plots show that the observed SBP measurements are positively skewed, i.e., with a right tail. Data transformation such as Box-Cox transformation can be applied (Box and Cox, 1964). The square-root transformed observations are shown to satisfy the symmetry condition. Let $Y_{ij} = \sqrt{T_{ij} - 50}$. We assume that Y_{ij} follow a LMM with a random intercept

$$Y_{ij} = \beta_0 + \beta_{age} x_{ij1} + \beta_{smoke} x_{ij2} + \beta_{exam} x_{ij3} + b_i + \epsilon_{ij}, \quad j = 1, 2,\ i = 1, \ldots, n,$$

where x_{ij1} is the baseline age of subject i at exam #2, x_{ij2} is the indicator variable for baseline smoking status of subject i at exam #1, x_{ij3} is 1 if $j = 2$ and 0 otherwise, and b_i and ϵ_{ij} are assumed to be independently and normally distributed with means 0 and variances respectively given by σ_b^2 and σ_ϵ^2.

Because a person's SBP changes over time, the two individual SBP readings at each exam are regarded as replicated surrogates. Several measurement error models for SBP reading have been proposed by different researchers (see, e.g., Carroll et al., 1984; Wang et al., 1998; Hall and Ma, 2007). Let T^*_{ijr} be the rth observed SBP reading for subject i at time j, $i = 1, \ldots, n$, $j = 1, 2$, $r = 1, 2$. We consider an error model $\log(T^*_{ijr} - 50) = \log(T_{ij} - 50) + e_{ijr}$ suggested by Wang et al. (1998), where

Table 2.1: Simulation results for M1 (2000 simulations)

	NAI1				NAI2				PML			
	%RB	SD	ASE	CP	%RB	SD	ASE	CP	%RB	SD	ASE	CP
	scenario (i): $\boldsymbol{\eta}$ is known											
β_0	-164.516	0.011	0.010	< 0.001	16.676	0.051	0.046	0.050	-0.564	0.037	0.035	0.942
β_1	-67.742	0.013	0.014	< 0.001	15.972	0.064	0.060	0.529	-0.831	0.040	0.039	0.949
σ_b^2	-80.299	0.003	0.184	1.000	235.840	0.090	0.253	0.943	-2.134	0.016	0.018	0.938
σ_ϵ^2	17.589	0.003	0.035	1.000	2439.115	0.234	0.035	< 0.001	3.653	0.022	0.025	0.960
	scenario (ii): $\boldsymbol{\eta}$ is estimated from internal validation data											
β_0	-	-	-	-	13.722	0.072	0.042	0.198	-0.039	0.040	0.042	0.948
β_1	-	-	-	-	13.121	0.067	0.054	0.610	0.319	0.053	0.049	0.946
σ_b^2	-	-	-	-	183.353	0.079	0.242	0.964	4.571	0.024	0.028	0.957
σ_ϵ^2	-	-	-	-	1975.917	0.220	0.035	< 0.001	-3.142	0.017	0.014	0.933

[†] The NAI1 approach fits a linear mixed model to data with surrogate response S_{ij} ignoring measurement error. The NAI2 approach fits a linear mixed model to data with the transformed surrogate $\tilde{Y}_{ij}$. The PML approach accounts for measurement error.

Table 2.2: Simulation results for M2 (2000 simulations)

	NAI1				NAI2				PML			
	%RB	SD	ASE	CP	%RB	SD	ASE	CP	%RB	SD	ASE	CP
	scenario (i): $\boldsymbol{\eta}$ is known											
β_0	-100.021	0.010	0.010	< 0.001	-0.042	0.021	0.021	0.952	-0.035	0.021	0.022	0.958
β_1	-49.959	0.015	0.015	< 0.001	0.082	0.030	0.030	0.943	0.097	0.030	0.030	0.948
σ_b^2	-75.002	0.003	0.162	1.000	-0.006	0.012	0.162	1.000	-3.277	0.012	0.013	0.949
σ_ϵ^2	25.030	0.003	0.035	1.000	400.122	0.014	0.035	< 0.001	-0.960	0.014	0.014	0.951
	scenario (ii): $\boldsymbol{\eta}$ is estimated from internal validation data											
β_0	-	-	-	-	-0.133	0.037	0.021	0.753	-0.105	0.031	0.030	0.943
β_1	-	-	-	-	0.439	0.043	0.030	0.840	0.215	0.039	0.041	0.956
σ_b^2	-	-	-	-	0.963	0.013	0.161	1.000	-2.718	0.017	0.017	0.948
σ_ϵ^2	-	-	-	-	406.534	0.025	0.035	< 0.001	-2.015	0.018	0.022	0.957

[†] The NAI1 approach fits a linear mixed model to data with surrogate response S_{ij} ignoring measurement error. The NAI2 approach fits a linear mixed model to data with the transformed surrogate $\tilde{Y}_{ij}$. The PML approach accounts for measurement error.

e_{ijr} is normally distributed with mean 0 and variance σ_e^2. Let $S_{ijr} = \log(T_{ijr}^* - 50)$. Therefore, we have

$$S_{ijr} = 2\log(Y_{ij}) + e_{ijr},$$

which is a case of nonlinear measurement error in response. An estimate of the measurement error variance σ_e^2 is obtained using formula (2.20) and is given by 0.009(0.000316), where the value inside the brackets is the standard error.

Table 2.3 reports the analysis results from the proposed method and two naive approaches: one ignores measurement error, and the other uses the transformed surrogates. The estimated regression coefficients β_{age}, β_{smoke}, and β_{exam} from the likelihood approach are 0.027(0.003), -0.120(0.061), and -0.087(0.017), respectively. Age is statistically associated with increasing blood pressure at the 5% level. The negative coefficient for smoking status may suggest an effect of smoking on decreasing blood pressure. As expected, the results from the NAI2 approach are similar to those from the proposed method due to the small value of the measurement error variance. The NAI1 estimates, however, are not comparable to the other two approaches, possibly in part due to a different scale of responses.

Table 2.3: Analysis of data from the Framingham Heart Study

	NAI1[†]			NAI2			PML		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
β_0	4.117	0.030	< 0.001	7.727	0.140	< 0.001	7.729	0.156	< 0.001
β_{age}	0.006	0.001	< 0.001	0.029	0.003	< 0.001	0.027	0.003	< 0.001
β_{smoke}	-0.027	0.012	0.031	-0.122	0.057	0.032	-0.120	0.061	0.048
β_{exam}	-0.020	0.004	< 0.001	-0.086	0.018	< 0.001	-0.087	0.017	< 0.001
σ_b^2	0.036	0.021	0.083	0.782	0.020	< 0.001	0.754	0.040	< 0.001
σ_ϵ^2	0.013	0.018	0.474	0.248	0.018	< 0.001	0.120	0.007	< 0.001

[†] The NAI1 approach fits a linear mixed model to data with surrogate response S_{ij} ignoring measurement error. The NAI2 approach fits a linear mixed model to data with the transformed surrogate $\tilde{Y}_{ij}$. The PML approach accounts for measurement error.

2.7 Discussion

In this chapter, we considered generalized linear mixed models for clustered data with measurement error in the response variable. We mainly focused on regression models for continuous response. It is known that when response error follows the classical additive model, the induced error can be absorbed into the random error term in a linear or linear mixed model. Therefore, naively fitting a mixed model to clustered data gives consistent estimates of the fixed effects for linear models. The estimate of the conditional variance of the true response, however, is not valid. In other cases where the measurement error is nonlinear, naive analysis with error ignored may lead to biased estimates and invalid inference. For the example on exponential measurement error, we showed that naively fitting mixed model can lead to seriously biased estimates of fixed effects. Although standard methods can be naively applied to the transformed surrogates, the bias, however, may still be large depending on the measurement error variance.

We formulated the marginal likelihood of the observed data and proposed a two-stage pseudo maximum likelihood approach when the error parameters are estimated from validation data. Our simulation studies show that estimators from likelihood-based approaches are consistent.

It is worth pointing out that a major problem with likelihood-based approaches is that it may be computationally intensive. The accuracy of the estimates relies on the order of the quadrature approximations to the integrals involved in the likelihood function. We found in our simulation that a quadrature approximation with order 5 performs well enough for a single integral. However, as the number of random components increases, more quadrature points are required in order to obtain a good approximation.

2.8 Technical Details

2.8.1 Naive estimators under exponential measurement error model

Suppose a general LMM is given by (2.2). When measurement error process follows the exponential model $\mathrm{E}[S_{ij}|Y_{ij}] = \exp(\gamma Y_{ij})$ for some parameter γ, equation (2.8) becomes

$$\begin{aligned}
\mathbf{0} &= \mathrm{E}_{X,Z}\mathrm{E}_{b|X,Z}\mathrm{E}_{Y|X,Z,b}\left[\frac{\sum_{j=1}^{m}\left\{\exp(Y_{ij}\gamma) - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*\right\}\mathbf{X}_{ij}}{\sigma_\epsilon^{*2}}\right. \\
&\quad \left. - \frac{\sigma_b^{*2}\{\sum_{j=1}^{m}\left[\exp(Y_{ij}\gamma) - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*\right]Z_{ij}\}\sum_{j=1}^{m}\mathbf{X}_{ij}Z_{ij}}{\sigma_\epsilon^{*2}(\sigma_b^{*2}\sum_{j=1}^{m}Z_{ij}^2 + \sigma_\epsilon^{*2})}\right] \\
&= \mathrm{E}_{X,Z}\mathrm{E}_{b|X,Z}\left[\frac{\sum_{j=1}^{m}\left\{\exp\left[(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + Z_{ij}b_i)\gamma + \sigma_\epsilon^2\gamma^2/2\right] - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*\right\}\mathbf{X}_{ij}}{\sigma_\epsilon^{*2}}\right. \\
&\quad \left. - \frac{\sigma_b^{*2}\{\sum_{j=1}^{m}(\exp\left\{(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + Z_{ij}b_i)\gamma + \sigma_\epsilon^2\gamma^2/2\right\} - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*)Z_{ij}\}\sum_{j=1}^{m}\mathbf{X}_{ij}Z_{ij}}{\sigma_\epsilon^{*2}\left(\sigma_b^{*2}\sum Z_{ij}^2 + \sigma_\epsilon^{*2}\right)}\right] \\
&= \mathrm{E}_{X,Z}\left[\frac{\sum_{j=1}^{m}\left[\exp(\gamma^2 Z_{ij}^2\sigma_b^2/2 + \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\gamma + \sigma_\epsilon^2\gamma^2/2) - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*\right]\mathbf{X}_{ij}}{\sigma_\epsilon^{*2}}\right. \\
&\quad \left. - \frac{\sigma_b^{*2}\{\sum_{j=1}^{m}\left[\exp(\gamma^2 Z_{ij}^2\sigma_b^2/2 + \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\gamma + \sigma_\epsilon^2\gamma^2/2) - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*\right]Z_{ij}\}\sum_{j=1}^{m}\mathbf{X}_{ij}Z_{ij}}{\sigma_\epsilon^{*2}\left(\sigma_b^{*2}\sum_{j=1}^{m}Z_{ij}^2 + \sigma_\epsilon^{*2}\right)}\right] \\
&= \sum_{j=1}^{m}\mathrm{E}_{X,Z}\left[\frac{\exp(\gamma^2 Z_{ij}^2\sigma_b^2/2 + \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\gamma + \sigma_\epsilon^2\gamma^2/2) - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^*}{\sigma_\epsilon^{*2}}\right. \\
&\qquad \left. \times \left(\mathbf{X}_{ij} - \frac{\sigma_b^{*2}Z_{ij}\sum_k \mathbf{X}_{ik}Z_{ik}}{\sigma_b^{*2}\sum_k Z_{ik}^2 + \sigma_\epsilon^{*2}}\right)\right]. \qquad (2.23)
\end{aligned}$$

One can see that there is generally no closed form for the relationship between $\boldsymbol{\beta}$ and the naive estimator $\boldsymbol{\beta}^*$.

2.8.2 Adjusted variance of $\hat{\boldsymbol{\theta}}_p$

Let $\mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{i=1}^n \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})$. A Taylor series expansion up to order one about the point $(\boldsymbol{\theta}, \boldsymbol{\eta})$ for $\mathbf{U}^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}})$ is given by

$$\mathbf{U}^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}}) \approx \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta}) + \frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}) + \frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}). \tag{2.24}$$

But $\mathbf{U}^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}}) = \mathbf{0}$. It is easy to show that the covariance matrix for $\hat{\boldsymbol{\theta}}_p$ is approximately

$$\begin{aligned} \boldsymbol{\Sigma}^* \approx\ & \mathrm{E}\left[\frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right]^{-1} \mathrm{E}\left[\mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})\mathbf{U}^{*\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta})\right] \left\{\mathrm{E}\left[\frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right]^{\mathrm{T}}\right\}^{-1} \\ & + \mathrm{E}\left[\frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right]^{-1} \mathrm{E}\left[\frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right] \mathrm{E}\left[(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}\right] \\ & \cdot \mathrm{E}\left[\frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}\right]^{\mathrm{T}} \left\{\mathrm{E}\left[\frac{\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right]^{\mathrm{T}}\right\}^{-1}. \end{aligned}$$

But under suitable regularity conditions, $\mathrm{E}\left[\mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})\mathbf{U}^{*\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta})\right] = \mathrm{E}\left[-\partial \mathbf{U}^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right]$, and when $\hat{\boldsymbol{\eta}}$ is unbiased for $\boldsymbol{\eta}$, $\mathrm{E}\left[(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^{\mathrm{T}}\right] = \boldsymbol{\mathcal{J}}^{*-1}(\boldsymbol{\eta})$. This yields the asymptotic covariance matrix given by (2.19).

Chapter 3

Misclassification in Correlated Binary Responses: An Estimating Equations Approach

3.1 Introduction

Longitudinal studies or clustered studies are important tools in epidemiological, clinical and social science research. In longitudinal studies, the response variable, often associated with a set of covariates, is observed on individuals repeatedly over a certain period of time. In clustered studies, such as household surveys, responses are often recorded from all members of the same family. The repeated responses within the same cluster or subject are typically correlated. Various models have been developed for analysis of such data, and a wide variety of estimation techniques have been proposed. In contrast to conditional models (e.g., transition models or mixed effects models), marginal models characterize the dependence of responses on covariates at the population level without including unobserved random components or past outcomes in the linear predictor. One compelling feature of such methods lies in their minimal model assumptions. For example, generalized estimating equations (GEE), proposed by Liang and Zeger (1986), focus on estimation of mean parameters, with association parameters between outcomes treated as nuisance. Extensions of the

GEE approach can be found, for instance, in Miller et al. (1993) and Molenberghs and Lesaffre (1999), among many others.

In many epidemiological studies, association structures among repeated outcomes are of scientific interest. For example, understanding the correlation of disease status among household members is often of primary interest in familial studies. Prentice (1988), Carey et al. (1993) and Yi and Cook (2002) extended the GEE approach by specifying a second set of generalized estimating equations to estimate association parameters for binary data. Those methods are useful to conduct simultaneous inference about the mean and association parameters. The validity of these methods requires a critical condition: variables must be precisely measured. However, this requirement is often violated in practice. Misclassification commonly arises with categorical data collected from epidemiological studies or longitudinal surveys. For example, a disease infection status may be wrongly identified due to a poor diagnostic test. If a survey questionnaire is not well designed, such as ambiguous wording in an ordinal item, it may lead to wrong interpretation by respondents and hence results in an incorrect category for the response variable.

With covariates subject to error, there has been extensive research on studying error effects and developing valid inferential procedures under various models. It is known that naive analysis ignoring covariate error generally leads to biased estimates and invalid inference (Carroll et al., 2006). However, little attention has been directed to problems with error-contaminated outcomes, such as misclassified responses, although Neuhaus (1999, 2002) discussed this problem under generalized mixed effects models, where a simple scenario of misclassification is considered. In this chapter, we consider marginal regression models for correlated binary data in the presence of response misclassification. We propose estimating equations methods that can correct for misclassification effects under a variety of practical settings. The proposed methods have several appealing features. They accommodate simultaneous inference for both marginal mean and association parameters; they can handle various misclassification scenarios, including cases with validation subsamples or replicates. Furthermore, the proposed methods are robust to model misspecification in a sense that no full distributional assumptions are required.

The rest of the chapter is organized as follows. Section 3.2 describes basic notation and model assumptions for the response and misclassification processes. Section 3.3 presents the proposed method for the case where misclassification parameters are known. In Sections 3.4 and 3.5, we develop inference methods that can handle unknown parameters associated with the misclassification process. Simulation studies and applications to real data are respectively presented in sections 3.6 and 3.7. Concluding remarks are presented in Section 3.8.

3.2 Notation and Model Formulation

3.2.1 The response process

Let Y_{ij} be the binary response for the jth subject in cluster i (or the jth measurement of subject i) and $\mathbf{X}_{ij}$ be the corresponding covariate vector, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, where n is the number of clusters, and m_i is the number of subjects in cluster i. Denote $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$ and $\mathbf{X}_i = (\mathbf{X}_{i1}^{\mathrm{T}}, \ldots, \mathbf{X}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Let $\mu_{ij} = \mathrm{E}[Y_{ij}|\mathbf{X}_i]$ be the marginal mean of the response, and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$. A generalized regression model is used to link μ_{ij} to the covariates, where $\mathrm{E}[Y_{ij}|\mathbf{X}_i] = \mathrm{E}[Y_{ij}|\mathbf{X}_{ij}]$ is assumed (e.g., Pepe and Anderson, 1994). That is,

$$g(\mu_{ij}) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, and $g(\cdot)$ is a monotone link function. Typical choices of $g(\cdot)$ include logit, probit, and complementary log-log functions. The variance of the response Y_{ij} is specified as $\mathrm{var}(Y_{ij}|\mathbf{X}_i) = \mu_{ij}(1 - \mu_{ij})$ accordingly.

When the mean parameters are of primary interest and association parameters are treated as nuisance, the GEE method discussed by Liang and Zeger (1986) is well suited for parameter estimation. However, to facilitate inference for association parameters that are often of interest for clustered data analysis, one needs to derive a second set of estimating functions to feature association structures. Here we assume that Y_{ij} and $Y_{i'j'}$ are independent when $i \neq i'$, but Y_{ij} and $Y_{ij'}$ may be correlated for $j \neq j'$. Let $C_{ijj'} = Y_{ij}Y_{ij'}$, $\mathbf{C}_i = (C_{ijj'}, j < j')^{\mathrm{T}}$, $\mu_{ijj'} = \mathrm{E}[C_{ijj'}|\mathbf{X}_i]$, and $\boldsymbol{\xi}_i = (\mu_{ijj'}, j < j')^{\mathrm{T}}$.

For $j < j'$, let the odds ratio for Y_{ij} and $Y_{ij'}$ be

$$\psi_{ijj'} = \frac{\Pr(Y_{ij} = 1, Y_{ij'} = 1|\boldsymbol{X}_i) \cdot \Pr(Y_{ij} = 0, Y_{ij'} = 0|\mathbf{X}_i)}{\Pr(Y_{ij} = 1, Y_{ij'} = 0|\mathbf{X}_i) \cdot \Pr(Y_{ij} = 0, Y_{ij'} = 1|\mathbf{X}_i)},$$

which is commonly used as an association measure for binary data. It is often assumed that $\Pr(Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}|\boldsymbol{X}_i) = \Pr(Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}|\mathbf{X}_{ij}, \mathbf{X}_{ij'})$. The odds ratios are customarily modeled as

$$\log \psi_{ijj'} = \mathbf{u}_{ijj'}^{\mathrm{T}}\boldsymbol{\alpha},$$

where $\mathbf{u}_{ijj'}$ is a set of pair-specific covariates featuring various association structures such as autoregressive or exchangeable structure between Y_{ij} and $Y_{ij'}$. The relationship between $\mu_{ijj'}$ and $\psi_{ijj'}$ is given by

$$\mu_{ijj'} = \begin{cases} \dfrac{a_{ijj'} - [a_{ijj'}^2 - 4(\psi_{ijj'} - 1)\psi_{ijj'}\mu_{ij}\mu_{ij'}]^{1/2}}{2(\psi_{ijj'} - 1)}, & \text{if } \psi_{ijj'} \neq 1, \\ \mu_{ij}\mu_{ij'}, & \text{if } \psi_{ijj'} = 1, \end{cases}$$

where $a_{ijj'} = 1 - (1 - \psi_{ijj'})(\mu_{ij} + \mu_{ij'})$ (e.g., Lipsitz et al., 1991; Yi and Cook, 2002).

3.2.2 Marginal model for the misclassification process

When the response Y_{ij} is subject to misclassification, a surrogate version S_{ij} is observed instead of Y_{ij}. Let $H_{ij} = \mathrm{I}(S_{ij} = Y_{ij})$ be the indicator variable for misclassification, $\mathbf{H}_i = (H_{i1}, \ldots, H_{im_i})^{\mathrm{T}}$, and $\mathbf{S}_i = (\ S_{i1}, \ldots, S_{im_i})^{\mathrm{T}}$. The marginal probability of misclassifying Y_{ij} is assumed to depend only on the information concerning the jth subject in cluster i, i.e., $\Pr(S_{ij} = 1|\mathbf{Y}_i, \mathbf{X}_i) = \Pr(S_{ij} = 1|Y_{ij}, \mathbf{X}_i)$. Let $\tau_{0ij} = \Pr(H_{ij} = 1|Y_{ij} = 0, \mathbf{X}_i)$ and $\tau_{1ij} = \Pr(H_{ij} = 1|Y_{ij} = 1, \mathbf{X}_i)$ be misclassification probabilities. Alternatively, if we let $\tau_{ij}(y_{ij}) = \Pr(H_{ij} = 1|Y_{ij} = y_{ij}, \mathbf{X}_i)$, then $\tau_{ij}(y_{ij}) = (1 - y_{ij})\tau_{0ij} + y_{ij}\tau_{1ij}$.

Logistic models may be employed to characterize these probabilities:

$$\begin{aligned} \text{logit}(\tau_{0ij}) &= \mathbf{L}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_0, & (3.1)\\ \text{logit}(\tau_{1ij}) &= \mathbf{L}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_1, & (3.2)\end{aligned}$$

where $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are vectors of associated regression parameters, and $\mathbf{L}_{ij}$ is a set of covariates that reflects various misclassification mechanisms. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^{\mathrm{T}}, \boldsymbol{\gamma}_1^{\mathrm{T}})^{\mathrm{T}}$. Covariates $\mathbf{L}_{ij}$ may be specified as various forms to feature different misclassification processes. It may contain the entire covariate vector $\mathbf{X}_{ij}$ in some situations; while in extreme cases, $\mathbf{L}_{ij}$ can be constant 1, that is, two parameters γ_0 and γ_1 are sufficient to describe the misclassification mechanism. The latter scenario corresponds to a homogeneous misclassification across all observations and clusters, with misclassification independent of covariates and the other outcomes: $\tau_{0ij} = \tau_0 = \text{expit}(\gamma_0)$, and $\tau_{1ij} = \tau_1 = \text{expit}(\gamma_1)$, where $\text{expit}(u) = \exp(u)/\{1 + \exp(u)\}$.

3.2.3 Association model for the misclassification process

When observations in the same cluster are measured using similar defective devices or by the same person, misclassifications on two observations within the same cluster are typically correlated with each other. In the same manner of characterizing the association structure for response process, we measure the dependence between H_{ij} and $H_{ij'}$ using odds ratios

$$\begin{aligned} \lambda_{ijj'}(y_{ij}, y_{ij'}) &= \frac{\Pr(H_{ij} = 1, H_{ij'} = 1|\mathbf{Y}_i = \mathbf{y}_i, \mathbf{X}_i)}{\Pr(H_{ij} = 1, H_{ij'} = 0|\mathbf{Y}_i = \mathbf{y}_i, \mathbf{X}_i)} \\ &\qquad \times \frac{\Pr(H_{ij} = 0, H_{ij'} = 0|\mathbf{Y}_i = \mathbf{y}_i, \mathbf{X}_i)}{\Pr(H_{ij} = 0, H_{ij'} = 1|\mathbf{Y}_i = \mathbf{y}_i, \mathbf{X}_i)}, \end{aligned}$$

where it is assumed that $\Pr(H_{ij} = h_{ij}, H_{ij'} = h_{ij'}|\mathbf{Y}_i = \mathbf{y}_i, \mathbf{X}_i) = \Pr(H_{ij} = h_{ij}, H_{ij'} = h_{ij'}|Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, \mathbf{X}_i)$. The odds ratio $\lambda_{ijj'}(y_{ij}, y_{ij'})$ can be modeled by

$$\log\{\lambda_{ijj'}(y_{ij}, y_{ij'})\} = \mathbf{u}_{ijj'}^{*\mathrm{T}}\boldsymbol{\nu}_{y_{ij}, y_{ij'}},$$

where $\mathbf{u}^*_{ijj'}$ is a vector of covariates that features various types of dependence, and $\boldsymbol{\nu}_{y_{ij},y_{ij'}}$ is a vector of regression coefficients that may vary with the values of y_{ij} and $y_{ij'}$. Let $\boldsymbol{\nu} = (\boldsymbol{\nu}^{\mathrm{T}}_{11}, \boldsymbol{\nu}^{\mathrm{T}}_{10}, \boldsymbol{\nu}^{\mathrm{T}}_{01}, \boldsymbol{\nu}^{\mathrm{T}}_{00})^{\mathrm{T}}$. Let $\boldsymbol{\eta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$ be the vector of parameters associated with the misclassification process.

For $j < j'$, let $F_{ijj'} = H_{ij}H_{ij'}$, and $\mathbf{F}_i = (F_{ijj'}, j < j')^{\mathrm{T}}$. Let $\zeta_{ijj'}(y_{ij}, y_{ij'}) = \mathrm{E}[F_{ijj'}|Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, \mathbf{X}_i]$, and $\boldsymbol{\zeta}_i = \mathrm{E}[\mathbf{F}_i|\mathbf{Y}_i,\mathbf{X}_i]$. Again, we assume that $\mathrm{E}[F_{ijj'}|\mathbf{Y}_i, \mathbf{X}_i] = \mathrm{E}[F_{ijj'}|Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, \mathbf{X}_i]$. The relationship between $\zeta_{ijj'}(y_{ij}, y_{ij'})$ and $\lambda_{ijj'}(y_{ij}, y_{ij'})$ is given by

$$\zeta_{ijj'}(y_{ij}, y_{ij'}) = \begin{cases} \Big\{a^*_{ijj'}(y_{ij}, y_{ij'}) - \big[a^{*2}_{ijj'}(y_{ij}, y_{ij'}) - 4\left\{\lambda_{ijj'}(y_{ij}, y_{ij'}) - 1\right\} \\ \qquad \times \lambda_{ijj'}(y_{ij}, y_{ij'})\tau_{ij}(y_{ij})\tau_{ij'}(y_{ij'})\big]^{1/2}\Big\}/\{2[\lambda_{ijj'}(y_{ij}, y_{ij'}) - 1]\}, \\ \qquad\qquad\qquad\qquad \text{if } \lambda_{ijj'}(y_{ij}, y_{ij'}) \neq 1, \\ \tau_{ij}(y_{ij})\tau_{ij'}(y_{ij'}), \qquad\qquad \text{if } \lambda_{ijj'}(y_{ij}, y_{ij'}) = 1, \end{cases}$$

where $a^*_{ijj'}(y_{ij}, y_{ij'}) = 1 - \{1 - \lambda_{ijj'}(y_{ij}, y_{ij'})\}\{\tau_{ij}(y_{ij}) + \tau_{ij'}(y_{ij'})\}$.

Let $\mu^S_{ijj'} = \mathrm{E}[S_{ij}S_{ij'}|\mathbf{X}_i]$ be the marginal mean of $S_{ij}S_{ij'}$ given covariates. In Section 3.9.1 we show that $\mu^S_{ijj'} \neq \mu_{ijj'}$. Even under the assumption that misclassifications of paired responses are independent of each other, i.e., $\Pr(S_{ij} = s_{ij}, S_{ij'} = s_{ij'}|\mathbf{Y}_i, \mathbf{X}_i) = \Pr(S_{ij} = s_{ij}|\mathbf{Y}_i, \mathbf{X}_i)\Pr(S_{ij'} = s_{ij'}|\mathbf{Y}_i, \mathbf{X}_i)$, $\mu^S_{ijj'}$ is not equal to $\mu_{ijj'}$. As a consequence, replacing Y_{ij} with S_{ij} in the marginal analysis (to be discussed in Section 3.3) often leads to biased inference.

3.3 Estimating Equations

3.3.1 Estimating equations under the true model

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ be the vector of response parameters, $\mathbf{D}_{1i} = \partial\boldsymbol{\mu}^{\mathrm{T}}_i/\partial\boldsymbol{\beta}$, and $\mathbf{B}_{1i} = \mathrm{diag}\{\mu_{i1}(1 - \mu_{i1}), \ldots, \mu_{im_i}(1 - \mu_{im_i})\}$. When the response variable is free of misclassification, estimates of mean parameters $\boldsymbol{\beta}$ can be obtained by solving the first-order estimating equations

$$\sum_{i=1}^{n} \mathbf{U}_{1i}(\boldsymbol{\theta}) = \mathbf{0}, \tag{3.3}$$

where $\mathbf{U}_{1i}(\boldsymbol{\theta}) = \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\boldsymbol{\epsilon}_{1i}$, $\boldsymbol{\epsilon}_{1i} = \mathbf{Y}_i - \boldsymbol{\mu}_i$, $\mathbf{V}_{1i} = \mathrm{cov}(\mathbf{Y}_i) = \mathbf{B}_{1i}^{1/2}\mathbf{R}_{1i}(\boldsymbol{\theta})\mathbf{B}_{1i}^{1/2}$, and $\mathbf{R}_{1i}(\boldsymbol{\theta})$ is the correlation matrix of $\mathbf{Y}_i$ with off-diagonal entries given by

$$\rho_{ijj'} = \frac{\mu_{ijj'} - \mu_{ij}\mu_{ij'}}{\sqrt{\mu_{ij}(1-\mu_{ij})}\,\sqrt{\mu_{ij'}(1-\mu_{ij'})}}, \quad j \neq j'.$$

Let $\mathbf{D}_{2i} = \partial\boldsymbol{\xi}_i^{\mathrm{T}}/\partial\boldsymbol{\alpha}$. Then the second-order estimating equations (Prentice 1988) for association parameters $\boldsymbol{\alpha}$ can be written as

$$\sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\theta}) = \mathbf{0}, \tag{3.4}$$

where $\mathbf{U}_{2i}(\boldsymbol{\theta}) = \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\boldsymbol{\epsilon}_{2i}$, $\boldsymbol{\epsilon}_{2i} = \mathbf{C}_i - \boldsymbol{\xi}_i$, and $\mathbf{V}_{2i}$ is a working covariance matrix for $\mathbf{C}_i$. Because the correlation between C_{ij} and $C_{ij'}$ typically involves third and fourth moments, one often uses an independent working matrix $\mathbf{V}_{2i} = \mathrm{diag}(\mu_{ijj'}(1 - \mu_{ijj'}); j < j')$ in order to avoid modeling higher order moments (e.g., Lipsitz et al., 1991; Yi and Cook, 2002).

3.3.2 Estimating equations in the presence of misclassification

When responses are subject to misclassification, the estimating functions in (3.3) and (3.4) with Y_{ij} replaced by the observed surrogate S_{ij} are no longer unbiased. In other words, naive analysis ignoring misclassifications usually yields inconsistent estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. In this section we construct modified estimating equations to correct the bias caused by misclassification.

Our proposed strategy is to construct modified estimating functions $\mathbf{U}_{1i}^{*}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i)$

and $\mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i)$ based on the observed data so that

$$\mathrm{E}[\mathbf{U}^*_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i)|\mathbf{Y}_i, \mathbf{X}_i] \quad = \quad \mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i), \tag{3.5}$$

and

$$\mathrm{E}[\mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i)|\mathbf{Y}_i, \mathbf{X}_i] \quad = \quad \mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i). \tag{3.6}$$

It can be shown that under mild regularity conditions, solving

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{U}^*_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i) \\ \mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i) \end{pmatrix} = \mathbf{0} \tag{3.7}$$

gives a consistent estimator for $\boldsymbol{\theta}$.

To highlight the proposed method, we first assume that the parameters $\boldsymbol{\eta}$ associated with the misclassification process have a known value $\boldsymbol{\eta}_0$. An unbiased surrogate of Y_{ij}, which is a function of S_{ij} and the misclassification probabilities, can be formulated as

$$Y^*_{ij} = \frac{S_{ij} - 1 + \tau_{0ij}}{\tau_{0ij} + \tau_{1ij} - 1},$$

with $\mathrm{E}[Y^*_{ij}|\mathbf{Y}_i, \mathbf{X}_i] = Y_{ij}$. Although Y^*_{ij} is unbiased for Y_{ij} for all j, $Y^*_{ij}Y^*_{ij'}$ is not necessarily an unbiased surrogate for $C_{ijj'}$ except for cases where misclassifications are independent. Therefore, for $C_{ijj'}$ we construct an unbiased surrogate as follows

$$C^*_{ijj'} = \frac{b_0 + (S_{ij} - b_1)(S_{ij'} - b_2)}{b_3},$$

where

$$\begin{aligned}
b_0 &= (1 - b_1)\tau_{0ij'} + (1 - b_2)\tau_{0ij} - \zeta_{ijj'}(0,0) - (1 - b_1)(1 - b_2), \\
b_1 &= \{\tau_{0ij} + \tau_{0ij'} + \tau_{1ij'} - 1 - \zeta_{ijj'}(0,1) - \zeta_{ijj'}(0,0)\} / (\tau_{1ij'} + \tau_{0ij'} - 1), \\
b_2 &= \{\tau_{0ij'} + \tau_{0ij} + \tau_{1ij} - 1 - \zeta_{ijj'}(1,0) - \zeta_{ijj'}(0,0)\} / (\tau_{1ij} + \tau_{0ij} - 1), \quad \text{and} \\
b_3 &= b_0 + b_1 b_2 - b_1\tau_{1ij'} - b_2\tau_{1ij} + \zeta_{ijj'}(1,1).
\end{aligned}$$

In Section 3.9.2 we outline the proof that $\mathrm{E}[Z^*_{ijj'}|\mathbf{Y}_i, \mathbf{X}_i] = C_{ijj'}$ for $j \neq j'$.

Let $\mathbf{Y}^*_i = (Y^*_{i1}, \ldots, Y^*_{im_i})^{\mathrm{T}}$, and $\mathbf{C}^*_i = (C^*_{ijj'}, j < j')^{\mathrm{T}}$. Define

$$\begin{pmatrix} \mathbf{U}^*_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}_0; \mathbf{S}_i, \mathbf{X}_i) \\ \mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}_0; \mathbf{S}_i, \mathbf{X}_i) \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\boldsymbol{\epsilon}^*_{1i} \\ \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\boldsymbol{\epsilon}^*_{2i} \end{pmatrix}, \tag{3.8}$$

where the modified residual vectors $\boldsymbol{\epsilon}^*_{1i}$ and $\boldsymbol{\epsilon}^*_{2i}$ are given by $\mathbf{Y}^*_i - \boldsymbol{\mu}_i$ and $\mathbf{C}^*_i - \boldsymbol{\xi}_i$, respectively. It is straightforward to verify that $\mathbf{U}^*_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}_0; \mathbf{S}_i, \mathbf{X}_i)$ and $\mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}_0; \mathbf{S}_i, \mathbf{X}_i)$ satisfy (3.5) and (3.6).

Let $\mathbf{U}^*_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = (\mathbf{U}^*_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)^{\mathrm{T}}, \mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}_0)^{\mathrm{T}})^{\mathrm{T}}$, and $\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\beta}}_0^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}_0^{\mathrm{T}})^{\mathrm{T}}$ be the solution of estimating equation $\sum_{i=1}^n \mathbf{U}^*_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \mathbf{0}$. Define $\boldsymbol{\Gamma}^*_0(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \mathrm{E}\left[\partial \mathbf{U}^*_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)/\partial \boldsymbol{\theta}^{\mathrm{T}}\right]$, and $\boldsymbol{\Sigma}^*_0(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \mathrm{E}\left[\mathbf{U}^*_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)\mathbf{U}^*_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)^{\mathrm{T}}\right]$. Under suitable regularity conditions, it can be shown that $n^{1/2}\left(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}\right)$ has an asymptotic normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}_0^{*-1}\boldsymbol{\Sigma}^*_0\left[\boldsymbol{\Gamma}_0^{*-1}\right]^{\mathrm{T}}$.

At the end of this section we note that there exist alternative approaches to correcting estimation bias induced by misclassification. A straightforward correction is given by

$$\begin{aligned} \mathbf{U}^\dagger_i(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i) &= \mathbf{U}_i(\boldsymbol{\theta}; \mathbf{S}_i, \mathbf{X}_i) - \mathrm{E}\left[\mathrm{E}\left\{\mathbf{U}_i(\boldsymbol{\theta}; \mathbf{S}_i, \mathbf{X}_i)|\mathbf{Y}_i, \mathbf{X}_i\right\}|\mathbf{X}_i\right] \\ &= \begin{pmatrix} \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\boldsymbol{\epsilon}^\dagger_{1i} \\ \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\boldsymbol{\epsilon}^\dagger_{2i} \end{pmatrix}, \end{aligned} \tag{3.9}$$

where $\boldsymbol{\epsilon}^\dagger_{1i} = \mathbf{S}_i - \boldsymbol{\mu}^S_i$, $\boldsymbol{\mu}^S_i = (\mu^S_{i1}, \ldots, \mu^S_{im_i})^{\mathrm{T}}$, $\boldsymbol{\epsilon}^\dagger_{2i} = \mathbf{C}^\dagger_i - \boldsymbol{\xi}^\dagger_i$, $\mathbf{C}^\dagger_i = (S_{ij}S_{ij'}, j < j')^{\mathrm{T}}$, and $\boldsymbol{\xi}^\dagger_i = \mathrm{E}\left[\mathbf{C}^\dagger_i|\mathbf{X}_i\right]$. Both this approach and our proposed approach use the naive covariance matrix $\mathbf{V}_{1i}$ that is for the underlying true $\mathbf{Y}_i$. One can see that the components in $\boldsymbol{\epsilon}^\dagger_{1i}$ and $\boldsymbol{\epsilon}^*_{1i}$ have relationship $\epsilon^\dagger_{1ij} = \epsilon^*_{1ij}(\tau_{0ij} + \tau_{1ij} - 1)$. If the misclassification process follows the simplest model and does not depend on covariates, i.e., $\tau_{0ij} = \tau_0$ and $\tau_{1ij} = \tau_1$, then the two approaches are equivalent, since the factor $(\tau_0 + \tau_1 - 1)$ in the estimating equations can be canceled. However, the equivalence does not hold when the misclassification process involves covariates. Another alternative correction approach is given by

$$\mathbf{U}_i^{**}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{X}_i) = \begin{pmatrix} \mathbf{D}_{1i}\mathbf{V}_{1i}^{*-1}\boldsymbol{\epsilon}_{1i}^* \\ \mathbf{D}_{2i}\mathbf{V}_{2i}^{*-1}\boldsymbol{\epsilon}_{2i}^* \end{pmatrix}, \tag{3.10}$$

where $\mathbf{V}_{1i}^*$ is the correct covariance matrix for $\mathbf{Y}_i^*$, and $\mathbf{V}_{2i}^*$ is a working covariance matrix for $\mathbf{C}_i^*$. One may have efficiency gain by using correct covariance matrices. In following sections, however, we still use the estimating functions given by (3.8) with naive covariance matrices $\mathbf{V}_{1i}$ and $\mathbf{V}_{2i}$.

3.4 Inference Method with Validation Subsample Available

In order to use (3.7) to perform inference about $\boldsymbol{\theta}$, it is critical that parameter $\boldsymbol{\eta}$ associated with misclassification is known. In practice, however, this condition is often not satisfied. The parameter $\boldsymbol{\eta}$ must be estimated from an additional source of data. It is then important to accommodate induced variation in inferential procedures for $\boldsymbol{\theta}$. In this and next sections, we develop modified estimation algorithms to cover two practical situations - either a validation subsample or replicates of surrogates are available for estimation of $\boldsymbol{\eta}$.

3.4.1 Estimating equations for $\boldsymbol{\eta}$

When an internal validation subsample is available, one can develop estimating equations for the parameters associated with the misclassification process.

If the values of all misclassification indicators H_{ij}'s were observed, estimates of $\boldsymbol{\eta}$ could be obtained as the solution to estimating equations

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{D}_{\eta 1i}\mathbf{V}_{\eta 1i}^{-1} \cdot (\mathbf{H}_i - \boldsymbol{\tau}_i) \\ \mathbf{D}_{\eta 2i}\mathbf{V}_{\eta 2i}^{-1} \cdot (\mathbf{F}_i - \boldsymbol{\zeta}_i) \end{pmatrix} = \mathbf{0},$$

where $\mathbf{D}_{\eta 1i} = \partial \boldsymbol{\tau}_i^{\mathrm{T}} / \partial \boldsymbol{\gamma}$, $\mathbf{D}_{\eta 2i} = \partial \boldsymbol{\zeta}_i^{\mathrm{T}} / \partial \boldsymbol{\nu}$, $\mathbf{V}_{\eta 1i} = \mathbf{B}_{\eta 1i}^{1/2} \mathbf{R}_{\eta 1i} \mathbf{B}_{\eta 1i}^{1/2}$, $\mathbf{B}_{\eta 1i} = \mathrm{diag}\{\tau_{i1}(y_{i1})[1 - \tau_{i1}(y_{i1})], \ldots, \tau_{im_i}(y_{im_i})[1 - \tau_{im_i}(y_{im_i})]\}$, $\mathbf{R}_{\eta 1i}$ is the correlation matrix of $\mathbf{H}_i$, and $\mathbf{V}_{\eta 2i}$ is often assumed to be an independence working covariance matrix to avoid specifying third and higher order moments of responses.

However, we do not observe the value of H_{ij} unless subject j is in the validation subsample. Let $\delta_{ij} = 1$ if the jth subject in cluster i belongs to the validation subsample and $\delta_{ij} = 0$ otherwise. We assume that the selection of subjects may depend on the covariates but not on the observed surrogates. This assumption ensures that $\boldsymbol{\eta}$ can be estimated from fitting a prospective model to the validation data without adjusting for the sampling scheme. Therefore, indicators δ_{ij}'s can be treated as constants.

We add a superscript δ to each vector and matrix to indicate the components corresponding to the validation subsample. To be specific, let $\mathbf{Q}_{1i}(\boldsymbol{\eta}) = \mathbf{D}_{\eta 1i}^{\delta} \left[\mathbf{V}_{\eta 1i}^{\delta}\right]^{-1} \cdot (\mathbf{H}_i - \boldsymbol{\tau}_i)^{\delta}$, and $\mathbf{Q}_{2i}(\boldsymbol{\eta}) = \mathbf{D}_{\eta 2i}^{\delta} \left[\mathbf{V}_{\eta 2i}^{\delta}\right]^{-1} \cdot (\mathbf{F}_i - \boldsymbol{\zeta}_i)^{\delta}$. Therefore, $\boldsymbol{\eta}$ can be estimated by solving

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{Q}_{1i}(\boldsymbol{\eta}) \\ \mathbf{Q}_{2i}(\boldsymbol{\eta}) \end{pmatrix} = \mathbf{0}. \tag{3.11}$$

3.4.2 Estimation equations for $\boldsymbol{\theta}$

Because of the availability of true response measurements in the validation subsample, estimating functions of $\boldsymbol{\theta}$ can be improved in terms of efficiency gain by pooling the validation subsample and the primary data set, as opposed to using only the primary data set that contains surrogate values. The pooling of the two data sets results in reduced number of surrogate measurements that need to be corrected. Therefore, modified estimating equations from the pool sample will lead to more efficient estimator for $\boldsymbol{\theta}$.

To this end, we define

$$\tilde{Y}_{ij} \quad = \quad (1 - \delta_{ij}) Y_{ij}^* + \delta_{ij} Y_{ij},$$

and

$$\tilde{C}_{ijj'} \;=\; \{1-(1-\delta_{ij})(1-\delta_{ij'})\}\,\tilde{Y}_{ij}\tilde{Y}_{ij'} + (1-\delta_{ij})(1-\delta_{ij'})C^*_{ijj'}.$$

Thus, $\tilde{Y}_{ij} = Y_{ij}$ if the jth subject in cluster i is in the validation subsample, $\tilde{Y}_{ij} = Y^*_{ij}$ otherwise, $\tilde{C}_{ijj'} = \tilde{Y}_{ij}\tilde{Y}_{ij'}$ if either Y_{ij} or $Y_{ij'}$ or both are available, and $\tilde{C}_{ijj'} = C^*_{ijj'}$ otherwise. Let $\tilde{\mathbf{Y}}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{im_i})^{\mathrm{T}}$ and $\tilde{\mathbf{C}}_i = (\tilde{C}_{ijj'}, j < j')^{\mathrm{T}}$. Let $\tilde{\mathbf{U}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{Y}}_i, \mathbf{X}_i) = \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\tilde{\boldsymbol{\epsilon}}_{1i}$, and $\tilde{\mathbf{U}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{C}}_i, \mathbf{X}_i) = \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\tilde{\boldsymbol{\epsilon}}_{2i}$, where $\tilde{\boldsymbol{\epsilon}}_{1i} = \tilde{\mathbf{Y}}_i - \boldsymbol{\mu}_i$ and $\tilde{\boldsymbol{\epsilon}}_{2i} = \tilde{\mathbf{C}}_i - \boldsymbol{\xi}_i$. The augmented version of (3.7) is then given by

$$\sum_{i=1}^{n} \begin{pmatrix} \tilde{\mathbf{U}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{Y}}_i, \mathbf{X}_i) \\ \tilde{\mathbf{U}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{C}}_i, \mathbf{X}_i) \end{pmatrix} = \mathbf{0}. \tag{3.12}$$

3.4.3 Estimation and asymptotic distribution

When a validation subsample is available, the response parameter vector $\boldsymbol{\theta}$ and the misclassification parameter vector $\boldsymbol{\eta}$ can be estimated through a two-stage estimation procedure.

Stage 1. Update the estimate of $\boldsymbol{\eta}$ via the Fisher scoring algorithm

$$\boldsymbol{\eta}^{(t+1)} = \boldsymbol{\eta}^{(t)} + \begin{pmatrix} \left\{-\sum_{i=1}^{n} \mathbf{J}_{1i}(\boldsymbol{\eta}^{(t)})\right\}^{-1} \cdot \sum_{i=1}^{n} \mathbf{Q}_{1i}(\boldsymbol{\eta}^{(t)}) \\ \left\{-\sum_{i=1}^{n} \mathbf{J}_{2i}(\boldsymbol{\eta}^{(t)})\right\}^{-1} \cdot \sum_{i=1}^{n} \mathbf{Q}_{2i}(\boldsymbol{\eta}^{(t)}) \end{pmatrix}, \quad t = 0, 1, \ldots$$

where $\boldsymbol{\eta}^{(t)}$ is an estimate of $\boldsymbol{\eta}$ at the tth iteration, $\mathbf{J}_{1i}(\boldsymbol{\eta}) = -\mathbf{D}^{\delta}_{\eta 1i}\left[\mathbf{V}^{\delta}_{\eta 1i}\right]^{-1}\left[\mathbf{D}^{\delta}_{\eta 1i}\right]^{\mathrm{T}}$ and $\mathbf{J}_{2i}(\boldsymbol{\eta}) = -\mathbf{D}^{\delta}_{\eta 2i}\left[\mathbf{V}^{\delta}_{\eta 2i}\right]^{-1}\left[\mathbf{D}^{\delta}_{\eta 2i}\right]^{\mathrm{T}}$. Let $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\gamma}}^{\mathrm{T}}, \hat{\boldsymbol{\nu}}^{\mathrm{T}})^{\mathrm{T}}$ denote the estimates at convergence.

Stage 2. Replace $\boldsymbol{\eta}$ with its estimate $\hat{\boldsymbol{\eta}}$ and solve (3.12) for $\boldsymbol{\theta}$ via the Fisher scoring algorithm. Given an initial value $\boldsymbol{\theta}^{(0)}$, we iteratively update $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \begin{pmatrix} \left\{-\sum_{i=1}^{n} \mathbf{M}_{1i}(\boldsymbol{\theta}^{(t)})\right\}^{-1} \cdot \sum_{i=1}^{n} \tilde{\mathbf{U}}_{1i}(\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\eta}}) \\ \left\{-\sum_{i=1}^{n} \mathbf{M}_{2i}(\boldsymbol{\theta}^{(t)})\right\}^{-1} \cdot \sum_{i=1}^{n} \tilde{\mathbf{U}}_{2i}(\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\eta}}) \end{pmatrix}, \quad t = 0, 1, \ldots$$

where $\mathbf{M}_{1i}(\boldsymbol{\theta}) = -\mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\mathbf{D}_{1i}^{\mathrm{T}}$, and $\mathbf{M}_{2i}(\boldsymbol{\theta}) = -\mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\mathbf{D}_{2i}^{\mathrm{T}}$. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}})^{\mathrm{T}}$ be the estimate of $\boldsymbol{\theta}$ at convergence.

We conclude this section with the asymptotic distribution for $\hat{\boldsymbol{\theta}}$ which accounts for the variation induced by estimation of $\boldsymbol{\eta}$. Let $\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta}) = \left(\tilde{\mathbf{U}}_{1i}^{\mathrm{T}}(\boldsymbol{\theta},\boldsymbol{\eta}), \tilde{\mathbf{U}}_{2i}^{\mathrm{T}}(\boldsymbol{\theta},\boldsymbol{\eta})\right)^{\mathrm{T}}$, $\mathbf{Q}_i(\boldsymbol{\eta}) = \left(\mathbf{Q}_{1i}^{\mathrm{T}}(\boldsymbol{\eta}), \mathbf{Q}_{2i}^{\mathrm{T}}(\boldsymbol{\eta})\right)^{\mathrm{T}}$, $\tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta},\boldsymbol{\eta}) = \mathrm{E}\left[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right]$, and $\tilde{\boldsymbol{\Omega}}_i(\boldsymbol{\theta},\boldsymbol{\eta}) = \tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta}) - \mathrm{E}[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}] \cdot \left\{\mathrm{E}\left[\partial\mathbf{Q}_i(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right]\right\}^{-1} \cdot \mathbf{Q}_i(\boldsymbol{\eta})$. In Section 3.9.3 we show that $\tilde{\mathbf{U}}(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}) = n^{-1/2}\sum_{i=1}^{n}\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\hat{\boldsymbol{\eta}})$ and $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ are asymptotically normally distributed with mean $\mathbf{0}$ and asymptotic covariance matrices given by $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Gamma}}^{-1}\tilde{\boldsymbol{\Sigma}}\left[\tilde{\boldsymbol{\Gamma}}^{-1}\right]^{\mathrm{T}}$, respectively, where $\tilde{\boldsymbol{\Sigma}} = \mathrm{E}\left[\tilde{\boldsymbol{\Omega}}_i(\boldsymbol{\theta},\boldsymbol{\eta})\tilde{\boldsymbol{\Omega}}_i^{\mathrm{T}}(\boldsymbol{\theta},\boldsymbol{\eta})\right]$. In Section 3.9.3 we also outline the inferential procedures.

3.5 Joint Estimation and Inference with Replicates

In some circumstances a validation data set is not possible to obtain, but instead, replicates are available by the design of the study. Now we describe an inference procedure to accommodate this practical situation. Here we use notation slightly different from those in the previous sections for ease of exposition. Let S_{ijr} be the rth replicate measure for Y_{ij}, $r = 1, \ldots, d_{ij}$, where d_{ij} is the number of replicates for subject j in cluster i, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Let $\mathbf{S}_{ij} = (S_{ij1}, \ldots, S_{ijd_{ij}})^{\mathrm{T}}$, and $H_{ijr} = \mathrm{I}(S_{ijr} = Y_{ij})$ be the misclassification indicator variable. For $j \neq j'$, we assume independence between H_{ijr} and $H_{ij'r'}$ given $\mathbf{Y}_i$ and $\mathbf{X}_i$. For $r \neq r'$, we assume that H_{ijr} and $H_{ijr'}$ are independently identically distributed given $\mathbf{Y}_i$ and $\mathbf{X}_i$. Again the assumption $\Pr(H_{ijr} = h_{ijr}|\mathbf{Y}_i, \mathbf{X}_i) = \Pr(H_{ijr} = h_{ijr}|Y_{ij}, \mathbf{X}_i)$ is often made. Let $\tau_{1ijr} = \Pr(H_{ijr} = 1|Y_{ij} = 1, \mathbf{X}_i)$ and $\tau_{0ijr} = \Pr(H_{ijr} = 1|Y_{ij} = 0, \mathbf{X}_i)$. Suppose that τ_{1ijr} and τ_{0ijr} are modeled by (3.1) and (3.2), respectively.

Define $\mathcal{Y}_{ijr}^{*} = (S_{ijr} - 1 + \tau_{0ijr})/(\tau_{0ijr} + \tau_{1ijr} - 1)$. Then the average version $\mathcal{Y}_{ij}^{*} = \sum_{r=1}^{d_{ij}} \mathcal{Y}_{ijr}^{*}/d_{ij}$ is unbiased for Y_{ij}, i.e., $\mathrm{E}\left[\mathcal{Y}_{ij}^{*}|\mathbf{Y}_i, \mathbf{X}_i\right] = Y_{ij}$. Let $\boldsymbol{\mathcal{Y}}_i^{*} = (\mathcal{Y}_{i1}^{*}, \ldots, \mathcal{Y}_{im_i}^{*})^{\mathrm{T}}$, and $\boldsymbol{\mathcal{C}}_i^{*} = (\mathcal{Y}_{ij}^{*}\mathcal{Y}_{ij'}^{*}, j < j')^{\mathrm{T}}$. Define $\boldsymbol{\mathcal{U}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\boldsymbol{\varepsilon}_{1i}$, and $\boldsymbol{\mathcal{U}}_{2i}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\boldsymbol{\varepsilon}_{2i}$, where $\boldsymbol{\varepsilon}_{1i} = \boldsymbol{\mathcal{Y}}_i^{*} - \boldsymbol{\mu}_i$ and $\boldsymbol{\varepsilon}_{2i} = \boldsymbol{\mathcal{C}}_i^{*} - \boldsymbol{\xi}_i$ are residual vectors. It is readily seen that $\mathrm{E}\left[\boldsymbol{\mathcal{U}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})|\mathbf{Y}_i, \mathbf{X}_i\right] = \mathbf{U}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})$ and $\mathrm{E}\left[\boldsymbol{\mathcal{U}}_{2i}(\boldsymbol{\theta},\boldsymbol{\gamma})|\mathbf{Y}_i, \mathbf{X}_i\right] = \mathbf{U}_{2i}(\boldsymbol{\theta},\boldsymbol{\gamma})$. Let

$\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \big(\boldsymbol{\mathcal{U}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})^{\mathrm{T}},\boldsymbol{\mathcal{U}}_{2i}(\boldsymbol{\theta},\boldsymbol{\gamma})^{\mathrm{T}}\big)^{\mathrm{T}}$. Therefore, a consistent estimator of $\boldsymbol{\theta}$ can be obtained by solving

$$\sum_{i=1}^{n}\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) \;=\; \mathbf{0}, \tag{3.13}$$

provided $\boldsymbol{\gamma}$ is given.

However, $\boldsymbol{\gamma}$ is unknown here, and it must be estimated. In the case with replicates S_{ijr}, the true response measurements Y_{ij}'s are not available. Thus, one cannot derive a set of estimating equations for misclassification parameters $\boldsymbol{\gamma}$ analogous to (3.11). Hence, the two-stage estimation procedure described in Section 3.4 no longer applies here. With replicates, estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ typically interacts, and a joint estimation procedure is required to simultaneously estimate $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. In the sequel, we construct estimating equations for misclassification parameters which typically involve response parameters. Information about the misclassification process is captured by the heterogeneity in the replicates. We generalize the discussion in White et al. (2001) who considered univariate logistic regression models with a misclassified binary covariate.

Let $A_{ijk} = 1$ if $\sum_{r=1}^{d_{ij}} S_{ijr} = k$ and $A_{ijk} = 0$ otherwise, $k = 1,\ldots,d_{ij}$, $j = 1,\ldots,m_i$, $i = 1,\ldots,n$. Define $\mathbf{A}_{ij} = (A_{ij1},\ldots,A_{ijd_{ij}})^{\mathrm{T}}$, and $\mathbf{A}_i = (\mathbf{A}_{i1}^{\mathrm{T}},\ldots,\mathbf{A}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Let $\pi_{ijk} = \mathrm{E}[A_{ijk}|\mathbf{X}_i]$ be the marginal mean of A_{ijk}, $\boldsymbol{\pi}_{ij} = (\pi_{ij1},\ldots,\pi_{ijd_{ij}})^{\mathrm{T}}$, and $\boldsymbol{\pi}_i = (\pi_{i1}^{\mathrm{T}},\ldots,\pi_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Apparently, π_{ijk} involves both $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

Now we describe estimating functions of $\boldsymbol{\gamma}$. For ease of exposition, we consider cases with $d_{ij} = 2$. The method can be easily extended to cases with $d_{ij} \geq 3$. Noting that

$$\begin{aligned}\Pr(A_{ij1} = 1|\mathbf{Y}_i,\mathbf{X}_i) \;&=\; \{(1-\tau_{1ij1})\tau_{1ij2} + (1-\tau_{1ij2})\tau_{1ij1}\}\,Y_{ij}\\ &\qquad + \{(1-\tau_{0ij1})\tau_{0ij2} + (1-\tau_{0ij2})\tau_{0ij1}\}\,(1-Y_{ij}),\end{aligned}$$

and

$$\Pr(A_{ij2} = 1|\mathbf{Y}_i,\mathbf{X}_i) \;=\; \tau_{1ij1}\tau_{1ij2}Y_{ij} + (1-\tau_{0ij1})(1-\tau_{0ij2})(1-Y_{ij}),$$

we write marginal means π_{ijr} $(r = 1, 2)$ as follows:

$$\begin{aligned}\pi_{ij1} \;&=\; \left\{(1-\tau_{1ij1})\tau_{1ij2} + (1-\tau_{1ij2})\tau_{1ij1}\right\}\mu_{ij} \\ &\qquad\quad + \left\{(1-\tau_{0ij1})\tau_{0ij2} + (1-\tau_{0ij2})\tau_{0ij1}\right\}(1-\mu_{ij}),\end{aligned}$$

and

$$\pi_{ij2} \;=\; \tau_{1ij1}\tau_{1ij2}\mu_{ij} + (1-\tau_{0ij1})(1-\tau_{0ij2})(1-\mu_{ij}).$$

Define $\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \boldsymbol{\mathcal{D}}_{\gamma i}\boldsymbol{\mathcal{V}}_{\gamma i}^{-1}(\mathbf{A}_i - \boldsymbol{\pi}_i)$, where $\boldsymbol{\mathcal{D}}_{\gamma i} = \partial\boldsymbol{\pi}_i^{\mathrm{T}}/\partial\boldsymbol{\gamma}$, and $\boldsymbol{\mathcal{V}}_{\gamma i}$ is the covariance matrix for $\mathbf{A}_i$ conditional on $\mathbf{X}_i$. For a given $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ can be estimated from estimating equations

$$\sum_{i=1}^{n}\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \;=\; \mathbf{0}. \tag{3.14}$$

In contrast to the two-stage estimation algorithm in Section 3.4, we must simultaneously employ (3.13) and (3.14) to iteratively update the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. To be specific, let $\boldsymbol{\mathcal{J}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -\boldsymbol{\mathcal{D}}_{\gamma i}\boldsymbol{\mathcal{V}}_{\gamma i}^{-1}\boldsymbol{\mathcal{D}}_{\gamma i}^{\mathrm{T}}$, $\boldsymbol{\Delta}_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) = -\boldsymbol{\mathcal{D}}_{\gamma i}\boldsymbol{\mathcal{V}}_{\gamma i}^{-1}\cdot\left(\partial\boldsymbol{\pi}_i/\partial\boldsymbol{\theta}^{\mathrm{T}}\right)$, and

$$\boldsymbol{\Lambda}_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\cdot\left(\partial\boldsymbol{\mathcal{Y}}_i^*/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right) \\ \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\cdot\left(\partial\boldsymbol{\mathcal{C}}_i^*/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right) \end{pmatrix}.$$

Given initial estimates $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$, we update the estimates via

$$\begin{aligned}\begin{pmatrix}\boldsymbol{\theta}^{(t+1)} \\ \boldsymbol{\gamma}^{(t+1)}\end{pmatrix} \;=\; \begin{pmatrix}\boldsymbol{\theta}^{(t)} \\ \boldsymbol{\gamma}^{(t)}\end{pmatrix} - \begin{pmatrix} \sum_{i=1}^{n}\mathbf{M}_i(\boldsymbol{\theta}^{(t)}) & \sum_{i=1}^{n}\boldsymbol{\Lambda}_i(\boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \\ \sum_{i=1}^{n}\boldsymbol{\Delta}_i(\boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}) & \sum_{i=1}^{n}\boldsymbol{\mathcal{J}}_{1i}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \end{pmatrix}^{-1} \\ \cdot\begin{pmatrix} \sum_{i=1}^{n}\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \\ \sum_{i=1}^{n}\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \end{pmatrix}, \quad t = 0, 1, \ldots\end{aligned}$$

until convergence. Let $\hat{\boldsymbol{\theta}}_{RS}$ and $\hat{\boldsymbol{\gamma}}_{RS}$ denote the final solutions to (3.13) and (3.14).

Now we conclude this section with the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{RS}$. Define

$$\boldsymbol{\Omega}_i^*(\boldsymbol{\theta},\boldsymbol{\gamma}) = \boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) - \mathrm{E}\left[\partial\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right] \cdot \left\{\mathrm{E}\left[\partial\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right]\right\}^{-1} \cdot \boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma}),$$

and

$$\begin{aligned}\boldsymbol{\Gamma}^*(\boldsymbol{\theta},\boldsymbol{\gamma}) &= \mathrm{E}\left[\partial\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right] - \mathrm{E}\left[\partial\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right] \\ &\quad \cdot \left\{\mathrm{E}\left[\partial\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right]\right\}^{-1} \cdot \mathrm{E}\left[\partial\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right].\end{aligned}$$

In Section 3.9.4 we establish that $n^{1/2}\left(\hat{\boldsymbol{\theta}}_{RS} - \boldsymbol{\theta}\right)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Gamma}^{*-1}\boldsymbol{\Sigma}^*\left[\boldsymbol{\Gamma}^{*-1}\right]^{\mathrm{T}}$, where $\boldsymbol{\Sigma}^* = \mathrm{E}[\boldsymbol{\Omega}_i^*(\boldsymbol{\theta},\boldsymbol{\gamma})$ $\boldsymbol{\Omega}_i^*(\boldsymbol{\theta},\boldsymbol{\gamma})^{\mathrm{T}}]$.

3.6 Numerical Assessment of the Proposed Methods

3.6.1 Design of simulation studies

We conduct simulation studies to assess the performance of the proposed methods in contrast to the naive method which ignores misclassification. We focus on the case with $m_i = m = 3$ for $i = 1,\ldots,n$, where the sample sizes are $n = 200$ and 400 for cases of known and unknown misclassification parameters, respectively. Two thousand simulations are run for each parameter configuration. The mean response model is given by

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3},$$

where X_{ij1} is 1 if the ith subject is randomized to the treatment group and 0 otherwise, X_{ij2} is 1 if $j = 2$ and 0 otherwise, and X_{ij3} is 1 if $j = 3$ and 0 otherwise. An exchangeable association structure is considered, which is given by

$$\log\psi_{ijj'} = \alpha. \qquad (3.15)$$

The regression parameters are specified by $\exp(\beta_0) = 2$, $\exp(\beta_1) = 0.5$, $\exp(\beta_2) = 2/3$ and $\exp(\beta_3) = 1/3$, and the association parameter is specified by $\alpha = \log(3.0)$. The binary response vector is generated from the probability function

$$\begin{aligned} &\Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}) \\ &\quad = \prod_{j=1}^{3} \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1-y_{ij}} \left\{ 1 + \sum_{1 \le j < j' \le 3} \rho_{ijj'} \frac{(y_{ij} - \mu_{ij})(y_{ij'} - \mu_{ij'})}{\sqrt{\mu_{ij}(1 - \mu_{ij})}\,\sqrt{\mu_{ij'}(1 - \mu_{ij'})}} \right\}. \end{aligned}$$

We consider both independent and correlated misclassification processes. For the independent case, the misclassification process is considered to be homogeneous and is characterized by two misclassification probabilities. Misclassification indicators H_{ij}'s are generated with probabilities given by a logistic model

$$\text{logit}(\tau_{ij}) = \begin{cases} \gamma_0, & \text{if } Y_{ij} = 0, \\ \gamma_1, & \text{if } Y_{ij} = 1. \end{cases} \tag{3.16}$$

Surrogate responses S_{ij} are then recorded as Y_{ij} if $H_{ij} = 1$ and $1 - Y_{ij}$ if $H_{ij} = 0$. Three settings for $\boldsymbol{\gamma}$ are considered: (i) $\gamma_0 = \text{logit}(0.95)$ and $\gamma_1 = \text{logit}(0.95)$; (ii) $\gamma_0 = \text{logit}(0.9)$ and $\gamma_1 = \text{logit}(0.9)$; and (iii) $\gamma_0 = \text{logit}(0.8)$ and $\gamma_1 = \text{logit}(0.8)$, which represent different levels of misclassification rates.

The performance of the proposed methods are assessed under three scenarios. For the first scenario where $\boldsymbol{\gamma}$ is known, each simulated sample contains $n = 200$ subjects. For the second scenario where $\boldsymbol{\gamma}$ is not known but an internal validation subsample is available, we take $n = 400$, and randomly select 30% of the subjects to be in the validation sample. For low misclassification rates as in setting (i), large sample size is usually necessary in order to obtain a valid estimate of $\boldsymbol{\gamma}$. For the third scenario where $\boldsymbol{\gamma}$ is not known but replicates are available, the sample size is set to be $n = 200$ and two replicate surrogates are used for each Y_{ij}.

For cases where misclassifications within the same subject are correlated, the mean model is also given by (3.16), while the association is modeled in the same manner as in the response process and is given by

$$\begin{aligned}\log\{\lambda_{ijj'}(y_{ij}, y_{ij'})\} &= \nu_1 \mathrm{I}(y_{ij} = y_{ij'}) + \nu_2 \mathrm{I}(y_{ij} \neq y_{ij'}), \\ &\qquad 1 \leq j < j' \leq 3,\ i = 1, \ldots, n.\end{aligned}$$

We set $\nu_1 = \log(2.0)$ and $\nu_2 = \log(1.5)$. The misclassification vector $\mathbf{H}_i$ is generated with the given probabilities. Two scenarios are considered. For the first scenario with known $\boldsymbol{\eta}$, sample size is $n = 200$. For the second scenario with unknown $\boldsymbol{\eta}$, sample size is increased to $n = 400$. Again 30% of subjects are randomly selected into the validation subsample.

3.6.2 Simulation results

Table 3.1 shows the simulation results of the first and second scenarios for cases where the misclassification process is independent. The column under each approach represents the percent relative bias (%RB), empirical variance (EV), average of model-based variance (AMV), and coverage rate of the 95% confidence intervals (CP). We first look at the results under known $\boldsymbol{\gamma}$. One can see that the naive analysis leads to downward biased estimates of response parameters even under a small proportion of misclassifications. Under setting (i) where misclassification proportion is 5%, for example, both the mean parameters and the association parameter are attenuated by a non-ignorable amount. As misclassification proportion increases, the attenuation increases. When misclassification proportion is increased to 20% in setting (iii), coverage probabilities for the naive estimates of mean parameters and association parameter are far below the nominal value 95%. In contrast, the proposed method performs reasonably well for all parameter configurations. The relative biases in mean parameters for settings (i) and (ii) with small and moderate misclassification rates are within an ignorable amount. The relative biases increase a little when the misclassification rate is relatively high. The coverage probability for α is slightly over the nominal value 95%. The variance estimates of the estimators are larger than those of the naive estimators and increase as the misclassification rate increases. For the case of estimated $\boldsymbol{\gamma}$, similar patterns are observed.

Table 3.1: Simulation results for the independent misclassification process (2000 simulations)

	Naive method				Proposed method							
	$(n = 200)$				known $\boldsymbol{\eta}$ $(n = 200)$				unknown $\boldsymbol{\eta}$ $(n = 400)$			
	%RB[†]	EV	AMV	CP	%RB	EV	AMV	CP	%RB	EV	AMV	CP
	(i) $\gamma_0 = \text{logit}(0.95)$, $\gamma_1 = \text{logit}(0.95)$											
β_0	-10.823	0.030	0.032	0.935	0.317	0.038	0.041	0.962	0.635	0.021	0.021	0.955
β_1	-9.593	0.038	0.040	0.938	1.943	0.049	0.051	0.952	0.367	0.024	0.025	0.955
β_2	-11.931	0.032	0.033	0.945	-1.180	0.040	0.042	0.957	-0.981	0.018	0.019	0.952
β_3	-10.621	0.035	0.036	0.903	0.776	0.046	0.047	0.956	0.672	0.021	0.022	0.958
α	-23.260	0.045	0.044	0.750	0.049	0.082	0.081	0.954	0.427	0.035	0.038	0.957
	(ii) $\gamma_0 = \text{logit}(0.9)$, $\gamma_1 = \text{logit}(0.9)$											
β_0	-21.130	0.029	0.031	0.866	1.005	0.050	0.052	0.957	0.200	0.027	0.027	0.956
β_1	-20.391	0.036	0.037	0.891	2.432	0.062	0.062	0.948	0.467	0.029	0.029	0.950
β_2	-21.869	0.035	0.035	0.923	-0.421	0.057	0.057	0.950	-1.630	0.023	0.025	0.965
β_3	-21.181	0.036	0.037	0.760	1.388	0.062	0.065	0.958	0.411	0.029	0.029	0.953
α	-41.768	0.041	0.040	0.366	0.875	0.140	0.138	0.954	1.040	0.061	0.061	0.955
	(iii) $\gamma_0 = \text{logit}(0.8)$, $\gamma_1 = \text{logit}(0.8)$											
β_0	-41.513	0.029	0.029	0.591	2.292	0.095	0.095	0.960	1.375	0.050	0.051	0.959
β_1	-41.297	0.032	0.032	0.636	3.508	0.103	0.104	0.953	1.240	0.045	0.045	0.958
β_2	-41.083	0.037	0.037	0.855	2.032	0.113	0.114	0.956	0.076	0.042	0.045	0.963
β_3	-41.572	0.038	0.039	0.365	3.091	0.131	0.133	0.958	1.870	0.053	0.056	0.961
α	-69.294	0.033	0.034	0.024	4.091	0.469	0.539	0.966	2.510	0.179	0.196	0.957

[†] %RB=relative bias in percent: $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})/\boldsymbol{\theta}$, EV=empirical variance, AMV=average of model-based variances, CP=coverage rate of the 95% CI

The simulation results for the case with replicates are shown in Table 3.2. The relative biases and coverage rates for the naive estimators are similar to those in Table 3.1. The proposed method performs well. The relative biases of the estimates are small for the first two settings where misclassification rates are low and moderate, and coverage rates are close to the nominal value 95%. For setting (iii) with higher misclassification rate, however, the relative biases in both the mean parameters and the association parameter are larger than those in settings (i) and (ii). The coverage rate is slightly over the nominal value 95% for the association parameter.

The results for correlated misclassifications are reported in Table 3.3. It can be seen that naively fitting GEE2 ignoring misclassifications leads to seriously biased estimates and low coverage rates. The corrected GEE2 approach gives reasonably good estimates of mean parameters. For the case where $\boldsymbol{\eta}$ is estimated from a validation subsample, estimates of $\boldsymbol{\theta}$ are greatly improved. Finally, we note that estimation of association parameters $\boldsymbol{\nu}$ involves larger variation when the size of a validation subsample becomes smaller, hence resulting in possibly more unstable estimates of $\boldsymbol{\theta}$ and large bias in the estimates of $\boldsymbol{\theta}$. In this situation, one possible resolution is to leave $\boldsymbol{\nu}$ not estimated by assuming misclassifications are independent. The method still works well for cases with higher misclassification rates.

3.7 Application

3.7.1 Analysis of the CCHS data

We apply the proposed method to analyze data from the Canadian Community Health Survey (CCHS) cycle 3.1 conducted in 2005. CCHS is a large scale on-going survey targeting individuals aged 12 and older in the Canadian population. Although the design of the survey is cross-sectional, the data can be viewed as clustered, since health status for subjects who live in the same neighborhood may be correlated.

The objective of our study is to explore the relationship between obesity status and some risk factors. We consider a subset of the data that contains 2699 respondents aged 18 and older in the Toronto health region who do not have missing response

Table 3.2: Simulation results for the independent misclassification process with replicates (2000 simulations)

	Naive method								Proposed method			
	1st replicates ($n = 200$)				2nd replicates ($n = 200$)				($n = 200$)			
	%RB[†]	EV	AMV	CP	%RB	EV	AMV	CP	%RB	EV	AMV	CP
	(i) $\gamma_0 = \text{logit}(0.95)$, $\gamma_1 = \text{logit}(0.95)$											
β_0	-9.912	0.032	0.032	0.926	-9.904	0.031	0.032	0.928	1.599	0.047	0.049	0.952
β_1	-9.979	0.040	0.040	0.929	-10.338	0.040	0.040	0.933	1.723	0.048	0.048	0.945
β_2	-10.466	0.034	0.033	0.941	-10.544	0.033	0.033	0.944	0.829	0.037	0.037	0.951
β_3	-10.646	0.036	0.036	0.907	-10.231	0.036	0.036	0.908	1.068	0.041	0.041	0.959
α	-22.419	0.043	0.044	0.765	-21.874	0.046	0.044	0.765	1.314	0.063	0.064	0.953
	(ii) $\gamma_0 = \text{logit}(0.9)$, $\gamma_1 = \text{logit}(0.9)$											
β_0	-20.675	0.032	0.031	0.860	-21.107	0.030	0.031	0.866	2.605	0.077	0.078	0.949
β_1	-20.955	0.038	0.037	0.871	-20.862	0.038	0.037	0.880	2.040	0.054	0.053	0.949
β_2	-20.970	0.037	0.035	0.912	-21.149	0.034	0.035	0.925	0.670	0.045	0.045	0.952
β_3	-21.738	0.040	0.037	0.743	-21.478	0.036	0.037	0.762	0.908	0.050	0.049	0.950
α	-41.459	0.039	0.040	0.366	-41.322	0.043	0.040	0.375	1.311	0.087	0.088	0.948
	(iii) $\gamma_0 = \text{logit}(0.8)$, $\gamma_1 = \text{logit}(0.8)$											
β_0	-40.886	0.029	0.029	0.608	-41.764	0.028	0.029	0.600	5.132	0.222	0.235	0.959
β_1	-41.454	0.033	0.032	0.633	-42.598	0.033	0.032	0.618	3.687	0.077	0.076	0.947
β_2	-41.726	0.037	0.037	0.854	-41.077	0.037	0.037	0.859	2.803	0.075	0.076	0.955
β_3	-42.349	0.041	0.039	0.351	-41.644	0.038	0.039	0.365	3.487	0.088	0.087	0.955
α	-69.273	0.037	0.034	0.026	-68.583	0.035	0.034	0.028	6.823	0.242	0.248	0.966

[†] %RB=relative bias in percent: $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})/\boldsymbol{\theta}$, EV=empirical variance, AMV=average of model-based variances, CP=coverage rate of the 95% CI

Table 3.3: Simulation results for the correlated misclassification process (2000 simulations)

	Naive method				Proposed method							
	$(n = 200)$				known $\boldsymbol{\eta}$ $(n = 200)$				unknown $\boldsymbol{\eta}$ $(n = 400)$			
	%RB[†]	EV	AMV	CP	%RB	EV	AMV	CP	%RB	EV	AMV	CP
	(i) $\gamma_0 = \text{logit}(0.95)$, $\gamma_1 = \text{logit}(0.95)$											
β_0	-10.449	0.030	0.032	0.932	0.797	0.035	0.037	0.954	0.144	0.019	0.019	0.952
β_1	-10.928	0.038	0.040	0.941	0.392	0.045	0.047	0.957	0.081	0.023	0.024	0.952
β_2	-9.989	0.032	0.033	0.944	1.088	0.036	0.037	0.951	0.168	0.018	0.018	0.945
β_3	-10.243	0.036	0.036	0.905	1.011	0.042	0.041	0.948	0.291	0.020	0.021	0.957
α	-22.189	0.043	0.044	0.766	-0.964	0.099	0.098	0.952	0.560	0.033	0.034	0.954
	(ii) $\gamma_0 = \text{logit}(0.9)$, $\gamma_1 = \text{logit}(0.9)$											
β_0	-21.514	0.015	0.015	0.773	0.206	0.021	0.021	0.954	0.225	0.022	0.023	0.954
β_1	-21.746	0.019	0.019	0.801	0.399	0.027	0.027	0.944	0.400	0.027	0.027	0.942
β_2	-21.908	0.017	0.017	0.894	-0.846	0.021	0.022	0.952	-0.821	0.021	0.022	0.953
β_3	-21.634	0.018	0.018	0.575	0.384	0.024	0.024	0.950	0.370	0.024	0.025	0.954
α	-36.455	0.021	0.020	0.204	0.770	0.089	0.085	0.949	1.148	0.054	0.053	0.942
	(iii) $\gamma_0 = \text{logit}(0.8)$, $\gamma_1 = \text{logit}(0.8)$											
β_0	-42.235	0.015	0.015	0.334	0.041	0.031	0.032	0.954	-0.064	0.035	0.036	0.959
β_1	-42.490	0.017	0.017	0.384	0.608	0.037	0.038	0.947	0.626	0.038	0.038	0.950
β_2	-42.605	0.018	0.018	0.742	-1.261	0.035	0.034	0.950	-1.245	0.035	0.035	0.954
β_3	-42.322	0.019	0.019	0.076	0.513	0.040	0.040	0.950	0.593	0.042	0.042	0.950
α	-56.991	0.017	0.018	0.003	0.721	0.246	0.251	0.964	0.753	0.169	0.171	0.959

[†] %RB=relative bias in percent: $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})/\boldsymbol{\theta}$, EV=empirical variance, AMV=average of model-based variances, CP=coverage rate of the 95% CI

or covariates. A total of 435 clusters with size varying from 2 to 15 are formed by postal codes. Covariates included in the analysis are age, sex, and physical activity index. Four age categories are defined: 18-34 (reference category), 35-49, 50-64, and 65+. There are three levels of physical activity index: active, moderate (reference category), and inactive. Let Y_{ij} denote the binary obesity status for subject j in cluster i. We assume Y_{ij} follows the logistic regression model

$$\text{logit}\ \mu_{ij} = \beta_0 + \sum_{k=1}^{3} \beta_k x_{ijk} + \beta_4 x_{ij4} + \sum_{k=5}^{6} \beta_k x_{ijk},$$

where x_{ijk} is 1 if subject j in cluster i belongs to the $(k+1)$th age category and 0 otherwise, $k = 1, \ldots, 3$, x_{ij4} is 1 if the subject is male and 0 otherwise, x_{ij5} is 1 if physical activity index is "active" and 0 otherwise, and x_{ij6} is 1 if physical activity index is "inactive" and 0 otherwise. The association between Y_{ij} and $Y_{ij'}$, measured by odds ratio $\psi_{ijj'}$, is modeled by equation (3.15). Because the surrogate responses are obtained from self-report interviews, obesity misclassifications are typically independent for different individuals and clusters. We assume that the misclassification process is modeled by (3.16). A subsample consisting of 150 subjects was selected, in which BMI was measured on each subject. We treat the derived obesity status for each subject in the subsample as the true binary response.

The analysis results for estimation of $\boldsymbol{\beta}$ and α from the proposed method are shown in Table 3.4 with comparison to results from naive analysis ignoring misclassifications. The estimates of misclassification parameters are given by $\hat{\gamma}_0 = 4.103$ and $\hat{\gamma}_1 = 0.693$. Therefore, about $1 - \text{expit}(\hat{\gamma}_0) = 1.63\%$ of the non-obese subjects self-reported as obese, and about $1 - \text{expit}(\hat{\gamma}_1) = 33.3\%$ of the obese subjects self-reported as non-obese. Note that the p-values from testing for no misclassifications are computed based on a one-sided alternative. The effect estimates of age categories 35-49, 50-64 and 65+ are 1.219, 1.558 and 1.483, respectively, indicating that subjects in these groups have a much higher probability of developing obesity compared to the baseline age group of 18-34. The estimate of the gender effect indicates that the probability of obesity in males is not significantly different from that in females. There is no evidence that a subject with a higher physical activity index has a smaller chance of developing obesity than that with a moderate index. For the inactive group, however, the result is significant at the 5% level. The odds of obesity in the inactive group is about twice compared to the moderately active group. The estimate of association parameter α is given by 0.104, which corresponds to an odds ratio of 1.11 between obesities of two subjects in the same cluster. However, there is no evidence for association between obesity among subjects in the neighborhood.

Table 3.4: Analysis results for obesity among adults in Toronto health region

		Naive method			Proposed method		
		Est.	SE	p-value	Est.	SE	p-value
Response models							
Intercept	(β_0)	-3.129	0.331	< 0.001	-3.189	0.655	< 0.001
Age	35-49 (β_1)	0.901	0.314	0.004	1.219	0.555	0.028
	50-64 (β_2)	1.204	0.316	< 0.001	1.558	0.568	0.006
	65+ (β_3)	1.145	0.337	0.001	1.483	0.581	0.011
Sex	male (β_4)	0.003	0.124	0.981	-0.001	0.152	0.997
PAI	active (β_5)	-0.401	0.191	0.036	-0.527	0.264	0.046
	inactive (β_6)	0.340	0.153	0.026	0.421	0.188	0.025
Association: (α)		0.073	0.114	0.522	0.104	0.169	0.539
Misclassification models							
expit(γ_0)					0.984	0.011	$0.076^\dagger$
expit(γ_1)					0.667	0.091	$< 0.001^\dagger$

† One-sided tests for no misclassification

3.7.2 Analysis of data from the Framingham Heart Study

Now we apply the proposed method to analyze a data set from the Framingham Heart Study, which is a longitudinal study consists of a series of examinations on the participants. The data we used here, as described in Carroll et al. (2006, p. 112), contains two measurements of systolic blood pressures (SBP) by different examiners at each of exams #2 and #3 for $n = 1615$ male subjects aged 31-65. One of the clinical interests is to understand what risk factors may be associated with high blood pressure (HBP). Potential risk factors include the smoking status recorded at exam #1, and age recorded at exam #2. Response variable Y_{ij} is the binary HBP indicator obtained from dichotomizing the true SBP at cut point 140 mmHg for subject i at the jth time point. Here true SBP is defined as the long term average of SBP measures. The mean model for HBP is given by

$$\text{logit}\ \mu_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}, \tag{3.17}$$

where x_{ij1} is the age of subject i at exam #2, x_{ij2} is 1 if subject i is a smoker and 0 otherwise, and x_{ij3} is 1 if $j = 2$ (i.e., exam #3) and 0 otherwise. Assume an exchangeable structure for association between HBPs at the two exams and the model is given by (3.15).

Since a single SBP measure is considered as an error-contaminated version for the true SBP, its dichotomized version may contain misclassifications (Carroll et al., 2006). Therefore, naively applying a standard method such as GEE2 to the data could lead to biased estimates of the effects of risk factors. Here we consider the marginal misclassification model given by (3.16), and the two replicates are conditionally independent given the true binary HBP.

The analysis results are shown in Table 3.5. The estimates of misclassification parameters are given by $\hat{\gamma}_0 = 2.850$ and $\hat{\gamma}_1 = 2.109$. Therefore, the rate of misclassifying a non-HBP subject into the HBP group is about $1 - \text{expit}(2.850) = 0.055$, and the rate of misclassifying an HBP subject into the non-HBP group is about $1 - \text{expit}(2.109) = 0.108$. The estimate of the age effect is highly significant, indicating that the probability of developing HBP increases with age for male adults. The smoking effect is not statistically significant at the 5% level. An estimate of the odds ratio between HBP at the two exams is given by $\exp(4.253) = 70.32$ with a p-value close to 0, indicating very strong associations. Along with the analysis from the proposed method, we also report in Table 3.5 the results from the naive analysis using the first or the second replicates at each time point. Although the trends in the estimates are similar to those from the proposed method, the estimates generally "shrink", especially in the association parameter.

Table 3.5: Analysis results for a data arising from the Framingham Heart Study

	Naive method						Proposed method		
	1st replicates			2nd replicates					
Parameter	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Response models									
Intercept (β_0)	-3.097	0.292	< 0.001	-3.566	0.317	< 0.001	-3.930	0.382	< 0.001
AGE (β_1)	0.051	0.006	< 0.001	0.056	0.006	< 0.001	0.066	0.007	< 0.001
SMOKE (β_2)	-0.114	0.114	0.313	-0.191	0.121	0.115	-0.188	0.137	0.168
EXAM (β_3)	-0.135	0.056	0.016	-0.092	0.058	0.113	-0.141	0.058	0.015
Association (α)	2.313	0.132	< 0.001	2.534	0.142	< 0.001	4.253	0.264	< 0.001
Misclassification models									
expit(γ_0)							0.945	0.021	0.005†
expit(γ_1)							0.892	0.142	0.013†

† One-sided tests for no misclassification

3.8 Discussion

In this chapter we propose semi-parametric methods based on estimating equations to handle misclassification in correlated binary responses. Since misclassification parameters are often unknown, additional information such as validation data or replicated measures is required in order to obtain estimates of these parameters. Proportion of validation subsample and cluster size play important roles in estimation of parameters governing the misclassification process, especially for the dependence structure. Our simulation studies demonstrate that the proposed methods perform well under various settings. In some situations, the size of the validation data is too small to effectively estimate the correlation between misclassifications. Assumptions about independent misclassifications may have to be made in order to obtain generic estimates of mean parameters of the misclassification process. For data with a rare outcome, estimators for misclassification parameters are often associated with large variances, as there may be sparse or zero counts in the classification table obtained from validation data. In circumstances where no validation data nor replicates are available to estimate misclassification parameters, one may conduct sensitivity analysis to evaluate the impact of misclassification on inference about response parameters. When misclassification rate is very small, i.e., below 1%, naive estimators for response parameters may be acceptable. Our methods are most useful for studies with moderate and serious misclassifications.

Our approach to modeling longitudinal data is a marginal regression one, which characterizes the dependence of the response on covariates but not on the history of outcomes. In contrast, a conditional regression model, or transition model (e.g., Azzalini, 1994; Heagerty, 2002; Diggle et al., 2002), may be employed to capture the serial dependence in some cases. If the category in the past observation is misclassified, inference results could be incorrect if misclassification effects are not properly accounted for. It would be interesting to modify the proposed methods to deal with the misclassification problem in transition models.

Our proposed methods can also be extended to further incorporate missing data. It is common in longitudinal studies that both measurement error and missing data may exist. Incomplete longitudinal binary data are often analyzed using marginal models such as the inverse probability weighted GEE (e.g., Yi and Cook, 2002). The weight matrix, however, may be dependent on the history of outcomes and can be problematic if misclassification effects are not adjusted for. Therefore, it is also interesting to develop statistical tools to simultaneously account for missing data and measurement error effects for correlated data analysis.

3.9 Technical Details

3.9.1 Marginal expectation of S_{ij} and $S_{ij}S_{ij'}$

The conditional expectation of S_{ij} given $(Y_{ij}, \mathbf{X}_i)$, is given by

$$\mathrm{E}[S_{ij}|Y_{ij}, \mathbf{X}_i] \;\; = \;\; 1 - \tau_{0ij} + (\tau_{0ij} + \tau_{1ij} - 1)Y_{ij}. \tag{3.18}$$

Let $\mu_{ij}^S = \mathrm{E}[S_{ij}|\mathbf{X}_i]$. Then $\mu_{ij}^S = 1 - \tau_{0ij} + (\tau_{0ij} + \tau_{1ij} - 1)\mu_{ij}$.

For $j \neq j'$, the conditional expectation of $S_{ij}S_{ij'}$, given Y_{ij}, $Y_{ij'}$ and $\mathbf{X}_i$, is given by

$$\begin{aligned}
&\mathrm{E}\left[S_{ij}S_{ij'}|Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, \mathbf{X}_i\right] \\
&\;= \Pr(S_{ij} = 1, S_{ij'} = 1|Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}, \mathbf{X}_i) \\
&\;= \begin{cases} \Pr(H_{ij} = 1, H_{ij'} = 1|Y_{ij} = 1, Y_{ij'} = 1, \mathbf{X}_i), & \text{if } y_{ij} = 1,\; y_{ij'} = 1, \\ \Pr(H_{ij} = 1, H_{ij'} = 0|Y_{ij} = 1, Y_{ij'} = 0, \mathbf{X}_i), & \text{if } y_{ij} = 1,\; y_{ij'} = 0, \\ \Pr(H_{ij} = 0, H_{ij'} = 1|Y_{ij} = 0, Y_{ij'} = 1, \mathbf{X}_i), & \text{if } y_{ij} = 0,\; y_{ij'} = 1, \\ \Pr(H_{ij} = 0, H_{ij'} = 0|Y_{ij} = 0, Y_{ij'} = 0, \mathbf{X}_i), & \text{if } y_{ij} = 0,\; y_{ij'} = 0, \end{cases} \\
&\;= y_{ij}y_{ij'}\Pr(H_{ij} = 1, H_{ij'} = 1|Y_{ij} = 1, Y_{ij'} = 1, \mathbf{X}_i) \\
&\quad + y_{ij}(1 - y_{ij'})\Pr(H_{ij} = 1, H_{ij'} = 0|Y_{ij} = 1, Y_{ij'} = 0, \mathbf{X}_i) \\
&\quad + (1 - y_{ij})y_{ij'}\Pr(H_{ij} = 0, H_{ij'} = 1|Y_{ij} = 0, Y_{ij'} = 1, \mathbf{X}_i) \\
&\quad + (1 - y_{ij})(1 - y_{ij'})\Pr(H_{ij} = 0, H_{ij'} = 0|Y_{ij} = 0, Y_{ij'} = 0, \mathbf{X}_i) \\
&\;= \zeta_{ijj'}(1,1)y_{ij}y_{ij'} \;+\; \{\tau_{1ij} - \zeta_{ijj'}(1,0)\}\, y_{ij}\,(1 - y_{ij'}) \;+\; \{\tau_{1ij'} - \zeta_{ijj'}(0,1)\} \\
&\quad \times\; (1 - y_{ij})\, y_{ij'} \;+\; \{1 - \tau_{0ij} - \tau_{0ij'} + \zeta_{ijj'}(0,0)\}\,(1 - y_{ij})\,(1 - y_{ij'}),
\end{aligned} \tag{3.19}$$

where again we assume $\Pr(S_{ij} = 1|\mathbf{Y}_i, \mathbf{X}_i) = \Pr(S_{ij} = 1|Y_{ij}, \mathbf{X}_i)$ and $\Pr(S_{ij} = 1|Y_{ij}, Y_{ij'}, \mathbf{X}_i) = \Pr(S_{ij} = 1|Y_{ij}, \mathbf{X}_i)$ for $j \neq j'$. Let $\mu^S_{ijj'} = \mathrm{E}\left[\mathrm{E}\left\{S_{ij}S_{ij'}|\mathbf{Y}_i, \mathbf{X}_i\right\}|\mathbf{X}_i\right]$. We then have

$$\begin{aligned}
\mu^S_{ijj'} &= \mathrm{E}[\zeta_{ijj'}(1,1)Y_{ij}Y_{ij'} + \{\tau_{1ij} - \zeta_{ijj'}(1,0)\}Y_{ij}(1-Y_{ij'}) \ + \ \{\tau_{1ij'} - \zeta_{ijj'}(0,1)\} \\
&\quad \times (1-Y_{ij})Y_{ij'} + \ \{1 - \tau_{0ij} - \tau_{0ij'} + \zeta_{ijj'}(0,0)\}(1-Y_{ij})(1-Y_{ij'})|\mathbf{X}_i] \\
&= \zeta_{ijj'}(1,1)\mu_{ijj'} + (\tau_{1ij} - \zeta_{ijj'}(1,0))(\mu_{ij} - \mu_{ij'j'}) \ + \ \{\tau_{1ij'} - \zeta_{ijj'}(0,1)\} \\
&\quad \times (\mu_{ij'} - \mu_{ijj'}) \ + \ \{1 - \tau_{0ij} - \tau_{0ij'} + \zeta_{ijj'}(0,0)\}(1 - \mu_{ij} - \mu_{ij'} + \mu_{ijj'}) \\
&= \{\tau_{0ij} + \tau_{1ij} - 1 + \tau_{0ij'} - \zeta_{ijj'}(1,0) - \zeta_{ijj'}(0,0)\}\mu_{ij} \\
&\quad + \ \{\tau_{0ij} - 1 + \tau_{1ij'} + \tau_{0ij'} - \zeta_{ijj'}(0,1) - \zeta_{ijj'}(0,0)\}\mu_{ij'} \\
&\quad + \{1 - \tau_{0ij} - \tau_{1ij} - \tau_{0ij'} - \tau_{1ij'} + \zeta_{ijj'}(1,1) + \zeta_{ijj'}(1,0) \\
&\quad + \zeta_{ijj'}(0,1) + \zeta_{ijj'}(0,0)\}\mu_{ijj'} \ + \ \{1 - \tau_{0ij} - \tau_{0ij'} + \zeta_{ijj'}(0,0)\}.
\end{aligned}$$

3.9.2 Derivation of unbiased surrogate for $C_{ijj'}$

Under the assumptions that $\mathrm{E}[H_{ij}|\mathbf{Y}_i, \mathbf{X}_i] = \mathrm{E}[H_{ij}|Y_{ij}, \mathbf{X}_i]$ and $\mathrm{E}[H_{ij}H_{ij'}|\mathbf{Y}_i, \mathbf{X}_i] = \mathrm{E}[H_{ij}H_{ij'}|Y_{ij}, Y_{ij'}, \mathbf{X}_i]$ for $j \neq j'$, we have

$$\begin{aligned}
&\mathrm{E}[C^*_{ijj'}|\mathbf{Y}_i, \mathbf{X}_i] \\
&\quad = \mathrm{E}\left[\frac{b_0 + (S_{ij} - b_1)(S_{ij'} - b_2)}{b_3}|\mathbf{Y}_i, \mathbf{X}_i\right] \\
&\quad = \frac{1}{b_3}\mathrm{E}[b_0 + (S_{ij} - b_1)(S_{ij'} - b_2)|\mathbf{Y}_i, \mathbf{X}_i] \\
&\quad = \frac{1}{b_3}\left\{b_0 + \mathrm{E}[S_{ij}S_{ij'}|\mathbf{Y}_i, \mathbf{X}_i] - b_1\mathrm{E}[S_{ij'}|\mathbf{Y}_i, \mathbf{X}_i] - b_2\mathrm{E}[S_{ij}|\mathbf{Y}_i, \mathbf{X}_i] + b_1b_2\right\} \\
&\quad = \frac{1}{b_3}\left\{b_0 + \mathrm{E}[S_{ij}S_{ij'}|Y_{ij}, Y_{ij'}, \mathbf{X}_i] - b_1\mathrm{E}[S_{ij'}|Y_{ij}, Y_{ij'}, \mathbf{X}_i]\right. \\
&\qquad \left. - b_2\mathrm{E}[S_{ij}|Y_{ij}, Y_{ij'}, \mathbf{X}_i] + b_1b_2\right\}.
\end{aligned}$$

Now applying the expressions (3.18) and (3.19), we obtain

$$
\begin{aligned}
&\mathrm{E}[C^*_{ijj'}|\mathbf{Y}_i,\mathbf{X}_i] \\
&= \frac{1}{b_3}\Bigg[b_0 + \zeta_{ijj'}(1,1)Y_{ij}Y_{ij'} \;+\; \{\tau_{1ij} - \zeta_{ijj'}(1,0)\}\,Y_{ij}\,(1-Y_{ij'}) \\
&\qquad +\; \{\tau_{1ij'} - \zeta_{ijj'}(0,1)\}\,(1-Y_{ij})\,Y_{ij'} \;+\; \{1-\tau_{0ij}-\tau_{0ij'}+\zeta_{ijj'}(0,0)\} \\
&\qquad \times (1-Y_{ij})\,(1-Y_{ij'}) \;-\; b_1\,\{1-\tau_{0ij'}+(\tau_{0ij'}+\tau_{1ij'}-1)Y_{ij'}\} \\
&\qquad - b_2\,\{1-\tau_{0ij}+(\tau_{0ij}+\tau_{1ij}-1)Y_{ij}\} \;+ b_1 b_2 \Bigg] \\
&= \frac{1}{b_3}\Bigg[b_0 + 1 - \tau_{0ij} - \tau_{0ij'} + \zeta_{ijj'}(0,0) - b_1(1-\tau_{0ij'}) - b_2(1-\tau_{0ij}) + b_1 b_2 \\
&\qquad +\; \{\tau_{1ij}+\tau_{0ij}+\tau_{0ij'}-1-\zeta_{ijj'}(1,0)-\zeta_{ijj'}(0,0)-b_2(\tau_{0ij}+\tau_{1ij}-1)\}\,Y_{ij} \\
&\qquad +\; \{\tau_{1ij'}+\tau_{0ij'}+\tau_{0ij}-1-\zeta_{ijj'}(0,1)-\zeta_{ijj'}(0,0)-b_1(\tau_{0ij'}+\tau_{1ij'}-1)\}\,Y_{ij'} \\
&\qquad + \{1-\tau_{0ij}-\tau_{1ij}-\tau_{0ij'}-\tau_{1ij'}+\zeta_{ijj'}(1,1)+\zeta_{ijj'}(1,0) \\
&\qquad + \zeta_{ijj'}(0,1)+\zeta_{ijj'}(0,0)\}Y_{ij}Y_{ij'} \Bigg] \\
&= \frac{1}{b_3}\,(0 + 0\cdot Y_{ij} + 0\cdot Y_{ij'} + b_3\cdot Y_{ij}Y_{ij'}) \\
&= Y_{ij}Y_{ij'}
\end{aligned}
$$

3.9.3 Consistency and asymptotic normality for $\hat{\boldsymbol{\theta}}$

Because of the unbiasedness of $\mathbf{Q}_i(\boldsymbol{\eta})$ and $\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})$, the estimators $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\theta}}$ are consistent for $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, respectively. By first-order Taylor series approximation, we have

$$
\begin{aligned}
n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\theta}}-\boldsymbol{\theta} \\ \hat{\boldsymbol{\eta}}-\boldsymbol{\eta} \end{pmatrix} \;&=\; -\begin{pmatrix} \mathrm{E}\left[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right] & \mathrm{E}\left[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right] \\ \mathbf{0} & \mathrm{E}\left[\partial\mathbf{Q}_i(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right] \end{pmatrix}^{-1} \\
&\qquad \cdot\, n^{-1/2}\begin{pmatrix} \sum_{i=1}^n \tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta}) \\ \sum_{i=1}^n \mathbf{Q}_i(\boldsymbol{\eta}) \end{pmatrix} + o_p(1).
\end{aligned}
$$

It follows that

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= -\, n^{-1/2} \left\{ \mathrm{E} \left[\partial \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\theta}^{\mathrm{T}} \right] \right\}^{-1} \cdot \left\{ \sum_{i=1}^{n} \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \right. \\ & \quad \left. -\, \mathrm{E} \left[\partial \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}^{\mathrm{T}} \right] \cdot \left\{ \mathrm{E} \left[\partial \mathbf{Q}_i(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}^{\mathrm{T}} \right] \right\}^{-1} \cdot \sum_{i=1}^{n} \mathbf{Q}_i(\boldsymbol{\eta}) \right\} + o_p(1) \\ &= -\, n^{-1/2} \tilde{\boldsymbol{\Gamma}}^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta}) \cdot \sum_{i=1}^{n} \tilde{\boldsymbol{\Omega}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) + o_p(1). \end{aligned}$$

Then applying the Central Limit Theorem establishes the asymptotic distribution.

Let

$$\mathbf{M}_i(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{M}_{1i}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{M}_{21i}(\boldsymbol{\theta}) & \mathbf{M}_{2i}(\boldsymbol{\theta}) \end{pmatrix},$$

where $\mathbf{M}_{21i}(\boldsymbol{\theta}) = -\mathbf{D}_{2i} \mathbf{V}_{2i}^{-1} \left(\partial \boldsymbol{\xi}_i / \partial \boldsymbol{\beta}^{\mathrm{T}} \right)$. The matrix $\tilde{\boldsymbol{\Gamma}}$ can be consistently estimated by, as $n \to \infty$,

$$\widehat{\boldsymbol{\Gamma}} = n^{-1} \sum_{i=1}^{n} \mathbf{M}_i(\hat{\boldsymbol{\theta}}).$$

Define

$$\begin{pmatrix} \tilde{\boldsymbol{\Lambda}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ \tilde{\boldsymbol{\Lambda}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}) \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{1i} \mathbf{V}_{1i}^{-1} \left(\partial \tilde{\mathbf{Y}}_i / \partial \boldsymbol{\eta}^{\mathrm{T}} \right) \\ \mathbf{D}_{2i} \mathbf{V}_{2i}^{-1} \left(\partial \tilde{\mathbf{C}}_i / \partial \boldsymbol{\eta}^{\mathrm{T}} \right) \end{pmatrix},$$

and

$$\mathbf{J}_i(\boldsymbol{\eta}) = \begin{pmatrix} \mathbf{J}_{1i}(\boldsymbol{\eta}) & \mathbf{0} \\ \mathbf{J}_{21i}(\boldsymbol{\eta}) & \mathbf{J}_{2i}(\boldsymbol{\eta}) \end{pmatrix},$$

where $\mathbf{J}_{21i}(\boldsymbol{\eta}) = -\mathbf{D}^{\delta}_{\eta 2i} \left[\mathbf{V}^{\delta}_{\eta 2i} \right]^{-1} \left(\partial \boldsymbol{\zeta}_i^{\delta} / \partial \boldsymbol{\nu}^{\mathrm{T}} \right)$. Similarly, as $n \to \infty$, $\mathrm{E}[\partial \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}^{\mathrm{T}}]$

and $\mathrm{E}\left[\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right]$ can be consistently estimated by

$$\tilde{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} \tilde{\boldsymbol{\Lambda}}_{1i}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) \\ \tilde{\boldsymbol{\Lambda}}_{2i}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) \end{pmatrix},$$

and

$$\mathbf{J}(\hat{\boldsymbol{\eta}}) = n^{-1} \sum_{i=1}^{n} \mathbf{J}_i(\hat{\boldsymbol{\eta}}),$$

respectively. Therefore, the matrix $\tilde{\boldsymbol{\Sigma}}$ can be consistently estimated by

$$\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{\Omega}}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) \tilde{\boldsymbol{\Omega}}_i^{\mathrm{T}}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}),$$

where $\tilde{\boldsymbol{\Omega}}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) = \tilde{\mathbf{U}}_i(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) - \tilde{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) \mathbf{J}^{-1}(\hat{\boldsymbol{\eta}}) \mathbf{Q}_i(\hat{\boldsymbol{\eta}})$. A consistent estimator for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by the empirical version $n^{-1} \widehat{\boldsymbol{\Gamma}}^{-1} \widehat{\boldsymbol{\Sigma}} \left[\widehat{\boldsymbol{\Gamma}}^{-1}\right]^{\mathrm{T}}$.

3.9.4 Consistency and asymptotic normality for $\hat{\boldsymbol{\theta}}_{RS}$

The asymptotic distribution of $\hat{\boldsymbol{\theta}}_{RS}$ can be established in a similar manner to that in Section 3.4. However, there is an important difference arising from the interplay of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ in both $\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma})$. Specifically, applying the Taylor series expansion, we obtain

$$\begin{aligned} n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}}_{RS} - \boldsymbol{\theta} \\ \hat{\boldsymbol{\gamma}}_{RS} - \boldsymbol{\gamma} \end{pmatrix} &= -\begin{pmatrix} \mathrm{E}\left[\partial \boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right] & \mathrm{E}\left[\partial \boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^{\mathrm{T}}\right] \\ \mathrm{E}\left[\partial \boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right] & \mathrm{E}\left[\partial \boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^{\mathrm{T}}\right] \end{pmatrix}^{-1} \\ &\quad \cdot n^{-1/2} \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) \\ \sum_{i=1}^{n} \boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \end{pmatrix} + o_p(1). \end{aligned}$$

It follows that

$$
\begin{aligned}
n^{1/2}\left(\hat{\boldsymbol{\theta}}_{RS}-\boldsymbol{\theta}\right) &= n^{-1/2}\Bigg(\mathrm{E}\left[\partial\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right]-\mathrm{E}\left[\partial\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right] \\
&\quad \cdot\left\{\mathrm{E}\left[\partial\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right]\right\}^{-1}\cdot\mathrm{E}\left[\partial\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right]\Bigg)^{-1} \\
&\quad \cdot\Bigg\{\sum_{i=1}^{n}\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})\;-\;\mathrm{E}\left[\partial\boldsymbol{\mathcal{U}}_i(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right] \\
&\quad \cdot\left\{\mathrm{E}\left[\partial\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right]\right\}^{-1}\cdot\;\sum_{i=1}^{n}\boldsymbol{\mathcal{Q}}_{1i}(\boldsymbol{\theta},\boldsymbol{\gamma})\Bigg\}+o_p(1) \\
&= n^{-1/2}\boldsymbol{\Gamma}^{*-1}(\boldsymbol{\theta},\boldsymbol{\gamma})\cdot\sum_{i=1}^{n}\boldsymbol{\Omega}_i^{*}(\boldsymbol{\theta},\boldsymbol{\gamma})+o_p(1).
\end{aligned}
$$

Thus, the Central Limit Theorem yields the results.

As $n\to\infty$, $\boldsymbol{\Gamma}^{*}$ and $\boldsymbol{\Sigma}^{*}$ can be consistently estimated by their empirical counterparts given by

$$
\widehat{\boldsymbol{\Gamma}}^{*}=n^{-1}\sum_{i=1}^{n}\left[\mathrm{M}_i(\hat{\boldsymbol{\theta}}_{RS})-\boldsymbol{\Lambda}_i^{*}(\hat{\boldsymbol{\theta}}_{RS},\hat{\boldsymbol{\gamma}}_{RS})\cdot\left\{\boldsymbol{\mathcal{J}}_{1i}(\hat{\boldsymbol{\theta}}_{RS},\hat{\boldsymbol{\gamma}}_{RS})\right\}^{-1}\cdot\boldsymbol{\Delta}_i^{*}(\hat{\boldsymbol{\theta}}_{RS},\hat{\boldsymbol{\gamma}}_{RS})\right]
$$

and

$$
\widehat{\boldsymbol{\Sigma}}^{*}=n^{-1}\sum_{i=1}^{n}\boldsymbol{\Omega}_i^{*}(\hat{\boldsymbol{\theta}}_{RS},\hat{\boldsymbol{\gamma}}_{RS})\boldsymbol{\Omega}_i^{*}(\hat{\boldsymbol{\theta}}_{RS},\hat{\boldsymbol{\gamma}}_{RS})^{\mathrm{T}},
$$

respectively. A consistent estimator for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_{RS}$ is given by $\widehat{\boldsymbol{\Gamma}}^{*-1}\widehat{\boldsymbol{\Sigma}}^{*}\left[\widehat{\boldsymbol{\Gamma}}^{*-1}\right]^{\mathrm{T}}$.

Chapter 4

Estimating Equations for Analysis of Longitudinal Ordinal Data with Misclassified Responses and Covariates

4.1 Introduction

Longitudinal studies often involve repeated measurements of categorical outcomes with a set of covariates on each subject. The response can be nominal, i.e., there is no particular ordering of the categories, or ordinal, i.e., there is a natural ordering of the levels. Ordinal data are commonly seen in surveys and medical studies, where the categories are general representations of an underlying continuous variable, such as the measure of severity of a health condition. There are three widely used approaches for modeling repeated categorical data: transition models, mixed models, and marginal models. Transition models describe the probability distribution of a subject's outcome at a time point given the history of outcomes. The interest focuses on how covariates influence the transition intensity or transition probability from one response level to another. Mixed models take into account the correlations between repeated measurements by specifying some cluster-level random components.

In marginal models, the focus is on the relationship between covariates and the response variable at the population level. Various methodological strategies are utilized to account for the correlation. Liang and Zeger (1986) proposed a generalized estimating equation (GEE) approach, which assumes that the marginal distribution of the response follows a generalized linear model and uses a working correlation structure to account for correlation between the repeated measurements. Prentice (1988) further extended this approach for repeated binary data by specifying an additional estimating equation for the second-order association parameters, in which an independent working correlation matrix is assumed for the pairwise products of responses. Lipsitz et al. (1991) suggested using odds ratio as a measure of association between binary responses.

In regression analyses, however, measurement errors or misclassifications in both the response variable and covariates often arise due to non-perfect measuring systems or due to the designs of the studies. Neuhaus (1999, 2002) considered misclassification in binary response in generalized linear models and generalized linear mixed models. Carroll et al. (2006) provided a comprehensive summary of the development of statistical methods for dealing with measurement error in nonlinear models, mostly error in covariates. The approaches for correcting covariate error can be separated into two major classes: structural modeling, and functional modeling. In structural modeling, full parametric assumption on the distribution of the mismeasured covariate is made. In contrast, functional modeling leaves the probability distribution of the covariate completely unspecified, which is particularly attractive for handling covariate error problems. One typical example for functional modeling is the SIMEX approach proposed by Cook and Stefanski (1994), which uses a resampling method to establish a relationship between the bias and the variance of the measurement error and then extrapolate back to the case where there is no measurement error. Another example is the corrected score method (Nakamura, 1990, 1992), in which consistent estimators can be obtained by solving a set of estimating equations. Several authors also have used the corrected score methods for analysis of survival data with covariate measurement error and misclassification (e.g., Yi and Lawless, 2006; Zucker and Spiegelman, 2008). For repeated measurements, likelihood-based methods may not be available due to the lack of full probability model assumptions. Therefore, functional modeling for correcting covariate error is especially appealing in marginal methods that only reply on the least model assumptions,.

In this chapter, we consider marginal modeling for longitudinal or clustered ordinal data, where the response and a categorical covariate are subject to misclassifications. Extensions to handling multiple misclassified covariates can be established in a similar spirit. We adapt the approach of Akazawa et al. (1998) to formulate unbiased estimating functions using constructed unbiased surrogates for misclassified variables.

The remainder of this chapter is organized as follows. In Section 4.2, we give detailed model formulation for the response process as well as for both the response and the covariate misclassification processes. In Section 4.3, unbiased estimating functions of response parameters are constructed, which account for both response and covariate misclassifications. In Section 4.4, we propose a two-stage estimation approach for cases where a validation subsample with true categories of response and covariates is available. Asymptotic properties of the estimators are also derived. Section 4.5 presents simulation studies to investigate the performance of the proposed method under a variety of settings. We demonstrate the method by applying it to a data set from the Framingham Heart Study in Section 4.6. Section 4.7 includes some final remarks and discussion.

4.2 Model Formulation

4.2.1 The response process

We are interested in modeling the relationship between a categorical ordinal response and some covariates that can be either continuous or categorical. Let Y_{ij} denote the response that has $(K+1)$ distinct levels, say, $0, 1, \ldots, K$, for subject i at time point j, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Dummy variables are often used to represent a categorical variable in estimation of parameters. Let $Y_{ijk} = 1$ if $Y_{ij} = k$, and $Y_{ijk} = 0$ otherwise, $k = 0, \ldots, K$. We treat level 0 as the reference category. Therefore, it is sufficient to describe the response using the vector $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})^{\mathrm{T}}$. Let $\mathbf{Y}_i = (\mathbf{Y}_{i1}^{\mathrm{T}}, \ldots, \mathbf{Y}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$.

We assume that the covariates for subject i at time point j include a vector of precisely measured covariates $\mathbf{Z}_{ij}$ and a categorical variable X_{ij} with $(K^x + 1)$ levels, say, $0, 1, \ldots, K^x$, that may be subject to misclassification. Let $X_{ijq} = 1$ if $X_{ij} = q$, and $X_{ijq} = 0$ otherwise, $q = 0, \ldots, K^x$. Similarly, we treat level 0 as the reference category. Let $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijK^x})^{\mathrm{T}}$. Let $\mathbf{Z}_i = (\mathbf{Z}_{i1}^{\mathrm{T}}, \ldots, \mathbf{Z}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{X}_i = (\mathbf{X}_{i1}^{\mathrm{T}}, \ldots, \mathbf{X}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$.

Mean model

For marginal modeling, the interest is in studying the effects of covariates at the population level. Let $\mu_{ijk} = \mathrm{E}[Y_{ijk}|\mathbf{X}_i, \mathbf{Z}_i]$, $k = 0, \ldots, K$. We have $\sum_{k=0}^{K} \mu_{ijk} = 1$. It is often assumed that $\mathrm{E}[Y_{ijk}|\mathbf{X}_i, \mathbf{Z}_i] = \mathrm{E}[Y_{ijk}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}]$ (e.g., Pepe and Anderson, 1994). For ordinal data, it is common to use cumulative probabilities $\lambda_{ijk} = \Pr(Y_{ij} \geq k|\mathbf{X}_i, \mathbf{Z}_i)$, $k = 1, \ldots, K$, as alternatives to the marginals (e.g., Agresti, 2002). Proportional odds models are then employed to relate the response to the covariate effects (e.g., Miller et al., 1993), which are given by

$$\operatorname{logit} \lambda_{ijk} = \beta_{0k} + \mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_x + \mathbf{Z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_z, \qquad k = 1, \ldots, K, \tag{4.1}$$

where $\operatorname{logit}(u) = \log\{u/(1-u)\}$, $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_z$ are vectors of regression parameters associated with the effects of the misclassified covariate and the error-free covariates, and β_{0k} is the intercept in the kth logit model. It is easy to see that the marginals can be calculated from the cumulative probabilities as

$$\begin{cases} \mu_{ij1} = \lambda_{ij1} - \lambda_{ij2}, \\ \vdots \\ \mu_{ij(K-1)} = \lambda_{ij(K-1)} - \lambda_{ijK}, \\ \mu_{ijK} = \lambda_{ijK}. \end{cases}$$

The variance of Y_{ijk} can be given by $\operatorname{var}(Y_{ijk}) = \mu_{ijk}(1 - \mu_{ijk})$, $k = 1, \ldots, K$. Let $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{0K}, \boldsymbol{\beta}_x^{\mathrm{T}}, \boldsymbol{\beta}_z^{\mathrm{T}})^{\mathrm{T}}$. Let $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \ldots, \mu_{ijK^x})^{\mathrm{T}}$ and $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$.

Association model

The second-order dependence between two ordinal responses is often characterized by the bivariate cumulative probability. Let $\varsigma_{i;jk;j'k'} = \Pr(Y_{ij} \geq k, Y_{ij'} \geq k' | \mathbf{X}_i, \mathbf{Z}_i)$, $k, k' = 1, \ldots, K$. We use global odds ratio (e.g., Williamson et al., 1995) as an association measure for ordinal responses, which is given by

$$\begin{aligned}\psi_{i;jk;j'k'} &= \frac{\Pr(Y_{ij} \geq k, Y_{ij'} \geq k' | \mathbf{X}_i, \mathbf{Z}_i) \cdot \Pr(Y_{ij} < k, Y_{ij'} < k' | \mathbf{X}_i, \mathbf{Z}_i)}{\Pr(Y_{ij} \geq k, Y_{ij'} < k' | \mathbf{X}_i, \mathbf{Z}_i) \cdot \Pr(Y_{ij} < k, Y_{ij'} \geq k' | \mathbf{X}_i, \mathbf{Z}_i)} \\ &= \frac{\varsigma_{i;jk;j'k'} \left\{1 - \lambda_{ijk} - \lambda_{ij'k'} + \varsigma_{i;jk;j'k'}\right\}}{\left\{\lambda_{ijk} - \varsigma_{i;jk;j'k'}\right\} \left\{\lambda_{ij'k'} - \varsigma_{i;jk;j'k'}\right\}}, \qquad j < j',\ k, k' = 1, \ldots, K.\end{aligned}$$

A log-linear model is commonly employed for the global odds ratio, which is given by (Williamson et al., 1995)

$$\log \psi_{i;jk;j'k'} = \phi + \phi_k + \phi_{k'} + \phi_{kk'} + \mathbf{u}_{ijj'}^{\mathrm{T}} \boldsymbol{\alpha}_1, \quad k, k' = 1, \ldots, K, \tag{4.2}$$

where ϕ is a global intercept term, ϕ_k is the effect of category k, $\phi_{kk'}$ is the interaction effect between categories k and k' (with $\phi_{kk'} = \phi_{k'k}$), and $\mathbf{u}_{ijj'}$ is a vector of pair-specific covariates, the effects of which are quantified by a vector of regression parameters $\boldsymbol{\alpha}_1$. Identifiability constraints must be placed on the regression parameters. Let $\phi_1 = 0$, $\phi_{1k} = \phi_{k1} = 0$ for $k = 1, \ldots, K$. Let $\boldsymbol{\alpha} = (\phi, \{\phi_k,\ k = 2, \ldots, K\}^{\mathrm{T}}, \{\phi_{kk'},\ 2 \leq k \leq k' \leq K\}^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$ be a vector of all second-order association parameters.

The bivariate cumulative probability can be expressed in terms of the global odds ratio and the two marginal cumulative probabilities as

$$\varsigma_{i;jk;j'k'} = \begin{cases} \left\{a_{i;jk;j'k'} - \sqrt{b_{i;jk;j'k'}}\right\} / \left\{2\left(\psi_{i;jk;j'k'} - 1\right)\right\}, & \text{if } \psi_{i;jk;j'k'} \neq 1, \\ \lambda_{ijk}\lambda_{ij'k'}, & \text{if } \psi_{i;jk;j'k'} = 1, \end{cases}$$

where $a_{i;jk;j'k'} = 1 - (1 - \psi_{i;jk;j'k'})\left(\lambda_{ijk} + \lambda_{ij'k'}\right)$, $b_{i;jk;j'k'} = a_{i;jk;j'k'}^2 - 4\left(\psi_{i;jk;j'k'} - 1\right) \times \psi_{i;jk;j'k'} \lambda_{ijk} \lambda_{ij'k'}$.

For $j < j'$, let $C_{i;jk;j'k'} = Y_{ijk} Y_{ij'k'}$, $k, k' = 1, \ldots, K$, and let $\mathbf{C}_{ijj'} = (C_{i;j1;j'1}, C_{i;j1;j'2}, \ldots, C_{i;jK;j'K})^{\mathrm{T}}$. Therefore, $\mathbf{C}_{ijj'}$ contains all pairwise products of indicator variables for the jth and the j'th responses for subject i. Let $\mu_{i;jk;j'k'} =$

$\mathrm{E}[C_{i;jk;j'k'}|\mathbf{X}_i, \mathbf{Z}_i]$, $k, k' = 1, \ldots, K$, which can be rewritten as

$$\begin{aligned}
\mu_{i;jk;j'k'} &= \Pr(Y_{ij} = k, Y_{ij'} = k'|\mathbf{X}_i, \mathbf{Z}_i) \\
&= \begin{cases}
\varsigma_{i;jk;j'k'} - \varsigma_{i;j(k+1);j'k'} & \\
\qquad -\varsigma_{i;jk;j'(k'+1)} + \varsigma_{i;j(k+1);j'(k'+1)}, & \text{if } 1 \leq k, k' < K, \\
\varsigma_{i;jK;j'k'} - \varsigma_{i;jK;j'(k'+1)}, & \text{if } k = K,\ 1 \leq k' < K, \\
\varsigma_{i;jk;j'K} - \varsigma_{i;j(k+1);j'K}, & \text{if } 1 \leq k < K,\ k' = K, \\
\varsigma_{i;jK;j'K}, & \text{if } k = k' = K.
\end{cases}
\end{aligned}$$

Let $\boldsymbol{\xi}_{ijj'} = \{\mu_{i;j1;j'1},\ \mu_{i;j1;j'2},\ \ldots,\ \mu_{i;jK;j'K}\}^{\mathrm{T}}$. Let $\mathbf{C}_i = (\mathbf{C}_{ijj'}^{\mathrm{T}}, j < j')^{\mathrm{T}}$ and $\boldsymbol{\xi}_i = (\boldsymbol{\xi}_{ijj'}^{\mathrm{T}}, j < j')^{\mathrm{T}}$.

We further let

$$\begin{aligned}
\mu_{i;j0;j'k'} &= \Pr(Y_{ij} = 0, Y_{ij'} = k'|\mathbf{X}_i, \mathbf{Z}_i) \\
&= \mathrm{E}\left[\left(1 - \sum_{k=1}^{K} Y_{ijk}\right) Y_{ij'k'}|\mathbf{X}_i, \mathbf{Z}_i\right] \\
&= \mu_{ij'k'} - \sum_{k=1}^{K} \mu_{i;jk;j'k'}, \qquad \text{for } k' \neq 0,
\end{aligned}$$

and

$$\begin{aligned}
\mu_{i;j0;j'0} &= \Pr(Y_{ij} = 0, Y_{ij'} = 0|\mathbf{X}_i, \mathbf{Z}_i) \\
&= \mathrm{E}\left[\left(1 - \sum_{k=1}^{K} Y_{ijk}\right)\left(1 - \sum_{k'=1}^{K} Y_{ij'k'}\right)|\mathbf{X}_i, \mathbf{Z}_i\right] \\
&= 1 - \sum_{k=1}^{K} \mu_{ijk} - \sum_{k'=1}^{K} \mu_{ij'k'} + \sum_{k=1}^{K}\sum_{k'=1}^{K} \mu_{i;jk;j'k'}.
\end{aligned}$$

The correlation between two indicator variables is given by

$$\rho_{i;jk;jl} = \frac{-\mu_{ijk}\mu_{ijl}}{\sqrt{\mu_{ijk}(1 - \mu_{ijk})}\sqrt{\mu_{ijl}(1 - \mu_{ijl})}},$$

for $0 \leq k < l \leq K$ at the same time point j, and is given by

$$\rho_{i;jk;j'k'} = \frac{\mu_{i;jk;j'k'} - \mu_{ijk}\mu_{ij'k'}}{\sqrt{\mu_{ijk}(1-\mu_{ijk})}\sqrt{\mu_{ij'k'}(1-\mu_{ij'k'})}},$$

for $j \neq j'$ and $k, k' = 0, \ldots, K$. Bahadur (1961) and Cox (1972) described the joint distribution of correlated binary responses in terms of their marginals and correlations. Prentice (1988) considered a special case of the presentation by setting the third and higher correlations to be zero. Because $(Y_{i1} = k_1, \ldots, Y_{im_i} = k_{m_i}) = (Y_{i1k_1} = 1, \ldots, Y_{im_ik_{m_i}} = 1)$, we can obtain the joint distribution of repeated categorical responses in terms of their marginals and pairwise correlations

$$\begin{aligned}
&\Pr\left(Y_{i1} = k_1, \ldots, Y_{im_i} = k_{m_i} | \mathbf{X}_i, \mathbf{Z}_i\right) \\
&\quad = \Pr\left(Y_{i1k_1} = 1, \ldots, Y_{im_ik_{m_i}} = 1 | \mathbf{X}_i, \mathbf{Z}_i\right) \\
&\quad = \prod_{j=1}^{m_i} \mu_{ijk_j} \left\{ 1 + \sum_{j<j'} \rho_{i;jk_j;j'k_{j'}} \frac{(1-\mu_{ijk_j})(1-\mu_{ij'k_{j'}})}{\sqrt{\mu_{ijk_j}(1-\mu_{ijk_j})}\sqrt{\mu_{ij'k_{j'}}(1-\mu_{ij'k_{j'}})}} \right\},
\end{aligned}$$

where $k_1, \ldots, k_{m_i} = 0, \ldots, K$.

4.2.2 The misclassification process for the response

We observe the surrogate version S_{ij} for the true response Y_{ij}. Let $\tau_{ijk,l} = \Pr(S_{ij} = l | Y_{ij} = k, \mathbf{X}_i, \mathbf{Z}_i)$ be the probability that the surrogate response falls into category l when the true category is k $(k, l = 0, \ldots, K)$. The $(K+1) \times (K+1)$ (mis)classification probability matrix is given by

$$\mathbf{P}_{ij} = \begin{pmatrix} \tau_{ij0,0} & \tau_{ij0,1} & \cdots & \tau_{ij0,K} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{ijK,0} & \tau_{ijK,1} & \cdots & \tau_{ijK,K} \end{pmatrix} = \begin{pmatrix} \tau_{ij0,0} & \boldsymbol{\tau}_{ij0}^{\mathrm{T}} \\ \vdots & \vdots \\ \tau_{ijK,0} & \boldsymbol{\tau}_{ijK}^{\mathrm{T}} \end{pmatrix}, \tag{4.3}$$

where $\boldsymbol{\tau}_{ijk}^{\mathrm{T}} = (\tau_{ijk,1}, \ldots, \tau_{ijk,K})$, $k = 0, \ldots, K$. Let $\mathbf{S}_{ij} = (S_{ij1}, \ldots, S_{ijK})^{\mathrm{T}}$, where $S_{ijl} = 1$ if $S_{ij} = l$, and $S_{ijl} = 0$ otherwise, $l = 1, \ldots, K$. It is easy to see that $\mathrm{E}[\mathbf{S}_{ij} | Y_{ij} = k, \mathbf{X}_i, \mathbf{Z}_i] = \boldsymbol{\tau}_{ijk}$ for $k = 0, \ldots, K$.

Generalized logit models are often employed for the misclassification process (e.g., Albert et al., 1997; Pfeffermann et al., 1998), which are given by

$$\log\left(\frac{\tau_{ijk,l}}{\tau_{ijk,0}}\right) = \mathbf{L}_{ij}^{\mathrm{T}}\boldsymbol{\gamma}_{kl}, \qquad k = 0, \ldots, K,\ l = 1, \ldots, K,$$

where $\mathbf{L}_{ij}$ is a covariate vector featuring the misclassification, and $\boldsymbol{\gamma}_{kl}$ is a vector of regression parameters in the logit model. For simplicity, we assume that $\mathbf{L}_{ij}$ do not include the misclassified covariate $\mathbf{X}_i$. Let $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k1}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{kK}^{\mathrm{T}})^{\mathrm{T}}$. Further let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_K^{\mathrm{T}})^{\mathrm{T}}$.

4.2.3 The misclassification process for the covariate

Instead of observing the categorical covariate X_{ij}, we obtain a surrogate version W_{ij}. Assume that the misclassification in covariate is independent of both the response process and the misclassification process for the response. Let $\pi_{ijq,r} = \Pr(W_{ij} = r|X_{ij} = q, \mathbf{Z}_i)$ be the probability that the surrogate covariate falls into category r when the true category is q $(q, r = 0, \ldots, K^x)$. The $(K^x+1)\times(K^x+1)$ (mis)classification probability matrix is given by

$$\mathbf{G}_{ij} = \begin{pmatrix} \pi_{ij0,0} & \pi_{ij0,1} & \cdots & \pi_{ij0,K^x} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{ijK^x,0} & \pi_{ijK^x,1} & \cdots & \pi_{ijK^x,K^x} \end{pmatrix} = \begin{pmatrix} \pi_{ij0,0} & \boldsymbol{\pi}_{ij0}^{\mathrm{T}} \\ \vdots & \vdots \\ \pi_{ijK^x,0} & \boldsymbol{\pi}_{ijK^x}^{\mathrm{T}} \end{pmatrix}, \tag{4.4}$$

where $\boldsymbol{\pi}_{ijq}^{\mathrm{T}} = (\pi_{ijq,1}, \ldots, \pi_{ijq,K^x})$, $q = 0, \ldots, K^x$. Let $\mathbf{W}_{ij} = (W_{ij1}, \ldots, W_{ijK^x})^{\mathrm{T}}$, where W_{ijr} is 1 if $W_{ij} = r$ and 0 otherwise. We have $\mathrm{E}[\mathbf{W}_{ij}|X_{ij} = q, \mathbf{Z}_i] = \boldsymbol{\pi}_{ijq}$ for $q = 0, \ldots, K^x$. Again, we use generalized logit models to characterize the misclassification process, which are given by

$$\log\left(\frac{\pi_{ijq,r}}{\pi_{ijq,0}}\right) = \mathbf{L}_{ij}^{x\mathrm{T}}\boldsymbol{\varphi}_{qr}, \qquad q = 0, \ldots, K^x,\ r = 1, \ldots, K^x,$$

where $\mathbf{L}_{ij}^{x}$ is a covariate vector associated with the misclassification, and $\boldsymbol{\varphi}_{qr}$ is a vector of the regression parameters in the logistic model. Let $\boldsymbol{\varphi}_q = (\boldsymbol{\varphi}_{q1}^{\mathrm{T}}, \ldots, \boldsymbol{\varphi}_{qK^x}^{\mathrm{T}})^{\mathrm{T}}$, and let $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_0^{\mathrm{T}}, \ldots, \boldsymbol{\varphi}_{K^x}^{\mathrm{T}})^{\mathrm{T}}$.

4.3 Estimating Equations

4.3.1 Estimating equations under the true model

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ be a vector of all response parameters. Let $\mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)$, where $\mathbf{D}_{1i} = \partial \boldsymbol{\mu}_i^{\mathrm{T}}/\partial \boldsymbol{\beta}$, $\mathbf{V}_{1i} = \mathbf{B}_{1i}^{1/2}\mathbf{R}_{1i}\mathbf{B}_{1i}^{1/2}$, $\mathbf{B}_{1i} = \mathrm{diag}\{\mu_{i11}(1-\mu_{i11}), \mu_{i12}(1-\mu_{i12}), \ldots, \mu_{im_iK}(1-\mu_{im_iK})\}$, $\mathbf{R}_{1i}$ is the correlation matrix of $\mathbf{Y}_i$ and involves both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Let $\mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}(\mathbf{C}_i - \boldsymbol{\xi}_i)$, where $\mathbf{D}_{2i} = \partial \boldsymbol{\xi}_i^{\mathrm{T}}/\partial \boldsymbol{\alpha}$, and $\mathbf{V}_{2i}$ is a working covariance matrix for $\mathbf{C}_i$. To avoid specifying third and higher-order moments, a block diagonal working matrix is often used for $\mathbf{V}_{2i}$. Specifically, the entries in the diagonal block matrices in $\mathbf{V}_{2i}$ involving only the jth and j'th time points are given by

$$\mathrm{cov}(C_{i;jk;j'k'}, C_{i;jl;j'l'}) = \begin{cases} \mu_{i;jk;j'k'}(1-\mu_{i;jk;j'k'}), & \text{for } (j,k;j',k') = (j,l;j',l'), \\ -\mu_{i;jk;j'k'}\mu_{i;jl;j'l'}, & \text{for } (j,k;j',k') \neq (j,l;j',l'). \end{cases}$$

In the absence of misclassifications, the original set of estimating equations for response parameters is given by

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) \\ \mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) \end{pmatrix} = \mathbf{0}. \tag{4.5}$$

Iterative estimation procedure such as Fisher's scoring algorithm can be used. Let $\mathbf{M}_{1i}(\boldsymbol{\theta}) = -\mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\mathbf{D}_{1i}^{\mathrm{T}}$, and $\mathbf{M}_{2i}(\boldsymbol{\theta}) = -\mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\mathbf{D}_{2i}^{\mathrm{T}}$. Given an initial estimate $\boldsymbol{\theta}^{(0)}$, we iteratively update the estimate of $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \begin{pmatrix} \left\{-\sum_{i=1}^{n} \mathbf{M}_{1i}(\boldsymbol{\theta}^{(t)})\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} \mathbf{U}_{1i}(\boldsymbol{\theta}^{(t)})\right\} \\ \left\{-\sum_{i=1}^{n} \mathbf{M}_{2i}(\boldsymbol{\theta}^{(t)})\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\theta}^{(t)})\right\} \end{pmatrix}, \quad t = 0, 1, \ldots$$

until convergence.

4.3.2 Estimating equations in the presence of covariate misclassification alone

In this subsection, we consider the case where X_{ij}'s are subject to misclassification, and Y_{ij}'s are correctly observed.

Akazawa et. al (1998) proposed a method to construct an unbiased surrogate for the vector $\mathbf{X}_{ij}$ from observed surrogate $\mathbf{W}_{ij}$. Define a $K^x \times K^x$ matrix

$$\mathbf{G}_{ij}^* = (\boldsymbol{\pi}_{ij1} - \boldsymbol{\pi}_{ij0}, \ldots, \boldsymbol{\pi}_{ijK^x} - \boldsymbol{\pi}_{ij0}),$$

and define

$$\mathbf{X}_{ij}^* = (X_{ij1}^*, \ldots, X_{ijK^x}^*)^{\mathrm{T}} = \mathbf{G}_{ij}^{*-1}(\mathbf{W}_{ij} - \boldsymbol{\pi}_{ij0}).$$

One can verify that $\mathrm{E}[\mathbf{X}_{ij}^*|\mathbf{X}_i, \mathbf{Z}_i] = \mathbf{X}_{ij}$. Let $X_{ij0}^* = 1 - \sum_{q=1}^{K^x} X_{ijq}^*$.

Denote by $\mathbf{e}_q$ the K^x-dimensional vector whose rth component is 1 if $r = q$ and 0 otherwise, $q = 1, \ldots, K^x$. Let $\mathbf{e}_0 = \mathbf{0}$ be a vector of all zeros. In Section 4.8.1 we generalize the result of Akazawa et. al (1998) to cases of an arbitrary vector of real-valued functions. By applying these results, we can obtain an unbiased surrogates for $\mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ and $\mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ from observed $(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i)$, which are given by

$$\begin{aligned}
\mathbf{U}_{1i}^*(\boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i) &= \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}_i, (\mathbf{e}_{q_1}^{\mathrm{T}}, \ldots, \mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} X_{ijq_j}^*, \\
\mathbf{U}_{2i}^*(\boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i) &= \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}_i, (\mathbf{e}_{q_1}^{\mathrm{T}}, \ldots, \mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} X_{ijq_j}^*.
\end{aligned}$$

Here the quantity $\prod_{j=1}^{m_i} X_{ijq_j}^*$ plays the role of weight for each of the $(K^x + 1)^{m_i}$ possibilities of the underlying true $\mathbf{X}_i$.

4.3.3 Estimating equations in the presence of response and covariate misclassifications

When the response variable is also subject to misclassification, an unbiased surrogate for the vector $\mathbf{Y}_{ij}$ can be constructed using similar techniques. Define

$$\mathbf{P}^*_{ij} = (\boldsymbol{\tau}_{ij1} - \boldsymbol{\tau}_{ij0}, \ldots, \boldsymbol{\tau}_{ijK} - \boldsymbol{\tau}_{ij0}),$$

and define

$$\mathbf{Y}^*_{ij} = (Y^*_{ij1}, \ldots, Y^*_{ijK})^{\mathrm{T}} = \mathbf{P}^{*-1}_{ij}(\mathbf{S}_{ij} - \boldsymbol{\tau}_{ij0}).$$

Then $\mathrm{E}\left[\mathbf{Y}^*_{ij} | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right] = \mathbf{Y}_{ij}$. Let $\mathbf{Y}^*_i = (\mathbf{Y}^{*\mathrm{T}}_{i1}, \ldots, \mathbf{Y}^{*\mathrm{T}}_{im_i})^{\mathrm{T}}$. Let $\boldsymbol{\eta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\varphi}^{\mathrm{T}})^{\mathrm{T}}$ denote a vector of all nuisance parameters. Therefore, unbiased estimating functions are given by

$$\begin{aligned}
&\mathbf{U}^{**}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{W}_i, \mathbf{Z}_i) \\
&\quad = \mathbf{U}^*_{1i}(\boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{Y}^*_i, \mathbf{W}_i, \mathbf{Z}_i) \\
&\quad = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}^*_i, (\mathbf{e}^{\mathrm{T}}_{q_1}, \ldots, \mathbf{e}^{\mathrm{T}}_{q_{m_i}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} X^*_{ijq_j}, \qquad (4.6) \\
&\mathbf{U}^{**}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{W}_i, \mathbf{Z}_i) \\
&\quad = \mathbf{U}^*_{2i}(\boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{Y}^*_i, \mathbf{W}_i, \mathbf{Z}_i) \\
&\quad = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}^*_i, (\mathbf{e}^{\mathrm{T}}_{q_1}, \ldots, \mathbf{e}^{\mathrm{T}}_{q_{m_i}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} X^*_{ijq_j}. \qquad (4.7)
\end{aligned}$$

With some algebra one can verify that $\mathrm{E}\left[\mathbf{U}^{**}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{W}_i, \mathbf{Z}_i) | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right] = \mathbf{U}_{1i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ and $\mathrm{E}\left[\mathbf{U}^{**}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \mathbf{S}_i, \mathbf{W}_i, \mathbf{Z}_i) | \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\right] = \mathbf{U}_{2i}(\boldsymbol{\theta}; \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$.

4.4 Inference Method with Validation Subsample Available

4.4.1 Estimating equations for $\boldsymbol{\eta}$

Equations (4.6) and (4.7) are constructed treating nuisance parameters $\boldsymbol{\eta}$ as known. In practice, however, $\boldsymbol{\eta}$ is often unknown. Instead, a validation subsample may be available (Carroll et al., 2006). Let $\delta_{ij} = 1$ if the jth observation for subject i is included in the validation subsample, and $\delta_{ij} = 0$ otherwise. In current stage of methodology development, we assume that both the true measurements of the response and the covariates are obtained for the jth observation if $\delta_{ij} = 1$. In reality, however, selecting a validation subsample for true response measurements may be independent of that for covariates, for which we need two sets of validation indicators. Our proposed methods can be easily extended to accommodate this situation.

It is often assumed that $\mathrm{E}[\mathbf{S}_{ij}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i] = \mathrm{E}[\mathbf{S}_{ij}|\mathbf{Y}_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}]$ (Pepe and Anderson, 1994). Therefore, we use notation $\boldsymbol{\tau}_{ij}(\mathbf{Y}_{ij}) = \mathrm{E}[\mathbf{S}_{ij}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i]$ to express the dependence of $\mathbf{S}_{ij}$ on $\mathbf{Y}_{ij}$. Under the assumption of independence between response misclassifications, the estimating function of $\boldsymbol{\gamma}$ contributed by subject i is given by

$$\mathbf{Q}_{\boldsymbol{\gamma}i}(\boldsymbol{\gamma}) = \sum_{j=1}^{m_i} \mathbf{D}_{\boldsymbol{\gamma}ij} \mathbf{V}_{\boldsymbol{\gamma}ij}^{-1} \left\{ \mathbf{S}_{ij} - \boldsymbol{\tau}_{ij}(\mathbf{Y}_{ij}) \right\} \delta_{ij},$$

where $\mathbf{D}_{\boldsymbol{\gamma}ij} = \partial \boldsymbol{\tau}_{ij}^{\mathrm{T}}(\mathbf{Y}_{ij})/\partial \boldsymbol{\gamma}$, and $\mathbf{V}_{\boldsymbol{\gamma}ij}$ is the covariance matrix of $\mathbf{S}_{ij}$ conditional on $\mathbf{Y}_{ij}$.

Similarly, assumption $\mathrm{E}[\mathbf{W}_{ij}|\mathbf{X}_i, \mathbf{Z}_i] = \mathrm{E}[\mathbf{W}_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}]$ is often made. Let $\boldsymbol{\pi}_{ij}(\mathbf{X}_{ij}) = \mathrm{E}[\mathbf{W}_{ij}|\mathbf{X}_i, \mathbf{Z}_i]$. Under the assumption of independence between misclassifications, the estimating function of $\boldsymbol{\varphi}$ contributed by subject i is given by

$$\mathbf{Q}_{\boldsymbol{\varphi}i}(\boldsymbol{\varphi}) = \sum_{j=1}^{m_i} \mathbf{D}_{\boldsymbol{\varphi}ij} \mathbf{V}_{\boldsymbol{\varphi}ij}^{-1} \left\{ \mathbf{W}_{ij} - \boldsymbol{\pi}_{ij}(\mathbf{X}_{ij}) \right\} \delta_{ij},$$

where $\mathbf{D}_{\boldsymbol{\varphi}ij} = \partial \boldsymbol{\pi}_{ij}^{\mathrm{T}}(\mathbf{X}_{ij})/\partial \boldsymbol{\varphi}$, and $\mathbf{V}_{\boldsymbol{\varphi}ij}$ is the covariance matrix of $\mathbf{W}_{ij}$ conditional on $\mathbf{X}_{ij}$.

4.4.2 Estimating equations for $\boldsymbol{\theta}$

The validation subsample can also be incorporated into the estimating functions to improve the efficiencies of the estimators for $\boldsymbol{\theta}$. Let $\tilde{\mathbf{Y}}_{ij} = (\tilde{Y}_{ij1}, \ldots, \tilde{Y}_{ijK})^{\mathrm{T}}$, where $\tilde{Y}_{ijk} = Y_{ijk}$ if $\delta_{ij} = 1$, and $\tilde{Y}_{ijk} = Y^*_{ijk}$ otherwise. Let $\tilde{C}_{i;jk;j'k'} = \tilde{Y}_{ijk}\tilde{Y}_{ij'k'}$ for $j \neq j'$. Similarly, let $\tilde{\mathbf{X}}_{ij} = (\tilde{X}_{ij1}, \ldots, \tilde{X}_{ijK^x})^{\mathrm{T}}$, where $\tilde{X}_{ijq} = X_{ijq}$ if $\delta_{ij} = 1$, and $\tilde{X}_{ijq} = X^*_{ijq}$ otherwise.

By incorporating the validation data in subject i, the improved estimating functions are given by

$$\tilde{\mathbf{U}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{Y}}_i, \tilde{\mathbf{X}}_i, \mathbf{Z}_i) = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{U}_{1i}(\boldsymbol{\theta}; \tilde{\mathbf{Y}}_i, (\mathbf{e}_{q_1}^{\mathrm{T}}, \ldots, \mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} \tilde{X}_{ijq_j},$$

$$\tilde{\mathbf{U}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{Y}}_i, \tilde{\mathbf{X}}_i, \mathbf{Z}_i) = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{U}_{2i}(\boldsymbol{\theta}; \tilde{\mathbf{Y}}_i, (\mathbf{e}_{q_1}^{\mathrm{T}}, \ldots, \mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} \tilde{X}_{ijq_j}.$$

Therefore, a more efficient estimate of $\boldsymbol{\theta}$ can be obtained by solving estimating equations

$$\sum_{i=1}^{n} \begin{pmatrix} \tilde{\mathbf{U}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{Y}}_i, \tilde{\mathbf{X}}_i, \mathbf{Z}_i) \\ \tilde{\mathbf{U}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}; \tilde{\mathbf{Y}}_i, \tilde{\mathbf{X}}_i, \mathbf{Z}_i) \end{pmatrix} = \mathbf{0}. \tag{4.8}$$

4.4.3 Estimation and asymptotic distribution

We use a two-stage estimation procedure for the parameters.

Stage 1. Estimate all parameters associated with the misclassification processes. Specifically, solve

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{Q}_{\boldsymbol{\gamma}i}(\boldsymbol{\gamma}) \\ \mathbf{Q}_{\boldsymbol{\varphi}i}(\boldsymbol{\varphi}) \end{pmatrix} = \mathbf{0},$$

and obtain estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\varphi}}$. Under the independence assumption for misclassification processes, this is equivalent to fitting generalized logit models to the validation data, in which misclassification events are now treated as "responses".

Stage 2. Replace $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ with their estimates, and solve (4.8) via the Fisher scoring algorithm. Given an initial value $\boldsymbol{\theta}^{(0)}$, we iteratively update $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \begin{pmatrix} \left\{\sum_{i=1}^{n} \tilde{\mathbf{M}}_{1i}(\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\eta}})\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} \tilde{\mathbf{U}}_{1i}(\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\eta}})\right\} \\ \left\{\sum_{i=1}^{n} \tilde{\mathbf{M}}_{2i}(\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\eta}})\right\}^{-1} \cdot \left\{\sum_{i=1}^{n} \tilde{\mathbf{U}}_{2i}(\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\eta}})\right\} \end{pmatrix}, \quad t = 0, 1, \ldots$$

where

$$\tilde{\mathbf{M}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{M}_{1i}(\boldsymbol{\theta}; \tilde{\mathbf{Y}}_i, (\mathbf{e}_{q_1}^{\mathrm{T}}, \ldots, \mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} \tilde{X}_{ijq_j},$$

$$\tilde{\mathbf{M}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{M}_{2i}(\boldsymbol{\theta}; \tilde{\mathbf{Y}}_i, (\mathbf{e}_{q_1}^{\mathrm{T}}, \ldots, \mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} \tilde{X}_{ijq_j}.$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}})^{\mathrm{T}}$ denote the estimate of $\boldsymbol{\theta}$ at convergence.

We conclude this section with the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ which accounts for extra uncertainty induced by the estimation of $\boldsymbol{\eta}$. Let $\tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \left(\tilde{\mathbf{U}}_{1i}^{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta}), \tilde{\mathbf{U}}_{2i}^{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta})\right)^{\mathrm{T}}$, and $\mathbf{Q}_i(\boldsymbol{\eta}) = \left(\mathbf{Q}_{\boldsymbol{\gamma}i}^{\mathrm{T}}(\boldsymbol{\gamma}), \mathbf{Q}_{\boldsymbol{\varphi}i}^{\mathrm{T}}(\boldsymbol{\varphi})\right)^{\mathrm{T}}$. By first-order Taylor series approximation, we have

$$\begin{aligned} n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \\ \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \end{pmatrix} &= - \begin{pmatrix} \mathrm{E}\left[\partial \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\theta}^{\mathrm{T}}\right] & \mathrm{E}\left[\partial \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}^{\mathrm{T}}\right] \\ \mathbf{0} & \mathrm{E}\left[\partial \mathbf{Q}_i(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}^{\mathrm{T}}\right] \end{pmatrix}^{-1} \\ &\quad \cdot n^{-1/2} \begin{pmatrix} \sum_{i=1}^{n} \tilde{\mathbf{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ \sum_{i=1}^{n} \mathbf{Q}_i(\boldsymbol{\eta}) \end{pmatrix} + o_p(1) \end{aligned}$$

With some algebra, we obtain

$$
\begin{aligned}
n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= -n^{-1/2}\left\{\mathrm{E}\left[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right]\right\}^{-1}\cdot\left\{\sum_{i=1}^{n}\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})\right.\\
&\quad\left.-\,\mathrm{E}\left[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right]\cdot\left\{\mathrm{E}\left[\partial\mathbf{Q}_i(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right]\right\}^{-1}\cdot\sum_{i=1}^{n}\mathbf{Q}_i(\boldsymbol{\eta})\right\}+o_p(1)\\
&= -n^{-1/2}\tilde{\boldsymbol{\Gamma}}^{-1}(\boldsymbol{\theta},\boldsymbol{\eta})\cdot\sum_{i=1}^{n}\boldsymbol{\Omega}_i(\boldsymbol{\theta},\boldsymbol{\eta})+o_p(1),
\end{aligned}
$$

where $\boldsymbol{\Omega}_i(\boldsymbol{\theta},\boldsymbol{\eta}) = \tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta}) - \mathrm{E}[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}]\cdot\left\{\mathrm{E}\left[\partial\mathbf{Q}_i(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right]\right\}^{-1}\cdot\mathbf{Q}_i(\boldsymbol{\eta})$, and $\tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta},\boldsymbol{\eta}) = \mathrm{E}\left[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\right]$. Therefore, $\tilde{\mathbf{U}}(\boldsymbol{\theta},\hat{\boldsymbol{\eta}})$ and $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ are asymptotically normally distributed with mean $\mathbf{0}$ and asymptotic covariance matrices given by $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Gamma}}^{-1}\tilde{\boldsymbol{\Sigma}}\left[\tilde{\boldsymbol{\Gamma}}^{-1}\right]^{\mathrm{T}}$, respectively, where $\tilde{\boldsymbol{\Sigma}} = \mathrm{E}\left[\boldsymbol{\Omega}_i(\boldsymbol{\theta},\boldsymbol{\eta})\boldsymbol{\Omega}_i^{\mathrm{T}}(\boldsymbol{\theta},\boldsymbol{\eta})\right]$.

Let

$$
\tilde{\mathbf{M}}_i(\boldsymbol{\theta},\boldsymbol{\eta}) = \begin{pmatrix} \tilde{\mathbf{M}}_{1i}(\boldsymbol{\theta},\boldsymbol{\eta}) & \mathbf{0} \\ \tilde{\mathbf{M}}_{21i}(\boldsymbol{\theta},\boldsymbol{\eta}) & \tilde{\mathbf{M}}_{2i}(\boldsymbol{\theta},\boldsymbol{\eta}) \end{pmatrix},
$$

where

$$
\tilde{\mathbf{M}}_{21i}(\boldsymbol{\theta},\boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_1=0}^{K^x}\mathbf{M}_{21i}(\boldsymbol{\theta};\tilde{\mathbf{Y}}_i,(\mathbf{e}_{q_1}^{\mathrm{T}},\ldots,\mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}},\mathbf{Z}_i)\prod_{j=1}^{m_i}\tilde{X}_{ijq_j},
$$

and $\mathbf{M}_{21i}(\boldsymbol{\theta};\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i) = -\mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\left(\partial\boldsymbol{\xi}_i/\partial\boldsymbol{\beta}^{\mathrm{T}}\right)$. As $n\to\infty$, the matrix $\tilde{\boldsymbol{\Gamma}}$ can be consistently estimated by

$$
\widehat{\boldsymbol{\Gamma}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{M}}_i(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}).
$$

Let

$$
\mathbf{J}_i(\boldsymbol{\eta}) = \begin{pmatrix} \mathbf{J}_{\gamma i}(\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\varphi i}(\boldsymbol{\varphi}) \end{pmatrix},
$$

where $\mathbf{J}_{\boldsymbol{\gamma}i}(\boldsymbol{\gamma}) = -\sum_{j=1}^{m_i} \mathbf{D}_{\boldsymbol{\gamma}ij}\mathbf{V}_{\boldsymbol{\gamma}ij}^{-1}\left(\partial\boldsymbol{\tau}_{ij}/\partial\boldsymbol{\gamma}^{\mathrm{T}}\right)\delta_{ij}$, and $\mathbf{J}_{\boldsymbol{\varphi}i}(\boldsymbol{\eta}) = -\sum_{j=1}^{m_i} \mathbf{D}_{\boldsymbol{\varphi}ij}\mathbf{V}_{\boldsymbol{\varphi}ij}^{-1}\cdot \left(\partial\boldsymbol{\pi}_{ij}/\partial\boldsymbol{\varphi}^{\mathrm{T}}\right)\delta_{ij}$. Define

$$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{\varphi}i}(\boldsymbol{\theta},\boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_1=0}^{K^x}\mathbf{U}_i(\boldsymbol{\theta};\tilde{\mathbf{Y}}_i,(\mathbf{e}_{q_1}^{\mathrm{T}},\ldots,\mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}},\mathbf{Z}_i)\left(\partial\prod_{j=1}^{m_i}\tilde{X}_{ijq_j}/\partial\boldsymbol{\varphi}^{\mathrm{T}}\right),$$

where $\partial\prod_{j=1}^{m_i}\tilde{X}_{ijq_j}/\partial\boldsymbol{\varphi}^{\mathrm{T}}$ is given in Section 4.8.2. Also define

$$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{\gamma}i}(\boldsymbol{\theta},\boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_1=0}^{K^x}\boldsymbol{\Lambda}_{\boldsymbol{\gamma}i}(\boldsymbol{\theta},\boldsymbol{\eta};\tilde{\mathbf{Y}}_i,(\mathbf{e}_{q_1}^{\mathrm{T}},\ldots,\mathbf{e}_{q_{m_i}}^{\mathrm{T}})^{\mathrm{T}},\mathbf{Z}_i)\prod_{j=1}^{m_i}\tilde{X}_{ijq_j},$$

where

$$\boldsymbol{\Lambda}_{\boldsymbol{\gamma}i}(\boldsymbol{\theta},\boldsymbol{\eta};\tilde{\mathbf{Y}}_i,\mathbf{X}_i,\mathbf{Z}_i) = \begin{pmatrix}\mathbf{D}_{1i}\mathbf{V}_{1i}^{-1}\partial\tilde{\mathbf{Y}}_i/\partial\boldsymbol{\gamma}^{\mathrm{T}}\\ \mathbf{D}_{2i}\mathbf{V}_{2i}^{-1}\partial\tilde{\mathbf{C}}_i/\partial\boldsymbol{\gamma}^{\mathrm{T}}\end{pmatrix},$$

and $\partial\tilde{\mathbf{Y}}_i/\partial\boldsymbol{\gamma}^{\mathrm{T}}$ and $\partial\tilde{\mathbf{C}}_i/\partial\boldsymbol{\gamma}^{\mathrm{T}}$ are given in Section 4.8.3. As $n\to\infty$, $\mathrm{E}[\partial\tilde{\mathbf{U}}_i(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}]$ and $\mathrm{E}\left[\partial\mathbf{Q}_i(\boldsymbol{\eta})/\partial\boldsymbol{\eta}^{\mathrm{T}}\right]$ can be consistently estimated by

$$\tilde{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) = n^{-1}\sum_{i=1}^{n}\begin{pmatrix}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{\gamma}i}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) & \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{\varphi}i}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\end{pmatrix}$$

and

$$\mathbf{J}(\hat{\boldsymbol{\eta}}) = n^{-1}\sum_{i=1}^{n}\mathbf{J}_i(\hat{\boldsymbol{\eta}}),$$

respectively. Therefore, the matrix $\tilde{\boldsymbol{\Sigma}}$ can be consistently estimated by

$$\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{\Omega}}_i(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\widehat{\boldsymbol{\Omega}}_i^{\mathrm{T}}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}),$$

where $\widehat{\boldsymbol{\Omega}}_i(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) = \tilde{\mathbf{U}}_i(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}}) - \tilde{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\eta}})\mathbf{J}^{-1}(\hat{\boldsymbol{\eta}})\mathbf{Q}_i(\hat{\boldsymbol{\eta}})$. A consistent estimator for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by $n^{-1}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Sigma}}\left[\widehat{\boldsymbol{\Gamma}}^{-1}\right]^{\mathrm{T}}$.

4.5 Simulation

4.5.1 Design of simulations

We conduct simulation studies to investigate the performance of the proposed method under different settings. We consider a longitudinal study in which three visits are planned for n patients, with the sample size $n = 1000$ for both cases with known and unknown $\boldsymbol{\eta}$. Half of the cohort is randomly assigned to the treatment group, while the other half is assigned to the placebo group. A 3-level categorical covariate X_{ij}, which takes value at 0,1, and 2 with proportions 0.5, 0.3, and 0.2, is generated for each patient at each visit independently. The kth logit model ($k = 1, 2$) for the response $\mathbf{Y}_{ij}$ are given by

$$\text{logit } \lambda_{ijk} = \beta_{0k} + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 Z_{ij1} + \beta_4 Z_{ij2} + \beta_5 Z_{ij3}, \quad k = 1, 2,$$

where X_{ij1} is 1 if $X_{ij} = 1$ and 0 otherwise, X_{ij2} is 1 if $X_{ij} = 2$ and 0 otherwise, Z_{ij1} is 1 if subject i is assigned to the treatment group and 0 otherwise, Z_{ij2} is 1 for visit #2 and 0 otherwise, and Z_{ij3} is 1 for visit #3 and 0 otherwise. The mean parameters are given by $\beta_{01} = \log(2)$, $\beta_{02} = \log(1/2)$, $\beta_1 = \log(2)$, $\beta_2 = \log(3)$, $\beta_3 = \log(1/2)$, and $\beta_4 = \log(3/4)$, and $\beta_5 = \log(1/2)$. Therefore, higher levels of X_{ij} are associated with increased probabilities of higher response levels, and the treatment has a positive effect on lowering response levels compared to the placebo. We consider two models for the second-order association structure:

(M1) A common global odds ratio is assumed for all pairs of responses, i.e.,

$$\log \psi_{i;jk;j'k'} = \phi, \quad k, k' = 1, 2, \tag{4.9}$$

where the single intercept is specified as $\phi = \log(3)$.

(M2) Global odds ratio is dependent of the response levels, i.e.,

$$\begin{aligned} \log \psi_{i;jk;j'k'} &= \phi + \phi_2 \cdot \mathrm{I}(k = 2) + \phi_2 \cdot \mathrm{I}(k' = 2) \\ &\quad + \phi_{22} \cdot \mathrm{I}(k = 2, k' = 2), \end{aligned} \tag{4.10}$$

where the association parameters are specified as $\phi = \log(3)$, $\phi_2 = \log(2/3)$, and $\phi_{22} = 2\log(3/2)$.

We generate the surrogate response S_{ij} under generalized logit models conditional on $Y_{ij} = k$, which are given by

$$\log\left(\frac{\tau_{ijk,l}}{\tau_{ijk,0}}\right) = \gamma_{kl}, \quad l = 1, 2.$$

Similarly, the surrogate categorical covariate W_{ij} is generated by, conditional on $X_{ij} = q$,

$$\log\left(\frac{\pi_{ijq,r}}{\pi_{ijq,0}}\right) = \varphi_{qr}, \quad r = 1, 2.$$

Three scenarios for the misclassification parameters are considered:

(i) $\gamma_{01} = \log(0.04/0.95)$, $\gamma_{02} = \log(0.01/0.95)$, $\gamma_{11} = \log(0.95/0.03)$, $\gamma_{12} = \log(0.02/0.03)$, $\gamma_{21} = \log(0.04/0.01)$, $\gamma_{22} = \log(0.95/0.01)$;
$\varphi_{01} = \log(0.04/0.95)$, $\varphi_{02} = \log(0.01/0.95)$, $\varphi_{11} = \log(0.95/0.03)$, $\varphi_{12} = \log(0.02/0.03)$, $\varphi_{21} = \log(0.04/0.01)$, and $\varphi_{22} = \log(0.95/0.01)$;

(ii) $\gamma_{01} = \log(0.08/0.90)$, $\gamma_{02} = \log(0.02/0.90)$, $\gamma_{11} = \log(0.90/0.06)$, $\gamma_{12} = \log(0.04/0.06)$, $\gamma_{21} = \log(0.08/0.02)$, $\gamma_{22} = \log(0.90/0.02)$;
$\varphi_{01} = \log(0.08/0.90)$, $\varphi_{02} = \log(0.02/0.90)$, $\varphi_{11} = \log(0.90/0.06)$, $\varphi_{12} = \log(0.04/0.06)$, $\varphi_{21} = \log(0.08/0.02)$, and $\varphi_{22} = \log(0.90/0.02)$;

(iii) $\gamma_{01} = \log(0.15/0.80)$, $\gamma_{02} = \log(0.05/0.80)$, $\gamma_{11} = \log(0.80/0.15)$, $\gamma_{12} = \log(0.05/0.15)$, $\gamma_{21} = \log(0.15/0.05)$, $\gamma_{22} = \log(0.80/0.05)$;
$\varphi_{01} = \log(0.15/0.80)$, $\varphi_{02} = \log(0.05/0.80)$, $\varphi_{11} = \log(0.80/0.15)$, $\varphi_{12} = \log(0.05/0.15)$, $\varphi_{21} = \log(0.15/0.05)$, and $\varphi_{22} = \log(0.80/0.05)$.

In scenario (i), the misclassification rate is about 5% for all categories of both the response variable and the covariate. The overall misclassification rate is increased to 10% and 20% in scenarios (ii) and (iii), respectively. The probability of a misclassification between non-adjacent categories is smaller than that for an adjacent misclassification. For example, an observation in category 0 has probabilities 0.15 and 0.05 of being recorded as category 1 and 2 in scenario (iii), respectively.

For the case of unknown $\boldsymbol{\eta}$, 30% of the observations are randomly selected into a validation subsample, in which the true categories for both the response variable and the covariate are obtained. A total of 2000 simulation runs are carried out for each single parameter configuration in each of the three scenarios.

4.5.2 Results of simulations

We present the results from the naive analysis, the proposed method with both known and unknown $\boldsymbol{\eta}$. Table 4.1 shows the results for simulation under setting M1. The four columns under each approach are the percent relative bias (%RB), empirical variance (EV), average of model-based variance (AMV), and coverage rate of the 95% confidence intervals (CP). One can see that the biases in the estimates from the naive approach are non-ignorable even under low misclassification rates. As the misclassification rates increase, the biases increase and the coverage rates decrease. The proposed method performs very well in reducing bias in the estimates of both mean and association parameters, and the coverage rates are close to the nominal value of 95%. As the misclassification rates increase, the variance associated with each estimator also increases. Similar patterns are observed for cases where $\boldsymbol{\eta}$ is unknown and is estimated from a validation subsample. We do not report in the table the estimates of the misclassification parameters due to the size of the table. The biases in the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ are ignorable for scenarios (ii) and (iii) but not scenario (i). The biases in the estimates of $\boldsymbol{\theta}$ for scenario (i) with low misclassification rates are a bit larger than those for scenarios (ii) and (iii), because some rare misclassification events (e.g., misclassification from the highest level to the lowest level, or vise versa) may not be present in the validation subsample. The convergence rate of the algorithm for scenario (i) is about 1860/2000, while those for scenarios (ii) and (iii) are about 1995/2000.

Table 4.2 shows the results for simulation under setting M2 with association structure given by (4.10). The estimates from the naive approach are all downward biased, particularly in β_1 and β_2 for the effects of the misclassified covariate as well as the association parameters. For the case of known $\boldsymbol{\eta}$, the proposed method performs well in correcting the induced biases by misclassification. We also observe that the associated variances increase as the misclassification rates increase. For the case of unknown $\boldsymbol{\eta}$, results are similar to those in Table 4.1. The biases are slightly larger for scenario (i). The proposed method performs reasonably well for scenarios (ii) and (iii) in terms of consistency. The variance estimators for the estimates of association parameters in scenario (iii) are upward biased, which leads to coverage rates slightly above the nominal value of 95%.

4.6 Application

4.6.1 Framingham Heart Study

We apply the proposed method to a data set containing $n = 1615$ male subjects aged 31-65 from the Framingham Heart Study (e.g., Carroll et al., 2006, p. 112). The cohort has been followed for morbidity and mortality, and participants have continued to return to the study every two years for a detailed medical history, physical examination, and laboratory tests. The data set includes exams #2 and #3. Our clinical interest is to study the relationship between blood pressure levels and its risk factors, as well as to understand the trend of the influence of the risk factors over time. In this example, the high blood pressure(HBP) status for subject i at time j ($i = 1, \ldots, n$, $j = 1$ for exam #2 and $j = 2$ for exam #3), is an ordinal variable with three levels: non-HBP, HBP Stage 1, and HBP Stage 2, which correspond to systolic pressure < 140 mmHg, $140 - 159$ mmHg, and ≥ 160 mmHg, respectively. Potential set of risk factors included in this study are serum cholesterol level (see, e.g., Ferrara et al., 2002), age, and smoking status. We consider cholesterol level as a categorical variable that can be normal, border-line, and hypercholesterolemia, which correspond to cholesterol measurement < 200 mg/dL, $200 - 239$ mg/dL, and ≥ 240 mg/dL, respectively (e.g., Grundy, 2000; Natarajan et al., 2002). The proportional

Table 4.1: Simulation results for Model 1 (2000 simulations)

	Naive method				Proposed method							
					known $\boldsymbol{\eta}$				unknown $\boldsymbol{\eta}$			
	%RB	EV	AMV	CP	%RB	EV	AMV	CP	%RB	EV	AMV	CP
	scenario (i): 5% misclassification rate											
β_{01}	4.14	0.006	0.007	0.943	0.19	0.007	0.008	0.956	-0.59	0.008	0.008	0.945
β_{02}	1.99	0.006	0.007	0.958	0.18	0.007	0.008	0.960	1.38	0.008	0.008	0.952
β_1	-13.20	0.006	0.006	0.764	-0.01	0.008	0.008	0.950	-5.74	0.020	0.013	0.932
β_2	-11.11	0.008	0.008	0.698	-0.01	0.011	0.010	0.948	1.58	0.015	0.010	0.921
β_3	-6.75	0.008	0.007	0.908	-0.18	0.009	0.008	0.946	0.26	0.009	0.008	0.945
β_4	-5.82	0.005	0.005	0.944	0.73	0.006	0.006	0.946	1.42	0.006	0.006	0.951
β_5	-6.49	0.005	0.005	0.912	0.20	0.006	0.006	0.954	0.65	0.006	0.006	0.957
ϕ	-13.07	0.010	0.010	0.672	0.20	0.014	0.015	0.956	1.10	0.016	0.016	0.957
	scenario (ii): 10% misclassification rate											
β_{01}	8.15	0.007	0.007	0.898	0.20	0.009	0.009	0.958	-0.12	0.009	0.010	0.966
β_{02}	4.19	0.006	0.007	0.942	0.40	0.009	0.009	0.958	0.68	0.010	0.010	0.958
β_1	-25.01	0.005	0.006	0.363	0.20	0.011	0.011	0.960	-0.73	0.012	0.011	0.957
β_2	-21.11	0.008	0.008	0.260	0.46	0.014	0.013	0.944	0.38	0.013	0.012	0.942
β_3	-12.71	0.007	0.007	0.812	0.23	0.009	0.009	0.954	0.16	0.009	0.009	0.947
β_4	-12.28	0.005	0.005	0.922	0.50	0.007	0.007	0.954	0.47	0.006	0.007	0.953
β_5	-12.72	0.005	0.006	0.788	0.42	0.007	0.007	0.959	0.35	0.006	0.007	0.963
ϕ	-24.69	0.009	0.009	0.188	0.52	0.022	0.022	0.948	0.61	0.020	0.022	0.959
	scenario (iii): 20% misclassification rate											
β_{01}	9.54	0.007	0.007	0.868	0.34	0.015	0.015	0.953	0.47	0.015	0.016	0.961
β_{02}	11.15	0.006	0.007	0.850	0.25	0.015	0.015	0.951	0.61	0.015	0.016	0.966
β_1	-48.17	0.006	0.006	0.004	0.68	0.030	0.029	0.942	1.07	0.025	0.023	0.939
β_2	-43.22	0.008	0.008	0.001	0.26	0.027	0.027	0.954	0.68	0.023	0.023	0.958
β_3	-26.32	0.006	0.006	0.372	0.42	0.012	0.012	0.952	0.52	0.011	0.011	0.956
β_4	-25.19	0.006	0.006	0.850	1.32	0.011	0.011	0.948	1.49	0.009	0.010	0.950
β_5	-26.58	0.006	0.006	0.319	0.42	0.011	0.012	0.948	0.62	0.010	0.011	0.954
ϕ	-45.03	0.008	0.008	0.001	1.19	0.054	0.052	0.942	0.90	0.040	0.048	0.964

Table 4.2: Simulation results for Model 2 (2000 simulations)

	Naive method				Proposed method							
					known $\boldsymbol{\eta}$				unknown $\boldsymbol{\eta}$			
	%RB	EV	AMV	CP	%RB	EV	AMV	CP	%RB	EV	AMV	CP
	scenario (i): 5% misclassification rate											
β_{01}	4.22	0.007	0.007	0.926	0.26	0.008	0.008	0.942	-0.50	0.009	0.008	0.944
β_{02}	1.93	0.007	0.007	0.944	0.16	0.008	0.008	0.945	1.46	0.010	0.008	0.937
β_1	-12.92	0.006	0.006	0.766	0.35	0.008	0.008	0.950	-5.65	0.021	0.013	0.932
β_2	-10.78	0.008	0.008	0.726	0.36	0.010	0.010	0.948	1.96	0.014	0.011	0.940
β_3	-6.06	0.007	0.007	0.912	0.56	0.008	0.008	0.944	0.72	0.008	0.008	0.946
β_4	-6.56	0.006	0.005	0.935	-0.07	0.006	0.006	0.942	0.77	0.006	0.006	0.945
β_5	-6.50	0.006	0.006	0.901	0.22	0.007	0.007	0.948	0.57	0.006	0.007	0.959
ϕ	-15.62	0.011	0.011	0.629	0.01	0.018	0.018	0.950	0.90	0.021	0.024	0.958
ϕ_2	-23.08	0.007	0.007	0.816	-0.24	0.011	0.012	0.952	0.73	0.011	0.014	0.960
ϕ_{22}	-20.44	0.013	0.013	0.682	-0.28	0.020	0.020	0.945	1.06	0.020	0.023	0.961
	scenario (ii): 10% misclassification rate											
β_{01}	8.57	0.007	0.007	0.893	0.67	0.009	0.009	0.948	0.47	0.009	0.010	0.951
β_{02}	3.98	0.007	0.007	0.938	0.22	0.010	0.009	0.944	0.37	0.010	0.010	0.942
β_1	-24.74	0.006	0.006	0.384	0.61	0.012	0.012	0.952	-0.18	0.013	0.011	0.948
β_2	-21.00	0.008	0.008	0.266	0.61	0.013	0.014	0.958	0.67	0.012	0.012	0.956
β_3	-12.21	0.007	0.007	0.802	0.82	0.009	0.009	0.942	0.80	0.009	0.008	0.943
β_4	-12.49	0.005	0.006	0.930	0.30	0.007	0.007	0.952	0.08	0.007	0.007	0.953
β_5	-12.49	0.006	0.006	0.779	0.74	0.008	0.008	0.950	0.51	0.007	0.007	0.956
ϕ	-28.82	0.011	0.010	0.134	0.35	0.026	0.025	0.948	0.29	0.023	0.025	0.957
ϕ_2	-41.04	0.007	0.007	0.481	0.77	0.018	0.018	0.944	0.53	0.015	0.017	0.961
ϕ_{22}	-36.65	0.013	0.012	0.258	0.48	0.031	0.030	0.943	0.35	0.026	0.030	0.966
	scenario (iii): 20% misclassification rate											
β_{01}	9.34	0.007	0.007	0.875	0.34	0.015	0.015	0.950	0.86	0.015	0.016	0.970
β_{02}	11.68	0.007	0.007	0.832	0.81	0.015	0.015	0.946	0.12	0.015	0.016	0.962
β_1	-47.59	0.006	0.006	0.009	1.03	0.030	0.029	0.951	0.76	0.024	0.023	0.948
β_2	-42.39	0.008	0.008	0.002	1.21	0.027	0.027	0.951	0.94	0.023	0.023	0.950
β_3	-26.13	0.006	0.006	0.350	0.78	0.012	0.012	0.946	0.95	0.011	0.011	0.950
β_4	-25.70	0.006	0.006	0.838	0.41	0.012	0.011	0.949	0.66	0.010	0.010	0.956
β_5	-26.48	0.006	0.006	0.348	0.70	0.013	0.012	0.939	0.86	0.011	0.011	0.953
ϕ	-53.34	0.008	0.009	0.000	1.69	0.074	0.069	0.950	1.72	0.053	0.067	0.971
ϕ_2	-72.70	0.007	0.006	0.044	0.35	0.053	0.048	0.944	1.27	0.039	0.050	0.973
ϕ_{22}	-59.97	0.011	0.011	0.004	0.44	0.076	0.075	0.948	1.49	0.057	0.080	0.978

odds models for the cumulative probabilities of the ordinal response are given by

$$\begin{aligned}\text{logit } \lambda_{ijk} &= \beta_{0k} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \beta_4 x_{ij4} + \beta_5 x_{ij5},\\ &\qquad\qquad\qquad i = 1, \ldots, n,\ j = 1, 2,\ k = 1, 2,\end{aligned}$$

where x_{ij1} is 1 if the cholesterol level for subject i at time j is border-line high and 0 otherwise, x_{ij2} is 1 if the cholesterol level is high and 0 otherwise, x_{ij3} is the age, x_{ij4} is 1 if the subject is a smoker and 0 otherwise, and x_{ij5} is 1 for exam #3 and 0 otherwise. We consider two association models given by (4.9) and (4.10).

Two measurements of systolic blood pressure by different examiners were obtained at each of the two exams. Because systolic blood pressure changes over time, a single measurement does not reflect the patient's long-time average of systolic blood pressure, which is usually regarded as the true measurement. Therefore, misclassification may be present in the HBP variable obtained from categorizing a single SBP measurement. Similarly, the observed cholesterol categories may also contain misclassifications, as the cholesterol serum measurements are subject to error. Since the data set does not contain a validation subsample, we conduct sensitivity analysis by assuming different misclassification rates described in scenarios (i) and (ii) in the Section 4.5.1.

Table 4.3 reports the results under the model with association structure given by (4.9). The proposed method is compared to the naive approach which ignores error in the SBP measurements. We report three analyses: the first one uses the first replicates in the exams, the second one uses the second replicates, and in the third one, the response category is obtained from the average of the two SBP replicates. One can see that the trend of the covariate effects is similar for both the naive approach and the proposed method. The estimates of mean and association parameters are boosted after accounting for misclassification, and the increments are getting larger as higher misclassification rates are assumed. This is consistent with the findings in the simulation studies in previous section. Note that the p-values from hypotheses testing for significant cholesterol effects (i.e., β_1 and β_2) are getting larger, since the standard errors are getting larger after accounting for misclassifications with higher rates.

Table 4.4 displays the results under the model with association structure given by (4.10). The patterns of the parameter estimates and their associated standard errors for the naive analysis and the proposed method are similar to those in Table 4.3. The association parameters ϕ_2 and ϕ_{22}, however, are not statistically significant at the 5% level in most cases. This suggests that a single global odds ratio is sufficient to describe the dependence of the response levels in the two exams.

4.7 Discussion

In this chapter we have proposed a marginal method for analysis of longitudinal or clustered ordinal data with response and covariate misclassifications. We constructed unbiased estimating functions of both the mean and the association parameters. The estimation bias induced by misclassifications can be reduced by solving these estimating equations.

The proposed method yields consistent estimators for the response parameters given the true misclassification parameters. Our simulation studies illustrate good performance of the proposed method under a variety of parameter configurations. For cases where misclassification parameters are unknown and a validation subsample is available, a two-stage estimation procedure is proposed. When the validation subsample is small and misclassification rate is low, estimates of the nuisance parameters associated with the misclassification process may have very large variation. Without validation data or replicate measures, one may conduct sensitivity analysis and see if the conclusion, e.g., significant effect of a covariate, changes for different misclassification settings.

Our future research work includes developing methods that can handle situations where replicates of the misclassified variables are available instead of a validation subsample. In this case, the two-stage estimation procedure can not be used. Joint estimation of the misclassification parameters and the response parameters may be

Table 4.3: Results of naive analysis and sensitivity analysis for Framingham data under Model 1

	1st replicates			2nd replicates			averaged replicates		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Naive method									
β_{01}	-3.315	0.299	< 0.001	-3.901	0.328	< 0.001	-3.869	0.326	< 0.001
β_{02}	-4.795	0.300	< 0.001	-5.256	0.329	< 0.001	-5.324	0.328	< 0.001
β_1	0.004	0.095	0.965	0.169	0.106	0.110	0.123	0.098	0.210
β_2	0.271	0.105	0.010	0.359	0.114	0.002	0.279	0.109	0.010
β_3	0.054	0.006	< 0.001	0.060	0.006	< 0.001	0.061	0.006	< 0.001
β_4	-0.139	0.115	0.228	-0.208	0.122	0.089	-0.236	0.122	0.052
β_5	-0.117	0.051	0.021	-0.109	0.053	0.038	-0.053	0.050	0.286
ϕ	2.670	0.162	< 0.001	2.717	0.170	< 0.001	2.991	0.174	< 0.001
Proposed method									
	scenario (i): 5% misclassification rate								
β_{01}	-3.827	0.361	< 0.001	-4.697	0.422	< 0.001	-4.637	0.416	< 0.001
β_{02}	-5.253	0.362	< 0.001	-5.968	0.422	< 0.001	-6.026	0.418	< 0.001
β_1	-0.019	0.133	0.888	0.232	0.159	0.144	0.168	0.145	0.245
β_2	0.331	0.145	0.022	0.470	0.169	0.005	0.375	0.160	0.019
β_3	0.061	0.007	< 0.001	0.071	0.008	< 0.001	0.072	0.007	< 0.001
β_4	-0.150	0.132	0.256	-0.241	0.144	0.095	-0.276	0.143	0.053
β_5	-0.137	0.058	0.019	-0.124	0.062	0.047	-0.062	0.059	0.289
ϕ	3.303	0.240	< 0.001	3.454	0.266	< 0.001	3.786	0.273	< 0.001
	scenario (ii): 10% misclassification rate								
β_{01}	-4.586	0.477	< 0.001	-6.116	0.646	< 0.001	-5.980	0.622	< 0.001
β_{02}	-5.930	0.478	< 0.001	-7.251	0.642	< 0.001	-7.263	0.621	< 0.001
β_1	-0.046	0.214	0.828	0.371	0.289	0.200	0.276	0.256	0.281
β_2	0.456	0.239	0.056	0.692	0.314	0.027	0.579	0.293	0.048
β_3	0.072	0.008	< 0.001	0.092	0.011	< 0.001	0.093	0.011	< 0.001
β_4	-0.163	0.157	0.300	-0.298	0.183	0.103	-0.344	0.178	0.054
β_5	-0.167	0.070	0.016	-0.136	0.077	0.078	-0.075	0.072	0.301
ϕ	4.113	0.425	< 0.001	4.533	0.542	< 0.001	4.910	0.557	< 0.001

Table 4.4: Results of naive analysis and sensitivity analysis for Framingham data under Model 2

	1st replicates			2nd replicates			averaged replicates		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
Naive method									
β_{01}	-3.288	0.298	< 0.001	-3.851	0.326	< 0.001	-3.828	0.326	< 0.001
β_{02}	-4.766	0.300	< 0.001	-5.204	0.326	< 0.001	-5.280	0.327	< 0.001
β_1	0.030	0.094	0.747	0.177	0.105	0.091	0.132	0.097	0.172
β_2	0.298	0.104	0.004	0.374	0.113	0.001	0.293	0.108	0.007
β_3	0.052	0.006	< 0.001	0.059	0.006	< 0.001	0.060	0.006	< 0.001
β_4	-0.131	0.115	0.251	-0.202	0.122	0.097	-0.231	0.121	0.056
β_5	-0.108	0.050	0.030	-0.113	0.052	0.029	-0.050	0.049	0.311
ϕ	2.167	0.135	< 0.001	2.353	0.145	< 0.001	2.578	0.150	< 0.001
ϕ_2	0.397	0.186	0.032	0.417	0.194	0.032	0.445	0.238	0.061
ϕ_{22}	0.097	0.279	0.728	-0.086	0.289	0.767	-0.095	0.378	0.802
Proposed method									
	scenario (i): 5% misclassification rate								
β_{01}	-3.816	0.361	< 0.001	-4.673	0.421	< 0.001	-4.601	0.416	< 0.001
β_{02}	-5.239	0.362	< 0.001	-5.943	0.421	< 0.001	-5.990	0.418	< 0.001
β_1	0.011	0.129	0.929	0.238	0.157	0.129	0.172	0.143	0.229
β_2	0.358	0.141	0.011	0.478	0.167	0.004	0.379	0.159	0.017
β_3	0.060	0.007	< 0.001	0.071	0.008	< 0.001	0.071	0.007	< 0.001
β_4	-0.145	0.132	0.271	-0.238	0.144	0.099	-0.273	0.143	0.055
β_5	-0.127	0.057	0.027	-0.126	0.062	0.042	-0.058	0.058	0.314
ϕ	2.834	0.211	< 0.001	3.268	0.256	< 0.001	3.621	0.279	< 0.001
ϕ_2	0.328	0.340	0.334	0.250	0.385	0.516	0.494	0.668	0.460
ϕ_{22}	0.052	0.522	0.920	-0.154	0.602	0.798	-0.582	1.172	0.619
	scenario (ii): 10% misclassification rate								
β_{01}	-4.607	0.478	< 0.001	-6.054	0.648	< 0.001	-5.922	0.628	< 0.001
β_{02}	-5.951	0.478	< 0.001	-7.194	0.644	< 0.001	-7.219	0.627	< 0.001
β_1	-0.048	0.215	0.822	0.279	0.317	0.378	0.207	0.290	0.475
β_2	0.461	0.237	0.052	0.595	0.349	0.089	0.517	0.339	0.127
β_3	0.073	0.008	< 0.001	0.093	0.011	< 0.001	0.093	0.011	< 0.001
β_4	-0.165	0.157	0.294	-0.302	0.182	0.098	-0.367	0.179	0.040
β_5	-0.174	0.070	0.014	-0.129	0.082	0.115	-0.092	0.078	0.240
ϕ	4.084	0.492	< 0.001	5.736	1.232	< 0.001	6.689	1.749	< 0.001
ϕ_2	-0.305	0.669	0.648	-0.994	1.191	0.404	0.102	0.756	0.893
ϕ_{22}	0.516	1.056	0.625	0.474	1.964	0.809	-2.148	0.274	< 0.001

feasible by simultaneously solving a set of estimating equations. Unlike correlated binary data with replicates, however, the minimum number of replicates required for model identifiability remains unclear for longitudinal ordinal data.

4.8 Technical Details

4.8.1 Extension of the results of Akazawa et. al (1998)

Akazawa et. al (1998) proved that, for an arbitrary real-valued function $f(\mathbf{X}_{ij})$, an unbiased surrogate can be given by $f^*(\mathbf{W}_{ij}) = \sum_{q=0}^{K^x} f(\mathbf{e}_q) X^*_{ijq}$. Here we generalize the result to multivariate case.

Lemma 1 Let $\mathbf{f}(\mathbf{X}_{ij}) = \{f_1(\mathbf{X}_{ij}), \ldots, f_p(\mathbf{X}_{ij})\}^{\mathrm{T}}$ be an arbitrary p-dimensional vector of real-valued functions of $\mathbf{X}_{ij}$. Define

$$\mathbf{f}^*(\mathbf{W}_{ij}) = \sum_{q=0}^{K^x} \mathbf{f}(\mathbf{e}_q) X^*_{ijq}.$$

Then $\mathrm{E}\left[\mathbf{f}^*(\mathbf{W}_{ij}) | \mathbf{X}_{ij}\right] = \mathbf{f}(\mathbf{X}_{ij})$.

Proof The lth component in $\mathbf{f}^*$ is given by $f_l^*(\mathbf{W}_{ij}) = \sum_{q=0}^{K^x} f_l(\mathbf{e}_q) X^*_{ijq}$. From Akazawa et. al (1998), $\mathrm{E}\left[f_l^*(\mathbf{W}_{ij}) | \mathbf{X}_{ij}\right] = f_l(\mathbf{X}_{ij})$. Therefore, the result holds.

Theorem 1 Let $\mathbf{f}(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i})$ be a vector of real-valued functions of $(\mathbf{X}_{i1}^{\mathrm{T}}, \ldots, \mathbf{X}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Assume that the misclassifications processes for $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i}$ are independent of each other. Define

$$\mathbf{f}^*(\mathbf{W}_{i1}, \ldots, \mathbf{W}_{im_i}) = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \mathbf{f}(\mathbf{e}_{q_1}, \ldots, \mathbf{e}_{q_{m_i}}) X^*_{i1q_1} \cdots X^*_{im_iq_{m_i}}.$$

Then

$$\mathrm{E}\left[\mathbf{f}^*(\mathbf{W}_{i1}, \ldots, \mathbf{W}_{im_i}) | \mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i}\right] = \mathbf{f}(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i}). \tag{4.11}$$

Proof We first look at the most inner expectation with respect to $\mathbf{W}_{i1}$ conditional on $\mathbf{X}_{i1}$. Under the assumption of independent misclassification, we have

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{W}_{i1}|\mathbf{X}_{i1}}\left[\mathbf{f}^*(\mathbf{W}_{i1},\ldots,\mathbf{W}_{im_i})\right] \\
&= \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_2=0}^{K^x}\mathrm{E}_{\mathbf{W}_{i1}|\mathbf{X}_{i1}}\left[\sum_{q_1=0}^{K^x}\mathbf{f}(\mathbf{e}_{q_1},\mathbf{e}_{q_2},\ldots,\mathbf{e}_{q_{m_i}})X^*_{i1q_1}\right]X^*_{i2q_2}\cdots X^*_{im_iq_{m_i}} \\
&= \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_2=0}^{K^x}\mathbf{f}(\mathbf{X}_{i1},\mathbf{e}_{q_2},\ldots,\mathbf{e}_{q_{m_i}})X^*_{i2q_2}\cdots X^*_{im_iq_{m_i}}.
\end{aligned}
$$

Therefore, the result (4.11) holds after applying all expectations with respect to $\mathbf{W}_{i1}$, $\ldots,\mathbf{W}_{im_i}$ conditional on $\mathbf{X}_{i1},\ldots,\mathbf{X}_{im_i}$.

4.8.2 Explicit form for $\partial\tilde{\mathbf{U}}_i/\partial\boldsymbol{\varphi}^{\mathrm{T}}$

Note that $\boldsymbol{\varphi}$ is involved in the constructed surrogates $\tilde{X}_{ijq_j}$ only. Therefore,

$$
\begin{aligned}
\frac{\partial\tilde{\mathbf{U}}_i}{\partial\boldsymbol{\varphi}^{\mathrm{T}}} &= \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_1=0}^{K^x}\mathbf{U}_i(\boldsymbol{\theta};\tilde{\mathbf{Y}}_i,(\mathbf{e}^{\mathrm{T}}_{q_1},\ldots,\mathbf{e}^{\mathrm{T}}_{q_{m_i}})^{\mathrm{T}},\mathbf{Z}_i)\frac{\partial\left(\prod_{j=1}^{m_i}\tilde{X}_{ijq_j}\right)}{\partial\boldsymbol{\varphi}^{\mathrm{T}}} \\
&= \sum_{q_{m_i}=0}^{K^x}\cdots\sum_{q_1=0}^{K^x}\mathbf{U}_i(\boldsymbol{\theta};\tilde{\mathbf{Y}}_i,(\mathbf{e}^{\mathrm{T}}_{q_1},\ldots,\mathbf{e}^{\mathrm{T}}_{q_{m_i}})^{\mathrm{T}},\mathbf{Z}_i)\left\{\sum_{j=1}^{m_i}\left(\prod_{j'\neq j}\tilde{X}_{ij'q_{j'}}\frac{\partial\tilde{X}_{ijq_j}}{\partial\boldsymbol{\varphi}^{\mathrm{T}}}\right)\right\},
\end{aligned}
$$

where

$$
\frac{\partial\tilde{X}_{ijq_j}}{\partial\boldsymbol{\varphi}^{\mathrm{T}}} = \begin{cases}\partial X^*_{ijq_j}/\partial\boldsymbol{\varphi}^{\mathrm{T}} & \text{if } \delta_{ij}=0,\\ \mathbf{0}^{\mathrm{T}} & \text{if } \delta_{ij}=1,\end{cases}
$$

For $q_j \neq 0$, the components in $\partial X^*_{ijq_j}/\partial \boldsymbol{\varphi}^{\mathrm{T}}$ are given by

$$\begin{aligned}
\frac{\partial X^*_{ijq_j}}{\partial \varphi_{qrv}} &= \frac{\partial}{\partial \varphi_{qrv}} \left\{ \mathbf{e}^{\mathrm{T}}_{q_j} \mathbf{G}^{*-1}_{ij} (\mathbf{W}_{ij} - \boldsymbol{\pi}_{ij0}) \right\} \\
&= \mathbf{e}^{\mathrm{T}}_{q_j} \left\{ \left(-\mathbf{G}^{*-1}_{ij}\right) \frac{\mathbf{G}^*_{ij}}{\partial \varphi_{qrv}} \mathbf{G}^{*-1}_{ij} (\mathbf{W}_{ij} - \boldsymbol{\pi}_{ij0}) - \mathbf{G}^{*-1}_{ij} \frac{\partial \boldsymbol{\pi}_{ij0}}{\partial \varphi_{qrv}} \right\}, \\
&\qquad\qquad q = 0, \ldots, K^x,\ r = 1, \ldots, K^x,\ v = 1, \ldots, \dim(\boldsymbol{\varphi}_{qr}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial X^*_{ijq_j}}{\partial \varphi_{qrv}} &= -\, \mathbf{e}^{\mathrm{T}}_{q_j} \mathbf{G}^{*-1}_{ij} \left\{ \frac{\mathbf{G}^*_{ij}}{\partial \varphi_{qrv}} \mathbf{X}^*_{ij} + \frac{\partial \boldsymbol{\pi}_{ij0}}{\partial \varphi_{qrv}} \right\} \\
&= -\, \mathbf{e}^{\mathrm{T}}_{q_j} \mathbf{G}^{*-1}_{ij} \left\{ X^*_{ij1} \left(\frac{\partial \boldsymbol{\pi}_{ij1}}{\partial \varphi_{qrv}} - \frac{\partial \boldsymbol{\pi}_{ij0}}{\partial \varphi_{qrv}} \right) + \ldots \right. \\
&\qquad\qquad \left. + X^*_{ijK^x} \left(\frac{\partial \boldsymbol{\pi}_{ijK^x}}{\partial \varphi_{qrv}} - \frac{\partial \boldsymbol{\pi}_{ij0}}{\partial \varphi_{qrv}} \right) + \frac{\partial \boldsymbol{\pi}_{ij0}}{\partial \varphi_{qrv}} \right\} \\
&= -\, \mathbf{e}^{\mathrm{T}}_{q_j} \mathbf{G}^{*-1}_{ij} \left\{ \sum_{q'=1}^{K^x} X^*_{ijq'} \frac{\partial \boldsymbol{\pi}_{ijq'}}{\partial \varphi_{qrv}} + \left(1 - \sum_{q'=1}^{K^x} X^*_{ijq'} \right) \frac{\partial \boldsymbol{\pi}_{ij0}}{\partial \varphi_{qrv}} \right\} \\
&= -\, \mathbf{e}^{\mathrm{T}}_{q_j} \mathbf{G}^{*-1}_{ij} \sum_{q'=0}^{K^x} X^*_{ijq'} \frac{\partial \boldsymbol{\pi}_{ijq'}}{\partial \varphi_{qrv}}.
\end{aligned}$$

However,

$$\frac{\partial \boldsymbol{\pi}_{ijq'}}{\partial \varphi_{qrv}} = \begin{cases} \mathbf{e}_r \pi_{ijqr}(1 - \pi_{ijqr}) L^x_{ijv} & \text{if } q' = q, \\ \mathbf{0} & \text{if } q' \neq q, \end{cases} \qquad q', q = 0, \ldots, K^x.$$

Therefore,

$$\begin{aligned}
\frac{\partial X^*_{ijq_j}}{\partial \varphi_{qrv}} &= -\, \mathbf{e}^{\mathrm{T}}_{q_j} \mathbf{G}^{*-1}_{ij} \mathbf{e}_r X^*_{ijq} \pi_{ijqr} (1 - \pi_{ijqr}) L^x_{ijv}, \qquad q = 0, \ldots, K^x, \\
&\qquad\qquad r = 1, \ldots, K^x. \qquad (4.12)
\end{aligned}$$

For $q_j = 0$,

$$
\begin{aligned}
\frac{\partial X^*_{ij0}}{\partial \varphi_{qrv}} &= -\sum_{q'=1}^{K^x} \frac{\partial X^*_{ijq'}}{\partial \varphi_{qrv}} \\
&= \mathbf{1}^{\mathrm{T}} \mathbf{G}^{*-1}_{ij} \mathbf{e}_r \pi_{ijqr}(1-\pi_{ijqr}) L^x_{ijv} X^*_{ijq}, \qquad q = 0, \ldots, K^x, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad r = 1, \ldots, K^x.
\end{aligned}
$$

4.8.3 Explicit form for $\partial \tilde{\mathbf{U}}_i / \partial \boldsymbol{\gamma}^{\mathrm{T}}$

Using similar argument, we can derive explicit form for $\partial \tilde{\mathbf{U}}_i / \partial \boldsymbol{\gamma}^{\mathrm{T}}$. We have

$$
\frac{\partial \tilde{\mathbf{U}}_i}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} = \sum_{q_{m_i}=0}^{K^x} \cdots \sum_{q_1=0}^{K^x} \boldsymbol{\Lambda}_{\gamma i}(\boldsymbol{\theta}; \tilde{\mathbf{Y}}_i, (\mathbf{e}^{\mathrm{T}}_{q_1}, \ldots, \mathbf{e}^{\mathrm{T}}_{q_{m_i}})^{\mathrm{T}}, \mathbf{Z}_i) \prod_{j=1}^{m_i} \tilde{X}_{ijq_j},
$$

where

$$
\boldsymbol{\Lambda}_{\gamma i}(\boldsymbol{\theta}; \tilde{\mathbf{Y}}_i, \mathbf{X}_i, \mathbf{Z}_i) = \begin{pmatrix} \mathbf{D}_{1i} \mathbf{V}^{-1}_{1i} \partial \tilde{\mathbf{Y}}_i / \partial \boldsymbol{\gamma}^{\mathrm{T}} \\ \mathbf{D}_{2i} \mathbf{V}^{-1}_{2i} \partial \tilde{\mathbf{C}}_i / \partial \boldsymbol{\gamma}^{\mathrm{T}} \end{pmatrix},
$$

and

$$
\frac{\partial \tilde{Y}_{ijk_j}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} = \begin{cases} \partial Y^*_{ijk_j} / \partial \boldsymbol{\gamma}^{\mathrm{T}} & \text{if } \delta_{ij} = 0, \\ \mathbf{0}^{\mathrm{T}} & \text{if } \delta_{ij} = 1, \end{cases} \qquad k_j = 1, \ldots, K.
$$

Similar to (4.12), the elements in $\partial Y^*_{ijk_j} / \partial \boldsymbol{\gamma}^{\mathrm{T}}$ are given by

$$
\begin{aligned}
\frac{\partial Y^*_{ijk_j}}{\partial \gamma_{klv}} &= -\,\mathbf{e}^{\mathrm{T}}_{k_j} \mathbf{P}^{*-1}_{ij} \sum_{k'=0}^{K} \frac{\partial \boldsymbol{\tau}_{ijk'}}{\partial \gamma_{klv}} Y^*_{ijk'} \\
&= -\,\mathbf{e}^{\mathrm{T}}_{k_j} \mathbf{P}^{*-1}_{ij} \mathbf{e}_l \tau_{ijkl}(1-\tau_{ijkl}) L_{klv} Y^*_{ijk}, \qquad k = 0, \ldots, K, \\
&\qquad\qquad l = 1, \ldots, K, \; v = 1, \ldots, \dim(\boldsymbol{\gamma}_{kl}).
\end{aligned}
$$

Furthermore, for $j \neq j'$, we have

$$\frac{\partial \tilde{C}_{i;jk_j);j'k_{j'})}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \quad = \quad \frac{\partial \tilde{\mathrm{Y}}_{ijk_j}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \tilde{\mathrm{Y}}_{ij'k_{j'}} + \frac{\partial \tilde{\mathrm{Y}}_{ij'k_{j'}}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \tilde{\mathrm{Y}}_{ijk_j}, \qquad k_j, k_{j'} = 1, \ldots, K.$$

Chapter 5

Regression Analysis of Binary Data from Complex Survey with Misclassification in an Ordinal Covariate

5.1 Introduction

Survey sampling has been a widely used method for collecting data. Auxiliary information is often collected and used for improving the estimation of population quantities for particular variables of interest, e.g., using model-calibration estimators (Wu and Sitter, 2001; Wu, 2003). On the other hand, analytic use of survey data has become more and more popular. Studying the relationship between a response variable and auxiliary variables in the target population can be among the primary objectives of a survey. Measurement error, however, arises frequently in data during the course of the collection. Many authors have considered correcting bias in the analysis of contaminated survey data, see, e.g., Ybarra and Lohr (2008) and Gregoire and Salas (2009) for small area estimation and ratio estimation with measurement error in auxiliary information.

Many variables collected from surveys are categorical and ordinal. These variables may be subject to misclassifications when the survey is based on self-report. In this chapter, we consider logistic regression analysis of data from complex surveys with misclassification in ordinal covariates. This problem arises often in health surveys, in which the objective is to investigate the association of some binary chronic conditions with categorical exposures that are collected with error. We first formulate the models for the response process and the misclassification process. We then discuss estimation and inference methods for the regression coefficients associated with the risk factors. An expected score approach is proposed for simultaneously accounting for misclassification and complex survey features. Results from a simulation study are reported to show the good performance of the proposed method. Finally, we apply the method to a data set from the Canadian Community Health Survey (CCHS) cycle 3.1.

5.2 Model Formulation

5.2.1 Response model

Different from Chapters $2-4$, in this chapter we focus the discussion on a univariate binary response variable. The interest of the study is to investigate the effects of certain risk factors on particular binary outcomes, such as the presence of any heart disease. Suppose a finite population consists of N individuals. Let Y_i denote the binary response variable for individual i ($i = 1, \dots, N$) such that $Y_i = 1$ if the outcome is present and $Y_i = 0$ otherwise. Let X_i be a $(K+1)$-level ordinal variable that takes values at $0, 1, \dots, K$ and is subject to misclassification. Let $X_{i0}, \dots, X_{iK}$ be indicators such that $X_{ik} = 1$ if $X_i = k$, and $X_{ik} = 0$ otherwise. Without loss of generality, we treat the lowest category as the reference. Therefore, the vector $\mathbf{X}_i = (X_{i1}, \dots, X_{iK})^{\mathrm{T}}$ is used to represent the original categorical X_i. Let $\mathbf{Z}_i$ be a vector of precisely measured covariates, including an intercept and possibly indicator variables for categorical covariates.

We assume that the finite population is generated from a superpopulation model

ζ. Let $\mu_i = \mathrm{E}_\zeta[Y_i|\mathbf{X}_i, \mathbf{Z}_i]$ be the conditional mean of Y_i under the superpopulation model. A logistic model is given by

$$\text{logit } \mu_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_x + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_z,$$

where $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_z$ are vectors of regression coefficients associated with the effects of $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^{\mathrm{T}}, \boldsymbol{\beta}_z^{\mathrm{T}})^{\mathrm{T}}$. If data on all N individuals were available, the population parameter $\boldsymbol{\beta}_N$ is then defined as the maximizer of the finite population log-likelihood

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) \\ &= \sum_{i=1}^{N} \left\{ Y_i \log(\frac{\mu_i}{1-\mu_i}) + (1-Y_i)\log(1-\mu_i) \right\}. \end{aligned}$$

Furthermore, $\boldsymbol{\beta}_N$ can be viewed as an estimate of the model parameter $\boldsymbol{\beta}$. It is equivalent to solving equations

$$\sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{Y_i - \mu_i}{V_i} = \mathbf{0},$$

where $V_i = \mu_i(1-\mu_i)$ is the conditional variance of Y_i under ζ.

Suppose a sample s consisting of n individuals is drawn from the finite population using a complex survey design p. Let d_i be the survey weights for individual i. The finite population parameter $\boldsymbol{\beta}_N$ and superpopulation model parameter $\boldsymbol{\beta}$ can be simultaneously estimated by maximizing a pseudo-likelihood $\sum_{i\in s} d_i \ell_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$. It can be shown that the resulting estimating function is unbiased for the finite population estimating function under survey design p, i.e.,

$$\mathrm{E}_p\left[\sum_{i\in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)\right] = \sum_{i=1}^{N} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i), \tag{5.1}$$

where E_p denotes expectation taken with respect to the sampling scheme.

5.2.2 Misclassification model

The simplest example is the misclassification of a binary variable, which involves only two error parameters. Misclassification of a categorical covariate with more than two levels are commonly seen in survey sampling, especially for measurements based on self-reporting. It is reasonable to assume that the misclassification of an ordinal covariate only occurs between adjacent categories (e.g., BMI categories, income levels). Furthermore, the misclassification process may depend on other covariates.

Let W_i be the observed surrogate for X_i. Correspondingly, let $W_{il} = 1$ if $W_i = l$, and $W_{il} = 0$ otherwise, $l = 0, \ldots, K$. Let $\pi_{ik,l} = \Pr(W_i = l | X_i = k, \mathbf{Z}_i)$ be the probability that the observed category is l given the true category is k for individual i, $k, l = 0, \ldots, K$. Based on the assumption of adjacent misclassifications, we have $\pi_{ik,l} = 0$ for $|k - l| \geq 2$. The probability of correctly classifying X_i into category k is then given by

$$\pi_{ik,k} = 1 - \pi_{ik,k-1}\mathrm{I}(k > 0) - \pi_{ik,k+1}\mathrm{I}(k < K),$$

where $\mathrm{I}(\cdot)$ is the indicator function.

We assume that the misclassification process is characterized by generalized (or multinomial) logit models (Pfeffermann et al., 1998)

$$\begin{aligned}
\log\left(\frac{\pi_{ik,k-1}}{\pi_{ik,k}}\right) &= \mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1}, \qquad k = 1, \ldots, K, \\
\log\left(\frac{\pi_{ik,k+1}}{\pi_{ik,k}}\right) &= \mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k+1}, \qquad k = 0, \ldots, K-1,
\end{aligned}$$

where $\mathbf{L}_i$ is a set of covariates (usually part of $\mathbf{Z}_i$) associated with the misclassification process, and $\boldsymbol{\varphi}_{k,k-1}$ and $\boldsymbol{\varphi}_{k,k+1}$ are vectors of regression parameters in the logit models for misclassification to a lower level and misclassification to a higher level, respectively. Let $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_{01}^{\mathrm{T}}, \ldots, \boldsymbol{\varphi}_{K,K-1}^{\mathrm{T}})^{\mathrm{T}}$. Therefore, the probability of misclassifying an observation into a lower category is given by

$$\pi_{ik,k-1} = \frac{\exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1})}{1 + \exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k-1}) + \exp(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\varphi}_{k,k+1})}, \quad k = 1, \ldots, K,$$

and the probability of misclassifying an observation into a higher category is given by

$$\pi_{ik.k+1} \quad = \quad \frac{\exp(\mathbf{L}_i^{\mathrm{T}} \boldsymbol{\varphi}_{k,k+1})}{1 + \exp(\mathbf{L}_i^{\mathrm{T}} \boldsymbol{\varphi}_{k,k-1}) + \exp(\mathbf{L}_i^{\mathrm{T}} \boldsymbol{\varphi}_{k,k+1})}, \quad k = 0, \ldots, K-1.$$

When both K and the number of covariates in $\mathbf{L}_i$ are large, the dimension of nuisance parameter vector $\boldsymbol{\varphi}$ can be very high. In some extreme cases, the misclassification process may be homogeneous, i.e., the probability of misclassifying the observation into the lower or higher category is consistent for all categories.

5.2.3 Model for the ordinal covariate

For covariate measurement error problems, the literatures distinguish structural modeling, which hypothesizes a distribution for the error-prone covariate, and functional modeling, which does not make any parametric assumptions for the marginal behavior of the covariate. Functional modeling may lose some efficiency. Often, the behavior of the precisely measured $\mathbf{Z}_i$ is not of interest, and its distribution can be left unspecified. When the behavior of the error-prone covariate is of interest (e.g., percentage distribution of BMI), however, it is convenient to hypothesize a marginal distribution for X_i.

For ordinal variables, cumulative probabilities are often used as alternatives to marginals. Let $\lambda_{ik} = \Pr(X_i \geq k | \mathbf{Z}_i)$, $k = 1, , \ldots, K$. The proportional odds models can be employed to characterize the distribution of X_i conditional on $\mathbf{Z}_i$ (e.g., Agresti, 2002). The kth model is given by

$$\text{logit } \lambda_{ik} = \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\alpha}_k, \quad k = 1, \ldots, K,$$

where $\boldsymbol{\alpha}_k = (\alpha_{0k}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$, α_{0k} is the intercept term in the kth logit model, and $\boldsymbol{\psi}$ is a vector of regression coefficients associated with $\mathbf{Z}_i$ and is common for all k. Let $\boldsymbol{\alpha} = (\alpha_{01}, \ldots, \alpha_{0K}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$ be a vector of all regression parameters associated with the distribution of X_i.

Similarly, the dimension of $\boldsymbol{\alpha}$ mainly depends on K and the dimension of $\mathbf{Z}_i$. When X_i and $\mathbf{Z}_i$ are independent, we only need to specify the marginal distribution of X_i, which is given by

$$\Pr(X_i = k) = \alpha_k, \quad k = 0, \ldots, K,$$

where $\sum_{k=0}^{K} \alpha_k = 1$. Under this assumption, the computational burden is dramatically reduced.

5.3 Parametric Estimation

5.3.1 Expected score for estimation of $\boldsymbol{\beta}$

If data were free of measurement error, $\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is unbiased under the sampling scheme and the superpopulation model. In the presence of misclassification, however, $\mathbf{X}_i$ is not available. Let $\mathbf{W}_i = (W_{i1}, \ldots, W_{iK})^{\mathrm{T}}$. Ignoring misclassification and naively solving a set of equations

$$\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}.$$

no longer yields valid estimate of $\boldsymbol{\beta}$. If there exists a set of estimating functions, say, $\mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, \mathbf{W}_i, \mathbf{Z}_i)$, that is close to $\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$, then solving

$$\sum_{i \in s} d_i \mathbf{U}_i^*(\boldsymbol{\beta}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}$$

may still lead to consistent estimator for $\boldsymbol{\beta}$.

We here construct an approximate version of $\sum_{i \in s} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ by taking conditional expectation with respect to the underlying unobserved variables given observed data $(Y_i, W_i, \mathbf{Z}_i)$. It can be seen that the conditional expectation is dependent on the response model, the measurement error model, as well as the covariate distributions. Without additional information, $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ cannot be estimated from the observed data. Therefore, validation data containing true values of the covariates are required so that it is possible to make inference about the measurement error process and the covariate distribution. Suppose internal validation data is available where the true error-prone covariate is partially observed. The original sample s can be divided into three subsets as follows:

$$\begin{aligned}
s_1 &= \{i : (Y_i, X_i, \mathbf{Z}_i)\},\\
s_2 &= \{i : (Y_i, W_i, X_i, \mathbf{Z}_i)\},\\
s_3 &= \{i : (Y_i, W_i, \mathbf{Z}_i)\}.
\end{aligned}$$

For $i \in s_3$, let $\mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathrm{E}_\zeta[\mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i)|Y_i, \mathbf{W}_i, \mathbf{Z}_i]$ be the expected score function of $\boldsymbol{\beta}$. Given $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, response parameter $\boldsymbol{\beta}$ can be estimated by solving

$$\sum_{i \in s_1 \cup s_2} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) + \sum_{i \in s_3} d_i \mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}. \qquad (5.2)$$

For $k = 1, \ldots, K$, let $\mathbf{e}_k$ denote a K-dimensional vector whose lth element is 1 if $l = k$ and 0 otherwise. Let $\mathbf{e}_0 = \mathbf{0}$. Let $\Omega_i(W_i) = \{k : \max(0, W_i - 1) \leq k \leq \min(W_i + 1, K)\}$ be a set of possible values for the underlying true covariate given W_i. The expected score function can be shown to be a weighted sum

$$\mathbf{U}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}; Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \sum_{k \in \Omega_i(W_i)} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{e}_k, \mathbf{Z}_i) \Pr(X_i = k | Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha}),$$

where $\Pr(X_i = k | Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})$ is the posterior weight of $(X_i = k)$ given observed data $(Y_i, W_i, \mathbf{Z}_i)$. With the properties of conditional distribution, we have

$$\begin{aligned}
&\Pr(X_i = k | Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})\\
&\quad = \frac{\Pr(Y_i, W_i, X_i = k | \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})}{\sum_{k' \in \Omega_i(W_i)} \Pr(Y_i, W_i, X_i = k' | \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})}\\
&\quad = \frac{\Pr(Y_i | W_i, X_i = k, \mathbf{Z}_i; \boldsymbol{\beta}) \Pr(W_i, X_i = k | \mathbf{Z}_i; \boldsymbol{\varphi}, \boldsymbol{\alpha})}{\sum_{k' \in \Omega_i(W_i)} \Pr(Y_i | W_i, X_i = k', \mathbf{Z}_i; \boldsymbol{\beta}) \Pr(W_i, X_i = k' | \mathbf{Z}_i; \boldsymbol{\varphi}, \boldsymbol{\alpha})}\\
&\quad = \frac{\Pr(Y_i | X_i = k, \mathbf{Z}_i; \boldsymbol{\beta}) \Pr(W_i | X_i = k, \mathbf{Z}_i; \boldsymbol{\varphi}) \Pr(X_i = k | \mathbf{Z}_i; \boldsymbol{\alpha})}{\sum_{k' \in \Omega_i(W_i)} \Pr(Y_i | X_i = k', \mathbf{Z}_i; \boldsymbol{\beta}) \Pr(W_i | X_i = k', \mathbf{Z}_i; \boldsymbol{\varphi}) \Pr(X_i = k' | \mathbf{Z}_i; \boldsymbol{\alpha})},
\end{aligned}$$

which involves the response model, misclassification model and covariate distribution.

If X_i and $\mathbf{Z}_i$ are independent, then

$$\begin{aligned}&\Pr(X_i = k|Y_i, W_i, \mathbf{Z}_i; \boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})\\ &\quad= \frac{\Pr(Y_i|X_i = k, \mathbf{Z}_i; \boldsymbol{\beta})\Pr(W_i|X_i = k, \mathbf{Z}_i; \boldsymbol{\varphi})\Pr(X_i = k; \boldsymbol{\alpha})}{\sum_{k' \in \Omega_i(W_i)} \Pr(Y_i|X_i = k', \mathbf{Z}_i; \boldsymbol{\beta})\Pr(W_i|X_i = k', \mathbf{Z}_i; \boldsymbol{\varphi})\Pr(X_i = k'; \boldsymbol{\alpha})}.\end{aligned}$$

For fixed $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, estimation of $\boldsymbol{\beta}$ can be performed through iteratively solving (5.2). We now describe the detailed steps of the algorithm as follows:

1. For $i \in s_3$, obtain the set of all possible values of X_i given W_i.
2. Given a current estimate $\hat{\boldsymbol{\beta}}^{(t)}$ and fixed $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, calculate the pseudo-survey weight for each enumerated possibility in the set $\Omega_i(W_i)$
$$d_{ik}^{(t)} = d_i \Pr(X_i = k|Y_i, W_i, \mathbf{Z}_i; \hat{\boldsymbol{\beta}}^{(t)}, \boldsymbol{\varphi}, \boldsymbol{\alpha})$$
3. Obtain new estimate $\hat{\beta}^{(t+1)}$ by solving
$$\sum_{i \in s_1 \cup s_2} d_i \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{X}_i, \mathbf{Z}_i) + \sum_{i \in s_3} \sum_{k \in \Omega_i(W_i)} d_{ik}^{(t)} \mathbf{U}_i(\boldsymbol{\beta}; Y_i, \mathbf{e}_k, \mathbf{Z}_i) = \mathbf{0}.$$
4. The algorithm iterates between steps 2 and 3 until it converges.

Let $\hat{\boldsymbol{\beta}}$ be the final estimate at convergence.

5.3.2 Estimation of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$

The estimation procedure for $\boldsymbol{\beta}$ requires knowledge of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, which can be estimated from the validation data. Estimate of $\boldsymbol{\varphi}$ can be obtained by fitting the misclassification model to subsample s_2, while estimate of $\boldsymbol{\alpha}$ can be obtained from the combined s_1 and s_2.

When the dimension of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are very high, the validation data may not be able to provide sufficient information for the estimation. In this situation, we may impose further assumptions such as simple misclassification process independent of $\mathbf{Z}_i$. For example, we can assume independence between X_i and $\mathbf{Z}_i$. Therefore, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, of which the entries can be estimated by

$$\hat{\alpha}_k = \frac{\sum_{i \in s_1 \cup s_2} d_i \mathrm{I}(X_i = k)}{\sum_{i \in s_1 \cup s_2} d_i}, \quad k = 1, \ldots, K.$$

5.3.3 Variance estimation

It is known that model-based variance matrix of the estimators $\hat{\boldsymbol{\beta}}$ is not preferred, as it does not take into account the complex sampling design. Therefore, we suggest using a resampling method such as bootstrap approach for variance estimation (e.g., Rao and Wu, 1988; Sitter, 1992). Because of the features of the complex survey and the non-response issue, the survey weight for each individual in each bootstrap sample need to be re-calculated in order to account for these features. Suppose $\hat{\boldsymbol{\beta}}_{(b)}$ is the estimate of $\boldsymbol{\beta}$ from an estimation procedure using the bth of B bootstrap samples. Given fixed $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, the approximate variance of $\hat{\boldsymbol{\beta}}$ is given by

$$BV(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\boldsymbol{\beta}}_{(b)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{(b)} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}. \tag{5.3}$$

When $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are estimated from internal validation data, the uncertainty in $(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\alpha}})$ need to be accounted for when calculating the variance of $\hat{\boldsymbol{\beta}}$. This can be done by re-estimating $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ in each bootstrap sample.

5.4 Simulation Study

5.4.1 Design of simulation

We conduct a simulation study to investigate the performance of the proposed method and compare it to the naive approach and the complete-cases approach. The configuration of the simulation is based on the data set from the CCHS Cycle 3.1 described in Chapter 1. Here we only consider simple random sampling from a superpopulation. We set the sample size to be $n = 100000$. Covariates include a three-level ordinal X_i

(valued at 1, 2 and 3) that is subject to misclassification, and a continuous Z_i free of measurement error. We first generate X_i with probabilities 0.2, 0.5, and 0.3 for levels 1, 2, and 3, respectively. We then generate Z_i independently under a standard normal distribution $\text{Normal}(0, 1)$ for all subjects. The binary response variable Y_i is generated under a logistic model

$$\text{logit}\ \mu_i = \beta_0 + \beta_1 \text{I}(X_i = 1) + \beta_2 \text{I}(X_i = 3) + \beta_z Z_i.$$

The parameters are specified by $\beta_0 = -3$, $\beta_1 = 0.3$, $\beta_2 = 0.5$, and $\beta_z = 0.5$. One can think of X_i as a categorical BMI variable, for which the levels represent underweight, normal weight, and overweight or obese categories. The coefficients β_1 and β_2 are specified in such a way that both level 1 and level 3 have positive effect on increasing the risk of developing the outcome compared to the normal level 2.

The surrogate W_i for X_i is generated under multinomial logit models given by

$$\log \{\pi_{ik,l}/\pi_{ik,k}\} = \varphi_{kl(0)} + \varphi_{kl(z)} Z_i \quad \text{for } |k - l| = 1.$$

The parameters associated with the misclassification process are specified by Table 5.1. The misclassification of X_i depends on Z_i in a sense that Z_i has a positive effect on increasing the probability of misclassifying a higher level into a lower one. The dependence is stronger for misclassification of $X_i = 3$ into $W_i = 2$ than for other cases.

Table 5.1: Values of $\boldsymbol{\varphi}$

X	W	$\boldsymbol{\varphi}_{kl(0)}$	$\boldsymbol{\varphi}_{kl(z)}$
1	2	-1.5	-0.05
2	1	-3.0	0.05
	3	-3.0	-0.05
3	2	-1.5	0.50

We obtain the final observed sample $s = \{(Y_i, W_i, Z_i), i = 1, \ldots, n\}$. Also, we obtain a validation subsample $s_2 = \{(Y_i, W_i, X_i, Z_i)\}$ by randomly selecting subjects from s with probability 0.04. Therefore, the size of the validation subsample is around

4000. The data is then analyzed using following approaches: a naive approach ignoring error, a complete-cases analysis using only the validation subsample, and the expected score method that accounts for misclassification in the whole sample. We replicate the simulation 500 times. At this stage, obtaining bootstrap variance matrix for $\hat{\boldsymbol{\beta}}$ in each simulation run will be time consuming. Therefore, we only include the empirical variance of the estimators using the 500 samples, from which we can see how variable each estimator is.

5.4.2 Simulation results

Table 5.2: Simulation results for the naive method, the complete-cases analysis, and the expected score method (500 simulations)

	Naive method		Complete-cases		Proposed method	
Parameter	%RB [†]	EV[‡]	%RB	EV	%RB	EV
β_0	-2.90	0.00036	-0.12	0.01148	-0.02	0.00068
β_1	-42.31	0.00110	-4.42	0.02926	-1.42	0.00200
β_3	-25.24	0.00085	-0.39	0.02244	-0.24	0.00157
β_z	1.71	0.00017	-0.03	0.00444	-0.01	0.00017

[†] %RB = $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\boldsymbol{\beta} \times 100$

[‡] Empirical variance based on 500 samples

Simulation results are shown in Table 5.2. All three estimators for β_z have small biases, although the naive estimator is slightly larger than the others. The estimates of β_1 and β_2 from the naive analysis, however, are attenuated by 42.3% and 25.2%, respectively, which are quite large compared to those from the other two approaches. In general, the complete-cases analysis and the expected score approach perform similarly regarding the relative bias, except that the bias in the estimate of β_1 from the complete-cases analysis is significantly larger than that from the expected score approach. The magnitude of the empirical variance of the estimators are similar for the naive approach and the expected approach, as both use the whole sample. The empirical variance of the estimators from the complete-cases analysis are relatively larger.

5.5 Data Analysis

In this section, we apply the developed method to data from the CCHS cycle 3.1 in 2005. Our interest is in the association of health conditions with risk factors including age, sex, physical activity, and body mass index (BMI). Based on Canadian guidelines, which are in line with those of the World Health Organization, BMI for adults is divided into six categories: underweight, normal weight, overweight, and three obese classes (see Table 5.3 for the range of each category). As BMI was derived from self-reported weight and height, the recorded category may be different from the true category for some subjects. The subsample contains both self-reported and measured weight and height and hence can be used as validation data. Five age groups are formed with 18-24 being the reference group. Physical activity index is a categorical variable with three levels: active, moderate, and inactive. Here the error-contaminated variable is the self-reported BMI category, and the true underlying variable is the measured BMI category. For this study, we exclude subjects who were less than 18 years old, as children are in a stage of development where weight and height may change over a short period of time. Women who were pregnant or breastfeeding were also excluded. Observations in the subsample with self-reported and measured BMI two categories apart are considered as outliers. Subjects with missing any of the error-free covariates or missing both the self-reported and the measured BMI were also excluded from the analysis. This left a sample of 114547 respondents with 4125 in the subsample.

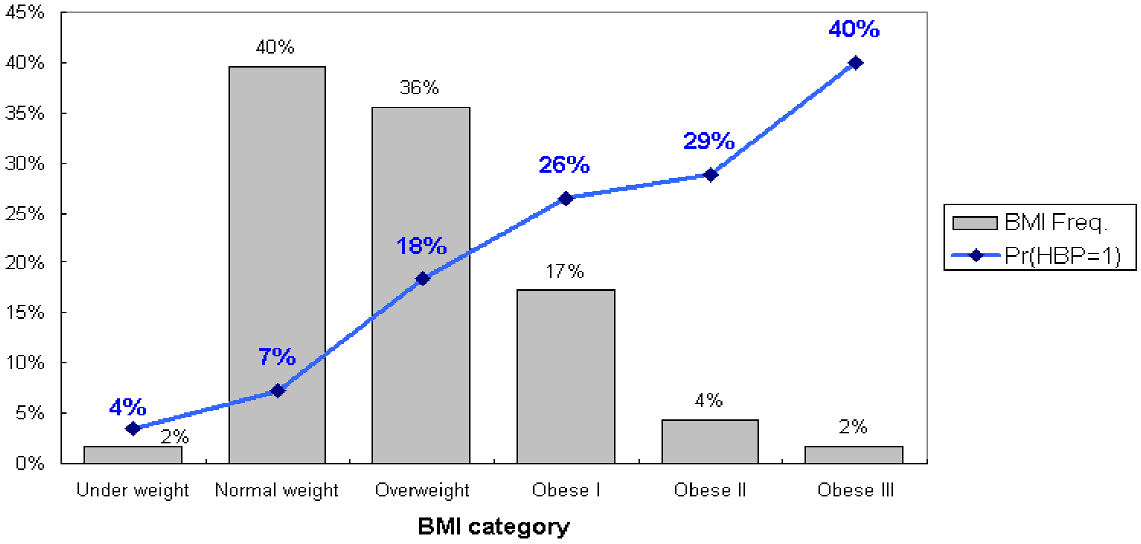
We first present some results from exploratory analysis using the validation subsample. Figures 5.1 and 5.2 show weighted estimates of population proportions for high blood pressure and heart disease in each BMI category. There is a clear trend of increasing proportion of subjects with high blood pressure as BMI category increases, indicating that obesity is a strong risk factor in developing high blood pressure. We observe a similar pattern in heart disease, except that the proportion is higher in the underweight category than in the normal-weight category. Table 5.4 reports the

Table 5.3: Body mass index categories

Category	BMI kg/m^2 range
Underweight (UW)	Less than 18.5
Normal weight (NW)	18.5 to 24.9
Overweight (OW)	25.0 to 29.9
Obese class I (OB I)	30.0 to 34.9
Obese class II (OB II)	35.0 to 39.9
Obese class III (OB III)	40.0 or more

sample percentages for BMI (mis)classifications. One can see that the normal weight subjects performed much better in BMI self-reporting than overweight or obese subjects did. In general, the proportion of subjects who correctly self-reported their BMI category decreases as BMI category increases. The subjects tended to under-report their BMI.

Figure 5.1: Population proportions for high blood pressure in each BMI category



Along with the expected score approach, we also include the naive analysis, which uses self-reported BMI except for subjects with measured BMI, and the complete-cases analysis, which uses only subjects in the subsample with available measured BMI. The normal-weight BMI category was treated as the reference group, and the relative risk of the other five BMI categories on the probabilities of having some

Figure 5.2: Population proportions for heart disease in each BMI category

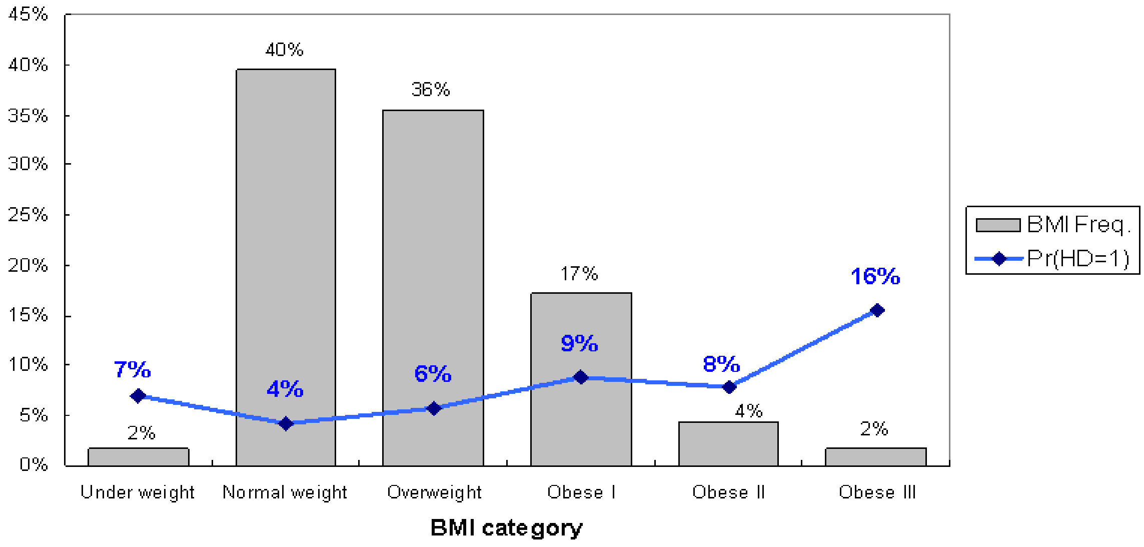


Table 5.4: Estimated BMI (mis)classification rates in the CCHS subsample

Measured	Self-reported BMI						
BMI	UW	NW	OW	OB I	OB II	OB III	Missing
UW	70.00%	27.50%					2.5%
NW	4.21%	90.06%	4.34%				1.38%
OW		29.33%	66.93%	2.00%			1.74%
OB I			46.42%	51.24%	0.58%		1.46%
OB II				57.79%	37.19%	3.02%	2.01%
OB III					32.43%	60.81%	6.76%

chronic conditions are of scientific interest. The variance estimations are based on 500 bootstrap samples with adjusted survey weights.

The result of parametric estimation and inference for high blood pressure is shown in Table 5.5. There are some interesting findings here. First,the expected score approach does not differ much from the naive approach in estimation of $\boldsymbol{\beta}_z$, namely the regression coefficients associated with the error-free covariates: age, sex and physical activity index. The estimates of $\boldsymbol{\beta}_z$ from complete-cases analysis, however, are not that close to those from the other two approaches with regard to the magnitude or even the direction. The three approaches do not quite agree in the risk estimates of BMI categories, although the trend of increasing risk across BMI categories is consistent. The direction of the risk estimate of the underweight category is positive for the expected score approach but is negative for the naive approach and the complete-cases approach. All three associated variance estimates, however, are very large compare to those for other BMI categories. This results in conclusion that the risk of having high blood pressure is not significantly higher in underweight people than in normal-weight people.

The result for heart disease is shown in Table 5.6. We observed similar patterns in the estimates. Based on the result from the expected score approach, the risk of having heart disease increases as BMI increases in general. However, subjects in underweight BMI category has relative higher risk than those in normal-weight category. In contrast, the risk for subjects in overweight category is not significantly different from those in norma-weight category. Due to the relatively smaller sample used in the complete-cases analysis, the variances associated with the BMI risk estimates are very large, resulting in conclusion of non-significant BMI effect on heart disease.

5.6 Discussion

In this chapter, we consider logistic regression analysis using survey data when an ordinal categorical covariate is subject to misclassification. We propose to use the expected score estimation method for analysis of this type of error-contaminated data. The implementation of the algorithm is relatively easy, as expectation over estimating

Table 5.5: Analysis results for high blood pressure

	Naive method † $(n = 114325)$			Complete-cases ‡ $(n = 4120)$			Expected score $(n = 114325)$		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-4.137	0.045	< 0.001	-4.670	0.375	< 0.001	-4.345	0.075	< 0.001
BMI									
Underweight	-0.099	0.106	0.349	-1.082	1.280	0.398	0.298	0.487	0.540
Normal weight	0.000	.	.	0.000	.	.	0.000	.	.
Overweight	0.660	0.036	< 0.001	0.645	0.173	< 0.001	0.787	0.052	< 0.001
Obese I	1.178	0.021	< 0.001	1.043	0.201	< 0.001	1.345	0.040	< 0.001
Obese II	1.638	0.042	< 0.001	1.488	0.316	< 0.001	1.849	0.100	< 0.001
Obese III	1.806	0.099	< 0.001	2.548	0.574	< 0.001	2.084	0.106	< 0.001
Age									
18-34	0.000	.	.	0.000	.	.	0.000	.	.
35-49	1.152	0.047	< 0.001	1.307	0.384	< 0.001	1.133	0.021	< 0.001
50-64	2.468	0.052	< 0.001	2.655	0.356	< 0.001	2.444	0.080	< 0.001
65+	3.431	0.063	< 0.001	3.812	0.355	< 0.001	3.369	0.058	< 0.001
Sex									
Male	-0.105	0.016	< 0.001	0.036	0.130	0.784	-0.115	0.067	0.084
Female	0.000	.	.	0.000	.	.	0.000	.	.
PAI									
Active	-0.119	0.043	0.006	0.149	0.217	0.494	-0.130	0.038	< 0.001
Moderate	0.000	.	.	0.000	.	.	0.000	.	.
Inactive	0.109	0.042	0.009	0.206	0.186	0.267	0.111	0.010	< 0.001

† Self-reported BMI is used except for subjects in the validation subsample

‡ Only the validation subsample is used

Table 5.6: Analysis results for heart disease

	Naive method [†]			Complete-cases [‡]			Expected score		
	($n = 114370$)			($n = 4123$)			($n = 114370$)		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-5.638	0.112	< 0.001	-5.052	0.646	< 0.001	-5.663	0.117	< 0.001
BMI									
Underweight	0.381	0.142	0.007	0.494	0.803	0.539	0.846	0.268	0.002
Normal weight	0.000	.	.	0.000	.	.	0.000	.	.
Overweight	0.118	0.044	0.007	-0.170	0.248	0.494	-0.005	0.092	0.957
Obese I	0.458	0.057	< 0.001	0.255	0.326	0.4339	0.480	0.079	< 0.001
Obese II	0.648	0.098	< 0.001	0.248	0.420	0.555	0.510	0.159	0.001
Obese III	0.850	0.130	< 0.001	0.883	0.578	0.126	0.955	0.152	< 0.001
Age									
18-34	0.000	.	.	0.000	.	.	0.000	.	.
35-49	0.918	0.130	< 0.001	0.156	0.735	0.832	0.930	0.131	< 0.001
50-64	2.382	0.107	< 0.001	1.859	0.686	0.007	2.403	0.108	< 0.001
65+	3.692	0.106	< 0.001	3.179	0.675	< 0.001	3.695	0.107	< 0.001
Sex									
Male	0.460	0.041	< 0.001	0.689	0.202	< 0.001	0.470	0.042	< 0.001
Female	0.000	.	.	0.000	.	.	0.000	.	.
PAI									
Active	-0.122	0.063	0.052	-0.206	0.333	0.536	-0.115	0.063	0.070
Moderate	0.000	.	.	0.000	.	.	0.000	.	.
Inactive	0.223	0.047	< 0.001	0.442	0.292	0.130	0.225	0.047	< 0.001

[†] Self-reported BMI is used except for subjects in the validation subsample

[‡] Only the validation subsample is used

functions with respect to a categorical variable can be written as summation over a few enumerated possible cases.

Expected score estimation calculates the posterior weights for all possible values of the unobserved true covariate, hence it relies on full parametric assumptions for the misclassification mechanism as well as covariate distribution. Robustness to model misspecification needs to be investigated. Also, the parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are estimated from the validation data and are treated as fixed in the estimation of $\boldsymbol{\beta}$. When calculating the bootstrap variance of $\hat{\boldsymbol{\beta}}$, one can account for the extra uncertainty by obtaining estimates of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ in each bootstrap sample. Otherwise, the standard error of $\hat{\boldsymbol{\beta}}$ would be underestimated in general. The main problem is that some bootstrap samples do not contain enough validation data to obtain stable estimators for $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$, especially for cases where the ordinal covariate has many levels, and the misclassification process involves large number of precisely measured covariates.

As mentioned before, the marginal distribution of X_i may be of interest, e.g., estimation of population frequency of each BMI category can be one objective of health surveys. When the dimensions of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$ are small, we can simultaneously estimate $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\alpha}$. Specifically, one can use the extended data with pseudo-survey weights $d_{ik}^{(t)}$ to update the estimates of $\boldsymbol{\varphi}$ and $\boldsymbol{\alpha}$. When X_i is independent of $\mathbf{Z}_i$, for instance, the estimate of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$ can be updated during each iteration by

$$\alpha_k^{(t+1)} = \frac{\sum_{i \in s_1 \cup s_2} d_i \mathrm{I}(X_i = k) + \sum_{i \in s_3} d_{ik}^{(t)}}{\sum_{i \in s} d_i}, \quad k = 1, \ldots, K.$$

In the data analysis example, the BMI variable is used as a risk factor for health conditions. However, BMI itself can be viewed as a response variable, and studying the association of obesity with some effects such as age, sex and physical activity index may be of interest. Misclassifications in both categorical response and categorical covariate are commonly seen in large scale surveys. Furthermore, data that arise from clustered and longitudinal studies are correlated. Full parametric models may not be available for the joint distribution of the clustered responses. This adds some difficulties in the direct application of the parametric approaches. Our future research will consider possible extension of existing methods to account for misclassification in both response and covariate.

Chapter 6

Concluding Remarks and Future Work

6.1 Response Measurement Error in Mixed Models for Correlated Data

In Chapter 2, we considered linear mixed models for clustered data with measurement error in the response variable. It is known that when response error follows the classical additive model, the induced error can be absorbed into the random error term in a linear or linear mixed model. Therefore, naively fitting a linear mixed model to clustered data gives consistent estimators for the fixed effects. The estimator for the conditional variance of the true response, however, is no longer correct. When measurement error is nonlinear, naive analysis with error ignored may lead to biased estimators and invalid inference. We showed by some examples that naively fitting a mixed model leads to seriously biased estimators for the fixed effects.

We considered another naive approach, which fits standard models to transformed data obtained from inverting the link function in the error process, provided that the link function is fully specified. Although the bias in the fixed-effect estimators can be reduced by a certain amount, the estimators for the variance parameters are seriously biased, because the transformed surrogate is not unbiased for the true underlying response. We proposed to use likelihood-based methods. For cases where the error parameters are unknown and validation data are available, the pseudo-likelihood approach, which uses a two-stage estimation procedure, give consistent estimators for both fixed-effects parameters and variance components. We conducted simulation studies and showed that the methods performed reasonably well under various settings.

As already pointed out, the likelihood-based approaches can be computational intensive, as the accuracy of the estimators relies on the order of the Gaussian quadrature approximations to the integrals involved in the likelihood function. We found in our simulation studies that an order of 10 quadrature approximation performs well enough for one-dimensional random effect. For mixed models with multi-dimensional random effects, the computation can be very slow.

6.2 Correlated Binary Responses with Misclassification

In Chapter 3 we considered correlated binary data with misclassified responses. There are many statistical models developed for analyzing correlated data arising from longitudinal studies and clustered studies. Misclassification, which is a special type of measurement error, is commonly seen in these studies. Naive analysis ignoring misclassification leads to biased estimates of model parameters. Neuhaus (2002) investigated the bias and efficiency loss due to the presence of misclassification in binary responses in a logistic mixed model. However, the approximate adjusting factor derived by the author is for simple models, e.g., only one covariate is involved and misclassifications are independent of each other.

We proposed marginal methods, in which only the marginal and second-order association models are specified for the clustered responses. We took an estimating equations approach for correcting the bias induced by misclassification. We also constructed unbiased second-order estimating functions when misclassifications are correlated. Several cases were discussed, including known error parameters, unknown error parameters but with validation data or replicated measures available. For replication studies, response and error parameters are required to be jointly estimated. The estimators from our developed approaches have good properties such as consistency and normality. Simulation studies showed that they performed very well under a variety scenarios for different cases. More scenarios for association structures in the response process as well as in the misclassification process need to be considered and more simulation studies need to be conducted.

Proportion of validation data and cluster size play important roles in making inference about the misclassification process, especially for the dependence structure. In situations where a small validation subsample does not provide enough information for estimating the correlation between misclassifications, assuming an independent misclassification process may be unavoidable. For studies where neither validation data nor replicated measures are available, which are very common in practice, the misclassification model can not be identified. The best one can do is to conduct sensitivity analysis.

6.3 Marginal Models for Longitudinal Ordinal Data with Misclassification in Responses and Covariates

Many health outcomes are ordinal, such as severity measure of a particular disease. These variables may be subject to misclassification when the measuring system is not gold standard or it is impossible to obtain accurate measurements. Similarly, many risk factors such as dietary intake and systolic blood pressure are measured with error. In Chapter 4 we developed marginal methods for analysis of longitudinal ordinal data with misclassification in both responses and covariates. Our simulation studies showed that the methods performed very well under a variety of scenarios.

In practice the misclassification processes for response and covariate are unknown. We assume that a validation subsample is available for making inference about the processes. The number of nuisance parameters involved in the processes, however, can be very large. Small validation subsample therefore may not provide enough information. In such situation, the best one can do is to assume very simple misclassification models as well as impose extra constraints such as adjacent misclassifications.

We have only considered independent misclassification processes for both the response and the covariate. Correlated misclassifications can occur in longitudinal clinical trails, in which same defected measuring devices are applied repeatedly, or family studies, in which self-report measures of family members may share a common bias.

Replicated measures instead of validation data may be available in some studies. While a joint estimation procedure must be employed, the minimum number of replicates required for valid inference about the misclassification process may be different for categorical or ordinal variables than for binary variables discussed in Chapter 3.

6.4 Future Work: Analysis of Correlated Data with Measurement Error, Incomplete Observations, and Complex Survey Designs

6.4.1 Marginal and association models with dropouts and measurement error

As mentioned in Chapter 3, marginal methods have been widely used for analysis of longitudinal and clustered data, where the marginal mean and association structure are of interest. Longitudinal categorical data often contain incomplete observations, e.g., dropouts, and non-response. Yi and Thompson (2005) described a likelihood-based approach to characterizing longitudinal binary data with drop-outs, in which marginal and dependence structures are specified as regression models to link the responses to the covariates. Estimating equation approaches such as inverse probability weighted (IPW) GEE are also widely employed (see, e.g., Yi and Cook, 2002; Chen et al., 2010). The weight matrix, which is constructed for each cluster that contains missing data, may be dependent on the history of response outcomes and/or covariates. Therefore, direct application of the IPWGEE approach may be hindered by the presence of misclassification in responses and/or covariates. Yi (2008) and Yi et al. (2010) considered correcting estimation bias induced by dropout and mismeasured covariates in longitudinal data. We will explore extending our developed marginal methods to simultaneously handle missing response and covariates as well response and covariate measurement error in data from longitudinal and clustered studies.

6.4.2 Transition models for longitudinal categorical data with misclassification

While a marginal regression model is used to characterize the dependence of the response on covariates, a conditional regression model, or transition model (Diggle et al., 2002), is used to capture the serial dependence in the response process. Azzalini (1994) described a first-order Markov chain model by assuming that the current state of a categorical response is dependent on the history only through the immediate previous response. Heagerty (2002) extended this marginalized transition model to allow pth-order serial dependence that is common in longitudinal data, in terms of the combination of a marginal regression model and a transition model. Chen et al. (2009) developed a Markov model for longitudinal categorical data which facilitates modelling both marginal and conditional structures. Pan et al. (2009) considered semiparametric transition models with one covariate measured with error and proposed an estimating equation approach, in which no distributional assumption was made for the underlying unobserved covariate. When the responses are subject to misclassification, however, naive inference about the dependence structure will lead to incorrect conclusions. Cook et al. (2000) described a latent Markov model for longitudinal binary data in the absence of a gold-standard reference test and adopted log-linear models for the dependence of the classifications of multiple diagnostic tests that are applied repeatedly over time. The case of correlated replicates is of particular relevance to physical examinations, diagnostic tests, as well as self-reported variables such as food intake in longitudinal studies. Similarly, Rosychuk and Thompson (2001, 2003) considered two-state Markov models with misclassified responses and proposed iterative biased-adjusted methods.

As mentioned before, categorical responses and covariates can be subject to misclassification at the same time. Some of our future research will focus on developing methods for analysis of categorical and ordinal data under transition models with misclassified responses and covariate measurement error.

6.4.3 Semi-parametric methods for correlated data with measurement error and incomplete observations

Semiparametric models combine both parametric models and non-parametric models, such as partially linear models and single index models (e.g., Ruppert et al., 2003). The effect of the error-prone covariate, which is often of interest, is usually modeled parametrically, while the effects of some, if not all, precisely measured covariates are modeled nonparametrically (e.g., Carroll et al., 2006). Some research work has been done for univariate data, e.g., Huang and Wang (2001) considered linear logistic regression with replicated error-prone covariates, and Liang (2000) proposed deconvolution methods for partially linear models with measurement error. Tsiatis and Ma (2004) proposed a class of semiparametric estimators in the general setting of functional measurement error models. Ma and Carroll (2006) constructed locally efficient semiparametric estimators for a general class of semiparametric models with measurement errors, in which a parametric model estimator and a local kernel estimator are combined through backfitting. Liu and Wu (2010) proposed and investigated the theoretical properties of a computationally efficient approximate method for a class of semiparametric nonlinear mixed-effects models with measurement error and incomplete data.

In contrast, not much work has been done for handling response measurement error (or misclassification) in the framework of semiparametric regression models. In our future work we will extend existing approaches and develop novel semiparametric methods for analysis of correlated data with measurement error and missing observations.

6.4.4 Analysis of data from complex surveys

In Chapter 5 we discussed covariate misclassification problems in data collected from surveys. We focused on logistic regression analysis of univariate binary data with misclassification in an ordinal covariate. It is well known that survey weights derived from the sampling design can be incorporated into estimating equations so that the population parameters and superpopulation model parameters can be simultaneously estimated (e.g., Godambe and Thompson, 1986). We proposed the expected score approach that can correct estimation bias induced by misclassifications. Data from large scale surveys often contain both measurement error and missing observations, which can occur during measuring, data recording, editing, etc. The main source of missing data is unit or item non-response. Our future research will include the development of statistical techniques to handle complex survey design, missing data, and measurement error in data from longitudinal surveys and family surveys.

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis.* Wiley-Interscience, 2nd edition.

[2] Akazawa, K., Kinukawa, N., and Nakamura, T. (1998). A note on the corrected score function adjusting for misclassification. *Journal of the Japan Statistical Society*, **28,** 115–123.

[3] Albert, P. S., Hunsberger, S. A., and Biro, F. M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *Journal of the American Statistical Association*, **92,** 1304–1311.

[4] Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81,** 767–775.

[5] Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction,* Solomon, H. (ed.), pp. 158-168, Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

[6] Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, **45,** 164–180.

[7] Binder, D. A. and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, **89,** 1035–1043.

[8] Bollinger, C. R. and David, M. H. (1997). Modeling discrete choice with response error: Food Stamp participation. *Journal of the American Statistical Association*, **92,** 827–835.

[9] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26,** 211 – 252.

[10] Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model. *Journal of the American Statistical Association*, **91,** 633–642.

[11] Buonaccorsi, J. P., Demidenko, E., and Tosteson, T. D. (2000). Estimation in longitudinal random effects models with measurement error. *Statistica Sinica*, **10,** 885–903.

[12] Buonaccorsi, J. P., Laake, P., and Veierød, M. B. (2005). On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, **61,** 831–836.

[13] Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80,** 517–526.

[14] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. London: Chapman and Hall/CRC, 2nd edition.

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

[16] Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, **53,** 573–585.

[17] Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference*, **63,** 39–54.

[18] Chen, B., Yi, G. Y., and Cook, R. J. (2009). Likelihood analysis of joint marginal and conditional models for longitudinal categorical data. *The Canadian Journal of Statistics*, **37,** 182–205.

[19] Chen, B., Yi, G. Y., and Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*, **105,** 336–353.

[20] Christopher, S. R. and Kupper, L. L. (1995). On the effects of predictor misclassification in multiple linear-regression analysis. *Communications in Statistics - Theory and Methods*, **24,** 13–37.

[21] Chua, T. C. and Fuller, W. A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, **82,** 46–51.

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

[23] Cook, R. J., Ng, E. T. M., and Meade, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics*, **56,** 1109–1117.

[24] Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B*, **50,** 225–265.

[25] Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics*, **21,** 113–120.

[26] Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, **82,** 407–410.

[27] Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society, Series B*, **63,** 55–62.

[28] Diggle, P. (1992). Discussion of 'Multivariate analysis of categorical data' by K.-Y. Liang, S. L. Zeger and B. Qaqish. *Journal of the Royal Statistical Society, Series B*, **45,** 28–29.

[29] Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* New York: Oxford University Press, 2nd edition.

[30] Ekholm, A. and Palmgren, J. (1987). Correction for misclassification using doubly sampled data. *Journal of Official Statistics*, **3,** 419–429.

[31] Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics*, **28,** 51–60.

[32] Ferrara, L. A., Guida, L., Iannuzzi, R., Celentano, A., and Lionello, F. (2002). Serum cholesterol affects blood pressure regulation. *Journal of Human Hypertension*, **16,** 337–343.

[33] Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80,** 141–151.

[34] Fuller, W. A. (1987). *Measurement Error Models.* New York: Wiley-Interscience.

[35] Fuller, W. A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, **63,** 121–141.

[36] Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, **54,** 127–138.

[37] Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, **112,** 564–569.

[38] Greenland, S. (1982). The effect of misclassification in matched-pair case-control studies. *American Journal of Epidemiology*, **116,** 402–406.

[39] Greenland, S. (1988). Statistical uncertainty due to misclassification: Implications for validation substudies. *Journal of Clinical Epidemiology*, **41,** 1167–1174.

[40] Greenland, S. (2008). Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification. *Journal of Statistical Planning and Inference*, **138,** 528–538.

[41] Gregoire, T. G. and Salas, C. (2009). Ratio estimation with measurement error in the auxiliary variate. *Biometrics*, **65,** 590–598.

[42] Griliches, Z. and Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics*, **31,** 93–118.

[43] Grundy, S. M. (2000). Early detection of high cholesterol levels in young adults. *Journal of the American Medical Association*, **284,** 365–367.

[44] Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* London: Chapman and Hall/CRC.

[45] Hall, P. and Ma, Y. Y. (2007). Semiparametric estimators of functional measurement error models with unknown error. *Journal of the Royal Statistical Society, Series B*, **69,** 429–446.

[46] Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72,** 320–338.

[47] Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, **58,** 342–351.

[48] Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, **15,** 1–19.

[49] Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis.* New York: Wiley-Interscience.

[50] Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors. *Journal of the American Statistical Association*, **72,** 914–921.

[51] Hu, Y. Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, **144,** 27–61.

[52] Huang, Y. and Wang, C. Y. (2001). Consistent functional methods fo logistic regression with error in covariates. *Journal of the American Statistical Association*, **96,** 1469–1482.

[53] Hwang, J. T. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association*, **81,** 680–688.

[54] Jaquet, F., Goldstein, I. B., and Shapiro, D. (1998). Effects of age and gender on ambulatory blood pressure and heart rate. *Journal of Human Hypertension*, **12,** 253–257.

[55] Jiang, W., Turnbull, B. W., and Clark, L. C. (1999). Semiparametric regression models for repeated events with random effects and measurement error. *Journal of the American Statistical Association*, **94,** 111–124.

[56] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38,** 963–974.

[57] Lee, L.-F. and Sepanski, J. H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, **90,** 429–440.

[58] Liang, H. (2000). Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part. *Journal of Statistical Planning and Inference*, **86,** 51 – 62.

[59] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73,** 13–22.

[60] Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54,** 3–40.

[61] Lin, X. and Carroll, R. J. (1999). SIMEX variance component tests in generalized linear mixed measurement error models. *Biometrics*, **55,** 613–619.

[62] Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78,** 153–160.

[63] Liu, W. and Wu, L. (2010). Some asymptotic results for semiparametric nonlinear mixed-effects models with incomplete data. *Journal of Statistical Planning and Inference*, **140,** 52–64.

[64] Lyles, R. H. and Kupper, L. L. (1997). A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics*, **53,** 1008–1025.

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

[66] Marques, T. A. (2004). Predicting and correcting bias caused by measurement error in line transect sampling using multiplicative error models. *Biometrics*, **60,** 757–763.

[67] McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley-Interscience.

[68] Miller, M. E., Davis, C. S., and Landis, J. R. (1993). The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics*, **49,** 1033–1044.

[69] Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18,** 2237–2255.

[70] Moore, J. C., Stinson, L. L., and Welniak, Edward J., J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, **16,** 331–361.

[71] Nakamura, T. (1990). Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*, **77,** 127–137.

[72] Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, **48,** 829–838.

[73] Natarajan, S., Lipsitz, S. R., and Nietert, P. J. (2002). Self-report of high cholesterol - Determinants of validity in US adults. *American Journal of Preventive Medicine*, **23,** 13–21.

[74] Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86,** 843–855.

[75] Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, **58,** 675–683.

[76] Neuhaus, J. M. and McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society, Series B*, **68,** 859–872.

[77] Novick, S. and Stefanski, L. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, **97,** 472–481.

[78] Pan, W., Zeng, D., and Lin, X. (2009). Estimation in semiparametric transition measurement error models for longitudinal data. *Biometrics*, **65,** 728–736.

[79] Paulino, C. D., Soares, P., and Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, **59,** 670–675.

[80] Pepe, M. and Anderson, G. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics*, **23,** 939–951.

[81] Pfeffermann, D., Skinner, C., and Humphreys, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, Series A*, **161,** 13–32.

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

[83] Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44,** 1033–1048.

[84] Prentice, R. L., Kakar, F., Hursting, S., Sheppard, L., Klein, R., and Kushi, L. H. (1988). Aspects of the rationale for the Women's Health Trial. *Journal of the National Cancer Institute*, **80,** 802–814.

[85] Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M. G., and Poulter, N. R. (2001). Association between smoking and blood pressure - Evidence from the Health Survey for England. *Hypertension*, **37,** 187–193.

[86] Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83,** 231–241.

[87] Rao, J. N. K., Yung, W., and Hidiroglou, M. A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā: The Indian Journal of Statistics, Series A*, **64,** 364–378.

[88] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90,** 106–121.

[89] Rosner, B. A. (1996). Measurement error models for ordinal exposure variables measured with error. *Statistics in Medicine*, **15,** 293–303.

[90] Rosychuk, R. J. and Thompson, M. E. (2001). A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics*, **29,** 395–404.

[91] Rosychuk, R. J. and Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine*, **22,** 2035–2055.

[92] Roy, S. and Banerjee, T. (2009). Analysis of misclassified correlated binary data using a multivariate probit model when covariates are subject to measurement error. *Biometrical Journal*, **51,** 420–432.

[93] Roy, S., Banerjee, T., and Maiti, T. (2005). Measurement error model for misclassified binary responses. *Statistics in Medicine*, **24,** 269–283.

[94] Ruppert, D., Wand, M. P., and Carroll, R. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

[95] Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78,** 719–727.

[96] Sitter, R. R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, **87,** 755–765.

[97] Song, P. X. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer.

[98] Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, **95,** 51–61.

[99] Statistics Canada (2005). *Canadian Community Health Survey (CCHS) Cycle 3.1 Public Use Microdata File User Guide*. Statistics Canada.

[100] Stefanski, L. A. and Buzas, J. S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association*, **90,** 541–550.

[101] Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, **90,** 1247–1256.

[102] Suh, E.-Y. and Schafer, D. W. (2002). Semiparametric maximum likelihood for nonlinear regression with measurement errors. *Biometrics*, **58,** 448–453.

[103] Sutradhar, B. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, **86,** 459–465.

[104] Tempelman, R. J. and Gianola, D. (1996). A mixed effects model for overdispersed count data in animal breeding. *Biometrics*, **52,** 265–279.

[105] Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, **65,** 1350–1361.

[106] Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. *Technometrics*, **14,** 187–202.

[107] Tsiatis, A. A. and Ma, Y. Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, **91,** 835–848.

[108] Veierød, M. B. and Laake, P. (2001). Exposure misclassification: bias in category specific Poisson regression coefficients. *Statistics in Medicine*, **20,** 771–784.

[109] Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, **83,** 447–452.

[110] Vonesh, E. F. and Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, **48,** 1–17.

[111] Wang, N. and Davidian, M. (1996). A note on covariate measurement error in nonlinear mixed effects models. *Biometrika*, **83,** 801–812.

[112] Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, **93,** 249–261.

[113] Wansbeek, T. (2001). GMM estimation in panel data models with measurement error. *Journal of Econometrics*, **104,** 259–268.

[114] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50,** 1–25.

[115] White, I., Frost, C., and Tokunaga, S. (2001). Correcting for measurement error in binary and continuous variables using replicates. *Statistics in Medicine*, **20,** 3441–3457.

[116] Williamson, J. M., Kim, K., and Lipsitz, S. R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association*, **90,** 1432–1437.

[117] Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, **80,** 791–795.

[118] Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, **90,** 937–951.

[119] Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96,** 185–193.

[120] Xiao, Z., Shao, J., and Palta, M. (2010). GMM in linear regression for longitudinal data with multiple covariates measured with error. *Journal of Applied Statistics*, **37,** 791–805.

[121] Yanez III, N. D., Kronmal, R. A., and Shemanski, L. R. (1998). The effects of measurement error in response variables and tests of association of explanatory variables in change models. *Statistics in Medicine*, **17,** 2597–2606.

[122] Ybarra, L. M. R. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95,** 919–931.

[123] Yi, G. Y. (2007). *Measurement Error and Missing Data Problems* (*Lecture Notes*). Department of Statistics and Actuarial Science, University of Waterloo.

[124] Yi, G. Y. (2008). A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*, **9,** 501–512.

[125] Yi, G. Y. and Cook, R. J. (2002). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, **97,** 1071–1080.

[126] Yi, G. Y. and Cook, R. J. (2005). Errors in the Measurement of Covariates. *The Encyclopedia of Biostatistics,* **3,** pp. 1741-1748. John Wiley and Sons Ltd., 2nd edition.

[127] Yi, G. Y. and Lawless, J, F. (2007). A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *Journal of Statistical Planning and Inference*, **137,** 1816–1828.

[128] Yi, G. Y., Liu, W., and Wu, L. (2010). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics* (to appear).

[129] Yi, G. Y. and Thompson, M. E. (2005). Marginal and association regression models for longitudinal binary data with drop-outs: a likelihood-based approach. *The Canadian Journal of Statistics*, **33,** 3–20.

[130] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77,** 642–648.

[131] Zucker, D. M. (2005). A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. *Journal of American Statistical Association*, **100,** 1264–1277.

[132] Zucker, D. M. and Spiegelman, D. (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Medicine*, **27,** 1911–1933.