Second-order Least Squares Estimation in Generalized Linear Mixed Models

By

He Li

A Thesis Submitted to the Faculty of Graduate Studies of
the University of Manitoba
in Partial Fulfilment of the Requirements of the Degree of
Doctor of Philosophy

Department of Statistics
University of Manitoba
Winnipeg

Abstract

Maximum likelihood is an ubiquitous method used in the estimation of generalized linear mixed model (GLMM). However, the method entails computational difficulties and relies on the normality assumption for random effects. We propose a second-order least squares (SLS) estimator based on the first two marginal moments of the response variables. The proposed estimator is computationally feasible and requires less distributional assumptions than the maximum likelihood estimator. To overcome the numerical difficulties of minimizing an objective function that involves multiple integrals, a simulation-based SLS estimator is proposed. We show that the SLS estimators are consistent and asymptotically normally distributed under fairly general conditions in the framework of GLMM.

Missing data is almost inevitable in longitudinal studies. Problems arise if the missing data mechanism is related to the response process. This thesis develops the proposed estimators to deal with response data missing at random by either adapting the inverse probability weight method or applying the multiple imputation approach.

In practice, some of the covariates are not directly observed but are measured with error. It is well-known that simply substituting a proxy vari-

able for the unobserved covariate in the model will generally lead to biased and inconsistent estimates. We propose the instrumental variable method for the consistent estimation of GLMM with covariate measurement error. The proposed approach does not need any parametric assumption on the distribution of the unknown covariates. This makes the method less restrictive than other methods that rely on either a parametric distribution of the covariates, or to estimate the distribution using some extra information.

In the presence of data outliers, it is a concern that the SLS estimators may be vulnerable due to the second-order moments. We investigated the robustness property of the SLS estimators using their influence functions. We showed that the proposed estimators have a bounded influence function and a redescending property so they are robust to outliers. The finite sample performance and property of the SLS estimators are studied and compared with other popular estimators in the literature through simulation studies and real world data examples.

Keywords: Bias reduction; Discrete response; Influence function; Instrumental variable; Least squares method; Longitudinal data; Measurement error; M-estimator; Mixed effects models; Outliers; Robustness; Simulation-based estimator.

Acknowledgements

I would like to offer my sincerest gratitude to Dr. Liqun Wang, my advisor, who has guided me throughout my Ph.D. work with patience and knowledge. He has been a true inspiration and a guiding force, without which this thesis would be possible. I also thank him for supporting, and making possible, my attendance at various statistical conferences and workshops through his Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Institute for Complex Data Structures (NICDS) research grants. I feel privileged and honored to have the opportunity to work with him over the past few years.

Besides my advisor, I thank the members of my advisory committee, who shared with me their knowledge and provided me with many suggestions to improve this thesis: Dr. James Fu, Dr. Saumen Mandal and Dr. Abba Gumel at University of Manitoba, and Dr. Joan Hu at Simon Fraser University. A special thanks goes to Dr. Taraneh Abarin, Post Doctoral Fellow Research at Samuel Lunenfeld Research Institute for her mentorship and encouragement during my PhD studies.

Last but not the least, I thank my family: my mum, for raising me

and supporting me financially up to early years of my education. My wife, Qin Chen, for her unconditional love and continuous support. Finally, my daughter, Kayla Li, for bringing endless joy and happy moments into our lives.

Dedication

This thesis is dedicated to my daughter, Kayla Li.

Contents

List of Tables

List of Figures

List of Acronyms

BF	Breast Feeding
BMI	Body Mass Index
CLT	Central Limit Theorem
DCT	Dominated Convergence Theorem
EXBF	Exclusive Breast Feeding
GEE	Generalized Estimation Equation
GLMM	Generalized Linear Mixed Models
GLMMeM	Generalized Linear Mixed Models with Measurement Error
IF	Influence Function
IV	Instrumental Variable
IPW	Inverse Probability Weighted
LMM	Linear Mixed Models
LMMeM	Linear Mixed Models with Measurement Error
MAR	Missing at Random
MCAR	Missing Completely at Random
ME	Measurement Error
MI	Multiple Imputation
MME	Method of Moments Estimator
MSE	Mean Squared Errors
MNAR	Missing Not at Random
MLE	Maximum Likelihood Estimator
MQLE	Marginal Quasi-Likelihood Estimator
PQLE	Penalized Quasi-Likelihood Estimator

RC	Regression Calibration
REML	Restricted Maximum Likelihood Estimator
RI	Random Intercept
RIS	Random Intercept and Slope
RMSE	Root Mean Squared Errors
SBE	Simulation-based Estimator
SBEIW	Independently Weighted Simulation-based Estimator
SIMEX	Simulation Extrapolation
SLSE	Second-order Least Squares Estimator
SNP	Single-nucleotide Polymorphism
ULLN	Uniform Law of Large Numbers
WSBE	Weighted Simulation-based Estimator

Chapter 1

Introduction

1.1 Longitudinal Data Analysis

In medical, biological, environmental and social sciences research, longitudinal data analysis is widely used and constitutes the most fundamental statistical research methodologies. Longitudinal data, by definition, is data collected from repeated observations of subjects over time. Typically, a fixed number of repeated observations are obtained at a set of common time points although they are not required to be distributed evenly throughout the duration of a study. The distinct feature of longitudinal data is that individual subjects are measured repeatedly across time and these measurements are likely to be correlated within the same individual. The scientific questions of interest in longitudinal studies, often involve not only the usual questions, such as how the mean response differs across treatments, but also how the change of subjects' responses over time (e.g., growth and aging) differs and

other issues concerning the relationship between responses and time.

There are several major advantages of collecting longitudinal data. First, longitudinal studies allow us to investigate how the variability of the response varies in time with covariates. For instance, a clinical trial designed to study time-varying drug efficacy in treating a disease, which cannot be examined by a cross-sectional study. Second, longitudinal studies have the capability to separate aging effects (changes over time within individuals) from cohort effects (differences between subjects at baseline). Third, longitudinal studies are more powerful to detect an association of interest compared to a cross-sectional study. The reason is that the repeated measurements from a single subject provide more independent information than a single measurement obtained from a single subject. Last, longitudinal studies can provide information about individual changes.

Conventional statistical methods require there to be an independence between observations. Longitudinal data, however, unlike cross-sectional data, is likely to violate this assumption because measurements within a subject may be correlated. Hence, the key challenge of longitudinal data analysis is to account for the dependency in the data using more sophisticated statistical methodologies. Although there have been extensive methodological developments for the analysis of longitudinal data in the last few decades (e.g. Molenberghs and Verbeke 2005; Carroll, Ruppert, Stefanski, and Crainiceanu 2006; Molenberghs and Kenward 2007; Fitzmaurice, Davidian, Molenberghs

and Verbeke 2008; McCulloch, Searle and Neuhasus 2008), there are still many emerging issues arising in practice which motivate further research in this area. In particular, the following problems are common in longitudinal studies:

- longitudinal data may either be continuous or categorical or a mixture of both;
- there are often missing data or dropouts;
- some variables may be measured with errors;
- data outliers are always present.

New statistical methods are required to address one or more of the above problems as standard methods are not directly applicable. Commonly used models for longitudinal data include: mixed models, marginal models and transition models. Each of these modeling approaches offers their own advantages and disadvantages.

Mixed models (Harville 1977; Laird and Ware 1982; Breslow and Clayton 1993), in which the regression coefficients are allowed to vary across subjects, are commonly used to incorporate both variations within and between subjects. They include a mixture of fixed effects, which are parameters associated with the entire population, and random effects which are associated with individual subjects. In general, the distribution of mixed effects

is usually assumed to be normal. Mixed models can not only describe the trend of data over time while taking account of the correlation that exists between successive measurements, but also describe the different variation for each subject over time. The mixed effects model is a powerful technique for the analysis of longitudinal data when the objective is to make inference about individuals rather than the population average.

In marginal models (Liang and Zeger 1986) the regression of the response on explanatory variables is modeled separately from within-subject correlation. These models focus on the mean structure, and more specifically on the regression parameters linked to the means. The within-subject dependence is treated as a nuisance, which needs to be accounted for since it affects the power of tests and the precision of the regression estimates. The estimation of parameters does not require full distributional assumptions, but rather only require specification of a regression model for the mean response. The primary objective of the marginal models is to estimate the effect of a set of covariates on the marginal expectation of response without explicitly accounting for subject to subject heterogeneity. Marginal models, are also referred to as population-average models due to the fact that they describe the average response in the population rather than an individual's responses (Zeger, Liang, and Albert 1988). A comprehensive discussion on the relation between marginal and random-effects models can be found in Heagerty and Zeger (2000) and Nelder and Lee (2004).

In transition models, the conditional mean of an outcome at the current time point is modeled as a function of its values at the previous time points and covariates (Diggle, Liang and Zeger 1994). These models are also known as conditional or Markov models. They are useful when one is interested in studying the effects of covariates and of past responses on the current response or predicting the future response given the past history. The within-subject correlation is easily accounted for by conditioning on the past responses, and the model can be easily fitted within the generalized linear model framework.

1.2 Data Examples

Two real world longitudinal data examples are presented in this section for illustration purposes.

1.2.1 Example 1: Framingham Study

In the Framingham study (Dawber, Moore and Mann 1957; Dawber 1980), 2634 participants' cholesterol level is measured every 2 years over a 10 year period. The objective is to study the change in cholesterol over time and examine the association with age at baseline and gender. Figure 1.1 shows cholesterol levels over time for 200 randomly selected individuals from the Framingham study and a glimpse of the raw data for illustration purposes is provided in Table 1.1. Figure 1.1 suggests all subjects seem to have a similar

trajectory and cholesterol levels increase linearly over time. However, each subject has his/her own trajectory line with a possibly different intercept and slope, which implies two sources of variations (within and between subject variations) exist in this dataset.

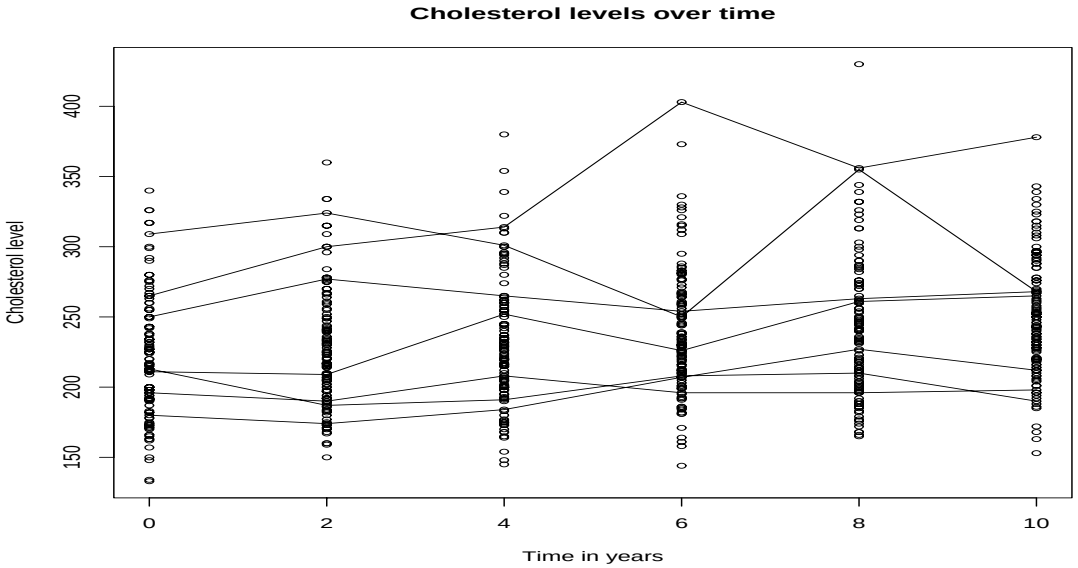
Table 1.1: Cholesterol levels for a subset of participants over time

Subject	Cholesterol	Sex	Age	Year
1	175	M	32	0
1	198	M	32	2
1	205	M	32	4
1	228	M	32	6
1	214	M	32	8
1	214	M	32	10
2	299	F	34	0
2	328	F	34	4
2	374	F	34	6
2	362	F	34	8
2	370	F	34	10
⋮	⋮	⋮	⋮	⋮

1.2.2 Example 2: Seizure Count Data

In a clinical trial, 59 epileptics who were randomized to receive either the antiepileptic drug progabide or a placebo , as an adjuvant to standard chemotherapy. The logarithm of a quarter of the number of epileptic seizures in the 8-week period preceding the trial (Base) and the logarithm of age (Age) were included as covariates in the analysis. For each individual, a multivariate re-

Figure 1.1: Trajectories of cholesterol levels for a subset of participants over time

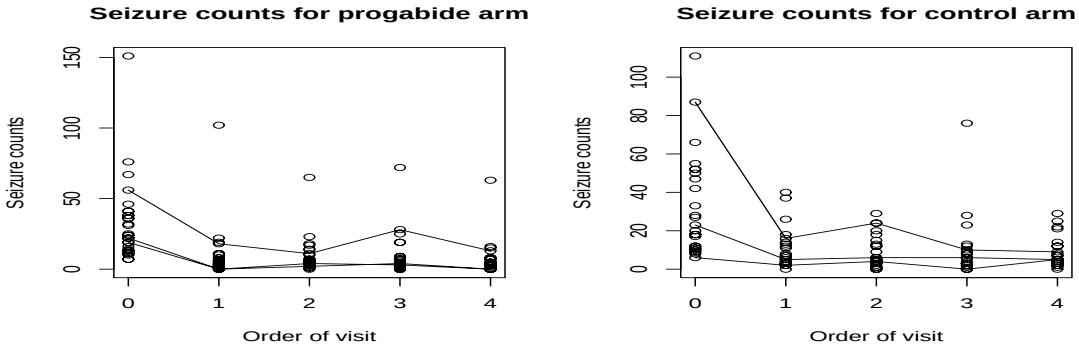


sponse variable consisted of the seizure counts during 2-week periods before each of four clinical visits (Visit, 1,2,3,4) was collected. This data was first analyzed by Thall and Vail (1990) to study whether the treatment effect is effective after adjusting for available covariates. The data set is shown in Table 1.2 and Figure 1.2. The response variable is count data which implies that Poisson regression model would be appropriate. All subjects that received the same treatment seem to have a similar trajectory but with noticeable intra-subject and inter-subject variabilities. Subjects that received different treatments may have possibly different intercepts and slopes. All the observations seem to be correlated within the same subject. In addition, there is a number of patients who seem to have irregularly large counts and may be potential data outliers.

Table 1.2: Epileptic seizure count data over time

Subject	Count	Treatment	Base	Age	Visit
1	5	placebo	11	31	1
1	3	placebo	11	31	2
1	3	placebo	11	31	3
1	3	placebo	11	31	4
⋮	⋮	⋮	⋮	⋮	⋮
29	11	progabide	76	18	1
29	14	progabide	76	18	2
29	9	progabide	76	18	3
29	8	progabide	76	18	4
⋮	⋮	⋮	⋮	⋮	⋮

Figure 1.2: Epileptic seizure counts over time



1.3 Overview of Work

Recently, Wang (2003, 2004) proposed a Second-order Least Squares Estimator (SLSE) for nonlinear measurement error models, and Wang and Leblanc (2008) compared the SLSE with the Ordinary Least Squares estimator in general nonlinear models. Wang (2007) extended this estimation method to nonlinear mixed effects models with homoscedastic errors. This estimation method is based on the first two marginal moments of the response variables given the covariates. He showed that under some regularity conditions the SLSE is consistent and asymptotically normally distributed. Li (2005) performed extensive simulation studies of the SLSE for nonlinear mixed effects models. Abarin (2008) applied the SLS method to cross-sectional regression model with application to measurement error. The focus of this thesis is to extend SLSE further to the Generalized Linear Mixed Models (GLMM), which have been widely used in the modeling of longitudinal data where the response is discrete.

This thesis contains some major extensions and studies of the second-order least squares methodology. First, we address the computational issues and implementation of the SLS estimators in practice. The finite sample performance has not been studied especially under different setups of the weight matrix. We conduct substantial numerical studies to investigate these in the GLMM framework. Furthermore, we relax the high-level regularity conditions in Wang (2007) to derive the asymptotic properties of the SLSE in

GLMM. Second, data outliers are common in longitudinal data. If no action is implemented to deal with these outliers, they may distort an analysis completely and lead to inappropriate conclusions. One of the concerns for SLSE is that the second moments used in the estimation procedure may enlarge the outlier impact. We investigate the robustness property of SLSE by means of the influence function, and show that the SLSE has a bounded influence function. Simulation studies are performed to confirm this robustness property. Third, our preliminary simulation studies, and the simulation studies in Wang (2007) indicate that there are some finite sample biases for the estimation of variance components. These biases are downward-oriented and diminish with increasing sample sizes. We study the source of this finite-sample bias and proposed a bias reduction technique by using independent weights. Forth, longitudinal studies often feature incomplete data. Problems arise if the missing data mechanism depends on the response process. We extend the SLSE to accommodate response data missing at random by either adapting the inverse probability weight method or applying the multiple imputation approach. Fifth, data measured with error are very common in longitudinal studies. Such data can cause significant difficulties in deriving correct results and interpretation. We propose the method of moment estimators for the generalized linear mixed models with measurement error using the instrumental variable approach.

The thesis is organized as follows. Chapter 2 focuses on the estimation of linear mixed model which is a special class of the GLMM. In Section 2.1,

we conduct a brief literature review on the existing estimation methodologies in linear mixed model. Section 2.2 introduces the SLSE and gives its consistency and asymptotic normality. We also discuss the implementation of SLSE and investigate its robust property against data outliers here. Numerical studies are examined to compare the finite sample performance of the proposed estimator with the maximum likelihood estimator under various scenarios in Section 2.3. The robustness property of the proposed method against data contamination is also demonstrated through simulation studies in this section. A real data application is illustrated in Section 2.4.

Chapter 3 proposes the simulation-based estimator (SBE) for the estimation of GLMM. In Section 3.1, we introduce the model and conduct a brief literature review on the estimation methodologies in GLMM. Section 3.2 discusses the model identifiability based on the first two marginal moments and introduces the simulation-based estimator. In Section 3.3, we conduct simulation studies to compare finite sample performances of the SBE with the quasi-likelihood estimator. A real data application is given in Section 3.4. Section 3.5 reviews the missing data problems in longitudinal data and proposes to accommodate response data missing at random by either adapting the inverse probability weight method or applying the multiple imputation approach. Monte Carlo simulation results are also reported in this section.

In Chapter 4, we introduce the linear mixed model with measurement error and review some existing estimation methods. We propose the method

of instrumental variable approach for the classical additive measurement error model estimation in Section 4.2. Here we establish theoretical results of the proposed estimator by assuming a known linear relationship between instrumental variables and measurement error variables. Section 4.3 examines an alternative model with a Berkson-type measurement on covariates. We investigate the finite sample performances of the proposed estimators in comparison with the naive maximum likelihood estimator in Section 4.4. Section 4.5 includes a simulation study based on a real data application.

In Chapter 5, we propose the method of moment estimators for the generalized linear mixed model with covariate measurement error using the instrumental variable approach. Section 5.2 introduces the model and the proposed estimation procedure. A nonlinear regression relationship between the instrumental variable and measurement error variables is assumed, and the asymptotic covariance matrix of the proposed estimator is derived by accounting for the estimation error of the regression/nuisance parameters. In Section 5.3, we construct the simulation-based estimator for the case where the closed forms of the marginal moments do not exist. In Section 5.4, we present simulation studies of finite sample performances of the proposed estimators. Chapter 6 briefly summarizes overall findings and outlines possible extensions for future work. The proofs of the theorems are given in the Appendices.

Chapter 2

Second-order Least Squares Estimation in Linear Mixed Models

2.1 Introduction

Linear mixed models (LMM, Laird and Ware 1982) are a common framework used to analyze repeatedly measured and clustered data which arise in many areas, such as medical and biological sciences, epidemiology, agriculture, social and environmental sciences. For subject i ($i = 1, \cdots, N$) being observed or measured repeatedly on n_i occasions, a linear mixed model can be expressed as

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{2.1}$$

where y_i is the $n_i \times 1$ vector of responses, β is a $p \times 1$ vector of the fixed population effects, and b_i is a $q \times 1$ vector of i^{th} subject's random effects

and follows a certain distribution with mean 0 and covariance $D(\theta)$. $D(\theta)$ is a $q \times q$ positive-definite covariance matrix depending on a $r \times 1$ vector of parameters θ. X_i and Z_i are the $n_i \times p$ and $n_i \times q$ design matrices to link β and b_i to y_i respectively. ϵ_i is the $n_i \times 1$ vector of residual error terms following a certain distribution with mean 0 and covariance $\sigma^2 I_{n_i}$. Also, all random vectors $\{b_i, \epsilon_i, i = 1, \cdots, N\}$ are assumed mutually independent.

For the estimation and inference of LMM, the most frequently employed approach is the maximum likelihood (ML) approach. Assume both the random effects and the residual errors are normally distributed. The marginal distribution of y_i is multivariate normal with mean $X_i\beta$ and variance $\sigma^2 I + Z_i D Z_i^T$. Assuming independence across subjects, the log-likelihood function is given by

$$l(\beta, \alpha) = c - \sum_{i=1}^{N} \frac{1}{2} \log(|\Lambda_i|) - \sum_{i=1}^{N} \frac{1}{2} (Y_i - X_i\beta)^T \Lambda_i^{-1} (Y_i - X_i\beta), \qquad (2.2)$$

where c is a constant and $\Lambda_i(\alpha) = \sigma^2 I + Z_i D Z_i^T$ depends on an unknown vector $\alpha = (\theta', \sigma^2)'$ of parameters. Estimation of $\psi = (\beta', \alpha')'$ requires joint maximization of (2.2) using numerical optimization technique such as the Newton-Raphson algorithm. In general, there is no analytic solutions available. However, if assume Λ_i is known, we can obtain the maximum likelihood estimator of β as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = (\sum_{i=1}^{N} X_i^T \Lambda_i^{-1} X_i)^{-1} \sum_{i=1}^{N} X_i^T \Lambda_i^{-1} Y_i. \qquad (2.3)$$

Since we usually do not know Λ_i, we typically estimate it from the data

using the MLE. In general, it is not possible to write down simple expressions for the ML estimate of Λ_i. The ML estimate of Λ_i has to be found by using numerical algorithms that maximize the likelihood. Once the ML estimate of Λ_i has been obtained, we simply substitute the estimate of Λ_i, say $\hat{\Lambda}_i$, to obtain the ML estimate of β. Because $\hat{\boldsymbol{\beta}}$ is estimated by maximum likelihood estimation method, the asymptotic covariance matrix of $\hat{\beta}$ is the inverse of the observed Hessian matrix at the optimum $-\partial l^2(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$, i.e,

$$\text{var}(\hat{\beta}) = (\sum_{i=1}^{N} X_i^T \Lambda_i^{-1} X_i)^{-1} \tag{2.4}$$

A criticism of the ML estimators for the variance components is that they are biased downward because they do not take into account the loss in degrees of freedom from the estimation of β. The method of residual or restricted maximum likelihood (REML) (Patterson and Thompson 1971) estimation was developed to address this problem. The main idea behind REML is to estimate the parameters of main interest without having to deal with the nuisance parameters. One possible way to obtain the restricted likelihood is to consider transformations of the data to a set of linear combinations of observations that have a distribution that does not depend on β. When the residual likelihood is maximized, we obtain estimates of Λ_i whose degrees of freedom are corrected for the reduction in degrees of freedom due to estimating α. That is, the extra determinant term effectively makes a correction or adjustments that is analogous to the correction to the denominator in Λ_i. If β is estimated by the MLE condition on α, then REML maximizes

the following slightly modified log-likelihood to obtain $\hat{\alpha}$

$$\begin{aligned} l(\hat{\beta}, \alpha) &= c - \sum_{i=1}^{N} \frac{1}{2} \log(|\Lambda_i|) - \sum_{i=1}^{N} \frac{1}{2} (Y_i - X_i\hat{\beta})^T \Lambda_i^{-1} (Y_i - X_i\hat{\beta}) \\ &\quad - \frac{1}{2} \log |\sum_{i=1}^{N} X_i^T \Lambda_i^{-1} X_i|. \end{aligned} \tag{2.5}$$

A comprehensive overview of the likelihood estimation algorithm and its properties can be found in Demidenko (2004) and Jiang (2007). In general, the computation of likelihood function is not simple and relies on Gaussian assumption for both random effects and residual error terms. Since the random effects are unobservable, it is not feasible to verify their distributional assumptions. It is thus natural to be concerned whether these methods yield reliable results when the Gaussian assumption is not appropriate. Several extensions of the LMM have been proposed to relax the Gaussian assumption for the random effects (e.g., Verbeke and Lesaffre 1997; Zhang and Davidian 2001; Lin and Lee 2008). However, these works still assume the distribution of residual errors to be normal, and impose certain parametric assumptions for random effects distribution, such as Student-t, mixture-normal or skew-normal. On the other hand, quasi-likelihood seems to be a viable solution since it does not require distributional assumptions on random effects or residual errors. However, since it is asymptotically equivalent to the ML method for LMM estimation (Wu, Gumpertz and Boos 2001; Jiang 2007), it suffers from lack of robustness against departure from Gaussian assumption just like the ML method.

Moreover, by assuming the distributions of random effects and residual errors to be Gaussian, it makes ML estimator vulnerable to data contamination or outliers (Pinheiro, Liu and Wu 2001). A few robust likelihood techniques have been proposed by implementing certain symmetric and long-tailed distributions, such as the Student-t distribution with low degrees of freedom (e.g., Lange, Little and Taylor 1989; Pinheiro, Liu and Wu 2001). However, to carry out this approach, one needs to know the degrees of freedom. Gill (2000) used the Huber function with a known c. The problem with this approach is the determination of c. Preisser and Qaqish (1996) suggested downweighting and deleting contaminated clusters for the generalized linear mixed models. Similarly, Christensen, Pearson and Johnson (1992) considered a case-deletion diagnostics for detecting influential observations in LMM. Both approaches require the identification of influential observations beforehand and remove them from data analysis. Richardson (1997) proposed a robust estimation in LMM with variance components only using the bounded influence estimator. Yau and Kuk (2002) proposed an approximate robust method based on the notion of ML for LMM. However, this method may lead to inefficient estimates of the regression coefficients and variance components.

2.2 Second-order Least Squares Estimation

2.2.1 Estimation and Inference

For subject i at a given occasion j, the LMM can be written as

$$y_{ij} = x'_{ij}\beta + z'_{ij}b_i + \epsilon_{ij}, \tag{2.6}$$

where x'_{ij} and z'_{ij} are the j^{th} rows of the design matrixes X_i and Z_i, respectively. The closed form of the first two marginal moments of the response in model (2.6) are

$$E(y_{ij}|X_i, Z_i) = x'_{ij}\beta, \tag{2.7}$$

$$E(y_{ij}y_{ik}|X_i, Z_i) = (x'_{ij}\beta)(x'_{ik}\beta) + z'_{ij}D(\theta)z_{ik} + \delta_{jk}\sigma^2, \tag{2.8}$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. Note that the derivation of the first two marginal moments dose not require any parametric assumption for the distribution of random effects or error terms.

Let $\psi = (\beta',\ \theta',\ \sigma^2)'$ and the parameter space $\Gamma = \Omega \times \Theta \times \Sigma \subset I\!R^{p+r+1}$. Following Wang (2007), the SLSE $\hat{\psi}_N$ for ψ is defined as the measurable function that minimizes

$$Q_N(\psi) = \sum_{i=1}^{N} \rho'_i(\psi) W_i \rho_i(\psi), \tag{2.9}$$

where $\rho_i(\psi) = (y_{ij} - \mu_{ij}(\psi),\ 1 \le j \le n_i,\ y_{ij}y_{ik} - \eta_{ijk}(\psi),\ 1 \le j \le k \le n_i)'$, $\mu_{ij}(\psi) = E(y_{ij}|X_i, Z_i)$, $\eta_{ijk}(\psi) = E(y_{ij}y_{ik}|X_i, Z_i)$ and $W_i = W(X_i, Z_i)$ is a nonnegative definite matrix of dimension $n_i(n_i + 3)/2$.

The following assumptions are used for the proof of the consistency and asymptotic properties of $\hat{\psi}_N$.

Assumption 2.2.1. (y_i, X_i, Z_i, n_i), $i = 1, ..., N$ are independent and identically distributed and satisfy $E\|W_i\|\left(y_{ij}^4 + \|x_{ij}\|^4 + \|z_{ij}\|^4 + 1\right) < \infty$, where $\|\cdot\|$ denotes the Euclidean norm.

Assumption 2.2.2. The parameter space $\Gamma \subset I\!R^{p+r+1}$ is compact.

Assumption 2.2.3. $E\left[(\rho_i(\psi) - \rho_i(\psi_0))' W_i (\rho_i(\psi) - \rho_i(\psi_0))\right] = 0$ if and only if $\psi = \psi_0$.

Assumption 2.2.4. The matrix $B = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right]$ is nonsingular.

These are common assumptions in the literature of linear models. In particular, assumptions 2.2.1 and 2.2.2 ensure that $Q_N(\psi)$ uniformly converges to $Q(\psi) = E\rho_i'(\psi) W_i \rho_i(\psi)$. Assumption 2.2.3 is a high-level identification condition to guarantee that $Q(\psi)$ attains a unique minimum at the true parameter value $\psi_0 \in \Gamma$. A sufficient condition for assumption 2.2.3 is that the matrix $\sum X_i' X_i$ is nonsingular with $\sum n_i > p$ and at least one matrix $Z_i' Z_i$ is positive definite with $\sum_{i=1}^{N} (n_i - q) > 0$, provided all random variables in the model are normally distributed (Demidenko 2004). Finally, assumption 2.2.4 is necessary for the existence of the variance of $\hat{\psi}_N$. In addition, the first partial derivative is given by

$$\frac{\partial \rho_i'(\psi)}{\partial \psi} = -\left(\frac{\partial \mu_{ij}(\psi)}{\partial \psi}, 1 \le j \le n_i, \frac{\partial \eta_{ijk}(\psi)}{\partial \psi}, 1 \le j \le k \le n_i\right),$$

with

$$\begin{aligned}\frac{\partial \mu_{ij}(\psi)}{\partial \psi} &= (x_{ij},\ 0,\ 0)', \\ \frac{\partial \eta_{ijk}(\psi)}{\partial \psi} &= \left((x_{ij}x'_{ik} + x_{ik}x'_{ij})\beta,\ \frac{\partial \mathrm{vec}(D)}{\partial \theta}\mathrm{vec}(z_{ij}z'_{ik}),\ \delta_{jk}\right)'.\end{aligned}$$

Theorem 2.2.1. *Under assumptions 2.2.1-2.2.3, as* $N \to \infty$, $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$.

Theorem 2.2.2. *Under assumptions 2.2.1-2.2.4, as* $N \to \infty$, $\sqrt{N}(\hat{\psi}_N - \psi_0) \xrightarrow{L} N(0, B^{-1}CB^{-1})$, *where*

$$B = E\left[\frac{\partial \rho'_i(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right] \tag{2.10}$$

and,

$$C = E\left[\frac{\partial \rho'_i(\psi_0)}{\partial \psi} W_i \rho_i(\psi_0)\rho'_i(\psi_0) W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right]. \tag{2.11}$$

Furthermore, with probability one,

$$B = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \rho'_i(\hat{\psi}_N)}{\partial \psi} W_i \frac{\partial \rho_i(\hat{\psi}_N)}{\partial \psi'}\right]$$

and

$$C = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \rho'_i(\hat{\psi}_N)}{\partial \psi} W_i \rho_i(\hat{\psi}_N)\rho'_i(\hat{\psi}_N) W_i \frac{\partial \rho_i(\hat{\psi}_N)}{\partial \psi'}\right].$$

2.2.2 Computation

In general, there is no explicit solution for the SLSE. The iterative Newton-Raphson algorithm could be used to compute SLSE, that is,

$$\hat{\psi}^{(t+1)} = \hat{\psi}^{(t)} - \left[\frac{\partial^2 Q_N(\hat{\psi}^{(t)})}{\partial \psi \partial \psi'}\right]^{-1} \frac{\partial Q_N(\hat{\psi}^{(t)})}{\partial \psi},$$

where $\hat{\psi}^{(t)}$ denotes the estimate of ψ at the t^{th} iteration,

$$\begin{aligned}
\frac{\partial Q_N(\hat{\psi}^{(t)})}{\partial \psi} &= 2\sum_{i=1}^{N} \frac{\partial \rho_i'(\hat{\psi}^{(t)})}{\partial \psi} W_i \rho_i(\hat{\psi}^{(t)}), \text{ and} \\
\frac{\partial^2 Q_N(\hat{\psi}^{(t)})}{\partial \psi \partial \psi'} &= 2\sum_{i=1}^{N} \left[\frac{\partial \rho_i'(\hat{\psi}^{(t)})}{\partial \psi} W_i \frac{\partial \rho_i(\hat{\psi}^{(t)})}{\partial \psi'} + (\rho_i'(\hat{\psi}^{(t)}) W_i \otimes I) \frac{\partial \text{vec}(\partial \rho_i'(\hat{\psi}^{(t)})/\partial \psi)}{\partial \psi'} \right].
\end{aligned}$$

In the above equation, since the term $(\rho_i'(\hat{\psi}^{(t)})W_i \otimes I)\frac{\partial \text{vec}(\partial \rho_i'(\hat{\psi}_t)/\partial \psi)}{\partial \psi'}$ has expectation zero, it can be ignored from the second derivative. Therefore, we have the following Newton-Raphson algorithm

$$\hat{\psi}^{(t+1)} = \hat{\psi}^{(t)} - \left[\sum_{i=1}^{N} \frac{\partial \rho_i'(\hat{\psi}^{(t)})}{\partial \psi} W_i \frac{\partial \rho_i(\hat{\psi}^{(t)})}{\partial \psi'} \right]^{-1} \sum_{i=1}^{N} \frac{\partial \rho_i'(\hat{\psi}^{(t)})}{\partial \psi} W_i \rho_i(\hat{\psi}^{(t)}). \quad (2.12)$$

For the choice of initial values in (2.12), we can use the so-called method of moments estimates or maximum likelihood estimates. To avoid the complexity of finding the derivatives of $Q_N(\psi)$, we can also choose the Nelder-Mead simplex method (Nelder and Mead 1965) to minimize the quadratic inference function $Q_N(\psi)$ to obtain $\hat{\psi}$. Another question is how to specify the form of weight W_i to carry out the SLSE. In theory, W_i only depends on X_i and Z_i, and any form of W_i satisfying the regularity conditions is valid for the SLS estimator. However, it would be desirable to make inferences based on the more precise estimator, so the optimal choice of W_i is the one which yields the minimum variance-covariance matrix of $\hat{\psi}_N$. This choice is given in the following theorem and has been proved in Abarin and Wang (2006).

Theorem 2.2.3. *Denote* $U_i = E[\rho_i(\psi_0)\rho_i'(\psi_0)|X_i, Z_i]$. *Then the minimum*

asymptotic variance-covariance matrix of $\hat{\psi}_N$ *is*

$$E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} U_i^{-1} \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right]$$

and this is obtained by setting $W_i = U_i^{-1}$.

In practice, the calculation of W_i is not feasible since it involves unknown parameters which need to be estimated first. One of the possible solution is using a two-stage procedure. First, minimize $Q_N(\psi)$ using a sub-optimal choice of W_i, such as an identity matrix, to obtain the first stage estimator $\hat{\psi}_{N1}$. Second, estimate U_i using $\hat{\psi}_{N1}$ and then minimize $Q_N(\psi)$ again with $W_i = \hat{U}_i^{-1}$ to obtain the second stage estimator $\hat{\psi}_{N2}$. In theory, $\hat{\psi}_{N2}$ is asymptotically more efficient than $\hat{\psi}_{N1}$ because $\hat{\psi}_{N2}$ has the minimum asymptotic variance-covariance matrix given in Theorem 2.2.3. In general, U_i can be estimated using any nonparametric method, such as kernel or spline estimators. However, in some cases, a simpler estimator of U_i would be

$$\hat{U} = \frac{1}{N} \sum_{i=1}^{N} \rho_i(\hat{\psi}_{N1}) \rho_i'(\hat{\psi}_{N1}). \tag{2.13}$$

In many real data applications, the subjects are clustered so that the values of X_i, Z_i are equal for all subjects within one cluster. In such cases each U_i can be estimated similarly to (2.13) using all the subjects within the same cluster. Since $\hat{U}_i$ is of dimension $n_i(n_i+3)/2$, numerical inversion of $\hat{U}_i$ may be difficult when n_i is large. In this case, one may consider using diagonal or certain block diagonal sub-matrix of U_i. In section 2.3, we conduct extensive simulation studies to investigate the sensitivity and efficiency of SLSE by using different specifications of the weight matrix.

2.2.3 Robustness

Outliers are common in experimental research data for reasons such as transcription error or technical equipment malfunction. If no action is implemented, such outliers may distort an analysis completely and lead to wrong conclusions. In mixed models, outliers may happen not only at the level of within-subject error but also at the level of within-subject variations. Sometimes they are referred to as e- and b-outliers respectively (Pinheiro, Liu and Wu 2001).

Here we study the robustness property of SLSE by means of the influence function (IF), which was introduced by Hampel, Ronchetti, Rousseeuw and Stahel (1986). The essential concept of IF is that one can use it to assess the asymptotical bias of the estimator caused by a certain degree of data contamination. The estimator is robust if the IF is bounded (Huber 2004). In principle, the SLSE is an M-estimator (Huber 2004) and minimizing the quadratic distance function (2.9) with optimal weight matrix in (2.13) is asymptotically equivalent to solving the equation

$$\sum_{i=1}^{N} \frac{\partial \rho_i'(\psi)}{\partial \psi} W_i \rho_i(\psi) = 0. \tag{2.14}$$

It follows from Hampel, Ronchetti, Rousseeuw and Stahel (1986) that when $N \to \infty$, the IF of the SLSE at point $v = (x_l, z_l)'$ is

$$\mathrm{IF}(v; \hat{\psi}_N, F) = -B(\hat{\psi}_N)^{-1} G(v; \hat{\psi}_N, F) \tag{2.15}$$

where F is the underlying distribution and B is given in (2.10), and

$$G(v; \hat{\psi}_N, F) = \frac{\partial \rho_l'(\hat{\psi}_N)}{\partial \psi} W_i \rho_l(\hat{\psi}_N). \tag{2.16}$$

If $\hat{\psi}_N$ is computed using the estimated optimal weight (2.13), we can show that as $\|v\| \to \infty$ $\left\|\text{IF}(v, \hat{\psi}_N)\right\| \to 0$. In particular, we have the following theorem:

Theorem 2.2.4. *If the SLSE $\hat{\psi}_N$ is computed using estimated optimal weight (2.13), then $\left\|IF(v, \hat{\psi}_N)\right\| \to 0$ as $\|v\| \to \infty$.*

The above result implies that the SLSE $\hat{\psi}_{mN}$ is a redescending M-estimator (Huber 2004). The implication of the redescending property means that the SLSE is able to reject extreme outliers completely. Intuitively, it is expected that the outlier will be automatically downweighted by the inverse of the optimal weight matrix U_l in the estimating equation (2.14). It does not require to screen data for outliers and make a subjective decision to exclude them from the analysis. This is practically meaningful because an outlier may be an indication of a problem with the data generation process but more importantly it may be a true unusual observation about reality.

2.3 Monte Carlo Simulation Studies

In this section, we carry out substantial simulation studies (1) to examine finite sample behavior of the SLSE; (2) to evaluate and compare the robustness of SLSE with restricted maximum likelihood (REML) estimator under

misspecified random effects and residual error distributions; (3) to investigate the sensitivity and efficiency of SLSE by using different specifications of the weight; and (4) to demonstrate the robustness of SLSE against outliers. We considered the following two linear mixed models commonly used to study the growth curves (Demidenko 2004; Jacqmin-Gadda et al. 2006):

1. random intercept (RI) model: $y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{i1} + \epsilon_{ij}$;
2. random intercept and slope (RIS) model: $y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{i1} + b_{i2} x_{ij} + \epsilon_{ij}$.

The following configurations are used for simulation:

- $N = 20, 50, 100, 200, 300, 400, 500$; $n = 4$ or 8; and $x_{ij} = j$, $j = 1, \cdots, n$;
- b_{i1}, b_{i2} and ϵ_{ij} are all generated independently from one of the following distribution: Gaussian, $\chi^2(3)$ and student's $t(4)$ distributions with mean 0 and variance θ_{11}, θ_{22} and σ^2 respectively;
- $\beta_1 = 8$, $\beta_2 = 2$, $\theta_{11} = 1.96$, $\theta_{22} = 1$ and $\sigma^2 = 1$.

All computations are done in R and the restricted maximum likelihood (REML) estimates are obtained from `lme` package. The SLSEs are computed using three different weight matrices:

1. identity weight (SLS1);

2. diagonal of the estimated optimal weight (2.13) (SLS2);

3. fully estimated optimal weight (2.13) (SLS3).

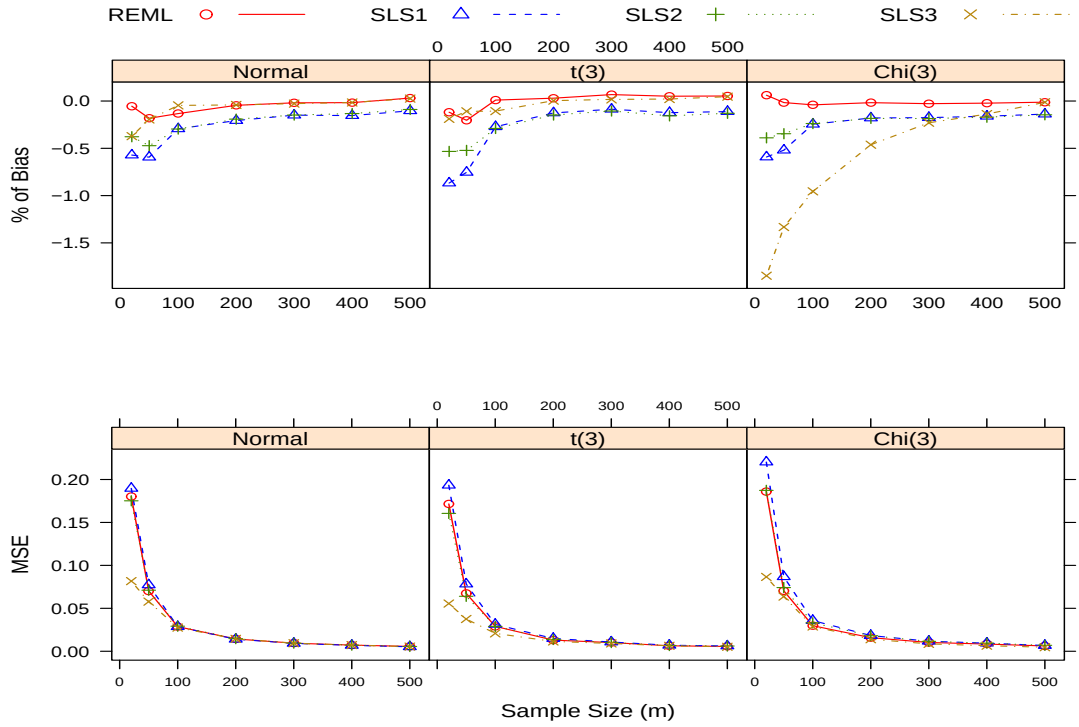
To determine how well the methods perform, we present the estimation bias and mean squared errors (MSE) of the estimators. For each model, 1000 Monte Carlo replications were carried out. For fair comparisons, the same dataset was used to obtain both REML estimates and SLS estimates, at each replication. To eliminate potential nonlinear numerical optimization problems on the selection of starting points, the true parameter values were used as starting values for the minimization and the optimal weight calculation for SLS method.

2.3.1 Robustness against Distribution Misspecification

The Monte Carlso simulation results are provided in Table 2.1 and Table 2.2. Since the relative performances of the estimates are similar for RI and RIS model, in consideration of space and clarity, we concentrate our discussion on the simulation results for the RI model. Overall simulation results in all sample sizes are summarized in Figure 2.1 - Figure 2.4. These figures contain the absolute value of estimate bias and MSE under correctly specified as well as misspecified models.

Fig 2.1 and Fig 2.2 depict the performance of SLS and REML methods for fixed effects. They show all Monte Carlo mean estimates are close to the true parameter values and no apparent biases are observed across all

Figure 2.1: Bias and MSE of β_1 from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors



methods. This is not surprising as a few simulations studies (e.g., Verbeke and Lesaffre 1997; Jacqmin-Gadda et al. 2006) have shown that maximum likelihood inference on fixed effects is robust to misspecified LMM. At relative small sample size ($N = 20, 50, 100$), SLS2 and SLS3 have lower MSE than REML and SLS1. As sample size increases from 200 to 500, all four methods behave very closely.

Figure 2.2: Bias and MSE of β_2 from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors

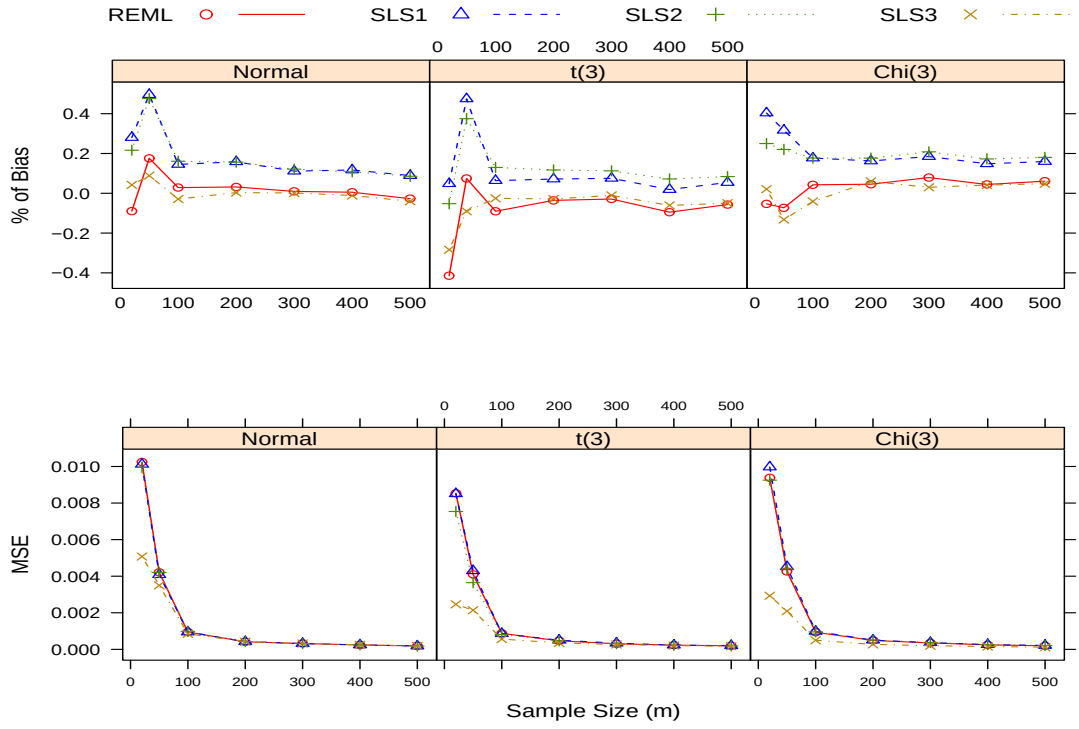


Fig 2.3 depicts the performance of estimators for the random effect. Under Gaussian assumption, the bias from REML is very trivial and much

smaller than the ones from all three SLS estimates; however, when the model is misspecified, there is a noticeably bias increase in REML estimates, especially in small sample sizes. In all cases, SLS estimates show much smaller MSE than REML, particulary when the model is misspecified. The variance and MSE reduction in the misspecified model can be as high as $70 - 80\%$ in some instances. Additionally, the simulation results suggest that SLS3 estimates have some downward bias, although this bias decrease with the increase of sample size. SLS1 appears to be biased at sample size 20 and 50 but this bias disappears with sample size increases to 100. All mean estimates from SLS2 are close to true parameter values and no apparent biases are observed. All SLS estimates have similar MSE when model is correctly specified, but SLS3 shows a relatively higher MSE than SLS1 and SLS2 when model is misspecified. Within SLS estimates, SLS1 and SLS2 seem to be more satisfactory in terms of both bias and MSE than SLS3. This may due to the numerical error for inversion of the optimal weight matrix in SLS3.

Figure 2.3: Bias and MSE of θ_{11} from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors

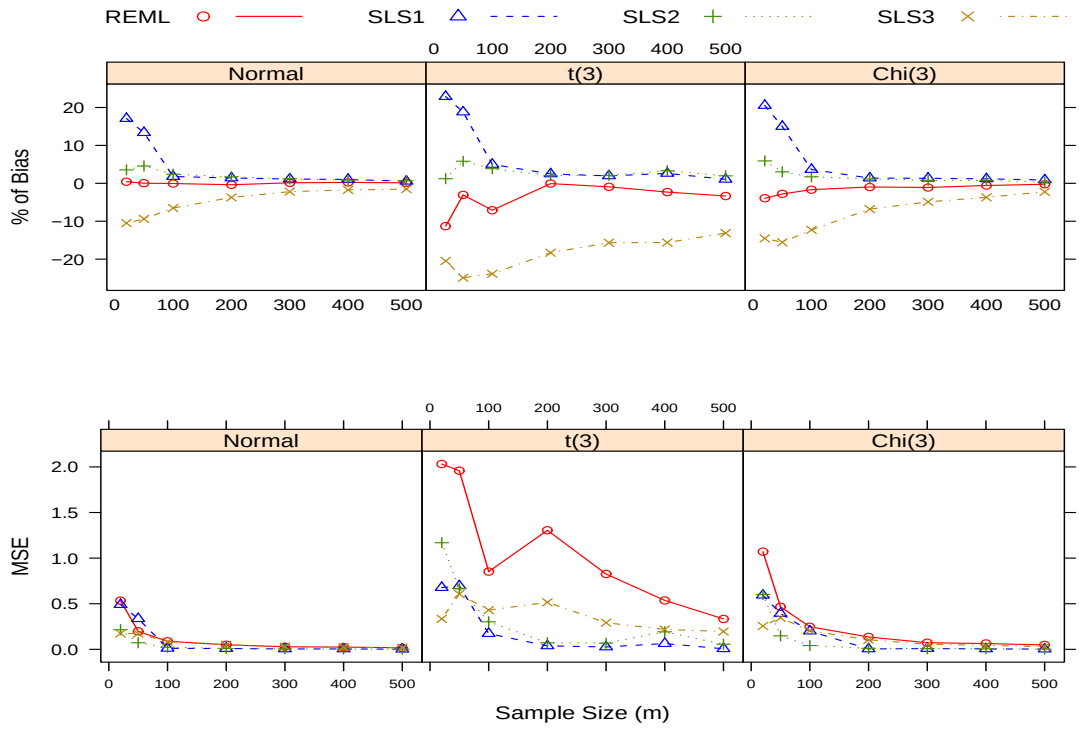


Table 2.1: Simulation results with normal and non-normal distributed random effect and residual errors based on the RI model

			RMEL			SLS1			SLS2			SLS3		
		N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
Normal	β_0	20	-0.004	0.180	0.180	-0.046	0.188	0.190	-0.030	0.174	0.175	-0.030	0.081	0.082
		50	-0.015	0.070	0.070	-0.047	0.075	0.077	-0.038	0.070	0.071	-0.016	0.058	0.058
		100	-0.011	0.029	0.029	-0.024	0.028	0.029	-0.024	0.029	0.029	-0.004	0.028	0.028
		200	-0.004	0.014	0.014	-0.017	0.014	0.014	-0.015	0.014	0.015	-0.003	0.014	0.014
		300	-0.002	0.009	0.009	-0.012	0.009	0.009	-0.012	0.009	0.010	-0.002	0.009	0.009
		400	-0.001	0.007	0.007	-0.012	0.007	0.007	-0.011	0.007	0.007	-0.001	0.007	0.007
		500	0.003	0.006	0.006	-0.008	0.005	0.005	-0.007	0.006	0.006	0.002	0.005	0.005
	β_1	20	-0.002	0.010	0.010	0.006	0.010	0.010	0.004	0.010	0.010	0.001	0.005	0.005
		50	0.004	0.004	0.004	0.010	0.004	0.004	0.010	0.004	0.004	0.002	0.003	0.003
		100	0.001	0.001	0.001	0.003	0.001	0.001	0.003	0.001	0.001	-0.001	0.001	0.001
		200	0.001	0.000	0.000	0.003	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000
		300	0.000	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000
		400	0.000	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000
		500	-0.001	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	-0.001	0.000	0.000

Continued on next page...

Table 2.1 – continued

			RMEL			SLS1			SLS2			SLS3		
		N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
	θ_{11}	20	0.009	0.533	0.534	0.335	0.378	0.490	0.069	0.210	0.215	-0.206	0.132	0.174
		50	0.001	0.197	0.197	0.262	0.267	0.336	0.089	0.064	0.072	-0.184	0.139	0.172
		100	-0.001	0.089	0.089	0.035	0.010	0.011	0.047	0.020	0.022	-0.128	0.068	0.084
		200	-0.007	0.049	0.049	0.027	0.008	0.009	0.034	0.009	0.010	-0.073	0.044	0.050
		300	0.003	0.028	0.028	0.022	0.004	0.004	0.023	0.004	0.005	-0.043	0.027	0.028
		400	0.005	0.024	0.024	0.020	0.004	0.004	0.017	0.003	0.003	-0.034	0.023	0.024
		500	0.003	0.017	0.017	0.011	0.001	0.001	0.013	0.002	0.002	-0.030	0.016	0.017
	ϕ	20	0.004	0.035	0.035	0.099	0.041	0.050	-0.018	0.040	0.041	-0.184	0.009	0.043
		50	0.003	0.013	0.013	0.083	0.031	0.038	0.024	0.011	0.012	-0.146	0.009	0.030
		100	0.002	0.005	0.005	0.011	0.001	0.001	0.014	0.002	0.002	-0.094	0.004	0.013
		200	0.003	0.002	0.002	0.009	0.001	0.001	0.010	0.001	0.001	-0.052	0.002	0.005
		300	0.001	0.002	0.002	0.006	0.000	0.000	0.007	0.000	0.000	-0.037	0.002	0.003
		400	0.001	0.001	0.001	0.006	0.000	0.000	0.005	0.000	0.000	-0.029	0.001	0.002
		500	0.002	0.001	0.001	0.003	0.000	0.000	0.004	0.000	0.000	-0.022	0.001	0.001
t(4)	β_0	20	-0.009	0.171	0.172	-0.069	0.189	0.193	-0.043	0.159	0.161	-0.015	0.055	0.056
		50	-0.016	0.067	0.067	-0.060	0.075	0.078	-0.042	0.062	0.064	-0.009	0.037	0.038

Continued on next page...

Table 2.1 – continued

		RMEL			SLS1			SLS2			SLS3		
	N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
	100	0.001	0.029	0.029	-0.022	0.031	0.031	-0.023	0.028	0.029	-0.009	0.021	0.021
	200	0.002	0.013	0.013	-0.010	0.015	0.015	-0.012	0.014	0.014	0.000	0.011	0.011
	300	0.005	0.010	0.010	-0.007	0.011	0.011	-0.008	0.010	0.010	0.001	0.009	0.009
	400	0.004	0.006	0.006	-0.010	0.007	0.007	-0.013	0.007	0.007	0.002	0.006	0.006
	500	0.004	0.006	0.006	-0.009	0.006	0.006	-0.010	0.006	0.006	0.004	0.005	0.005
β_1	20	-0.008	0.008	0.009	0.001	0.009	0.009	-0.001	0.008	0.008	-0.006	0.002	0.002
	50	0.001	0.004	0.004	0.009	0.004	0.004	0.008	0.004	0.004	-0.002	0.002	0.002
	100	-0.002	0.001	0.001	0.001	0.001	0.001	0.003	0.001	0.001	-0.001	0.001	0.001
	200	-0.001	0.000	0.000	0.001	0.000	0.001	0.002	0.000	0.000	-0.001	0.000	0.000
	300	-0.001	0.000	0.000	0.002	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000
	400	-0.002	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	-0.001	0.000	0.000
	500	-0.001	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.000	-0.001	0.000	0.000
θ	20	-0.221	1.982	2.031	0.448	0.476	0.677	0.024	1.168	1.169	-0.401	0.173	0.334
	50	-0.060	1.955	1.958	0.368	0.560	0.696	0.114	0.655	0.668	-0.488	0.368	0.606
	100	-0.139	0.833	0.852	0.098	0.163	0.172	0.076	0.296	0.302	-0.468	0.211	0.430
	200	-0.002	1.305	1.305	0.049	0.036	0.038	0.040	0.070	0.072	-0.359	0.387	0.516

Continued on next page...

Table 2.1 – continued

			RMEL			SLS1			SLS2			SLS3		
		N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
		300	-0.018	0.825	0.825	0.038	0.024	0.026	0.041	0.066	0.068	-0.307	0.198	0.292
		400	-0.045	0.535	0.537	0.051	0.061	0.063	0.065	0.188	0.192	-0.306	0.125	0.218
		500	-0.066	0.329	0.333	0.021	0.006	0.006	0.039	0.055	0.056	-0.257	0.129	0.195
	ϕ	20	-0.079	0.211	0.217	0.079	0.192	0.198	-0.146	0.129	0.150	-0.302	0.014	0.105
		50	-0.014	0.197	0.197	0.098	0.139	0.149	-0.032	0.053	0.054	-0.376	0.009	0.150
		100	-0.013	0.160	0.160	0.033	0.071	0.073	-0.012	0.027	0.027	-0.338	0.005	0.119
		200	-0.016	0.085	0.085	0.020	0.013	0.014	-0.004	0.005	0.005	-0.279	0.006	0.083
		300	-0.016	0.059	0.059	0.016	0.009	0.010	0.001	0.005	0.005	-0.247	0.006	0.067
		400	-0.004	0.059	0.059	0.020	0.016	0.016	0.008	0.005	0.005	-0.225	0.005	0.056
		500	-0.005	0.046	0.046	0.009	0.005	0.005	0.006	0.002	0.003	-0.209	0.004	0.048
Chi(3)	β_0	20	0.005	0.186	0.186	-0.047	0.218	0.220	-0.031	0.186	0.187	-0.148	0.065	0.087
		50	-0.001	0.070	0.070	-0.042	0.085	0.087	-0.028	0.073	0.074	-0.107	0.052	0.064
		100	-0.003	0.030	0.030	-0.020	0.035	0.036	-0.019	0.033	0.033	-0.076	0.023	0.029
		200	-0.001	0.016	0.016	-0.014	0.018	0.018	-0.015	0.018	0.018	-0.037	0.013	0.014
		300	-0.002	0.010	0.010	-0.014	0.012	0.012	-0.015	0.011	0.012	-0.018	0.008	0.009
		400	-0.002	0.008	0.008	-0.013	0.009	0.009	-0.014	0.009	0.009	-0.011	0.006	0.006

Continued on next page...

Table 2.1 – continued

		RMEL			SLS1			SLS2			SLS3		
	N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
	500	-0.001	0.006	0.006	-0.011	0.007	0.007	-0.012	0.007	0.007	-0.001	0.005	0.005
β_1	20	-0.001	0.009	0.009	0.008	0.010	0.010	0.005	0.009	0.009	0.000	0.003	0.003
	50	-0.001	0.004	0.004	0.006	0.004	0.005	0.004	0.004	0.004	-0.003	0.002	0.002
	100	0.001	0.001	0.001	0.004	0.001	0.001	0.004	0.001	0.001	-0.001	0.000	0.000
	200	0.001	0.000	0.000	0.003	0.000	0.001	0.004	0.001	0.001	0.001	0.000	0.000
	300	0.002	0.000	0.000	0.004	0.000	0.000	0.004	0.000	0.000	0.001	0.000	0.000
	400	0.001	0.000	0.000	0.003	0.000	0.000	0.003	0.000	0.000	0.001	0.000	0.000
	500	0.001	0.000	0.000	0.003	0.000	0.000	0.004	0.000	0.000	0.001	0.000	0.000
θ	20	-0.077	1.066	1.072	0.402	0.431	0.593	0.116	0.589	0.602	-0.285	0.174	0.255
	50	-0.055	0.463	0.466	0.294	0.307	0.393	0.059	0.144	0.148	-0.305	0.253	0.346
	100	-0.033	0.246	0.247	0.071	0.195	0.200	0.034	0.040	0.041	-0.241	0.146	0.204
	200	-0.019	0.135	0.135	0.027	0.003	0.004	0.023	0.011	0.011	-0.133	0.087	0.105
	300	-0.022	0.072	0.072	0.026	0.008	0.009	0.016	0.006	0.007	-0.096	0.047	0.056
	400	-0.011	0.063	0.063	0.023	0.004	0.004	0.015	0.006	0.006	-0.072	0.039	0.044
	500	-0.005	0.048	0.048	0.017	0.002	0.003	0.006	0.001	0.001	-0.044	0.026	0.028

Continued on next page...

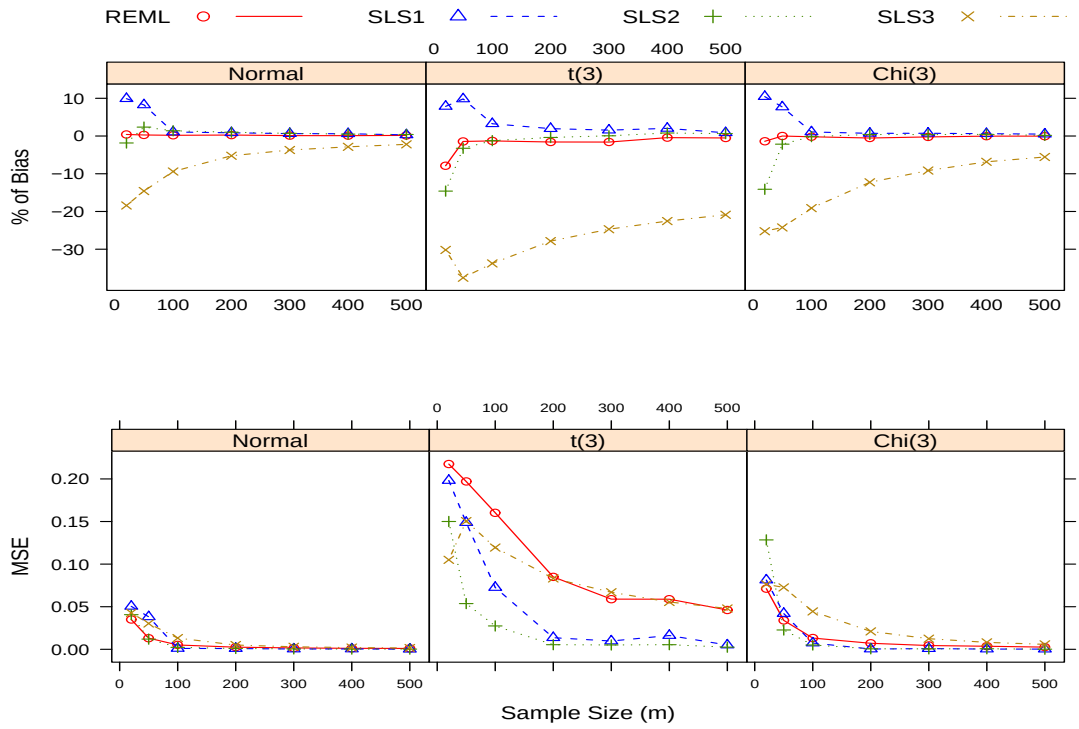
Table 2.1 – continued

		RMEL			SLS1			SLS2			SLS3		
	N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
ϕ	20	-0.014	0.071	0.071	0.104	0.070	0.081	-0.141	0.109	0.128	-0.252	0.013	0.076
	50	0.000	0.034	0.034	0.077	0.036	0.042	-0.022	0.022	0.023	-0.242	0.014	0.073
	100	-0.002	0.013	0.013	0.011	0.007	0.007	-0.002	0.005	0.005	-0.191	0.008	0.044
	200	-0.005	0.007	0.007	0.007	0.000	0.000	0.004	0.001	0.001	-0.123	0.006	0.021
	300	-0.002	0.004	0.004	0.007	0.001	0.001	0.004	0.000	0.001	-0.092	0.004	0.012
	400	0.000	0.004	0.004	0.006	0.000	0.000	0.004	0.001	0.001	-0.068	0.003	0.008
	500	0.000	0.003	0.003	0.005	0.000	0.000	0.002	0.000	0.000	-0.056	0.003	0.006

Fig 2.4 summarizes the performance of estimators for the residual error variance. Similar to the results from random effect, the bias from REML is smaller than the ones from all three SLS estimates in all models. SLS3 estimates have some downward bias in the correctly specified model and this bias gets bigger in the misspecified model, even though this bias decreases with the increase of sample size. The similar finite bias was also observed in the limited stimulation studies by Wang (2007). As a result, this finite bias contributes significantly to its higher MSE. When comparing SLS1 and SLS2, a similar pattern is observed as the random effects. In particular, SLS1 and SLS2 performs much better than REML in terms of MSE under misspecified models, and the variance and MSE reduction in the misspecified model can be more than 70% in some instances. SLS1 appears to have a slightly higher bias than SLS2 at sample size 20 to 100 but there is no clear pattern with the increase of sample size.

Overall, the simulation results demonstrate that all methods show their finite sample properties, as with the increasing number of sample size, their MSE decrease and precision increase. Moreover, the bias and/or variance from REML estimates of random effects has significant increase from Gaussian to non-Gaussian LMM; however, they remain relatively stable for SLS estimates. This confirms our assumptions that SLS estimator is superior to REML for misspecified models because it does not rely on any parametric assumptions of random effects or residual errors. In addition, SLS3 has shown smaller variance and MSE than REML even for Gaussian LMM. For

Figure 2.4: Bias and MSE of ϕ from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors



SLS estimates, SLS1-3 perform almost the same for fixed effects but SLS1 and SLS2 perform more satisfactory in terms of both bias and variance than SLS3, especially when the model is misspecified. This is due to the fact that in SLS1 and SLS2, the weighting matrix depend on less the parameters that are poorly estimated because of misspecification. There is some finite bias observed in the SLS3 estimates for the random effect and residual error variance but neither observed in SLS1 nor SLS2. Intuitively, this phenomenon may due to the computational complexity of inverting the full optimal weight matrix in SLS3. In comparison of SLS1 and SLS2, SLS2 demonstrates more efficiency than SLS1, especially under small sample size. Thus, it is reasonable to conclude that in practice, the diagonal of optimal matrix should be used for SLS estimation without significant loss of efficiency.

Table 2.2: Simulation results with normal and non-normal distributed random effect and residual errors based on the RIS model

			RMEL			SLS1			SLS2			SLS3		
		N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
Normal	β_0	50	-0.047	0.449	0.452	-0.051	0.439	0.441	-0.061	0.441	0.456	-0.055	0.453	0.444
		100	0.001	0.035	0.035	-0.002	0.030	0.030	-0.012	0.034	0.037	-0.002	0.037	0.034
		200	0.002	0.017	0.017	-0.009	0.015	0.015	-0.011	0.017	0.018	-0.001	0.018	0.017
		300	0.001	0.012	0.012	-0.013	0.010	0.010	-0.014	0.012	0.012	-0.001	0.012	0.012
		400	0.001	0.008	0.008	-0.012	0.007	0.007	-0.013	0.008	0.009	-0.002	0.008	0.008
		500	0.000	0.007	0.007	-0.014	0.006	0.006	-0.013	0.007	0.007	-0.004	0.007	0.007
	β_1	50	-0.018	0.047	0.047	-0.036	0.044	0.046	-0.014	0.045	0.048	-0.018	0.048	0.045
		100	0.000	0.013	0.013	-0.013	0.012	0.012	0.004	0.013	0.013	-0.001	0.013	0.013
		200	-0.001	0.006	0.006	-0.006	0.006	0.006	0.004	0.006	0.006	-0.002	0.006	0.006
		300	0.000	0.004	0.004	-0.003	0.004	0.004	0.005	0.004	0.004	0.000	0.004	0.004
		400	0.000	0.003	0.003	0.000	0.003	0.003	0.006	0.003	0.003	0.000	0.003	0.003
		500	0.004	0.002	0.002	0.003	0.002	0.002	0.009	0.002	0.003	0.003	0.002	0.002
	θ_{11}	50	-0.006	0.559	0.560	0.102	0.157	0.168	0.031	0.345	0.090	-0.142	0.089	0.365
		100	0.015	0.257	0.257	0.053	0.029	0.032	0.046	0.191	0.040	-0.068	0.037	0.196

Continued on next page...

Table 2.2 – continued

		RMEL			SLS1			SLS2			SLS3		
	N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
	200	0.009	0.125	0.125	0.045	0.101	0.103	0.018	0.102	0.011	-0.027	0.011	0.103
	300	-0.006	0.085	0.085	0.028	0.016	0.017	0.012	0.071	0.005	-0.038	0.005	0.072
	400	-0.006	0.066	0.066	0.021	0.002	0.002	0.009	0.056	0.002	-0.027	0.002	0.057
	500	0.005	0.051	0.051	0.019	0.001	0.002	0.007	0.043	0.002	-0.013	0.002	0.043
θ_{22}	50	-0.019	0.064	0.065	0.123	0.148	0.163	-0.035	0.043	0.059	-0.106	0.058	0.054
	100	0.000	0.029	0.029	0.117	0.055	0.069	-0.007	0.024	0.025	-0.053	0.025	0.026
	200	-0.005	0.015	0.015	0.068	0.027	0.031	-0.008	0.013	0.009	-0.035	0.009	0.014
	300	0.003	0.009	0.009	0.054	0.017	0.020	-0.003	0.008	0.005	-0.021	0.005	0.008
	400	0.001	0.007	0.007	0.044	0.010	0.012	-0.002	0.006	0.003	-0.014	0.003	0.006
	500	-0.002	0.006	0.006	0.041	0.008	0.009	-0.002	0.005	0.002	-0.014	0.002	0.005
ϕ	50	-0.002	0.026	0.026	0.012	0.024	0.024	0.000	0.019	0.015	-0.143	0.015	0.039
	100	0.000	0.010	0.010	0.011	0.003	0.003	0.014	0.009	0.004	-0.085	0.004	0.016
	200	0.001	0.005	0.005	0.010	0.002	0.002	0.006	0.005	0.001	-0.046	0.001	0.007
	300	0.000	0.003	0.003	0.008	0.000	0.001	0.005	0.003	0.001	-0.032	0.001	0.004
	400	-0.002	0.002	0.002	0.007	0.000	0.000	0.004	0.002	0.000	-0.027	0.000	0.003
	500	0.001	0.002	0.002	0.006	0.000	0.000	0.004	0.002	0.000	-0.019	0.000	0.002

Continued on next page. . .

Table 2.2 – continued

			RMEL			SLS1			SLS2			SLS3		
		N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
t(4)	β_0	50	-0.002	0.076	0.076	-0.002	0.070	0.070	-0.023	0.062	0.079	-0.007	0.079	0.062
		100	-0.002	0.039	0.039	-0.007	0.037	0.037	-0.018	0.036	0.042	-0.012	0.042	0.036
		200	-0.001	0.018	0.018	-0.006	0.016	0.016	-0.016	0.017	0.019	-0.007	0.019	0.017
		300	0.002	0.012	0.012	-0.006	0.011	0.011	-0.011	0.012	0.013	-0.005	0.013	0.012
		400	0.004	0.009	0.009	-0.010	0.008	0.008	-0.010	0.009	0.009	-0.001	0.009	0.009
		500	0.001	0.007	0.007	-0.011	0.007	0.007	-0.012	0.007	0.008	-0.003	0.008	0.007
	β_1	50	-0.002	0.024	0.024	-0.025	0.025	0.025	0.005	0.018	0.025	0.000	0.025	0.018
		100	-0.007	0.012	0.012	-0.022	0.013	0.013	-0.003	0.012	0.012	-0.005	0.012	0.012
		200	0.001	0.006	0.006	-0.010	0.007	0.007	0.005	0.006	0.007	0.002	0.007	0.006
		300	-0.002	0.004	0.004	-0.010	0.005	0.005	0.002	0.004	0.004	-0.002	0.004	0.004
		400	-0.003	0.003	0.003	-0.005	0.004	0.004	0.002	0.003	0.003	-0.003	0.003	0.003
		500	-0.003	0.002	0.002	-0.006	0.003	0.003	0.001	0.002	0.002	-0.004	0.002	0.002
	θ_{11}	50	0.087	1.240	1.248	0.125	0.243	0.259	0.099	0.507	0.321	-0.246	0.311	0.567
		100	0.031	0.714	0.715	0.078	0.053	0.059	0.068	0.401	0.130	-0.186	0.125	0.436
		200	0.013	0.424	0.424	0.041	0.013	0.015	0.049	0.206	0.106	-0.130	0.103	0.223

Continued on next page...

Table 2.2 – continued

			RMEL			SLS1			SLS2			SLS3		
		N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
		300	0.019	0.270	0.271	0.037	0.016	0.017	0.041	0.148	0.050	-0.074	0.048	0.153
		400	-0.001	0.180	0.180	0.024	0.004	0.005	0.022	0.119	0.014	-0.071	0.014	0.124
		500	0.013	0.141	0.141	0.024	0.004	0.004	0.025	0.099	0.020	-0.048	0.019	0.101
	θ_{22}	50	0.003	0.250	0.250	0.160	0.219	0.244	-0.046	0.078	0.228	-0.202	0.226	0.119
		100	0.005	0.189	0.189	0.126	0.096	0.112	-0.008	0.056	0.171	-0.136	0.171	0.074
		200	-0.009	0.058	0.058	0.090	0.057	0.065	-0.003	0.040	0.046	-0.080	0.046	0.046
		300	-0.006	0.046	0.046	0.079	0.046	0.052	0.003	0.028	0.038	-0.069	0.038	0.032
		400	0.004	0.050	0.050	0.053	0.026	0.029	0.004	0.024	0.040	-0.045	0.040	0.026
		500	0.005	0.033	0.033	0.056	0.032	0.035	0.009	0.020	0.025	-0.040	0.025	0.021
	ϕ	50	-0.021	0.057	0.058	0.009	0.028	0.028	0.008	0.016	0.034	-0.262	0.034	0.085
		100	-0.013	0.029	0.029	0.015	0.007	0.007	0.010	0.012	0.014	-0.194	0.014	0.050
		200	-0.001	0.024	0.024	0.015	0.004	0.004	0.013	0.009	0.004	-0.135	0.004	0.027
		300	-0.003	0.014	0.014	0.011	0.002	0.002	0.012	0.007	0.004	-0.106	0.003	0.018
		400	-0.004	0.010	0.010	0.008	0.001	0.001	0.010	0.005	0.003	-0.091	0.003	0.013
		500	-0.004	0.008	0.008	0.010	0.002	0.002	0.009	0.004	0.002	-0.080	0.002	0.011
chi(3)														

Continued on next page...

Table 2.2 – continued

		RMEL			SLS1			SLS2			SLS3		
	N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
β_0	50	-0.106	0.956	0.968	-0.106	0.964	0.975	-0.121	0.924	0.978	-0.193	0.963	0.961
	100	-0.015	0.159	0.159	-0.021	0.160	0.160	-0.031	0.155	0.165	-0.064	0.164	0.159
	200	0.002	0.011	0.011	-0.011	0.011	0.012	-0.016	0.011	0.013	-0.009	0.013	0.011
	300	0.003	0.008	0.009	-0.010	0.008	0.008	-0.013	0.007	0.010	-0.003	0.010	0.007
	400	0.003	0.007	0.007	-0.011	0.007	0.007	-0.013	0.006	0.008	0.004	0.008	0.006
	500												
		-0.022	0.079	0.079	-0.048	0.081	0.083	-0.018	0.074	0.080	-0.066	0.080	0.078
β_1	50	-0.005	0.021	0.021	-0.018	0.023	0.024	0.000	0.019	0.022	-0.033	0.022	0.020
	100	-0.007	0.014	0.014	-0.017	0.016	0.017	-0.001	0.014	0.015	-0.022	0.015	0.014
	200	0.001	0.004	0.004	-0.003	0.005	0.005	0.007	0.004	0.004	-0.010	0.004	0.004
	300	0.001	0.003	0.003	-0.002	0.004	0.004	0.006	0.003	0.004	-0.007	0.004	0.003
	400	0.000	0.002	0.002	-0.001	0.003	0.003	0.006	0.002	0.003	-0.006	0.003	0.002
	500												
	50	0.054	1.003	1.006	0.139	0.324	0.343	0.014	0.480	0.141	-0.221	0.141	0.529
θ_{11}	100	0.002	0.439	0.439	0.071	0.118	0.123	0.034	0.297	0.065	-0.159	0.064	0.322
	200	-0.025	0.214	0.215	0.049	0.043	0.045	0.024	0.170	0.031	-0.115	0.030	0.184
	300	-0.039	0.139	0.140	0.038	0.054	0.055	0.023	0.115	0.017	-0.083	0.017	0.122
	400	-0.023	0.112	0.112	0.024	0.003	0.003	0.017	0.096	0.008	-0.058	0.008	0.099

Continued on next page. . .

Table 2.2 – continued

		RMEL			SLS1			SLS2			SLS3		
	N	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
	500	-0.014	0.091	0.092	0.020	0.001	0.002	0.010	0.079	0.003	-0.028	0.003	0.080
	50	-0.018	0.132	0.132	0.160	0.313	0.338	-0.046	0.072	0.125	-0.173	0.123	0.102
θ_{22}	100	-0.009	0.068	0.068	0.096	0.095	0.104	-0.022	0.050	0.058	-0.119	0.058	0.064
	200	-0.008	0.041	0.041	0.082	0.048	0.055	-0.027	0.032	0.028	-0.073	0.028	0.038
	300	0.001	0.021	0.021	0.062	0.028	0.032	0.003	0.018	0.013	-0.043	0.013	0.020
	400	-0.004	0.017	0.017	0.051	0.015	0.018	-0.004	0.015	0.012	-0.036	0.012	0.016
	500	-0.003	0.013	0.013	0.042	0.009	0.011	-0.004	0.012	0.008	-0.029	0.008	0.013
	50	-0.030	0.055	0.056	-0.012	0.055	0.055	-0.007	0.028	0.033	-0.241	0.033	0.086
ϕ	100	-0.009	0.024	0.024	0.017	0.011	0.012	0.008	0.015	0.010	-0.154	0.010	0.039
	200	-0.010	0.013	0.013	0.009	0.004	0.004	0.002	0.011	0.006	-0.092	0.006	0.020
	300	0.000	0.007	0.007	0.012	0.004	0.004	0.009	0.006	0.002	-0.063	0.002	0.010
	400	0.001	0.006	0.006	0.007	0.000	0.000	0.008	0.005	0.002	-0.048	0.002	0.007
	500	0.002	0.004	0.004	0.007	0.000	0.000	0.004	0.004	0.001	-0.038	0.001	0.005

2.3.2 Robustness against Outliers

We conducted some simulation studies to compare the estimates of REML with SLS when outliers exist. The RI model is used, and we generated 100 subjects ($N = 100$) with 8 measurements per subject ($n = 8$). 1000 Monte Carlo replications were carried out. In the first simulation study, we randomly contaminated one measurement within some subjects (corresponding to a single e-outlier). The proportions of contaminated subjects were chosen as 0%, 5%, 10%, 15% 20%, 25%, 30% and 35%. In the second simulation study, we contaminate the distributions of both b_i and e_i with the mixed normal model of the form $(1-p)\cdot N(0,\theta_{11}) + p\cdot f\cdot N(0,\theta_{11})$ and $(1-p)\cdot N(0,\sigma^2) + p\cdot f\cdot N(0,\sigma^2)$. The expected percentage of outliers p was selected as $0, 0.1, 0.2, 0.3$ and 0.4, and the contamination factor f was selected as 10.

Table 2.3 reports the Monte Carlo mean estimates and MSE in simulation study one. For the sake of saving space, we only present the simulation results with 0%, 5%, 15% and 30%, since similar pattern of results are observed. The influence of the outliers is clearly unbounded for REML estimates because the estimation bias and MSE increase as the percentage of data contamination increases. The magnitude of increase is especially dramatic for the random effect and residual error variances. The same phenomenon is observed in SLS1 estimates. In particular, SLS1 shows extremely lack of robustness against outlying measurements. This is not surprising be-

cause no downweight is applied in SLSE by using identity weight matrix, and the marginal second moments enlarge the affect of outliers. SLS2 is relatively more robust than SLS1 and REML with a smaller MSE, especially for moderate percentages of outliers. In contrast, SLS3 is clearly bounded and provides consistent mean and MSE estimates regardless of the percentage of data contamination. Thus, we demonstrate SLSE using the optimal weight matrix is robust against irregular measurements.

Table 2.4 reports the Monte Carlo mean estimates and MSE in simulation study two. For the fixed effects, means from all estimators remain unbias and the corresponding MSE stay stable regardless of the percentage of data contamination. It indicates that b- and e-outliers does not affect the precision of all estimators for the fixed effects. The robustness of ML estimation for fixed effects in this situation was also found in the simulation studies by Pinheiro, Liu and Wu (2001). For the random effects and residual errors, both REML and SLS1 result in considerable bias and MSE increase with the increase of data contamination. SLS2 is relatively more robust than SLS1 and REML, even though there are some higher bias and MSE at larger percentages of data contamination. In contrast, SLS3 performs much more satisfactory and produces the smallest bias and MSE among all estimators, which represent substantial efficiency gains.

Table 2.3: Simulation results for different percentage contaminations of a single response in the RI model at $N = 100$ and $n = 8$

	RMEL		SLS1		SLS2		SLS3	
%	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
				$\beta_1 = 8$				
0	8	0	8	0	8	0	7.999	0.02
5	8.5	1.4	5.5	68	8.1	0.1	7.9897	0.02
10	9.1	3.9	4.1	213	8.7	1.5	8.0055	0.02
15	9.8	7.8	5.5	377	9.4	3.9	7.9985	0.01
20	10.6	13.2	5	579	10	6.6	8.0046	0.01
25	10.9	17.7	2	704	10.1	7.7	7.9982	0.01
30	11.9	27.7	1.8	958	10.4	9.7	7.9945	0.02
35	12.5	37	-1	1369	10.4	9.8	7.9996	0.01
				$\beta_2 = 2$				
0	2	0	2	0	2	0	1.9992	0
5	2.1	0.1	2.6	4	2	0	1.9999	0
10	2.2	0.2	3.1	15	1.9	0	1.9989	0
15	2.3	0.4	3.2	26	2	0.2	1.9997	0
20	2.4	0.5	3.5	40	2.1	0.2	1.9991	0
25	2.6	1	4.2	53	2.3	0.4	1.999	0
30	2.6	1.2	4.2	66	2.4	0.5	1.9994	0
35	2.8	1.7	4.9	94	2.5	0.6	1.9988	0
				$\sigma_b^2 = 1.96$				
0	1.97	0	1.97	0	2	0	1.8473	0.05
5	2.27	11	35	3396	2	0	1.8427	0.05
10	3.99	134	47	6968	2.2	1	1.8404	0.05
15	6.44	260	56	14374	3.4	9	1.8547	0.04
20	8.65	458	81	35765	5.1	31	1.8585	0.04
25	14	4600	100	42393	6.6	61	1.8704	0.04

Continued on next page...

Table 2.3 – continued

	RMEL		SLS1		SLS2		SLS3	
%	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
30	19	6165	125	64188	9.3	140	1.877	0.04
35	20	2765	157	112301	12	229	1.8871	0.04
				$\sigma^2 = 1$				
0	1	0.0028	1	0.0009	1	0	0.8984	0.01
5	253	552545	162	135993	1	0.01	0.9038	0.01
10	699	2230175	318	420054	7.9	1218	0.9152	0.01
15	1087	3804815	460	849707	85	18598	0.9774	0.12
20	1601	6864968	649	1378208	150	42230	1.017	0.21
25	2281	11602110	980	2869012	195	66261	1.0726	0.57
30	3061	17463520	1426	5416065	248	101369	1.0763	0.54
35	3952	27098810	1803	7547520	287	132365	1.1332	1.18

Table 2.4: Simulation results for different percentage contaminations of b-outliers in the RI model at $N = 100$ and $n = 8$

	RMEL		SLS1		SLS2		SLS3	
%	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
				$\beta_1 = 8$				
0	8.0017	0.0257	8.0133	0.0253	8.0024	0.0256	7.999	0.0168
10	7.9872	0.0481	8.0696	0.0525	8.0418	0.0478	7.991	0.018
20	7.9895	0.0697	8.1159	0.0815	8.0898	0.0714	8.0064	0.0342
30	8.0216	0.1014	8.181	0.1317	8.1352	0.1048	8.0035	0.0514
40	7.9967	0.1159	8.1667	0.1597	8.141	0.1243	7.994	0.0712
				$\beta_2 = 2$				
0	2.0002	0.0002	1.9974	0.0002	1.9992	0.0002	1.9992	0.0002
10	2.0008	0.0004	1.9935	0.0005	1.9956	0.0005	2.0004	0.0002
20	1.9998	0.0006	1.9903	0.0007	1.9911	0.0007	1.9981	0.0003
30	1.9992	0.0009	1.9879	0.0012	1.9893	0.001	1.9982	0.0005
40	1.9996	0.0011	1.9874	0.0015	1.9871	0.0013	1.999	0.0006
				$\sigma_b^2 = 1.96$				
0	1.9667	0.0826	1.9666	0.0113	1.9654	0.002	1.8473	0.047
10	3.6709	4.0814	2.0688	0.4002	2.3579	1.2344	2.1161	0.1727
20	5.4817	14.5464	2.7656	5.7757	3.0425	4.9856	2.6766	1.3226
30	7.2821	31.6571	3.7218	22.1528	4.5153	16.4061	3.389	3.7958
40	9.0106	54.0235	5.2967	50.5112	5.5028	28.6985	4.2571	8.0611
				$\sigma^2 = 1$				
0	0.9993	0.0028	1.002	0.0009	1.001	0.0001	0.8984	0.0118
10	1.9145	0.87	1.0895	0.114	1.1263	0.1158	1.0872	0.0183
20	2.7897	3.2767	1.6067	1.3319	1.4233	0.7226	1.3808	0.1817
30	3.7245	7.5413	2.175	3.7683	2.1298	2.9917	1.7729	0.6809
40	4.5896	13.0129	2.6615	6.4271	2.5511	5.2334	2.1595	1.4584

%: Percentage of Contaminations

2.4 Application

The proposed estimator is applied to the longitudinal data on cholesterol levels collected as part of the famed Framingham heart study introduced in Chapter 1. In the study, 2634 participants' cholesterol level was measured every 2 years over 10 year period. The objective is to study change in cholesterol over time and examine the association with age at baseline and gender. This dataset is widely used in the linear mixed model literature, partly because many studies conclude that the distribution of subject-specific intercept is non-Gaussian. See, e.g., Zhang and Davidian (2001) and Lin and Lee (2008). For illustration, we select a sample of 133 participants (60 men and 73 women) whose cholesterol measurements as well as covariates of interest are completely observed at the duration of follow-up time. In general, the following linear mixed effect model is well accepted to fit the data:

$$y_{ij} = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Age}_i + \beta_3 \text{t}_{ij} + b_{0i} + b_{1i}\text{t}_{ij} + \epsilon_{ij}, \;\; i = 1, \cdots, 133, \;\; j = 1, \cdots, 6,$$

where y_{ij} is the cholesterol level for the i^{th} subject at the j^{th} time point, and y_{ij} was divided by 100 for numerical calculation stability; t_{ij} (in years) was taken as (time - 5)/10 measured from the baseline; Sex_i is a gender indicator (0 = female, 1 = male); and Age_i is age at baseline. $(b_{0i}, b_{1i})'$ is assumed to be normally distributed with mean zero and covariance $D = (\theta_{11}, \theta_{12}, \theta_{22})'$, and ϵ_{ij} is assumed to be normally distributed with mean zero and variance σ^2.

Table 2.5 includes the estimates and the corresponding 95% confidence interval. For fixed effects, SLS estimates are highly agree with ML, but with slightly tighter confidence intervals. Regarding the random effects and the residual errors, the estimates are quite different between these two methods. This finding is not surprising because the estimates of variance components are usually more difficult to estimate and known to have fairly large variabilities. However, the confidence intervals from SLSE are much smaller, which may due to the non-normality distributed random effects. Thus, SLSE provides more precise estimates than ML in this example.

Table 2.5: SLS and ML estimation of Framingham cholesterol data

	SLS		ML	
Parameter	Estimate	95% Confidence Interval	Estimate	95% Confidence Interval
β_0	1.5380	(1.3028, 1.7732)	1.5740	(1.2343, 1.9137)
β_1	-0.0369	(-0.1178, 0.0440)	-0.0338	(-0.1564, 0.0889)
β_2	0.0193	(0.0138, 0.0248)	0.0186	(0.0107, 0.0265)
β_3	0.2745	(0.2341, 0.3149)	0.2787	(0.2248, 0.3326)
θ_{11}	0.1033	(0.0731, 0.1335)	0.1259	(0.0934, 0.1584)
θ_{12}	0.0077	(0.0000, 0.0236)	0.0218	(0.0005, 0.0430)
θ_{22}	0.0418	(0.0208, 0.0628)	0.0390	(0.0136, 0.0644)
σ^2	0.0329	(0.0280, 0.0378)	0.0432	(0.0380, 0.0484)

Chapter 3

Simulation-Based Estimation in Generalized Linear Mixed Models

3.1 Introduction

Generalized linear mixed models (GLMM) have been widely used in the modeling of longitudinal data where the response is discrete. They can be viewed as a natural combination of linear mixed models (Laird and Ware 1982) and generalized linear models. In contrast to marginal or generalized estimating equations (GEE) models (Zeger, Liang, and Albert 1988), GLMM emphasize on the regression coefficients as well as the variance components of random effects.

3.1.1 Model Formulation

Suppose subject i is measured repeatedly on n_i occasions. For a GLMM, it is assumed that the response variable $y_{ij} \in I\!R$ is conditionally independent, given the covariates and random effects $b_i \in I\!R^q$, and have conditional distributions from the exponential family

$$f(y_{ij}|b_i, X_i, Z_i) = \exp\left\{\frac{\omega_{ij}y_{ij} - a(\omega_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \ i = 1, \cdots, N, \ j = 1, \cdots, n_i \tag{3.1}$$

where ϕ is a dispersion parameter, ω_{ij} is the canonical parameter and $a(\cdot)$ and $c(\cdot)$ are known functions. Let $X_i = (x'_{i1}, x'_{i2}, \ldots, x'_{in_i})'$ and $Z_i = (z'_{i1}, z'_{i2}, \ldots, z'_{in_i})'$. The conditional mean and variance

$$\mu^c_{ij} = E(y_{ij}|b_i, X_i, Z_i) = a^{(1)}(\omega_{ij}) \tag{3.2}$$

$$v^c_{ij} = Var(y_{ij}|b_i, X_i, Z_i) = \phi a^{(2)}(\omega_{ij}) \tag{3.3}$$

satisfy $g^{-1}\{\mu^c_{ij}\} = x'_{ij}\beta + z'_{ij}b_i$ and $v^c_{ij} = \phi\nu(\mu^c_{ij})$, where $a^{(d)}$ denotes the k^{th} derivatives against ω_{ij}, $g^{-1}(\cdot)$ and $\nu(\cdot)$ are known link and variance functions, respectively. The random effects are assumed to have mean zero and distribution $f_b(u;\theta)$ with unknown parameters $\theta \in I\!R^r$. In this model, the parameter of interest is $\psi = (\beta', \ \theta', \ \phi)'$.

3.1.2 Maximum Likelihood Estimation

For estimation and inference in GLMM, the most frequently employed approach is likelihood-based. In general, the log-likelihood function has the

following form:

$$\begin{aligned} l(\beta, \theta, \phi) &= \sum_{i=1}^{N} f(y_i|\beta, \theta, \phi) \\ &= \sum_{i=1}^{N} \int_{\mathbb{R}_q} \prod_{j=1}^{n_i} f(y_{ij}|u_i, \beta, \phi) f_b(u; D(\theta)) du_i \end{aligned}$$

which is N integrals over the $q-$dimensinal random effects b_i. Except in some special cases (e.g, identity link), these integrals are intractable. The analysis is even more difficult when the dimension of random effects is high or there are crossed random effects. To overcome this numerical difficulty, several methods have been proposed to approximate the integrals in the likelihood function, e.g., marginal quasi-likelihood and penalized quasi-likelihood (PQL) estimation (Breslow and Clayton 1993), adaptive quadrature (Rabe-Hesketh, Skrondal, and Pickles 2002), and maximum simulated likelihood (Durbin and Koopman 1997). In the following, we provide a brief review of PQL and adaptive quadrature methods.

3.1.3 Penalized Quasi-likelihood (PQL) Estimation

PQL method (Breslow and Clayton 1993) is based on a decomposition of the data into the mean and an error term, with a first-order Taylor series expansion of the mean which is a non-linear function of the linear predictor. It is analogous to iteratively reweighted least squares for linear models in that the model is linear in each iteration (Fitzmaurice, Davidian, Molenberghs and

Verbeke 2008). More specifically, one considers the decomposition

$$y_{ij} = g(x'_{ij}\beta + z'_{ij}b_i) + \epsilon_{ij} \tag{3.4}$$

where ϵ_{ij} have the appropriate distribution with variance equal to (3.3). Using the current estimates $\hat{\beta}^k$ and $\hat{b}_i^k$, the models is linearized by expanding $g(x'_{ij}\beta + z'_{ij}b_i)$ as a first-order Taylor series around current estimates. This yields

$$\begin{aligned} y_{ij} \quad \approx \quad & g(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k) + g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)x'_{ij}(\beta - \hat{\beta}^k) + \\ & g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)z'_{ij}(b_i - \hat{b}_i^k) + \epsilon_{ij} \end{aligned} \tag{3.5}$$

Re-ordering the above expression gives

$$y_{ij} = \zeta_{ij} + g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)x'_{ij}\beta + \xi_{ij} \tag{3.6}$$

where $\zeta_{ij} = g(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k) - g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)x'_{ij}\beta^k - g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)z'_{ij}\hat{b}_i^k$ is the sum of terms involving current estimates and $\xi_{ij} = g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)z'_{ij}b_i + \epsilon_{ij}$ is the terms involving random effects b_i and the residual error terms ϵ_{ij}. Note that (3.6) is of the linear mixed models form, $\hat{\beta}^{k+1}$ can be obtained by using generalized least squares method based on the model-implied covariance matrix Σ^k of the total residuals ξ_{ij}. The parameters of the random part are updated by fitting the model-implied covariance matrix Σ^{k+1} to the sample covariance matrix of the estimated total reburials (Goldstein 2003; Fitzmaurice, Davidian, Molenberghs and Verbeke 2008). The model fitting is done by iterating between the calculation of the pseudo-data based on current

estimates (i.e., $\hat{\beta}^k$ and $\hat{b}_i^k$) and the fitting of the approximate linear mixed model for these pseudo-data to obtain the next estimates (i.e., $\hat{\beta^{k+1}}$ and $\hat{b}_i^{k+1}$). This iteration continues until convergence is reached. The resulting estimates are called penalized quasi-likelihood estimates because they can be obtained from optimizing a quasi-likelihood function which only involves first and second-order condition moments, augmented with a penalty terms on the random effects (Molenberghs and Verbeke 2005).

A variant of the PQL algorithm is the marginal quasi-likelihood (MQL) method (Goldstein 1991, 2003), which is based on a linear Taylor expansion of the mean $g(x'_{ij}\beta + z'_{ij}b_i)$ around the current estimates $\hat{\beta}^k$ for the fixed effects and around $b_i = 0$ for the random effects. This yields a very similar expression as (3.5) and (3.6) expect $g'(x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k)$ is replaced by $g'(x'_{ij}\hat{\beta}^k)$. The resulting estimates are called MQL because they are obtained by evaluation of the marginal linear predictor $x'_{ij}\hat{\beta}^k$ instead of the conditional linear predictor $x'_{ij}\hat{\beta}^k + z'_{ij}\hat{b}_i^k$.

In general, MQL only performs well if the variance of random-effects is small and both methods perform bad for dichotomous outcomes with small cluster size. With increasing number of measurements per subject, MQL remains biased but PQL is consistent. The algorithms can be improved considerably by using a second-order Taylor expansion in the random effects (Goldstein and Rasbash 1996).

3.1.4 Gaussian Quadrature Estimation

Gaussian quadrature is a numerical integration technique that approximates any integral of the form

$$\int f(u)\exp(-u^2)du$$

by a weighted sum, namely

$$\int f(u)\exp(-u^2)du \approx \sum_{i=1}^{Q} w_q f(u_q).$$

Here Q is the order of the approximation, the u_q are the solutions of the Q^{th} order Hermite polynomial and the w_q are corresponding weights. The higher the value Q, the more accurate the approximation will be. The nodes (or quadrature points) u_q and the weights w_q are reported in Abramowitz and Stegun (1972). Alternatively, Press, Teukolsky, Vetterling and Flannery (1992) proposed an algorithm for calculating all u_q and w_q for any value Q. Gaussian quadrature approximates the likelihood by picking optimal subdivisions at which to evaluate the integrand. However, in practice a large number of quadrature points is required to approximate the likelihood and the integrand can have a very sharp peak between adjacent quadrature points. Adaptive Gaussian quadrature overcomes these problems by rescaling and shifting the nodes such that the integrand is sampled in a suitable range, however, adaptive Gaussian quadrature is much more time consuming. For a detailed discussion on Adaptive Gaussian quadrature, one can refer to (Rabe-Hesketh, Skrondal and Pickles 2002; Pinheiro and Chao 2006). In general,

the method can only deal with a small number of random effects (at most 2-3 random effects) which limits its general applicability.

A comprehensive evaluation and comparison of these approximate methods is unavailable in statistical literature. However, some limited studies have shown that the analytical simplification may not be always satisfactory and may produce biased and highly inefficient estimates (Lin and Breslow 1996; Joe 2008). Furthermore, the likelihood methods rely on normal assumption for random effects. Since the random effects are unobservable, it is not feasible to verify their distributional assumptions. It is thus natural to be concerned whether these methods yield reliable results when the normality assumption is violated. In addition, it is also known that likelihood-based methods are sensitive to data outliers. On the other hand, there are many works extending the GEE-type or quasi-likelihood to the estimation of GLMM (Zeger, Liang and Albert 1988; Jiang 1998; Sutradhar 2004). However, these methods are usually inefficient and require the simulation size S to go to infinity to obtain consistent estimators. In practice, since S has to be fixed, these methods only produce approximate consistent estimates.

3.2 Simulation-Based Estimation

3.2.1 Model Identifiability

Based on the conditional moments in (3.2) and (3.3), and assuming conditional independence of y_{ij}, the second-order condition moments can be

express as $\eta^c_{ijk} = \mu^c_{ij}\mu^c_{ik} + \delta_{jk}v^c_{ij}$, where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. Therefore, the first and second marginal moments can be expressed as

$$\mu_{ij}(\psi) \;=\; E(y_{ij}|X_i, Z_i) = \int g(x'_{ij}\beta + z'_{ij}u) f_b(u;\theta)du \tag{3.7}$$

and

$$\begin{aligned}\eta_{ijk}(\psi) = E(y_{ij}y_{ik}|X_i, Z_i) \;&=\; \int g(x'_{ij}\beta + z'_{ij}u)g(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)du \\ &\quad +\delta_{jk}\phi \int \nu(g(x'_{ij}\beta + z'_{ij}u)) f_b(u;\theta)du. \end{aligned} \tag{3.8}$$

Throughout this chapter, all integrals are taken over the space $\mathbb{R}^q$. It is straightforward to show that ψ can be estimated using nonlinear least squares method provided they are identifiable by (3.7) and (3.8) (Wang 2004, 2007).

In the following, we motivate our approach using two most popular GLMM as examples to demonstrate that ψ can indeed be identified and consistently estimated using the first two marginal moments (3.7) and (3.8).

Example 3.2.1. Consider a mixed Poisson model for counts, where $V(y_{ij}|b_i) = E(y_{ij}|b_i)$ and $\log E(y_{ij}|b_i) = x'_{ij}\beta + z'_{ij}b_i$. Assuming $b_i \sim N(0, D(\theta))$, we have

$$\mu_{ij}(\psi) \;=\; \exp(x'_{ij}\beta + z'_{ij}D(\theta)z_{ij}/2), \;\text{ and} \tag{3.9}$$

$$\eta_{ijk}(\psi) \;=\; \mu_{ij}(\psi)\mu_{ik}(\psi)\exp[z'_{ij}D(\theta)z_{ik})] + \;\delta_{jk}\phi\mu_{ij}(\psi). \tag{3.10}$$

All unknown parameters in this model are identifiable because they can be consistently estimated by (3.9) and (3.10) which are usual nonlinear regression equations in observed variables.

Example 3.2.2. Consider a mixed logistic model for a binary response y_{ij}, where $\phi = 1$ and $\text{logit}\{Pr(y_{ij} = 1|b_i)\} = x'_{ij}\beta + z'_{ij}b_i$. For this model we find

$$\mu_{ij}(\psi) = E(y_{ij}^2|X_i, Z_i) = \int \left(\frac{e^{x'_{ij}\beta + z'_{ij}u}}{1 + e^{x'_{ij}\beta + z'_{ij}u}} \right) f_b(u;\theta)du, \text{ and} \quad (3.11)$$

$$\eta_{ijk}(\psi) = \int \left(\frac{e^{x'_{ij}\beta + z'_{ij}u}}{1 + e^{x'_{ij}\beta + z'_{ij}u}} \right) \left(\frac{e^{x'_{ik}\beta + z'_{ik}u}}{1 + e^{x'_{ik}\beta + z'_{ik}u}} \right) f_b(u;\theta)du, \text{ for } j < k. \quad (3.12)$$

The integrals in (3.11) and (3.12) are intractable but can be approximated using Monte Carlo simulation techniques. Therefore, all parameters can be consistently estimated by (3.11) and (3.12) through the nonlinear least squares method.

Example 3.2.3. Consider a mixed Probit model for a binary response y_{ij}, where $\phi = 1$ and $\Phi^{-1}\{Pr(y_{ij} = 1|b_i)\} = x'_{ij}\beta + z'_{ij}b_i$. Assuming $b_i \sim N(0, D(\theta))$, we find

$$E(y_{ij}|X_i, Z_i) = E(y_{ij}^2|X_i, Z_i) = \Phi(x'_{ij}\beta \cdot |D(\theta)z_{ij}z'_{ij} + I|^{-q/2}); \quad (3.13)$$

and

$$E(y_{ij}y_{ik}|X_i, Z_i) = \int \Phi(x'_{ij}\beta + z'_{ij}b_i)\Phi(x'_{ik}\beta + z'_{ik}b_i)f_b(u;\theta)du. \quad (3.14)$$

In this model, the first marginal moment admits an analytic form, while the second marginal moment does not.

3.2.2 Estimation and Inference

Even though the model is identifiable through the first two marginal moments, closed forms of these moments are usually not available in GLMM. In addition, the density $f_b(u;\theta)$ is usually unknown. Therefore, we propose a simulation-based approach to overcome these two difficulties simultaneously. As it is well known, simulation-based estimation is computationally convenient when moment functions cannot be evaluated directly (Pakes and Pollard 1989; Gouriéroux and Monfort 1997). The basic idea is to form unbiased estimators of integrals in moment equations with their Monte Carlo simulators. In particular, we propose a simulation-by-parts (Wang 2004) technique to construct two sets of moments. First, generate random points $u_{is}, s = 1, 2, \ldots, 2S$ from a known density $h(u)$, and construct

$$\mu_{ij,1}(\psi) = \frac{1}{S}\sum_{s=1}^{S}\frac{g(x'_{ij}\beta + z'_{ij}u_{is})f_b(u_{is};\theta)}{h(u_{is})}, \tag{3.15}$$

$$\begin{aligned}\eta_{ijk,1}(\psi) &= \frac{1}{S}\sum_{s=1}^{S}\frac{g(x'_{ij}\beta + z'_{ij}u_{is})g(x'_{ik}\beta + z'_{ik}u_{is})f_b(u_{is};\theta)}{h(u_{is})} \\ &\quad + \frac{\delta_{jk}\phi}{S}\sum_{s=1}^{S}\frac{\nu(g(x'_{ij}\beta + z'_{ij}u_{is}))f_b(u_{is};\theta)}{h(u_{is})}\end{aligned} \tag{3.16}$$

using the first half of the points $u_{is}, s = 1, 2, \ldots, S$. Then construct $\mu_{ij,2}(\psi)$ and $\eta_{ijk,2}(\psi)$ similarly using the second half of the points $u_{is}, s = S+1, S+2, \ldots, 2S$. It is obvious that the simulated moments are unbiased estimate of the true moments, since $E(\mu_{ij,t}(\psi)|X_i, Z_i) = \mu_{ij}(\psi)$ and $E(\eta_{ijk,t}(\psi)|X_i, Z_i) = \eta_{ijk}(\psi)$, $t = 1, 2$. We denote the parameter space by $\Gamma = \Omega \times \Theta \times \Sigma \in I\!R^{p+r+1}$,

and the true parameter value by $\psi_0 = (\beta_0', \ \theta_0', \ \phi_0)' \in \Gamma$. Finally, the SBE $\hat{\psi}_{N,S}$ for ψ is defined as

$$\hat{\psi}_{N,S} = \underset{\psi \in \Gamma}{\operatorname{argmin}} \, Q_{N,S}(\psi) = \underset{\psi \in \Gamma}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_{i,1}'(\psi) W_i \rho_{i,2}(\psi),$$

where $\rho_{i,t}(\psi) = (y_{ij} - \mu_{ij,t}(\psi), 1 \leq j \leq n_i, y_{ij}y_{ik} - \eta_{ijk,t}(\psi), 1 \leq j \leq k \leq n_i)'$ and $W_i = W(X_i, Z_i)$ is a nonnegative definite matrix which may depend on X_i and Z_i. By using two different sets of independent simulated points, $Q_{N,S}(\psi)$ is an unbiased estimator of $Q_N(\psi)$ because $\rho_{i,1}(\psi)$ and $\rho_{i,2}(\psi)$ are conditionally independent given (Y_i, X_i, Z_i) and hence

$$\begin{aligned} E\left[\rho_{i,1}(\psi) W_i \rho_{i,2}(\psi)\right] &= E\left[E\left(\rho_{i,1}(\psi) | Y_i, X_i, Z_i\right) W_i E\left(\rho_{i,2}(\psi) | Y_i, X_i, Z_i\right)\right] \\ &= E\left(\rho_i(\psi) W_i \rho_i(\psi)\right) \end{aligned} \tag{3.17}$$

where $\rho_i(\psi) = (y_{ij} - \mu_{ij}(\psi), \ 1 \leq j \leq n_i, \ y_{ij}y_{ik} - \eta_{ijk}(\psi), \ 1 \leq j \leq k \leq n_i)'$.

To construct simulated moments in (3.15) and (3.16), it only requires the random effects distribution to have a known parametric form. Hence, instead of relying on normality assumption on b_i, we can use more flexible distributions. For example, one can follow Davidian and Gallant (1993) and Zhang and Davidian (2001) to represent the density of b_i by the standard seminonparametric density which includes normal, skewed, multi-modal, fat- or thin-tailed densities. One can also impose the Tukey(g, h) family distribution (Field and Genton 2006) for b_i as well which is generated by a single transformation of the standard normal and covers a variety of distributions.

To establish the consistency and asymptotic normality of $\hat{\psi}_{N,S}$ we make

the following assumptions.

Assumption 3.2.1. $g(\cdot)$ and $\nu(\cdot)$ are continuous functions; $f_b(u;\theta)$ is continuous in $\theta \in \Theta$ for all u.

Assumption 3.2.2. (y_i, X_i, Z_i, n_i), $i = 1, ..., N$ are independent and identically distributed and satisfy $E\left[\|W_i\|\left(y_{ij}^4 + 1\right)\right] < \infty$; $g^2(x'\beta + z'u)f_b(u;\theta)$ and $|\nu(g(x'\beta + z'u))|f_b(u;\theta)$ are bounded by a positive function $G(x, z, u)$ satisfying $E\left[\|W_i\|\left(\int G(X_i, Z_i, u)du\right)^2\right] < \infty$.

Assumption 3.2.3. The parameter space $\Gamma \subset I\!R^{p+r+1}$ is compact.

Assumption 3.2.4. $E[\rho_i(\psi) - \rho_i(\psi_0)]'W_i[\rho_i(\psi) - \rho_i(\psi_0)] = 0$ if and only if $\psi = \psi_0$.

Assumption 3.2.5. $g(\cdot)$ and $\nu(\cdot)$ are twice continuously differentiable and $f_b(u;\theta)$ is twice continuously differentiable w.r.t to θ in an open subset $\theta_0 \in \Theta_0 \subset \Theta$. Furthermore, all first and second order partial derivatives of $g(x'\beta + z'u)f_b(u;\theta)$ and $\nu(g(x'\beta + z'u))f_b(u;\theta)$ w.r.t $(\beta', \theta')'$ are bounded absolutely by the positive function $G(x, z, u)$ given in assumption 3.2.2.

Assumption 3.2.6. The matrix

$$B = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right] \tag{3.18}$$

is nonsingular.

Theorem 3.2.4. *Suppose that* $Supp(h) \supseteq Supp(f_b(\ \cdot\ ;\theta))$ *for all* $\theta \in \Theta_0$. *Then for any fixed* $S > 0$, *as* $N \to \infty$,

1. *under assumptions 3.2.1-3.2.4,* $\hat{\psi}_{N,S} \xrightarrow{a.s.} \psi_0$*;*

2. *under assumptions 3.2.1-3.2.6,* $\sqrt{N}(\hat{\psi}_{N,S} - \psi_0) \xrightarrow{L} N(0, B^{-1}C_S B^{-1})$*, where*

$$\begin{aligned} 2C_S &= E\left[\frac{\partial \rho'_{i,1}(\psi_0)}{\partial \psi} W_i \rho_{i,2}(\psi_0) \rho'_{i,2}(\psi_0) W_i \frac{\partial \rho_{i,1}(\psi_0)}{\partial \psi'}\right] \\ &\quad + E\left[\frac{\partial \rho'_{i,1}(\psi_0)}{\partial \psi} W_i \rho_{i,2}(\psi_0) \rho'_{i,1}(\psi_0) W_i \frac{\partial \rho_{i,2}(\psi_0)}{\partial \psi'}\right]. \quad (3.19) \end{aligned}$$

Note the above asymptotic results do not require the simulation size S tends to infinity because we use the simulation-by-parts technique to approximate moments. This is fundamentally different from other simulation-based methods which require S goes to infinity to obtain consistent estimators (Zeger, Liang, and Albert 1988; Jiang 1998; Sutradhar 2004). In general, the simulation approximation of the integrals will result in certain efficiency loss but this loss decreases at the rate $O(1/S)$ (Wang 2004). Therefore, the efficiency loss due to the simulations can be made small by increasing S. For the choice of $h(u)$, in theory, it has no impact on the asymptotic efficiency of the estimator, as long as it has sufficiently large support. However, the choice of $h(u)$ will affect the finite sample variances of the simulated moments. It is well known that the finite sample variances will be minimized when $h(u) \propto |g(x'_{ij}\beta + z'_{ij}u) f_b(u;\theta)|$ and $h(u) \propto |g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)|$.

When closed forms of moments exist such as in Example 3.2.1, the SBE becomes M-estimator (Huber 2004) $\hat{\psi}_N$. We can shown $\hat{\psi}_N$ is consistent and asymptotic normal distributed. In particular, we have the following corollary.

Corollary 3.2.5. *As $N \to \infty$,*

1. *under assumptions 3.2.1-3.2.4, $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$;*

2. *under assumptions 3.2.1-3.2.6, $\sqrt{N}(\hat{\psi}_N - \psi_0) \xrightarrow{L} N(0, B^{-1}CB^{-1})$, where B and C are given in (3.18) and*

$$C = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} W_i \rho_i(\psi_0) \rho_i'(\psi_0) W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right]$$

.

Remark 3.2.1. Since random effects are usually assumed to have zero mean it is more convenient to define $b_i = D(\theta)^{1/2}\xi_i$ where the random variable ξ has mean zero and covariance matrix I_q. Hence alternatively, we can re-write (3.15)-(3.16) as

$$\begin{aligned}
\mu_{ij,1}(\psi) &= \frac{1}{S}\sum_{s=1}^{S} \frac{g(x_{ij}'\beta + z_{ij}' D(\theta)^{1/2} u_{is}) f_\xi(u_{is})}{h(u_{is})}, \\
\eta_{ijk,1}(\psi) &= \frac{1}{S}\sum_{s=1}^{S} \frac{g(x_{ij}'\beta + z_{ij}' D(\theta)^{1/2} u_{is}) g(x_{ik}'\beta + z_{ik}' D(\theta)^{1/2} u_{is}) f_\xi(u_{is})}{h(u_{is})} \\
&\quad + \frac{\delta_{jk}\phi}{S}\sum_{s=1}^{S} \frac{\nu(g(x_{ij}'\beta + z_{ij}' D(\theta)^{1/2} u_{is})) f_\xi(u_{is})}{h(u_{is})}.
\end{aligned}$$

In this case, there is no parameter of interest in $f_\xi(u_{is})$.

Remark 3.2.2. For binary responses y_{ij}, $E(y_{ij}|X_i, Z_i) = E(y_{ij}^2|X_i, Z_i)$ with probability one. Therefore, the terms $y_{ij}^2 - E(y_{ij}^2|X_i, Z_i)$ in $\rho_{i,1}(\psi)$ and $\rho_{i,2}(\psi)$ are redundant and do not need to be included.

Remark 3.2.3. For certain GLMM such as a probit model with normal distributed random effects, the first marginal moment admits an analytical form but not the second marginal moments. In this case, only the second moments need to be simulated.

3.2.3 Computation

In general, the SBE does not admit an explicit solution and can be computed using Newton-Raphson algorithm as

$$\hat{\psi}^{(\tau+1)} = \hat{\psi}^{(\tau)} - \left(\frac{\partial^2 Q_{N,S}(\hat{\psi}^{(\tau)})}{\partial \psi \partial \psi'} \right)^{-1} \frac{\partial Q_{N,S}(\hat{\psi}^{(\tau)})}{\partial \psi},$$

where $\hat{\psi}^{(\tau)}$ denotes the estimate of ψ at the τ^{th} iteration, and

$$\frac{\partial Q_{N,S}(\hat{\psi}^{(\tau)})}{\partial \psi} = \sum_{i=1}^{N} \left[\frac{\partial \rho'_{i,1}(\hat{\psi}^{(\tau)})}{\partial \psi} W_i \rho_{i,2}(\hat{\psi}^{(\tau)}) + \frac{\partial \rho'_{i,2}(\hat{\psi}^{(\tau)})}{\partial \psi} W_i \rho_{i,1}(\hat{\psi}^{(\tau)}) \right] \tag{3.20}$$

$$\begin{aligned} \frac{\partial^2 Q_{N,S}(\hat{\psi}^{(\tau)})}{\partial \psi \partial \psi'} &= \sum_{i=1}^{N} \left[\frac{\partial \rho'_{i,1}(\hat{\psi}^{(\tau)})}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\hat{\psi}^{(\tau)})}{\partial \psi'} + (\rho'_{i,2}(\hat{\psi}^{(\tau)}) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho'_{i,1}(\hat{\psi}^{(\tau)})/\partial \psi)}{\partial \psi'} \right] \\ &+ \sum_{i=1}^{N} \left[\frac{\partial \rho'_{i,2}(\hat{\psi}^{(\tau)})}{\partial \psi} W_i \frac{\partial \rho_{i,1}(\hat{\psi}^{(\tau)})}{\partial \psi'} + (\rho'_{i,1}(\hat{\psi}^{(\tau)}) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho'_{i,2}(\hat{\psi}^{(\tau)})/\partial \psi)}{\partial \psi'} \right] \end{aligned} \tag{3.21}$$

The terms $(\rho'_{i,1} W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho'_{i,2}/\partial \psi)}{\partial \psi'}$ and $(\rho'_{i,2} W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho'_{i,1}/\partial \psi)}{\partial \psi'}$ are $o_p(1)$ so they can be omitted from the second derivative for computational convenience. Here, we use $g^{(d)}(\cdot)$ and $\nu^{(d)}(\cdot)$, $d = 0, 1, 2$, to denote their d^{th} order derivatives, and use $f_b^{(d)}(u; \theta)$ to denote its d^{th} order partial derivative w.r.t. θ. By assumption 3.2.5 and the dominated convergence theorem, the nonzero

first order derivatives in (3.20) and (3.20) can be expressed in the following form:

$$\begin{aligned}
\frac{\partial \mu_{ij,1}(\psi)}{\partial \beta} &= \frac{x_{ij}}{S}\sum_{s=1}^{S} g^{(1)}(x'_{ij}\beta + z'_{ij}u_{is})\frac{f_b(u_{is};\theta)}{h(u_{is})},\\
\frac{\partial \mu_{ij,1}(\psi)}{\partial \theta} &= \frac{1}{S}\sum_{s=1}^{S} \frac{g(x'_{ij}\beta + z'_{ij}u_{is})}{h(u_{is})} f_b^{(1)}(u_{is};\theta),\\
\frac{\partial \eta_{ijk,1}(\psi)}{\partial \beta} &= \frac{x_{ij}}{S}\sum_{s=1}^{S} g^{(1)}(x'_{ij}\beta + z'_{ij}u_{is})\frac{g(x'_{ik}\beta + z'_{ik}u_{is})f_b(u_{is};\theta)}{h(u_{is})}\\
&\quad + \frac{x_{ik}}{S}\sum_{s=1}^{S} g(x'_{ij}\beta + z'_{ij}u_{is})g^{(1)}(x'_{ik}\beta + z'_{ik}u_{is})\frac{f_b(u_{is};\theta)}{h(u_{is})}\\
&\quad + \delta_{ik}\phi\frac{x_{ij}}{S}\sum_{s=1'}^{S} \nu^{(1)}(g(x'_{ij}\beta + z'_{ij}u))g^{(1)}(x'_{ij}\beta + z'_{ij}u)\frac{f_b(u;\theta)}{h(u_{is})}\\
\frac{\partial \eta_{ijk,1}(\psi)}{\partial \theta} &= \frac{1}{S}\sum_{s=1}^{S} \frac{g(x'_{ij}\beta + z'_{ij}u)g(x'_{ik}\beta + z'_{ik}u)}{h(u_{is})} f_b^{(1)}(u;\theta)\\
&\quad + \delta_{ik}\phi\frac{1}{S}\sum_{s=1}^{S} \frac{\nu(g(x'_{ij}\beta + z'_{ij}u))}{h(u_{is})} f_b^{(1)}(u;\theta),\\
\frac{\partial \eta_{ijk}(\psi)}{\partial \phi} &= \delta_{ik}\frac{1}{S}\sum_{s=1}^{S} \frac{\nu(g(x'_{ij}\beta + z'_{ij}u))f_b(u;\theta)}{h(u_{is})}.
\end{aligned}$$

Another important question is how to specify the form of weight W_i to compute $\hat{\psi}_{N,S}$ in an optimal way, such that $AV\left(\hat{\psi}_N(W_i)\right) - AV\left(\hat{\psi}_N(W_i^{opt})\right)$ is nonnegative definite for all possible W_i. It can be shown that W_i^{opt} is approximately equal to

$$A_i^{-1} = E[\rho_{i,1}(\psi_0)\rho'_{i,2}(\psi_0)|X_i, Z_i]^{-1}. \tag{3.22}$$

The proof is analogous to that in Abarin and Wang (2006) and is therefore omitted. In practice, A_i is not feasible since it involves unknown parame-

ters to be estimated. One possible solution is using a two-stage procedure. First, minimize $Q_{N,S}(\psi)$ using a sub-optimal choice of W_i, such as an identity weight matrix, to obtain the first stage estimator $\hat{\psi}_{N1,S}$. Second, estimate $W_i = \hat{A}_i^{-1}$ using $\hat{\psi}_{N1,S}$ and then minimize $Q_{N,S}(\psi)$ again with $\hat{A}_i^{-1}$ to obtain the second stage estimator $\hat{\psi}_{N2,S}$. In general, the computation of A_i in (3.22) is difficult since it requires the specification of third- and fourth-order moments of y_{ij}. However, these high order moments can be easily approximated using the Monte Carlo simulation method introduced in this section. Alternatively, A_i can be estimated using any nonparametric method such as kernel or spline estimators. In some cases, a simple consistent estimator of A_i would be

$$A(\hat{\psi}) = \frac{1}{N}\sum_{i=1}^{N} \rho_{i,1}(\hat{\psi}_{N1})\rho'_{i,2}(\hat{\psi}_{N1}). \tag{3.23}$$

In many real data applications, the subjects are clustered so that the values of X_i, Z_i are equal or close for all subjects within one cluster. In such cases, each A_i can be estimated similarly to (3.23) using all the subjects within the same cluster.

3.2.4 Robustness

Let v be the subset of observations (X_l, Y_l) under investigation, and the IF of SBE at point v takes the form

$$\mathrm{IF}(v; \hat{\psi}_{N,S}, F) = -B(\hat{\psi}_N(F))^{-1}\frac{\partial \rho'_{l,1}(v; \hat{\psi}_{N,S}(F))}{\partial \psi}\hat{A}^{-1}\rho_{l,2}(v; \hat{\psi}_{N,S}(F)), \tag{3.24}$$

where F is the underlying distribution and B is given in (3.18).

Corollary 3.2.6. *If the SLSE $\hat{\psi}_N$ is computed using the estimated optimal weight (3.23), then $\left\| IF(v; \hat{\psi}_N, F) \right\| \to 0$ as $\|v\| \to \infty$.*

The implication of above corollary is that the influence function of $\hat{\psi}_N$ is bounded and $\hat{\psi}_N$ has a redescending property (Huber 2004). It is expected that data outliers in either x or y directions will be automatically downweighted by the inverse of the estimated optimal weight matrix. It does not require detection for outliers beforehand to implement downweighting strategy. The proof is analogous to that of Theorem 2.2.4 and is therefore omitted.

3.2.5 Bias Reduction

It is noticed in the simulation studies by Wang (2007) and our preliminary simulation studies, there are some finite sample biases for the estimation of variance components by the SBE. These biases are downward-oriented and diminish with increasing sample sizes. The source of this bias lies in the fact that the optimal weight in (3.22) is replaced by a root-N estimate given in (3.23) for the second stage minimization. Asymptotically this replacement has no impact on the properties of SBE. However, it does make a difference in finite samples because $A_i(\hat{\psi})$ depends on y_i and causes the correlation with $\rho_{i,1}(\psi)$ and $\rho_{i,2}(\psi)$. Note in the setup of the SBE, we require W_i may only depend on X_i and Z_i. Evaluating this bias analytically is not easy. Instead,

we extend the independently weighted method proposed by Altonji and Segal (1996) for the bias reduction. The basic idea is to break the correlation between $A_i(\hat{\psi})$ and $\rho_{i,t}(\psi)$ by designing the weighting matrix using observations other than used to construct the sample moments. We randomly split the sample into K groups with N_k subjects in each group and the independently weighted SBE (SBEIW) $\hat{\psi}_{N,S}^{IW}$ for ψ is defined as the measurable function that minimizes

$$Q_{N,S}(\psi) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N_k}(\rho_{i,1}^{k}(\psi))' A_i^{-k}(\hat{\psi})\rho_{i,2}^{k}(\psi), \tag{3.25}$$

where $\rho_{i,t}^{k}(\psi)$ is constructed for the k^{th} group and $A_i^{-k}(\hat{\psi})$ is constructed using all but the k^{th} group. Intuitively, this estimator is less biased because the statistical dependence between the weight matrix and sample moments were broken. However, splitting the sample causes efficiency loss due to the loss in degrees of freedom. Since $\text{cov}(\hat{\psi}_{N,S}^{k}, \hat{\psi}_{N,S}^{k+l}) = 0$ for $l \neq 0$ by design, it can be easily shown that

$$\text{cov}(\hat{\psi}_{N,S}^{IW}) = \frac{1}{K^2}\sum_{k=1}^{K}\text{cov}(\hat{\psi}_{N,S}^{k}),$$

where $\hat{\psi}_{N,S}^{k}$ is obtained by minimizing $\sum_{i=1}^{N_k}(\rho_{i,1}^{k}(\psi))' A_i^{-k}(\hat{\psi})\rho_{i,2}^{k}(\psi)$. In the simulation studies presented in section 3.3.1, we select $K = 2$ and observe significant improvement in estimation bias over SBE with negligible efficiency loss.

3.3 Numerical Studies

3.3.1 Monte Carlo Simulation Studies

In this section, we evaluate the finite sample behavior the proposed estimator, and compare them with the penalized quasi-likelihood estimator (PQLE) by Breslow and Clayton (1993). We conducted substantial numerical studies by using different generalized linear mixed models and parameter configurations. We carried out 500 Monte Carlo replications in each simulation study and reported the average biases $((1/500)\sum_{i=1}^{500}\hat{\psi}_i - \psi_0)$ and the root mean square errors (RMSE; $(1/500)\sum_{i=1}^{500}(\hat{\psi}_i - \psi_0)^2$). All computations are done in R and PQL estimates are obtained from `glmmPQL` package.

The first simulation study is designed based on Example 3.2.1. In particular, we simulated the model $\log E(y_{ij}|b_i) = \beta_0 + \beta_1 x_{ij} + b_i,\ \ j = 1, \cdots, 4$, where $x_{ij} = 0.1j$, $\beta = (3, -1)'$ and $b_i \sim N(0, 0.25)$. In the present simulation, we set $N = 50, 100, 200, 300, 400$ and chose the density $N(0, 1)$ to be $h(u)$ and generated $S = 1000$ independent u_{is} for SBE. For comparison purpose, we also computed the ψ_N by using the two marginal moments from equations (3.9) and (3.10). Table 3.1 reports the biases and the root mean square errors (RMSE). Fig. 3.1 visually summarizes the performance of all estimators at various sample sizes in terms of RMSE and percentage of bias. From Table 3.1 and Fig 3.1, we see that all estimators perform satisfactorily and show clearly their asymptotic proprieties, i.e., the estimated RMSE decrease with the increase of sample size. For fixed effects, both estimated RMSE

and biases from the proposed estimators are very close to each other and are comparable to the PQLE, although ψ_N^{IW} and $\psi_{N,S}^{IW}$ have slightly higher RMSE for β_1. For the random effect parameter θ, all estimators present similar estimated RMSE; PQLE, ψ_N and $\psi_{N,S}$ show some downward bias while ψ_N^{IW} and $\psi_{N,S}^{IW}$ show some upward bias. From Fig 3.1, significant higher percent $(10-20\%)$ bias is observed in ψ_N as well as in $\psi_{N,S}$; however, it is worth noting this bias gradually reduces with the increase of sample size. In contrast, ψ_N^{IW} and $\psi_{N,S}^{IW}$ have less than 5% bias which demonstrates bias reduction by using the proposed independent weight methodology. In addition, we use histograms to show how close the distribution of SBE estimates are to the normal distributions and compare with PQL estimates. From Fig 3.2, we have found that when $N = 200$ the distribution is already fairly close to normal for all estimators; thus, the asymptotic normality properties of the proposed estimates are justified.

Table 3.1: Biases (RMSE) of the parameter estimates

N	PQLE	SLSE	SLSIW	SBE	SBEIW
			$\beta_0 = 3$		
50	0.006 (0.082)	-0.086 (0.115)	0.001 (0.162)	-0.069 (0.109)	0.012 (0.168)
100	0.012 (0.060)	-0.053 (0.077)	-0.009 (0.090)	-0.039 (0.075)	0.007 (0.103)
200	0.010 (0.040)	-0.029 (0.052)	-0.009 (0.058)	-0.022 (0.055)	0.005 (0.061)
300	0.006 (0.033)	-0.021 (0.040)	-0.005 (0.040)	-0.016 (0.047)	-0.003 (0.052)
400	0.009 (0.031)	-0.017 (0.035)	-0.005 (0.034)	-0.010 (0.044)	-0.003 (0.043)
			$\beta_1 = -1$		
50	-0.007 (0.152)	0.007 (0.143)	0.020 (0.341)	0.009 (0.130)	-0.005 (0.329)
100	-0.004 (0.109)	0.006 (0.106)	0.013 (0.180)	0.008 (0.107)	0.007 (0.195)
200	0.000 (0.073)	0.002 (0.077)	0.015 (0.109)	0.004 (0.074)	0.013 (0.115)
300	-0.001 (0.064)	0.003 (0.061)	0.007 (0.081)	0.000 (0.058)	0.003 (0.081)
400	-0.001 (0.056)	0.001 (0.054)	0.003 (0.067)	0.003 (0.054)	0.006 (0.065)
			$\theta = 0.25$		
50	-0.010 (0.053)	-0.043 (0.060)	0.011 (0.105)	-0.054 (0.076)	0.012 (0.122)
100	-0.007 (0.040)	-0.043 (0.056)	0.004 (0.066)	-0.045 (0.069)	0.001 (0.081)
200	-0.004 (0.026)	-0.030 (0.042)	0.012 (0.059)	-0.036 (0.059)	0.000 (0.060)
300	-0.003 (0.023)	-0.024 (0.035)	0.006 (0.048)	-0.027 (0.051)	0.002 (0.055)
400	-0.004 (0.019)	-0.022 (0.032)	0.002 (0.033)	-0.025 (0.048)	0.005 (0.048)

An addition simulation study was conducted based on the RIS model in Chapter 2 to confirm the performance of SLSIW. We have similar findings as the first simulation study and the results are presented in Table 3.2.

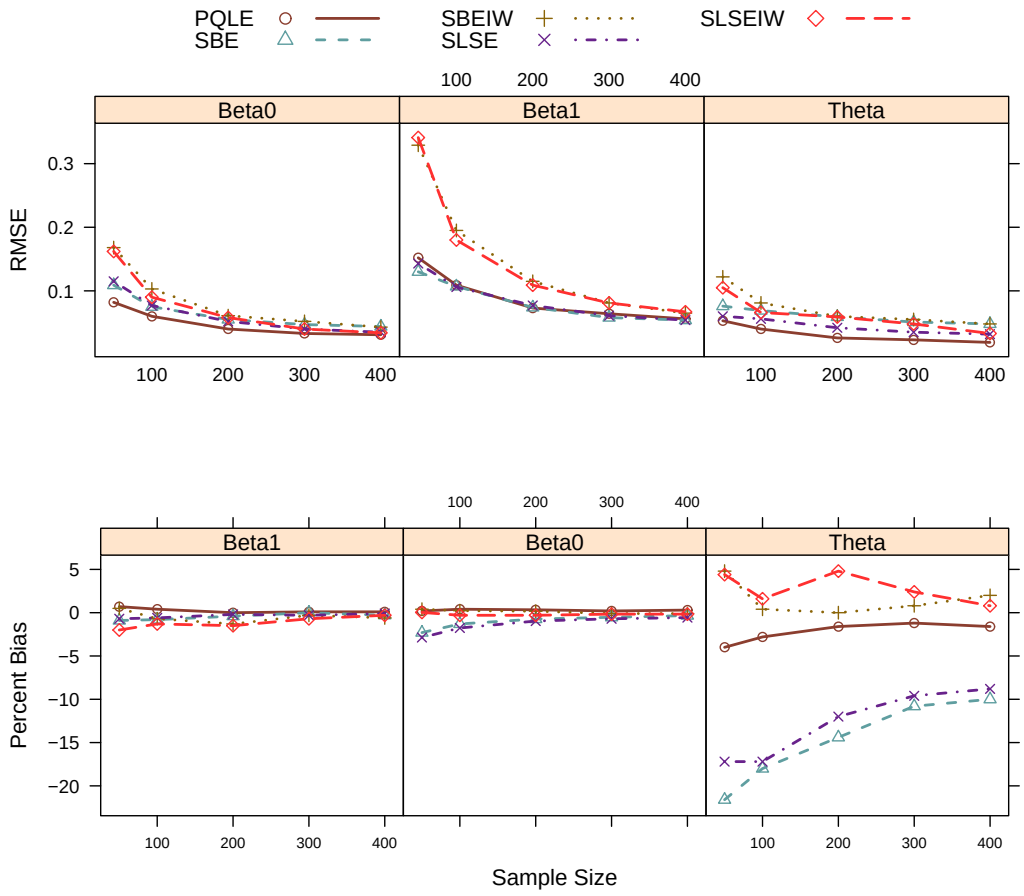


Figure 3.1: RMSE and percentage of bias of parameter estimates for model at various sample sizes.

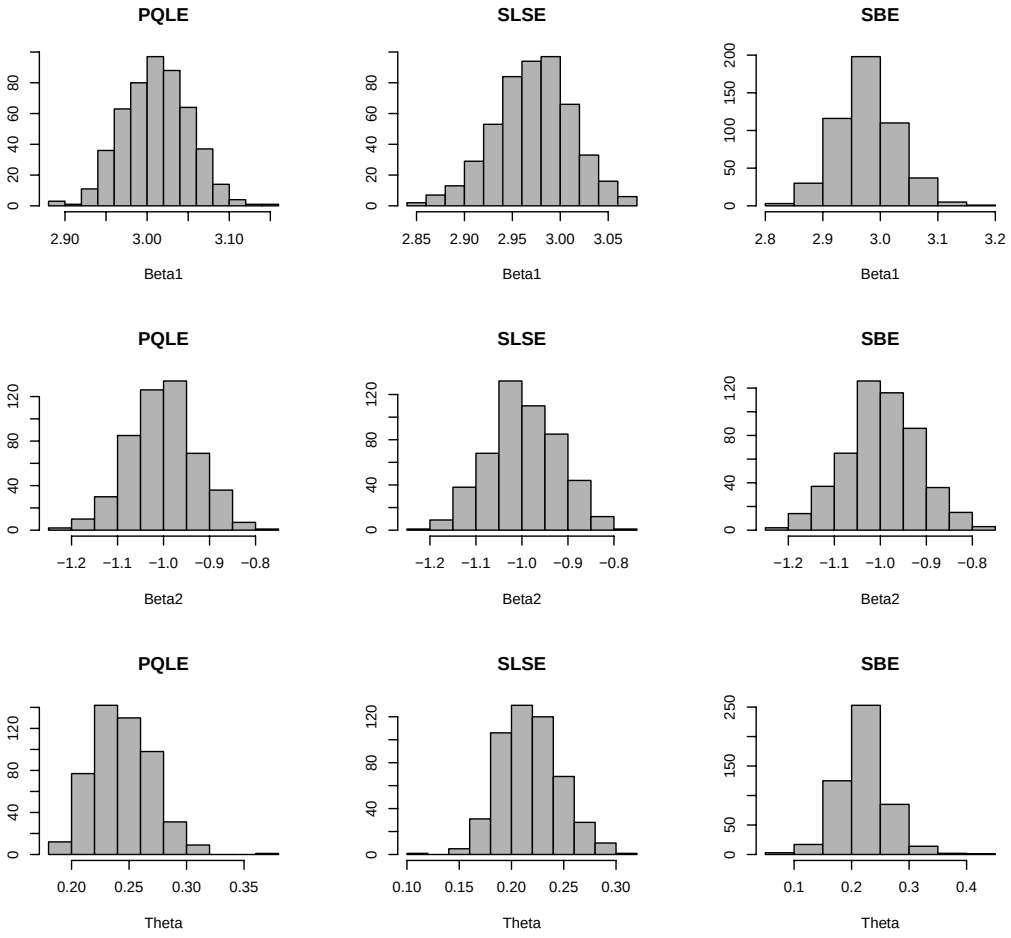


Figure 3.2: Histograms of PQLE, SLSE and SBE for model with $N = 200$.

Table 3.2: Simulation results with normal distributed random effect and residual errors based on the RIS model

		SLS		SLSIW	
	N	Bias	RMSE	Bias	RMSE
β_0	50	-0.055	0.666	-0.023	0.470
	100	-0.002	0.184	0.004	0.257
	200	-0.001	0.131	-0.005	0.153
	300	-0.001	0.108	-0.005	0.114
	400	-0.002	0.092	-0.005	0.099
	500	-0.004	0.084	-0.005	0.090
β_1	50	-0.018	0.212	-0.003	0.293
	100	-0.001	0.112	-0.010	0.157
	200	-0.002	0.075	0.002	0.089
	300	0.000	0.065	0.000	0.071
	400	0.000	0.056	0.001	0.060
	500	0.003	0.050	0.008	0.053
θ_{11}	50	-0.142	0.604	0.028	1.185
	100	-0.068	0.442	0.089	0.656
	200	-0.027	0.321	0.062	0.422
	300	-0.038	0.268	0.034	0.308
	400	-0.027	0.238	0.020	0.273
	500	-0.013	0.209	0.032	0.234
θ_{22}	50	-0.106	0.232	-0.025	0.394
	100	-0.053	0.162	0.035	0.216
	200	-0.035	0.119	0.013	0.141
	300	-0.021	0.090	0.011	0.100
	400	-0.014	0.077	0.015	0.087

Continued on next page...

Table 3.2 – continued

		SLS		SLSIW	
	N	Bias	RMSE	Bias	RMSE
	500	-0.014	0.071	0.009	0.079
ϕ	50	-0.143	0.197	-0.011	0.228
	100	-0.085	0.127	-0.001	0.130
	200	-0.046	0.083	0.013	0.088
	300	-0.032	0.064	0.007	0.067
	400	-0.027	0.054	0.004	0.053
	500	-0.019	0.048	0.006	0.049

In the second simulation study, we consider a logistic model: $\text{logit}(Pr(y_{ij} = 1|b_i)) = \beta_0 + \beta_1 \times trt_i + \beta_2 x_{ij} + b_{i0} + b_{i1} x_{ij}$, where $b_i \sim N[(0,0)', \text{diag}(\theta_0, \theta_1)]$. In the present simulation, we selected $N = 200, 300$ and $n = 5$; covariates $trt_i = 1$ for half the sample and 0 for the remainder, $x_{ij} = (j-3)/2$; $\beta = (-1.0,\ \ 0.5,\ \ 0.5)'$; $\theta_0 = 1$ and $\theta_1 = 0.5$. To compute the SBE, we chose the density of $N[(0,0)', \text{diag}(2,2)]$ to be $h(u)$, and generated independent points $u_{is},\ s = 1, ..., 2S$ using $S = 500, 1000$ and 2000 respectively. Table 3.3 reports the simulation results. Overall, it is clear that the SBE results in smaller bias than the PQLE for fixed effects as well as the random effect θ_0, while the SBE has slightly bigger bias only for the random effect θ_1. The finding is not surprising, as it is known that the PQLE may have severe bias in the estimates of the fixed effects and variance components of

random effects, when repeated measured data are binary. As the sample size N increases from 200 to 300, the RMSE for all parameters from all methods decrease. For the SBE, as the number of simulated values S decreases from 2000 to 500, RMSE become slightly bigger but the estimates stays relatively stable. It implies that even at a relative small sample size of simulated values $S = 500$, the SBE still produces reasonable estimates. Comparing the PQLE with the SBE computed using $S = 2000$, the PQLE seems to have smaller RMSE, especially for the random effect estimates.

Table 3.3: Biases (RMSE) of the parameter estimates with different number of the simulated points S for SBE

		SBE		
	PQLE	$S = 2000$	$S = 1000$	$S = 500$
N=200				
$\beta_0 = -1$	0.109 (0.180)	-0.071 (0.188)	-0.070 (0.200)	-0.049 (0.191)
$\beta_1 = 0.5$	-0.054 (0.189)	0.029 (0.217)	0.040 (0.218)	0.032 (0.174)
$\beta_2 = 0.5$	-0.057 (0.124)	0.030 (0.141)	0.030 (0.139)	0.024 (0.109)
$\theta_0 = 1$	-0.108 (0.258)	0.103 (0.332)	0.112 (0.375)	0.063 (0.358)
$\theta_1 = 0.5$	0.074 (0.279)	0.082 (0.402)	0.107 (0.392)	0.061 (0.366)
N=300				
$\beta_0 = -1$	0.113 (0.164)	-0.030 (0.135)	-0.045 (0.154)	-0.033 (0.178)
$\beta_1 = 0.5$	-0.067 (0.176)	0.021 (0.170)	0.024 (0.169)	0.027 (0.183)
$\beta_2 = 0.5$	-0.058 (0.109)	0.022 (0.109)	0.022 (0.107)	0.013 (0.108)
$\theta_0 = 1$	-0.116 (0.210)	0.055 (0.255)	0.071 (0.298)	0.051 (0.345)
$\theta_1 = 0.5$	0.088 (0.241)	0.074 (0.319)	0.073 (0.324)	0.045 (0.334)

The third simulation study is to demonstrate the robustness of the proposed estimator in the presence of outliers, we conducted simulation studies

on random intercept Poisson and logistic models with one covariate, and the parameter values β=(1, 1)' and θ=0.25. We generated $N = 100$ subjects with $n = 5$ measurements per subject. The values of the covariate $x_{ij} = (j-3)/2$ in the Poisson mixed model, and one random measurement within five different subjects was contaminated by using $100y_{ij}$ (i.e., 5% subjects with one outlier). For the logistic model, x_{ij} was generated from $N(0, 1)$. Since the response variable y_{ij} is binary in the logistic model, outliers arise in x. To create outliers, we followed Sinha (2004, 2006) to replace one randomly chosen x value within five different subjects by $x + 3$ (i.e., 5% subjects with one outlier). For comparison, we also present the simulation results without outliers. Table 3.4 summarizes the simulation results. In the case of Poisson mixed model, the SBE stays almost the same as outliers increase from 0% to 5% while a significant increase from PQLE. For the logistic model, the SBE shows smaller biases for the estimation of β_1 and θ in the presence of outliers. For the estimation of fixed effects β_0 and β_1, the SBE provides smaller RMSE than the PQLE. However, the PQLE of θ appears to have smaller RMSE. This interesting and counterintuitive phenomenon was also found in the similar simulation study conducted in Sinha (2004) and Noh and Lee (2007) when they compared their proposed robust estimation methods with the classical likelihood-based method. Similarly, we can argue that the RMSE of the PQLE of θ underestimates because of the relatively larger biases observed in the PQLE of the fixed effects.

Table 3.4: Biases(RMSE) for the parameter estimates with and without outliers

	No Outliers		With Outliers	
	PQLE	SLSE/SBE	PQLE	SLSE/SBE
Poisson Model				
$\beta_0 = 1$	0.021 (0.060)	-0.082 (0.103)	0.162 (0.205)	-0.057 (0.082)
$\beta_1 = 1$	-0.001 (0.038)	0.017 (0.043)	-0.004 (0.163)	0.011 (0.041)
$\theta = 0.25$	-0.013 (0.043)	-0.047 (0.062)	0.097 (1.029)	-0.040 (0.059)
Logistic Model				
$\beta_0 = 1$	0.020 (0.212)	0.066 (0.306)	-0.059 (0.412)	0.074 (0.365)
$\beta_1 = 1$	0.051 (0.229)	0.117 (0.317)	-0.108 (0.433)	0.073 (0.301)
$\theta = 0.25$	0.017 (0.320)	-0.021 (0.571)	-0.013 (0.295)	0.028 (0.643)

3.3.2 Application

In this section, we apply the proposed methods to analyze the popular epilepsy seizure count data presented in Chapter 1. The data come from a clinical trial of 59 epileptics who were randomized to receive either the antiepileptic drug progabide (TRT = 1) or a placebo (TRT = 0), as an adjuvant to standard chemotherapy. The logarithm of a quarter of the number of epileptic seizures in the 8-week period preceding the trial (BASE) and the logarithm of age (Age) were included as covariates in the analysis. For each individual, a multivariate response variable consisted of the seizure counts during 2-week periods before each of four clinical visits (VISIT, coded -0.3,-0.1, 0.1 and 0.3) was collected. By a thorough investigation, Thall and Vail (1990) identified a number of patients as outliers, who has irregular large counts. Recently, the data were further analyzed by Sinha (2006) using his

proposed robust quasi-likelihood estimator (RQLE). Here we consider the following model used by Sinha (2006):

$$\log E(y_{ij}|b_i) = x'_{ij}\beta + b_{i0} + b_{i1}\text{VISIT}_{ij}, \tag{3.26}$$

where $b_{i0} \sim N(0, \theta_0)$ and $b_{i1} \sim N(0, \theta_1)$ are the independent random effects, and x_{ij} represents the vector of the predictors BASE, TRT, AGE, VISIT, and the interaction between BASE and TRT. Following (3.9) and (3.10), the first two moments are

$$\begin{aligned}
\mu_{ij} &= \exp(x'_{ij}\beta + \theta_0/2 + (\text{VISIT}_{ij})^2\theta_1/2) \\
\nu_{ijk} &= \exp\left[(x_{ij} + x_{ik})'\beta + 2\theta_0 + (\text{VISIT}_{ij} + \text{VISIT}_{ik})^2\theta_1/2\right] + \delta_{ik}\mu_{ij}.
\end{aligned}$$

To calculate the standard error of $\hat{\Upsilon}_N$, we have trivially

$$\begin{aligned}
\frac{\partial \mu_{ij}}{\partial \psi} &= \left(x_{ij}\mu_{ij}, \frac{\mu_{ij}}{2}, \frac{(\text{VISIT}_{ij})^2}{2}\mu_{ij}\right)' \\
\frac{\partial \nu_{ijk}}{\partial \psi} &= \left((x_{ij} + x_{ik})\nu_{ijk}, 2\nu_{ijk}, \frac{(\text{VISIT}_{ij} + \text{VISIT}_{ik})^2\nu_{ijk}}{2}\right)' + \delta_{ik}\frac{\partial \mu_{ij}}{\partial \psi}.
\end{aligned}$$

Table 3.5 reports the fixed and random effect estimates by the SBE, the RQLE and the classical marginal quasi-likelihood estimator (MQLE). The estimates of the fixed effects are very similar and the covariate BASE is highly significant by all three approaches. However, we observe a significant difference in the estimates of the random effects. In particular, the SBE estimates highly agree with the RQL estimates, but quite different from those obtained by the MQL method. The standard errors (SE) of θ_0^2 from all approaches are relatively close but the SBE results in a SE reduction of

50% for θ_1^2 in comparison with the MQLE. Since Sinha (2006) concludes that the RQL method appears to be successful in handling outliers in the epilepsy data, we confirm that the SBE has the same property.

Table 3.5: Comparison of parameter estimates and their standard errors (SE) for the seizure count data

	SLSE	RQLE*	MQLE*
Parameter	Estimates (SE)	Estimates (SE)	Estimates (SE)
INTERCEPT	-1.324 (1.672)	-1.330 (0.928)	-1.388 (1.248)
BASE	0.915 (0.117)	0.895 (0.083)	0.890 (0.141)
TRT	-0.758 (0.627)	-0.795 (0.446)	-0.849 (0.424)
TRT × Base	0.397 (0.205)	0.260 (0.238)	0.324 (0.216)
AGE	0.453 (0.485)	0.462 (0.277)	0.463 (0.365)
VISIT/10	-0.230 (0.268)	-0.230 (0.156)	-0.253 (0.241)
θ_0	0.135 (0.093)	0.130 (0.050)	0.257 (0.083)
θ_1	0.117 (0.709)	0.116 (0.357)	1.904 (1.386)

*Obtained from Sinha (2006).

3.4 Incomplete Longitudinal Data

3.4.1 Missing Data Mechanism

Incomplete longitudinal data are almost inevitable in longitudinal studies due to various reasons (e.g. dropout or noncompliance in clinical trials). Problems arise if the mechanism leading to the missing data depends on response process. It is known that ignoring missing data or the use of naive methods may introduce bias and lead to misleading inferences (Little and Rubin 2002). To obtain valid inferences from incomplete longitudinal data,

it is essential to consider the reason for missingness, which is usually referred to as the missing data mechanism. The missing data mechanism attempts to answer, from a statistical perspective, the question of why data is missing. The central issue is to properly characterize the probabilistic relationship between the value that should have been observed but it was not observed. This relationship is defined statistically in terms of the conditional distribution of the missing data indicator matrix given the observed data. Little and Rubin (2002) gave a general treatment of statistical analysis of missing data mechanisms, which includes a useful hierarchy of missing-value models. Let $R_i = (R_{i1}, R_{i2}, \cdots, R_{in})'$ be the vector of missing data indicators for Y_i, such that $R_{ij} = 1$ if response Y_{ij} is observed, and 0 otherwise. We partition Y_i into Y_i^O and Y_i^M, where Y_i^O contains those Y_{ij} for which $R_{ij} = 1$ and Y_i^M contains the remaining components. Assuming X_i is always observed, the three classifications of missing data mechanisms are:

(i). Missing Completely At Random (MCAR): Data are said to be MCAR if the probability of failure to observe a value is unrelated to any observed or unobserved data. The MCAR assumption is often too strong to be plausible in practical situations, except in the case where data is missing by design. Under the MCAR assumption, the conditional distribution of the missing data mechanism given the data Y is given by

$$P(R_i|Y_i, X_i) = P(R_i|X_i).$$

(ii). Missing At Random (MAR): Data are said to be MAR if the probability

of an observation being missing only depends on observed data. MAR is a weaker and more plausible assumption than MCAR. Under the MAR assumption, the conditional distribution of the missing data mechanism given the data Y is given by

$$P(R_i|Y_i, X_i) = P(R_i|Y_i^O, X_i).$$

(iii). Missing Not At Random (MNAR): Data are said to be MNAR if the probability of an observation being missing depends on both observed and unobserved data. Under the MNAR assumption, the conditional distribution of the missing data mechanism given the data Y is given by

$$P(R_i|Y_i, X_i) = P(R_i|Y_i^O, Y_i^M, X_i).$$

3.4.2 Missing Data Patterns

There are two broad classes of missing data pattern: intermittent missing and dropout. Intermittent missing pattern refers to the scenario that a subject completes the study but skips a few occasions in the middle of the study period. Dropout (attrition, lost of follow-up) is a particular example of monotone pattern of missingness, which means if one observation is missing, then all the observation after it will be unobserved. Intermittent missing is often easier to deal with because the subject is still participating the study and the reason of missing values can be ascertained. Dropout is often more serious because the subject is no longer available and we cannot be certain that the

dropout is or is not related to the observed or unobserved outcome. MAR mechanisms are commonly assumed when the interest lies on estimation of parameters, especially when monotone missing data patterns are under consideration. The rationale of using such mechanisms is discussed by authors including Robins, Rotnitzky and Zhao (1995), and Lindsey (2000). In the following, we focus on the discussion of modeling monotone MAR longitudinal data.

3.4.3 Estimation of Missing Data Process

A transition model is considered for modeling the dropout data process. Let $\lambda_{ij} = P(R_{ij} = 1|R_{i,j-1} = 1, X_i, Z_i, Y_i^O)$ be the conditional probability that subject is observed at visit j, given that subject is present at time $j-1$; and $\pi_{ij} = P(R_{ij} = 1|X_i, Z_i, Y_i^O)$ be the marginal probability subject i is present at time j which is equal to $\prod_{t=2}^{j} \lambda_{it}$. Generally it is assumed that all individuals are observed on the first occasion, that is $R_{i1} = 1$ and $\lambda_{i1} = 1$. Let $\pi_{ijk} = P(R_{ij} = 1, R_{ik} = 1|X_i, Z_i, Y_i^O)$ be the probability for observing both Y_{ij} and Y_{ik} given the response history and covariates. For monotone missing data pattern, $R_{ij} = 0$ implies $R_{ik} = 0$ for all $j < k$ so $\pi_{ijk} = \pi_{ik}$. Let $H_{ij} = \{y_{i1}, y_{i2}, \cdots, y_{i,j-1}\}$ to be the history of observed responses for subject i up to (but not including) time point j.

Usually, λ_{ij} is unknown and must be estimated from the observed data. Since the missing data indicator variable R_{ij} is binary, we consider a logistic

regression model for the drop-out process

$$\text{logit}\lambda_{ij} = A'_{ij}\alpha \tag{3.27}$$

where A_{ij} is the vector consisting of information of X_i, Z_i and H_{ij}, and α is the vector of regression parameters. This model has been widely used in modeling the drop-out process (e.g., Diggle and Kenward 1994; Fitzmaurice, Laird and Zahner 1996; Molenberghs, Kenward and Lesaffre 1997; Yi and Cook 2002). Estimation of parameters α can be proceeded by running a logistic regression analysis using the likelihood method. In particular, we let D_i be the random dropout time for subject i and d_i be a realization, $i = 1, \cdots, N$. Define

$$L_i(\alpha) = (1 - \lambda_{id_i}) \prod_{t=2}^{d_i - 1} \lambda_{it}$$

where λ_{it} is given in (3.27). Then the corresponding score equation of subject i is $S_i(\alpha) = \sum_{i=1}^{N} \partial \log L_i(\alpha)/\partial\beta$, which yields unbiased estimate of α if model (3.27) is correctly specified. The resulting estimator is denoted by $\hat{\alpha}$. Moreover, the marginal probabilities $\pi_{ij}(\hat{\alpha})$ can be consistently estimated.

3.4.4 Weighted SBE

The SBE based on the observed data is given by

$$\hat{\psi}^o_{N,S} = \underset{\psi\in\Gamma}{\text{argmin}}\, Q^o_{N,S} = \underset{\psi\in\Gamma}{\text{argmin}} \sum_{i=1}^{N} (\Delta_i \rho_{i,1}(\psi))' W_i (\Delta_i \rho_{i,2}(\psi)),$$

where $\Delta_i = \text{diag}(R_{ij}, 1 \le i \le n_i, R_{ij}R_{ik}, 1 \le j \le k \le n_i)$. Under MCAR assumption, the SBE remains valid because

$$\frac{1}{N} Q^o_{N,S}(\psi) \xrightarrow{a.s.} E[(\Delta_i \rho_{i,1}(\psi))' W_i (\Delta_i \rho_{i,2}(\psi))] = E(\Delta_i) Q(\psi),$$

where $E(\Delta_i) = \text{diag}(\pi_{ij}, 1 \le i \le n_i, \pi_{ij}\pi_{ik}, 1 \le j \le k \le n_i)$ and $Q(\psi) = E\rho_i'(\psi) W_i \rho_i(\psi)$. Since $Q(\psi)$ attains a unique minimum at $\psi_0 \in \Gamma$ (see section A.4) and $E(\Delta_i)$ does not depend on ψ, it is straightforward to show that $\hat{\psi}^o_{N,S} \xrightarrow{a.s.} \psi_0$ as $N \to \infty$. Also, it is easy to show the estimator of B and C_s in (3.18) and (3.19) based on the observed data are consistent by similar manipulations. However, under MAR assumption, Δ_i depends on Y_i and the SBE is no long valid because $E[(\Delta_i \rho_{i,1}(\psi))' W_i (\Delta_i \rho_{i,2}(\psi))] \neq E(\Delta_i) Q(\psi)$.

We consider modifying the proposed method to adjust MAR type of monotone missingness through the inverse probability weighted (IPW) method. It is known that IPW is a general methodology for constructing estimators of smooth parameters under non- or semi-parametric models for the full data and a semi-parametric or parametric model for the missingness mechanism (Horvitz and Thompson 1952; Robins and Rotnitzky 1995; Yi and Cook 2002). The idea is to weight each subject's contribution in the estimation by the inverse probability that a subject drops out at the time he dropped out. The weights are obtained from models for the missing data process, and these models must be correctly specified for the resulting estimators to be consistent.

Let $\tilde{\Delta}_i = \text{diag}(R_{ij}/\pi_{ij}, 1 \le i \le n_i, R_{ij}R_{ik}/\pi_{ijk}, 1 \le j \le k \le n_i)$ be

the weight matrix accommodating missingness. We define the weighted SBE (WSBE) as

$$\hat{\tilde{\psi}}_{N,S} = \underset{\tilde{\psi}\in\Gamma}{\operatorname{argmin}}\, \tilde{Q}^{o}_{N,S}(\psi) = \underset{\psi\in\Gamma}{\operatorname{argmin}} \sum_{i=1}^{N} \tilde{\rho}'_{i,1}(\psi)\tilde{W}_i\tilde{\rho}_{i,2}(\psi),$$

where $\tilde{\rho}_{i,t}(\psi) = \tilde{\Delta}_i\rho_{i,t}(\psi)$. By model assumptions and the law of iterated expectation, we can show that

$$\begin{aligned} E[\tilde{Q}^{o}_{N,S}(\psi)] &= E[\sum_{i=1}^{N} \tilde{\Delta}_i\rho_{i,1}(\psi)'W_i\tilde{\Delta}_i\rho_{i,2}(\psi)] \\ &= E[\sum_{i=1}^{N} E[\tilde{\Delta}_i\rho_{i,1}(\psi)'W_i\tilde{\Delta}_i\rho_{i,2}(\psi)|X_i, Z_i, Y_i]] \\ &= E[\sum_{i=1}^{N} E[\tilde{\Delta}_i|X_i, Z_i, Y_i]\rho_{i,1}(\psi)'W_i\rho_{i,2}(\psi)] \\ &= E[Q_{N,S}(\psi)], \end{aligned}$$

where the last equality holds due to the fact that $E[\tilde{\Delta}_i|X_i, Z_i, Y_i]$ is an identity matrix if the probabilities π_i are correctly specified. Under two additional assumptions

Assumption 3.4.1. Given the past history of observed responses and covariates, the probability that individual i is still in the study at time j is bounded away from zero or $P(R_{ij} = 1|R_{i,j-1} = 1, X_i, Y_{ij}) > 0$.

Assumption 3.4.2. The probability of dropout model must be correctly specified. i.e. $\lambda_{ij} = P(R_{ij} = 1|R_{i,j-1} = 1, X_i, Z_i, Y_i)$.

Theorem 3.2.4 holds for $\hat{\tilde{\psi}}_{N,S}$ except that $\rho_{i,t}(\psi)$ is replaced by $\tilde{\rho}_{i,t}(\psi)$ in (3.2.6) and (3.19). The optimal weight $\tilde{A}_i$ becomes $E[\tilde{\rho}_{i,1}(\psi_0)\tilde{\rho}'_{i,2}(\psi_0)|X_i, Z_i]$.

For the computation of optimal weight, the moment estimator defined in (3.23) is not conformable because the length of $\tilde{\rho}_i(\psi)$ is different across subjects. Therefore, we propose to construct the optimal weight using the simulation method. The second-order marginal moments can be calculated using (3.16), and the third- and fourth-order moments will be calculated using the same method. In particular, we first construct the third conditional moments for y_{ij}, i.e,

$$Cov(y_{ij}, y_{ik}y_{il}|b_i, X_i, Z_i) = \begin{cases} E(y_{ij}^3|b_i, X_i, Z_i) - \mu_{ij}^c\eta_{ijj}^c, & \text{if } j = k = l, \\ \eta_{ijj}^c\mu_{ik}^c - \mu_{ij}^c\eta_{ikl}^c, & \text{if } j = k \neq l, \\ \eta_{ijj}^c\mu_{il}^c - \mu_{ij}^c\mu_{ikl}^c, & \text{if } j = l \neq k, \\ \mu_{ij}^c\mu_{ik}^c\mu_{il}^c - \mu_{ij}^c\eta_{ikl}^c, & \text{if } j = k \neq l. \end{cases}$$

For the fourth conditional moments, we have

$$Cov(y_{ij}y_{ik}, y_{il}y_{it}|b_i, X_i, Z_i) = \begin{cases} E(y_{ij}^4|b_i, X_i, Z_i) - (\eta_{ijj}^c)^2, & \text{if } j = k = l = t, \\ E(y_{ij}^3|b_i, X_i, Z_i)\mu_{it}^c - \eta_{ijj}^c\eta_{ijt}^c, & \text{if } j = k = l \neq t, \\ \eta_{ijj}^c\mu_{il}^c\mu_{it}^c - \eta_{ijj}^c\eta_{ilt}^c, & \text{if } j = k \neq l \neq t, \\ \mu_{ij}^c\mu_{ik}^c\mu_{il}^c\mu_{it}^c - \eta_{ijk}^c\eta_{ilt}^c, & \text{if } j = k \neq l, \end{cases}$$

and other elements can be computed similarly. To compute $E(y_{ij}^3|b_i, X_i, Z_i)$ and $E(y_{ij}^4|b_i, X_i, Z_i)$, we can use the following results by McCullagh and Nelder (1989). That is,

$$\begin{aligned} E[(y_{ij} - \mu_{ij}^c)^3|b_i, X_i, Z_i] &= \phi^2 a^{(3)}(\omega_{ij}), \\ E[(y_{ij} - \mu_{ij}^c)^4|b_i, X_i, Z_i] &= \phi^3 a^{(4)}(\omega_{ij}) + 3(\phi a^{(2)}(\omega_{ij}))^2. \end{aligned}$$

Then we can use the simulation method introduced in section 3.2.2 to construct the marginal moments, and thus obtain the optimal weight. Alternatively, we can adopt the idea of working variance matrix in Prentice and

Zhao (1991) and Vonesh, Wang, Nie and Majumdar (2002) to construct optimal working weight. For example, assuming y_i is from a multivariate normal distribution, the third moment of y_i is

$$\text{cov}(y_{ij}, y_{ik}y_{il}) = \mu_{il}\sigma_{ijk} + \mu_{ik}\sigma_{ijl}, \text{ for all } j,\ k,\ l,$$

where $\sigma_{ijk} = E(y_{ij} - u_{ij})(y_{ik} - u_{ik})$, and the forth moment of y_i, for all $j,\ k,\ l,\ t$ is

$$\text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) = \sigma_{ijl}\sigma_{ikt}+\sigma_{ijt}\sigma_{ikl}+\mu_{ik}\mu_{il}\sigma_{ijt}+\mu_{ij}\mu_{il}\sigma_{ikt}+\mu_{ik}\mu_{it}\sigma_{ijl}+\mu_{ij}\mu_{it}\sigma_{ikl}.$$

Thus, both third and forth moments can be specified by using only the first and second moments. Alternatively, we can assume independence among the elements of y_i. Then the third moment of y_i is given by

$$\text{cov}(y_{ij}, y_{ik}y_{il}) = \begin{cases} E[(y_{ij} - u_{ij})^3] + 2\mu_{ij}\sigma_{ijj} - 2\mu_{ij}^3 & \text{if } j = k = l, \\ \sigma_{ijj}\mu_{ik} & \text{if } j = l \neq k, \\ \sigma_{ijj}\mu_{il} & \text{if } j = k \neq l, \\ 0 & \text{otherwise.} \end{cases}$$

The forth moment of y_i is

$$\text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) = \begin{cases} E[y_{ij}^4] - \mu_{ij}^2 - \sigma_{ijj} & \text{if } j = k = l = t, \\ E[(y_{ij} - u_{ij})^3]\mu_{it} + 2\mu_{ij}\mu_{it}\sigma_{ijj} & \text{if } j = k = l \neq t, \\ E[(y_{ij} - u_{ij})^3]\mu_{il} + 2\mu_{ij}\mu_{il}\sigma_{ijj} & \text{if } j = k = t \neq l, \\ 0 & \text{otherwise.} \end{cases}$$

If we further assume the underlying distribution of y_i is symmetric, we have $E[(y_{ij} - u_{ij})^3] = 0$.

3.4.5 Multiple Imputation

An alternative for SBE to handle missing data is by the means of multiple imputation. Multiple imputation (MI) was first proposed by Rubin (1977) and was described in detail by Rubin (1987) and Schafer (1997). The key idea of this procedure is to fill out each missing value with a set of M plausible values that represent the uncertainty about the right value to impute. Multiple imputation inference involves three distinct stages:

(i). Fill out missing values, Y^M, M times to generate M complete data sets. Here MAR assumption is key to the validity of MI because Y_M are generally sampled from a condition distribution $f(Y^M|Y^O, \psi)$. Commonly used imputation methods include regression method (Rubin 1987), predictive mean matching method (Heitjan and Little 1991; Schenker and Taylor 1996), propensity score method (Rosenbaum and Rubin 1983; Lavori, Dawson, and Shera 1995) and MCMC Method (Schafer 1997).

(ii). Each imputed data set is analyzed by using standard procedure, and the resulting estimates and the corresponding sampling covariances (within-imputation variance) are denoted by $\tilde{\psi}_k^M$ and V_k, $k = 1, \cdots, M$.

(iii). The results from the M analyses are combined to produce a single MI estimator, $\tilde{\psi}_{MI}$, and to draw inferences based on 'Rubin's rules' for MI. The MI estimator of ψ is the average of the individual estimators

$$\tilde{\psi}_{MI} = \frac{1}{M} \sum_{k=1}^{M} \tilde{\psi}_k^M.$$

The estimated variance of this combines between- and within-imputation variability as follows

$$V_{MI} = \frac{1}{M}\sum_{i=1}^{M} V_k + \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{M}(\tilde{\psi}_k^M - \tilde{\psi}_{MI})^2.$$

Pros and cons of IPW methods with respect to MI have been the subject of some debate (Scharfstein, Rotnitzky and Robins 1999; Clayton, Spiegelhalter, Dunn and Pickles 1998; Carpenter, Kenward and Vansteelandt 2006). In the following section, we conduct some simulation studies to compare WSBE and MI-SBE.

3.4.6 Monte Carlo Simulation Studies

We conduct some simulation studies to assess the performance of the WSBE, MI-SBE and SBE under MCAR and MAR scenarios with moderate amount (10%-30%) of missing data. We carried out 500 Monte Carlo replications in each simulation study and reported the average biases $((1/500)\sum_{i=1}^{500}\hat{\psi}_i - \psi_0)$ and the root mean square errors (RMSE; $(1/500)\sum_{i=1}^{500}(\hat{\psi}_i - \psi_0)^2$). Two scenarios of continuous responses and count data are considered here. The continuous response y_{ij} is generated from a linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij},$$

and ϵ_{ij} is generated from a standard normal distribution. For count data, y_{ij} is generated from a mixed poisson model with

$$\log E(y_{ij}|b_i) = \beta_0 + \beta_1 x_{ij} + b_i.$$

We set $\beta = (1, 1)'$ and generate b_i from the normal distribution $N(0, \theta)$ with $\theta = 0.25$. The true covariate x_{ij} is simulated from the normal distribution $N(1, 1)$ for the linear model and $N(0.5, 1)$ for the Poisson model. Sample sizes are set at $N = 100$ and $N = 200$, and the number of observations per subject n_i is set to be four. The MI were carried out with the `R` package `MICE`, and employees the predictive mean matching method for the data imputation. Since MICE uses a fixed seed for random number generation, we vary this seed using the iteration time in each run. A review of this package and comparisons with other software is given by Horton and Lipsitz (2001). Further, we set the number of multiple imputations $M = 5$ which is generally sufficient to yield efficient results (Rubin 1987). Monotone missing data indicator R_{ij} is generated from the following logistic model

$$\text{logit}\lambda_{ij} = \alpha_0 + \alpha_1 y_{i,j-1}.$$

When considering a MAR missing mechanism, we set $\alpha = (3, 1)'$ and $\alpha = (3, 0.5)'$ for the continuous response; we set $\alpha = (0.5, 0.5)'$ and $\alpha = (0.5, 0.1)'$ for the count response. When considering a MCAR missing mechanism, we set $\alpha = (3, 0)'$ for both models. These parameter setups not only lead to difference missing data mechanism but also different percentage of missing data.

Table 3.6-3.7 summarize the simulation results. Overall, the results from the linear regression model have similar patterns to those of the Poisson regression model. It can be seen that the finite sample biases and RMSE are

reasonably small for WSBE and MI-SBE in all situations. When data MCAR (i.e., $\alpha_1 = 0$), all methods performs quite similarly. This not surprising as we show the naive SBE remains valid under MCAR. When data MAR (i.e., $\alpha_1 \neq 0$), we can see obvious superiority of WSBE and MI-SBE over SBE in terms of bias and RMSE, especially for the estimation of variance components. Although in some cases the naive SBE yields relative small bias for fixed effects, we notice there is a convergent issue for the naive method in the computation. When sample size increases from $N = 100$ to $N = 200$, both bias and RMSE decrease for WSBE and MI-SBE which suggest they produce consistent estimates. However, this is not the case for naive SBE under MAR data. In general, MI-SBE outperforms WSBE. This is not surprising as it is documented in the literature that IPW is generally less efficient (Robins et al. 1995). To improve efficiency, one may consider applying augmented inverse probability weight method (Robins, Rotnitzky and Zhao 1995; Rotnitzky, Robins and Scharfstein 1998; Scharfstein, Rotnitzky and Robins 1999; Bang and Robins 2005). Furthermore, we notice the numerical computation is more stable in MI-SBE.

Table 3.6: Simulation results for the liner regression model

Missingness		SBE		WSBE		MI-SBE	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
				N=100			
(3, 0)	β_0	-0.046	0.143	-0.047	0.144	0.008	0.100
	β_1	0.019	0.077	0.019	0.077	-0.015	0.071
	θ	0.117	0.265	0.121	0.273	0.047	0.115
	ϕ	-0.053	0.177	-0.053	0.177	-0.038	0.141
(3, 0.5)	β_0	-0.041	0.134	-0.058	0.139	-0.003	0.098
	β_1	0.019	0.075	0.012	0.075	-0.010	0.066
	θ	0.135	0.268	0.102	0.248	0.026	0.112
	ϕ	-0.074	0.179	-0.020	0.166	-0.033	0.135
(3, 1)	β_0	-0.035	0.130	-0.069	0.148	0.008	0.106
	β_1	0.017	0.073	0.016	0.077	-0.017	0.058
	θ	0.134	0.272	0.101	0.260	0.048	0.145
	ϕ	-0.073	0.178	-0.012	0.172	-0.008	0.131
				N=200			
(3, 0)	β_0	-0.020	0.100	-0.020	0.099	0.009	0.071
	β_1	0.009	0.059	0.009	0.059	-0.015	0.049
	θ	0.055	0.183	0.055	0.183	0.066	0.100
	ϕ	-0.026	0.124	-0.027	0.124	-0.019	0.102
(3, 0.5)	β_0	-0.027	0.101	-0.042	0.104	-0.003	0.069
	β_1	0.006	0.062	0.008	0.060	-0.007	0.045
	θ	0.085	0.209	0.045	0.187	0.044	0.081
	ϕ	-0.051	0.132	-0.002	0.123	-0.029	0.087
(3, 1)	β_0	-0.016	0.103	-0.037	0.102	-0.005	0.073
	β_1	0.023	0.157	0.011	0.058	-0.008	0.046
	θ	0.080	0.188	0.037	0.175	0.042	0.090
	ϕ	-0.053	0.125	0.010	0.120	-0.023	0.099

Table 3.7: Simulation results for the Poisson regression model

Missingness		SBE		WSBE		MI-SBE	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
				N=100			
(3,0)	β_0	-0.143	0.389	-0.153	0.403	0.041	0.090
	β_1	-0.015	0.252	-0.016	0.247	-0.081	0.150
	θ	0.137	0.339	0.148	0.350	-0.041	0.080
(0.5,0.5)	β_0	-0.262	0.488	-0.150	0.518	0.014	0.087
	β_1	-0.034	0.246	0.042	0.277	-0.075	0.164
	θ	0.337	0.494	0.043	0.404	-0.025	0.065
(0.5,0.1)	β_0	-0.299	0.504	-0.145	0.410	0.026	0.086
	β_1	-0.034	0.255	0.023	0.280	-0.149	0.202
	θ	0.362	0.532	0.068	0.328	0.036	0.081
				N=200			
(3,0)	β_0	-0.090	0.283	-0.087	0.276	0.034	0.069
	β_1	-0.026	0.194	-0.025	0.195	-0.076	0.123
	θ	0.089	0.251	0.084	0.241	-0.027	0.060
(0.5,0.5)	β_0	-0.205	0.392	-0.078	0.273	0.008	0.061
	β_1	-0.044	0.192	0.018	0.198	-0.071	0.130
	θ	0.285	0.421	-0.015	0.220	-0.003	0.047
(0.5,0.1)	β_0	-0.275	0.465	-0.093	0.333	0.016	0.061
	β_1	-0.044	0.216	0.008	0.228	-0.141	0.177
	θ	0.337	0.486	0.017	0.258	0.049	0.071

Chapter 4

Second-order Least Squares Estimation in Linear Mixed Models with Measurement Error on Covariates and Response

4.1 Introduction

Generalized linear mixed models have been widely used in the modeling of longitudinal data where the response can be either discrete or continuous. Various estimation methods for GLMM have been developed in the literature (e.g, Breslow and Clayton 1993; Durbin and Koopman 1997; Rabe-Hesketh, Skrondal and Pickles 2002). However, estimation and inference in a GLMM remain very challenging when some of the covariates are not directly observed but are measured with error.

It is well-known that simply substituting a proxy variable for the unobserved covariate in the model will generally lead to biased and inconsistent estimates of regression coefficients and variance components (e.g, Wang and Davidian 1996; Wang, Lin, Gutierrez, and Carroll 1998; Carroll, Ruppert, Stefanski, and Crainiceanu 2006). To account for the measurement error (ME) as well as the correlation in the longitudinal data, Wang, Lin, Gutierrez and Carroll (1998) proposed the simulation extrapolation (SIMEX) method to correct for the bias of the naive penalized quasi-likelihood estimator in a generalized linear mixed model with measurement error (GLMMeM), while Wang, Lin and Guittierrez (1999), and Bartlett, Stavola and Frost (2009) proposed a regression calibration (RC) approach. However, it is known that both RC and SIMEX approaches yield approximate but inconsistent estimators in general. Tosteson, Buonaccorsi, and Demidenko (1998) proposed a bias-corrected estimator but it was shown to be highly inefficient. Buonaccorsi, Demidenko and Tosteson (2000) proposed the likelihood based methods and Zhong, Fung, and Wei (2002) studied the corrected score approach. However, the ML methods rely strongly on Gaussian assumption for random effects, ME variables and residual error terms. In addition, the likelihood function for GLMMeM is generally intractable. Non- or semi-parametric approaches have also been considered for models with normally distributed measurement errors (Tsiatis and Davidian 2001; Pan, Zeng and Lin 2009). Instrumental variable method have been used by many researchers to overcome ME problems in cross-sectional data (Fuller 1987; Buzas and Stefanski 1996; Wang

and Hsiao 1995, 2010; Carroll, Ruppert, Stefanski, and Crainiceanu 2006; Schennach 2007). In practice, any variable that correlates with the error-prone true covariate can serve as a valid Instrumental variable, e.g., a second independently measurement. Furthermore, the assumption of instrumental variable is weaker than that of replicate data because Instrumental variables can be a biased observation for the true covariates (Carroll and Stefanski 1994; Carroll, Ruppert, Stefanski, and Crainiceanu 2006).

In this chapter, we follow Abarin et. al. (2010) consider the linear mixed models with measurement error (LMMeM) which can be regarded as a special class of GLMMeM. In this model, we not only allow covariates but also response subject to classical ME. Also, we consider both Berkson and classical measurement errors in covariates because it is well-known that the Berkson and classical measurement errors lead to fundamentally different statistical structures and therefore must be treated differently (Carroll, Ruppert, Stefanski, and Crainiceanu 2006). A nonlinear regression model with Berkson error is usually identifiable without extra information (Wang 2004). Classical ME models usually need extra information such as replicate measurements, validation data, instrumental variables, or knowledge of the measurement properties in order to be identifiable (Carroll, Ruppert, Stefanski, and Crainiceanu 2006; Schennach 2007; Wang and Hsiao 2010). Therefore, we propose an exact consistent estimation method for LMMeM based on the method of moments and instrumental variables.

4.2 Linear Mixed Effects Model with Measurement Error

4.2.1 Model Formulation

We define a linear mixed measurement error model (LMMeM) for the jth observation on the ith individual as

$$y_{ij} = X'_{ij}\beta_x + Z'_{ij}\beta_z + B'_{ij}b_i + \varepsilon_{ij}, \;\; i = 1, \ldots, N, \;\; j = 1, \ldots, n_i \quad (4.1)$$

where $y_{ij} \in I\!R$ is the jth response for the ith subject; $b_i \in I\!R^q$ is the random effect having mean zero, covariance $D(\theta)$ and distribution $f_b(t;\theta)$ with unknown parameters $\theta \in I\!R^{p_b}$; and ε_{ij}'s are mutually independent error terms with zero mean and equal variances σ^2_ε; $\beta_x \in I\!R^{p_x}$ and $\beta_z \in I\!R^{p_z}$ are vectors of fixed effects; $Z_{ij} \in I\!R^{p_z}$ and $B_{ij} \in I\!R^q$ are predictors observed without error; $X_{ij} \in I\!R^{p_x}$ is the unobserved predictors. Further, we observe W_{ij} defined as

$$W_{ij} = X_{ij} + \delta_{ij}, \quad (4.2)$$

where $\delta_{ij} \in I\!R^{p_x}$ is the vector of measurement errors. This is called a classical additive ME model (Carroll, Ruppert, Stefanski, and Crainiceanu 2006), which is the most common model for ME on covariates. We also suggest a classical ME for the response as

$$y^w_{ij} = y_{ij} + \xi_{ij}, \quad (4.3)$$

where ξ is a random measurement error with mean zero and covariance matrix $\sigma^2_\xi I$. This model also assumes that ξ is independent from y. Moreover, we

assume that W is surrogate, which means that given the true covariates, W does not provide any extra information about the distribution of the response. Further, ε, δ, and ξ are assumed to be independent from all random variables in the model, as well as from each other. Since there is no assumption concerning the functional forms of the distributions of X, δ, ε, and ξ, model (4.1)-(4.3) is semi-parametric.

4.2.2 Estimation and Inference

In order to overcome the ME problem on X, one might suggest to replace it with the observed variable, which is W. This is how naive procedures estimate the parameters. We will assess later how ignoring ME effect on covariates and/or response can affect the estimation procedure. Moreover, since in this model, W and δ are correlated, simply replacing X by $W - \delta$ violates the independency of the error term $\varepsilon - \delta$ and the covariates W. It is not obvious to determine for which parameters the naive estimator is inconsistent, unless we have more assumptions on the model. For example, Carroll, Ruppert, Stefanski, and Crainiceanu (2006) showed that if X is normally distributed, a classical additive error model holds, and X and Z are independent, then the naive estimator will be consistent only for the fixed and the random effects corresponding to Z. Unlike classical ME, it is straightforward to see how the ME affects the response under the classical additive ME model in (4.3). Since this model assumes that Y and ξ are independent, and ξ has mean zero, then the naive estimator that uses y_{ij}^w

instead of y_{ij} remains unbiased. However, ignoring ME effect on response variable and simply assuming that the error gets absorbed into the model error is a myth. Even an unbiased ME on response increases the variability of the fitted model (Carroll, Ruppert, Stefanski, and Crainiceanu 2006).

Here we assume that an instrumental variable V_{ij} is available and is related to X_{ij} through

$$X_{ij} = \gamma V_{ij} + U_{ij}, \tag{4.4}$$

where γ is a row full rank matrix of unknown parameters and U is independent from V and δ, has mean zero and variance covariance matrix αI. Substituting (4.4) into (4.2) results in a usual linear regression equation

$$E(W_{ij} \mid V_{ij}) = \gamma V_{ij}. \tag{4.5}$$

It is straightforward to estimate γ using (4.5), so here we assume that γ is known. In practice, one can estimate γ either using an external independent sample or a subset of the main sample and estimate the other parameters in the unused sample. Based on model assumptions, we can write three sets of marginal moments as

$$E(y_{ij}^w \mid V_i) = (\gamma V_{ij})'\beta_x + Z_{ij}'\beta_z, \tag{4.6}$$

$$\begin{aligned} E(y_{ij}^w y_{ik}^w | V_i) &= E(y_{ij}^w \mid V_i)E(y_{ik}^w \mid V_i) + B_{ij}D(\theta)B_{ik}' \\ &\quad +\varphi_{jk}\alpha\beta_x'\beta_x + \varphi_{jk}\sigma_\varepsilon^2 + \varphi_{jk}\sigma_\xi^2, \end{aligned} \tag{4.7}$$

and

$$E(y_{ij}^w W_{ik} \mid V_i) = E(y_{ij}^w \mid V_i)\gamma V_{ik} + \varphi_{jk}\alpha\beta_x \tag{4.8}$$

$\varphi_{jk} = 1$ if $j = k$, and zero otherwise. Following the convention of mixed modeling literature, throughout this chapter all expectations are taken conditional on B_i and Z_i implicitly. Shennach (2007), Wang and Hsiao (2010) have shown that a general model with independent cross-sectional data can be identified using instrumental variables and these moment equations, provided certain regularity conditions hold.

In this model, the observed variables are $(y_{ij}^w, W_{ij}', V_{ij}', Z_{ij}', B_{ij}')'$ and the parameter of interest is $\psi = (\beta_x', \beta_z', \theta', \alpha', \sigma_\varepsilon^2)'$. In practice, σ_ξ^2 is not usually the parameters of interest, and its estimation is straightforward, therefore it is assumed to be known in the following. Hence, the theoretical results may be regarded as conditional on the pre-estimate of σ_ξ^2 and γ. Let $\rho_i(\psi)$ as $(y_{ij}^w - E(y_{ij}^w \mid V_i), 1 \leq j \leq n_i, y_{ij}^w y_{ik}^w - E(y_{ij}^w y_{ik}^w | V_i), y_{ij}^w W_{ik} - E(y_{ij}^w W_{ik} \mid V_i), 1 \leq j \leq k \leq n_i)'$, then the method of moments estimator (MME) for ψ is defined as

$$\hat{\psi}_N = \underset{\psi \in \Omega_\psi}{\operatorname{argmin}}\, Q_N(\psi) = \underset{\psi \in \Omega_\psi}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_i'(\psi) A_i \rho_i(\psi), \tag{4.9}$$

where A_i is a nonnegative definite matrix which may depend on V, Z and B. We should mention in here that adding interaction terms between X and one or more variables in design matrix X does not affect our estimation procedure. Wang, Lin, Gutierrez, and Carroll (1998) showed that the naive ML estimates of the coefficients subject to ME are asymptotically biased.

Theorem 4.2.1. *Under some regularity conditions $\hat{\psi}_N$ is strongly consistent and $\sqrt{N}(\hat{\psi}_N - \psi_0) \xrightarrow{L} N(0, D_\psi^{-1} C D_\psi^{-1})$ as $N \to \infty$, where*

$$C = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} A_i \rho_i(\psi_0) \rho_i'(\psi_0) A_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right] \tag{4.10}$$

and,

$$D_\psi = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} A_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right]. \tag{4.11}$$

The theorem actually shows that MME gets closer to the true value of parameter, when the sample size increases. Therefore, the finite sample bias of this method decreases with the increase in the sample size. However, it is not the case for the naive estimator. The bias in the naive estimator does not decrease with the sample size since it is a function of the variability in ME. Therefore, we expect the MME to be more efficient. The above asymptotic covariance matrix depends on the weighting matrix A_i. It is of interest to choose an appropriate matrix A_i to obtain the most efficient estimator. It can be shown (Abarin and Wang 2006) that the most efficient choice of weight is $A_i^{opt} = E[\rho_i(\psi_0)\rho_i'(\psi_0)|V_i]^{-1}$.

4.3 Berkson Measurement Error Models for Covariates

4.3.1 Model Formulation

A Berkson measurement error model for X_{ij} is defined as

$$X_{ij} = W_{ij} + \delta_{ij}, \tag{4.12}$$

where δ is a random measurement error with mean zero and variance covariance matrix $\sigma_\delta^2 I$, and independent from W. Although (4.12) might look similar to (4.2), they are actually very different. In a Berkson model, the true covariate is assumed to have more variability than the observed covariate, and W_{ij} is reasonably assumed to be independent of δ_{ij}. Substituting (4.3) and (4.12) into (4.1), we have

$$y_{ij}^w = (W_{ij} + \delta_{ij})'\beta_x + Z_{ij}'\beta_z + B_{ij}'b_i + \varepsilon_{ij} + \xi_{ij}. \tag{4.13}$$

Comparing the parameters in (4.1) to those in (4.13), we can see that the naive estimator of fixed effects and the variance components of θ are consistent. The only variance component for which the naive estimator is inconsistent, is α. Since the error in (4.13) is $\varepsilon_{ij} + \delta_{ij}'\beta_x + \xi_{ij}$, the naive estimator is consistent for $\sigma_\varepsilon^2 + \sigma_\delta^2\beta_x'\beta_x + \sigma_\xi^2$, instead of σ_ε^2. Even if we estimate σ_ξ^2 either using an external sample or a subset of the main sample in advance, by

$$\sigma_\xi^2 = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \frac{(n_i - 1)(y_{ij}^w - \bar{y_{.j}^w})^2}{\sum_{i=1}^{N}(n_i - 1)},$$

σ_ε^2 is still not identifiable. This estimator is crucial for predicting the response using the true covariates. More specifically, in testing hypothesis, the presence of ME on some of the covariates, and as a result of that, overestimation of σ_ε^2 can cause "false negative" results. It can be estimated either by some assumptions on the distribution of X or δ, or using external or internal subset of the primary data.

4.3.2 Estimation and Inference

Now, we show that using only the first two moment equations, we can estimate all the parameters of the model. Based on the methodology in Wang (2004), under the model assumption, we have

$$E(y_{ij}^w|W_i) = W_{ij}'\beta_x + Z_{ij}'\beta_z, \tag{4.14}$$

and the moments of S_{ij} given W and Z are

$$\begin{aligned} E(y_{ij}^w y_{ik}^w|W_i) &= E(y_{ij}^w|W_i)E(y_{ik}^w|W_i) + B_{ij}D(\theta)B_{ik}' \\ &\quad +\varphi_{ik}\sigma_\delta^2\beta_x'\beta_x + \varphi_{ik}\sigma_\varepsilon^2 + \varphi_{ik}\sigma_\xi^2, \end{aligned} \tag{4.15}$$

For this case, we define $\rho_i(\psi) = (y_{ij}^w - E(y_{ij}^w|W_i), 1 \le j \le n_i, y_{ij}^w y_{ik}^w - E(y_{ij}^w y_{ik}^w|W_i), 1 \le j \le k \le n_i)'$, and the method of moments estimator (MME) for ψ is defined as

$$\hat{\psi}_N = \underset{\psi\in\Omega_\psi}{\operatorname{argmin}}\, Q_N(\psi) = \underset{\psi\in\Omega_\psi}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_i'(\psi)A_i\rho_i(\psi), \tag{4.16}$$

where A_i is a nonnegative definite matrix which may depend on W, Z and B. Similar to the classic model for X, it can be shown that $\hat{\gamma}_N$ is strongly consistent and asymptotically normally distributed, with mean zero and the covariance matrix is given in the same form as (4.10) and (4.11).

4.4 Monte Carlo Simulation Studies

In this section, we carry out some simulation studies with different scenarios to show the impact of ME on covariates only, or on both covariates and

responses, using the method of moment and the naive maximum likelihood estimators. We are also interested in examining the effect of the sample size on the estimators and their finite sample behavior. We examined these issues under both an additive classical and Berkson ME models. Moreover, we investigated the sensitivity of MME under misspecification of the ME model.

4.4.1 Design of Simulation Studies

We considered the following LMMeM with two different sample sizes $N = 100$ and $N = 300$.

$$y_{ij} \quad = \quad \beta_0 + \beta_1 X_{ij} + b_i + \varepsilon_{ij}, j = 1, ..., 4. \tag{4.17}$$

The random intercept b_i was generated from a normal distribution with mean zero and variance 0.25, and ε_{ij} was generated from a standard normal distribution. For each of the sample sizes, 1000 datasets were simulated. All computations were done in R and the maximum likelihood estimates were obtained from the `lmer` package. The MME was computed using fully estimated optimal weight. To determine how well the methods perform, we present the estimation bias and RMSE of the estimators. To eliminate some potential nonlinear numerical optimization problems in the determination of the starting points, the true parameter values were used as starting values for the minimization and the optimal weight calculation for the MME method.

For a classical ME model, the instrumental variable was generated from

a standard normal distribution. Therefore, we could generate X through an instrumental model that describes the relation between X and V according to

$$X_{ij} = 1.2 + 0.4 * V_{ij} + U_{ij}, \tag{4.18}$$

where U and V are generated from a standard normal distribution. We generated δ from a standard normal distribution. This variability of ME on X is considered quite large in ME literature. However, in the following simulation studies, we show that unlike some other methods such as RC, MME works quite satisfactory, even when ME has large variability. We also generated W according to a time variant version of model (4.2).

For the Berkson case, we considered X as a time variant variable. For both classical and Berkson ME models, the error-prone response was generated according to model (4.17).

To examine the sensitivity of the MME when we have misspecification in the ME model on covariates, we assumed that the true ME model for X is classical, when it was actually Berkson. A classical ME is the most frequently used model, so most likely to be chosen by default when one does not know the details of the design of a study. In order to ensure that all the relationships between the variables are satisfied, we generated U and δ independently from a standard normal distribution, and then generated W and V from a bivariate normal distribution with mean vector $(0.2, 0)'$ and variances of 3.96 and 1.4, respectively. The covariance between the two

variables can be easily calculated based on (4.18) and the classical ME for X. In the last step, we generated X from a Berkson model.

4.4.2 Simulation Results

Tables 4.1-4.5 summarize the results of the simulations. Tables 4.1 and 4.2 show the results for a classical model, when we have either ME on X only (Table 4.1) or ME on both X and Y (Table 4.2). As expected, the naive estimator for all the parameters (except θ) is more biased compared to the MME. We also notice that the bias in MLE is persistent even when we increase the sample size to 300 while the bias in MME reduces. MLE has a smaller bias than MME in estimating θ because the MLE of θ_0 is unbiased (Wang, Lin, Gutierrez, and Carroll 1998). This is not very surprising since MLE is using the strength of the full information on the distribution of X, ε, and θ. Both estimators have smaller bias on θ when the sample size increases. The large bias in the naive variance estimator of σ_ε^2 shows an overestimation of the variability of the model error term. This bias increases even more when MLE ignores the ME on both X and Y.

Tables 4.3 and 4.4 summarize the results for the Berkson case with either Berkson ME on X only (Table 4.3) or Berkson ME on X and classical ME on Y (Table 4.4). Although the MME shows a larger finite sample bias in estimating β_0, β_1, and θ, the MLE has a much larger bias in estimating σ_ε^2. The finite sample bias in MME reduces with increasing N but this is not

Table 4.1: Bias(RMSE) of the MLE and MME based on the classical ME model with ME on X

	$N = 100$		$N = 300$	
True Value	MLE	MME	MLE	MME
$\beta_0 = 8$	0.1031(0.1499)	0.0060(0.0921)	0.1003(0.1197)	-0.0001(0.0631)
$\beta_1 = 2$	-0.5068(0.5093)	-0.0406(0.0669)	-0.5050(0.5059)	-0.0176(0.0436)
$\theta = 0.25$	0.0078(0.1786)	0.0079(0.3487)	-0.0021(0.1113)	0.0063(0.2250)
$\sigma_\varepsilon^2 = 1$	2.9808(2.9963)	0.0121(0.4393)	2.9961(3.0019)	0.0240(0.3018)

Table 4.2: Bias(RMSE) of the MLE and MME based on the classical ME model with ME on both X and Y

	$N = 100$		$N = 300$	
True Value	MLE	MME	MLE	MME
$\beta_0 = 8$	0.0952(0.1534)	0.0024(0.1036)	0.0996(0.1214)	-0.0001(0.0690)
$\beta_1 = 2$	-0.5035(0.5066)	-0.0441(0.0759)	-0.5044(0.5055)	-0.0213(0.0511)
$\theta = 0.25$	0.0221(0.2166)	0.0352(0.4088)	0.0003(0.1323)	0.0194(0.2536)
$\sigma_\varepsilon^2 = 1$	3.9773(3.9960)	-0.0605(0.5110)	3.9874(3.9943)	0.0031(0.3433)

the case for MLE. This bias increases to a very large amount when we have ME on both X and Y. This indicates the large impact of ignoring ME on both covariate and response for the naive estimator.

Table 4.3: Bias(RMSE) of the MLE and MME based on the Berkson ME model with ME on X

	$N = 100$		$N = 300$	
True Value	MLE	MME	MLE	MME
$\beta_0 = 8$	0.0021(0.1587)	0.0009(0.1526)	-0.0006(0.0904)	-0.0062(0.0952)
$\beta_1 = 2$	0.0004(0.0801)	-0.0322(0.0782)	-0.0007(0.0456)	-0.0103(0.0487)
$\theta = 1.96$	0.0077(0.3510)	-0.0942(0.4650)	0.0123(0.1994)	-0.0197(0.2684)
$\sigma_\varepsilon^2 = 1$	0.9996(1.013)	-0.0795(0.2950)	1.0052(1.0095)	-0.0289(0.1760)

Table 4.4: Bias(RMSE) of the MLE and MME based on the Berkson ME model with ME on both X and Y

	$N = 100$		$N = 300$	
True Value	MLE	MME	MLE	MME
$\beta_0 = 8$	0.0062(0.1595)	-0.0169(0.1595)	-0.0013(0.0939)	-0.0107(0.0977)
$\beta_1 = 2$	0.0023(0.0962)	-0.0295(0.0959)	-0.0012(0.0555)	-0.0080(0.0585)
$\theta = 1.96$	0.0155(0.3973)	0.0180(0.5260)	0.0050(0.2243)	0.0088(0.2881)
$\sigma_\varepsilon^2 = 1$	1.9948(2.0098)	-0.1580(0.4084)	2.0024(2.0072)	-0.0485(0.2199)

Table 4.5 shows that under misspecified ME model for X, MME still provides quite satisfying estimators for fixed effects, even for a relatively small sample size. Although the estimators for the variance of the random

Table 4.5: Bias(RMSE) of the MME based on the misspecified ME model with ME on X

True Value	Bias	RMSE
$\beta_0 = 8$	-0.0552	0.1740
$\beta_1 = 2$	-0.0238	0.0851
$\theta = 1.96$	0.3993	0.8383
$\sigma_\varepsilon^2 = 1$	2.5306	2.6248

effect and the model error term are biased, the results are encouraging, since fixed effects are often of more interest. Considering that MME does not use any distributional assumptions on any of the random variables in the model, it still provides satisfactory estimators for most of the parameters of interest in real applications. The large biases in θ and σ_ε^2 can be explained by two factors. Firstly, the ME model is a part of the full model. If the ME model is misspecified, the full model will be misspecified. Secondly, the correlation between V and X is weaker than the correlation between W and X in a Berkson case, so the estimates based on V are usually less accurate than those based on W. Comparing Tables 4.3 and 4.5, one might find that MLE is a better estimator, in the case of misspecification. Since the naive MLE provides unbiased estimators for both random and fixed effects in a Berkson ME models (Buonaccorsia and Lin 2002), it can be a better choice under the assumed misspecification case provided the distributions of all the variables in the model are correctly specified.

4.5 Example - A Birth and Child Cohort Study

The simulation studies in the previous section consider a relatively simple LMMeM. To better reflect the complexity of the LMMeMs generally applied to longitudinal studies, we generated another set of simulations where the model considered was based on The Western Australian birth and child (Raine) Cohort (Raine Study 2010; Abarin, Li, Wang and Briollais 2010), an ongoing health research study in which pregnant women were recruited between 16 and 18 weeks gestation, and their children followed up from birth to 18-years. Simulations were used instead of the real data directly because the true value of the variables is unknown in the real data. The LMMeM is used to model the children's body mass index (BMI) growth trajectories in this study as a function of the gene FTO (fat mass and obesity-associated) and more particularly the single-nucleotide polymorphism (SNP) rs9939609 in this gene. The purpose of our study is to test for an interaction between this SNP and breast feeding accounting for possible ME in BMI and breast feeding measurements.

In most longitudinal research studies, when BMI at a certain age is collected, the variable of interest for BMI is actually the long term average value of BMI for the person in that year. The reason why the true and observed BMI differ is that weight has daily, as well as seasonal variation. Moreover, since BMI only takes into consideration overall weight and height, it can cause an overestimation or underestimation of the true BMI. For the

ME in the response (error-prone BMI), a classical model seems reasonable, as according to Carroll, Ruppert, Stefanski, and Crainiceanu (2006), BMI is measured uniquely for an individual and it can also be replicated. Therefore, we generated the observed response according to (4.3), where ξ_{ij} was generated from a normal distribution of mean zero and variance 0.1.

Some epidemiological studies showed that self-reported information on duration of exclusive breast feeding tends to be biased (Rios, Neuhauser, Margen and Melnick 1992). The main reason is that generally breast feeding is mixed with other kind of milks and solids, which can mask the real effect of "exclusive" breast feeding (EXBF). In the modeling of BMI growth trajectories, our interest is therefore to propose a ME model for the duration of breast feeding (BF), considering EXBF as the true value. We select a classical model for the ME of BF, as it seems there is more variability in the observed (BF) than the true value (EXBF) (Rios, Neuhauser, Margen and Melnick 1992). Another motivation is that the duration of breast feeding measurements can be replicated. In some studies replicates are not available, such as measures of radiation exposure. We considered $EXBF$ as a time invariant variable, as it is observed once for every individual. In the ME model, δ is generated from a standard normal distribution. In the design of the study, if the same duration for breast feeding was assigned to a group of women and the true value changed around it (women can feed their babies longer or shorter time than what is assigned to them), the ME model could be Berkson. The instrumental variable V that we use for the study is the

minimum values of the age that women stopped breast feeding and the age at which mothers started to feed their babies with other kind of milks. Our study on Raine data shows that instrumental variable is related to $EXBF$ according to (4.18), where U is independent from V and δ, and has a standard normal distribution. We generated V from a uniform distribution with minimum and maximum value 3 and 10 months, respectively.

We simulated data from a previous LMM applied to this data, and considered an association between age, gender, the SNP in the FTO gene assuming a dominant genetic model (the homozygotes for the rare allele and heterozygotes have the same β parameters), duration of mother's (exclusive) breast feeding, and the interaction between the gene and the duration of exclusive breast feeding as covariates, and BMI as the response. Age was our only time-dependent variable. We generated 400 individuals with 7 observations at ages 2, 3, 5, 8, 10, 13, and 15 years. The vector of fixed effects includes an intercept, breast feeding (error-prone variable), age, age^2, age^3, gender, and the FTO SNP (carriers of the minor allele versus non-carriers the minor allele), and also its interaction with duration of breast feeding. We also assumed that we have a random intercept and a random effect on age (i.e. slope parameter) for this model, as it appears that the variability of BMI between individuals changes with age, and also at birth. We generated the independent random effects from a normal distribution with mean zero and a diagonal covariance matrix $(\theta_{11} = 5.5225, \theta_{22} = 0.1156)'$. ε_{ij} in this model was generated from a normal distribution with mean zero

and variance 1.03. Based on the RAINE data, the vector of fixed effects is $(16.8, 0.6, 0.055, -0.004, 0.4, -0.2, 0.27, -0.05)'$. For each of the sample sizes considered, we generated 1000 Monte Carlo simulations and the Monte Carlo mean estimates and root mean squared errors (RMSE) for the estimators were computed. All computations were done in R and the naive ML estimates are obtained from `lmer` package. We did not consider σ_ξ^2, σ_ε^2, and α as parameters of interest. Therefore, we treated them as known.

We computed the estimators of the vector of parameters using (4.9). We used the diagonal matrix form of the weighting matrix to compute A_i. Table 4.6 shows the estimates of parameters using both the maximum likelihood and the method of moment estimations, as well as bias and root mean squared error of the estimators. Although for most parameters of interest MLE has smaller RMSE, a closer look at the bias indicates that MLE is converging to a wrong target. More specifically, for the effects that are related to ME (like the effects for EXBF and the interaction), MME provides more accurate estimates. As we mentioned in the previous section, the naive estimator of the coefficients corresponding to BF are asymptotically biased. Since the naive estimator of the coefficients corresponding to BF is biased, we can say that the estimator of the gene-environment interaction term involving BF is also biased. For the estimate of the residual error variance, the naive MLE overestimates it by a large extent, while MME is nearly unbiased. This is mainly because the variance induced by ME is not accounted for, in the naive estimate and is subsequently attributed to the residual error .

Table 4.6: Bias and RMSE of the MLE and MME

		MLE		MME	
Effect	True Value	Bias	RMSE	Bias	RMSE
Intercept	16.8	0.2788	0.4724	0.1369	0.4492
Age	0.6	0.0028	0.0784	-0.0337	0.1845
Age^2	0.055	-0.0004	0.0103	0.0051	0.0255
Age^3	-0.004	0.0000	0.0004	-0.0002	0.001
$Gender$	0.4	0.0033	0.2599	-0.0160	0.2968
$EXBF$	-0.2	-0.0744	0.1092	-0.0154	0.1623
$Gene$	0.27	0.0528	1.0749	0.0154	1.0066
$Interaction$	-0.05	-0.0178	0.2559	-0.0069	0.2775
θ_{11}	5.5225	-0.0411	0.4391	-0.0035	0.4558
θ_{22}	0.1156	0.0000	0.0085	-0.0036	0.0668
σ^2_ε	1.03	-0.1004	0.1071	-0.0027	0.0945

For the effect of the FTO SNP, MME also provides a better estimator. Carroll, Ruppert, Stefanski, and Crainiceanu (2006) showed that the naive estimator of the effect on the accurately measured covariate that is dependent on the error-prone covariate, is biased. MME also shows a smaller bias on the estimates of intercept and θ_{11}. Wang, Lin, Gutierrez, and Carroll (1998) showed that the naive estimator of the intercept for Gaussian data is asymptotically biased. However, the result is surprising for θ_{11}, as theoretically, we do not expect to have much difference between the naive estimator and MME. Both estimators provide quite satisfactory estimators with no apparent bias for the effect of Age^2, Age^3, and θ_{22}. However, MLE performs better than MME in estimating the effects of Age and $Gender$. Overall, con-

sidering the fact that MLE actually benefits from the advantage of assuming a distribution for error terms and the ME variable, as well as random effects, the results are even more satisfactory for MME.

Chapter 5

Second-order Least Squares Estimation in Generalized Linear Mixed Models with Measurement Error

5.1 Introduction

In Chapter 4, we studied the method of moments estimators for the LMMeM (a special class of GLMMeM) and assumed a simple linear relationship between ME variables and instrumental variables. We derived the asymptotic variance matrix of the MME assuming the regression coefficients between ME variables and instrumental variables are known. This chapter further extends the method of moments for the GLMMeM using the instrumental variable approach. Here we only consider classical ME in covariates but assume a more general nonlinear regression relationship between ME variables and instrumental variables. We also derive the asymptotic covariance ma-

trix of the proposed estimators by accounting for the estimation error of the regression/nuisance parameters.

5.2 Generalized Linear Mixed Models with Covariate Measurement Error

5.2.1 Model Formulation

Consider the following generalized linear mixed model with measurement error (GLMMeM)

$$g^{-1}\left(E(y_{ij}|b_i, X_{ij})\right) = X'_{ij}\beta_x + Z'_{ij}\beta_z + B'_{ij}b_i, \tag{5.1}$$

$$V(y_{ij}|b_i, X_{ij}) = \phi\nu(g(X'_{ij}\beta_x + Z'_{ij}\beta_z + B'_{ij}b_i)) \tag{5.2}$$

where $i = 1, \ldots, N,\ \ j = 1, \ldots, n_i,\ y_{ij} \in I\!R$ is the jth response for the ith subject; $b_i \in I\!R^q$ is the random effect having mean zero and distribution $f_b(t;\theta)$ with unknown parameters $\theta \in I\!R^{p_b}$; $\beta_x \in I\!R^{p_x}$ and $\beta_z \in I\!R^{p_z}$ are vectors of fixed effects; $g^{-1}(\cdot)$ is a link function; $\nu(\cdot)$ is a known variance function and $\phi \in I\!R$ is a scalar parameter that may be known or unknown. It is assumed that y_{ij} given b_i are independent and belong to an exponential family. Further, $Z_{ij} \in I\!R^{p_z}$ and $B_{ij} \in I\!R^q$ are known predictors observed without error; and $X_{ij} \in I\!R^{p_x}$ is unobservable. Instead one observes

$$W_{ij} = X_{ij} + \delta_{ij}, \tag{5.3}$$

where δ_{ij} is the vector of measurement errors. Model (5.1) - (5.2) has been studied by various authors, e.g., Wang, Lin, Gutierrez, and Carroll (1998);

Buonaccorsi, Demidenko and Tosteson (2000); Zhong, Fung and Wei (2002); Carroll, Ruppert, Stefanski, and Crainiceanu (2006).

5.2.2 Model Identifiability

It is known that the parameters of classical ME models generally require extra information in order to be identified (Carroll, Ruppert, Stefanski, and Crainiceanu 2006; Schennach 2007). Moreover, even if certain ME models are identifiable, additional information is useful to improve the efficiency of estimation (Schneeweiss and Augustin 2005). The common source of additional data includes: replicate measurements, validation data, instrumental variables, or knowledge of the measurement error distributions. Here we assume that one observes a set of instrumental variables $V_{ij} \in I\!R^{p_V}$ that is related to the error-prone predictor X_{ij} through

$$X_{ij} = m(V_{ij}; \gamma) + U_{ij}, \tag{5.4}$$

where $m(\cdot)$ is a known function, $\gamma \in I\!R^{p_v}$ is a vector of unknown parameters, $U_{ij} \in I\!R^{p_x}$ is independent of V_{ij} and has mean zero and distribution $f_U(u; \alpha)$ with unknown parameters $\alpha \in I\!R^{p_u}$. Further, we assume that the ME δ_{ij} is independent of X_{ij}, V_{ij} and y_{ij}, $E(y_{ij}|X_{ij}, b_i) = E(y_{ij}|X_{ij}, V_{ij}, b_i)$ and $E(y_{ij}y_{ik}|X_{ij}, b_i) = E(y_{ij}y_{ik}|X_{ij}, V_{ij}, b_i)$ where $j \leq k$. Following the convention of mixed modeling literature, throughout this chapter all expectations are taken conditional on B_i and Z_i implicitly. There are no assumption on the functional forms of the distributions of X_{ij} and δ_{ij}. In this model, the

observed variables are $(y_{ij}, W'_{ij}, V'_{ij}, Z'_{ij}, B'_{ij})'$ and the parameter of interest is $\psi = (\beta'_x, \beta'_z, \theta', \alpha', \phi)'$.

To estimate all unknown parameters in the model, we first note that substituting (5.4) into (5.3) results in a usual regression equation

$$E(W_{ij}|V_{ij}) = m(V_{ij};\gamma) \tag{5.5}$$

which can be used to obtain consistent estimator for γ by least squares method. In practice, γ can be pre-estimated using an external sample or a subset of the main sample. We denote $X_i = (X'_{i1}, X'_{i2}, \ldots, X'_{in_i})'$, and denote W_i, V_i, Z_i, B_i and Y_i analogously. By model assumptions and the law of iterated expectation, we have the following moments

$$\begin{aligned}
\kappa_{1,ij}(\psi) &= E(y_{ij}|V_i) \qquad (5.6)\\
&= E\left[E(y_{ij}|b_i, X_i, V_i)|V_i\right]\\
&= E\left[E(y_{ij}|b_i, X_i)|V_i\right]\\
&= E\left[g(X'_{ij}\beta_x + Z'_{ij}\beta_z + B'_{ij}b_i)|V_i\right]\\
&= \int g\left[(m(V_{ij};\gamma)+u)'\beta_x + Z'_{ij}\beta_z + B'_{ij}t\right] f_b(t;\theta) f_U(u;\alpha)dtdu,
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\kappa_{2,ijk}(\psi) &= E(y_{ij}y_{ik}|V_i) \qquad (5.7)\\
&= E\left[E(y_{ij}|b_i,X_i)E(y_{ik}|b_i,X_i)|V_i\right]\\
&= \int g\left[(m(V_{ij};\gamma)+u)'\beta_x + Z'_{ij}\beta_z + B'_{ij}t\right]\times\\
&\quad g\left[(m(V_{ik};\gamma)+u)'\beta_x + Z'_{ik}\beta_z + B'_{ik}t\right] f_b(t;\theta)f_U(u;\alpha)dtdu +\\
&\quad \varphi_{jk}\phi\int \nu\{g\left[(m(V_{ij};\gamma)+u)'\beta_x + Z'_{ij}\beta_z + B'_{ij}t\right]\}f_b(t;\theta)f_U(u;\alpha)dtdu,
\end{aligned}$$

and

$$\begin{aligned}
\kappa_{3,ijk}(\psi) &= E(y_{ij}W_{ik}|V_i) \qquad (5.8)\\
&= E(y_{ij}X_{ik}|V_i)\\
&= E\left[X_{ik}E(y_{ij}|b_i,X_i)|V_i\right]\\
&= \int (m(V_{ik};\gamma)+u)g\left[(m(V_{ij};\gamma)+u)'\beta_x + Z'_{ij}\beta_z + B'_{ij}t\right] f_b(t;\theta)f_U(u;\alpha)dtdu,
\end{aligned}$$

where $\varphi_{jk}=1$ if $j=k$ and zero otherwise. In the following we consider three popular GLMMeM examples.

Example 5.2.1. Consider a linear mixed model with continuous responses and an identity link function $g(\cdot)$. Assuming U_{ij} has mean zero and variance matrix αI, and b_i has mean zero and covariance matrix Σ_b, we have the explicit form of the moments

$$\begin{aligned}
\kappa_{1,ij}(\psi) &= m(V_{ij};\gamma)'\beta_x + Z'_{ij}\beta_z,\\
\kappa_{2,ijk}(\psi) &= \kappa_{1,ij}(\psi)\kappa_{1,ik}(\psi) + B_{ij}\Sigma_b B'_{ik} + \varphi_{jk}\alpha\beta'_x\beta_x + \varphi_{jk}\sigma^2,\\
\kappa_{3,ijk}(\psi) &= \kappa_{1,ij}(\psi)m(V_{ik};\gamma) + \varphi_{jk}\alpha\beta_x.
\end{aligned}$$

It is worth noting that no distributional assumptions are required for U_{ij} and b_i to obtain these moments.

Example 5.2.2. Consider a random intercept mixed Poisson model for counts, where $\log E(y_{ij}|b_i, x_{ij}) = \beta_0 + \beta_x x_{ij} + \beta_z z_i + \beta_{xz} x_{ij} z_i + b_i$ and $\phi = 1$; x_{ij}, z_i and b_i are scalars. Assuming $b_i \sim N(0, \theta)$ and $u_{ij} \sim N(0, \alpha I)$, we can derive the explicit forms of the moments as

$$\begin{aligned}
\kappa_{1,ij}(\psi) &= \exp\left(\beta_0 + (\beta_x + \beta_{xz} z_i) m(v_{ij}; \gamma) + (\beta_x^2 + \beta_{xz}^2 z_i^2)\alpha/2 + \beta_z z_i + \theta/2\right), \\
\kappa_{2,ijk}(\psi) &= \kappa_{1,ij}(\psi)\kappa_{1,ik}(\psi) \exp\left((\beta_x^2 + \beta_{xz}^2 z_i^2)\alpha + \theta\right) + \varphi_{jk}\kappa_{1,ij}(\psi), \\
\kappa_{3,ijk}(\psi) &= m(v_{ik}; \gamma)\kappa_{1,ij}(\psi) + \varphi_{jk}(\beta_x + \beta_{xz} z_i)\alpha\kappa_{1,ij}(\psi).
\end{aligned}$$

Example 5.2.3. Consider a mixed logistic model for a binary response y_{ij}, where $\phi = 1$ and $g(\cdot)$ is the logistic distribution function. For this model we find

$$\begin{aligned}
\kappa_{1,ij}(\psi) &= \int g\left(m(V_{ij}; \gamma)'\beta_x + Z_{ij}'\beta_z + B_{ij}' t\right) f_b(t; \theta) f_U(u; \alpha) dt du, \\
\kappa_{2,ijk}(\psi) &= \int g\left(m(V_{ij}; \gamma)'\beta_x + Z_{ij}'\beta_z + B_{ij}' t\right) \\
&\quad \cdot g\left(m(V_{ik}; \gamma)'\beta_x + Z_{ik}'\beta_z + B_{ik}' t\right) f_b(t; \theta) f_U(u; \alpha) dt du, \\
\kappa_{3,ijk}(\psi) &= \int (m(V_{ik}; \gamma) + u) g\left((m(V_{ij}; \gamma) + u)'\beta_x + Z_{ij}'\beta_z + B_{ij}' t\right) f_b(t; \theta) f_U(u; \alpha) dt du.
\end{aligned}$$

The above integrals are intractable but can be approximated using Monte Carlo simulators. This case will be treated in the next section.

5.2.3 Estimation and Inference

Since γ is of secondary interest, it is treated as a nuisance parameter and is estimated by nonlinear least squares (NLS) method based on equation (5.5) as

$$\hat{\gamma}_N = \underset{\gamma\in\Omega_\gamma}{\operatorname{argmin}}\, \Psi_N(\gamma) = \underset{\gamma\in\Omega_\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} r_i'(\gamma) r_i(\gamma), \tag{5.9}$$

where $r_i'(\gamma) = (W_{ij} - m(V_{ij};\gamma), 1 \le j \le n_i)$. Under standard regularity conditions, $\hat{\gamma}_N - \gamma_0 = O_p(N^{-1/2})$. Then we replace γ in (5.6)-(5.8) by its least squares estimator $\hat{\gamma}_N$ and denote the moments as $\hat{\kappa}_{1,ij}, \hat{\kappa}_{2,ijk}$, and $\hat{\kappa}_{3,ijk}$ correspondingly. Throughout this chapter, we denote the parameter space of a parameter vector, say ψ, by Ω_ψ. In particular, the parameter spaces of β_x and β_z are denoted as Ω_x and Ω_z respectively. Then the method of moments estimator (MME) for ψ is defined as

$$\hat{\psi}_N = \underset{\psi\in\Omega_\psi}{\operatorname{argmin}}\, Q_N(\psi) = \underset{\psi\in\Omega_\psi}{\operatorname{argmin}} \sum_{i=1}^{N} \hat{\rho}_i'(\psi) A_i \hat{\rho}_i(\psi), \tag{5.10}$$

where $\hat{\rho}_i'(\psi) = (y_{ij} - \hat{\kappa}_{1,ij}(\psi), 1 \le j \le n_i, y_{ij}y_{ik} - \hat{\kappa}_{2,ijk}(\psi), y_{ij}W_{ik} - \hat{\kappa}_{3,ijk}(\psi), 1 \le j \le k \le n_i)$ and $A_i = A(V_i)$ is a nonnegative definite matrix that may depend on V_i.

To derive the consistency and asymptotic normality of $\hat{\psi}_N$, we make the following assumptions.

Assumption 5.2.1. $g(\cdot)$ and $\nu(\cdot)$ are continuously differentiable; $m(v;\cdot)$ is a Lebegue measurable function of v and is continuously differentiable with respect to γ.

Assumption 5.2.2. $(Y_i, W_i, V_i, Z_i, B_i, n_i)$, $i = 1, ..., N$ are independent and identically distributed and satisfy $E\left[\|A_i\|\left(y_{ij}^4 + \|y_{ij}W_{ij}\|^2 + 1\right)\right] < \infty$; Further, there exists a positive function $G(v, t, u)$ satisfying

$$E\left[\|A\|\left(\int G(V,t,u)(\|m(V,\gamma)+u\|+1)dtdu\right)^2\right] < \infty,$$

such that $g^2\left[(m(v,\gamma)+u)'\beta_x + Z_{ij}'\beta_z + B_{ij}'t\right] f_b(t;\theta)f_U(u;\alpha)$ and $\nu\left\{g\left[(m(v,\gamma)+u)'\beta_x + Z_{ij}'\beta_z + B_{ij}'t\right]\right\} f_b(t;\theta)f_U(u;\alpha)$ are bounded by $G(v,t,u)$.

Assumption 5.2.3. The parameter space $\Omega_\psi \subset I\!R^{p_x+p_z+p_b+p_u+1}$ is compact.

Assumption 5.2.4. $E[\rho_i(\psi) - \rho_i(\psi_0)]'A_i[\rho_i(\psi) - \rho_i(\psi_0)] = 0$ if and only if $\psi = \psi_0$.

Assumption 5.2.5. $g(\cdot)$ and $\nu(\cdot)$ are twice continuously differentiable; $f_b(t;\theta)$ and $f_U(u;\alpha)$ are twice continuously differentiable w.r.t to θ and α respectively in some open subsets $\theta_0 \in \Omega_{\theta 0} \subset \Omega_\theta$ and $\alpha \in \Omega_{\alpha 0} \in \Omega_\alpha$. Furthermore, all first and second order partial derivatives of $g\left[(m(V_{ij},\gamma)+u)'\beta_x + Z_{ij}'\beta_z + B_{ij}'t\right] f_b(t;\theta)f_U(u;\alpha)$ and $\nu\left\{g\left[(m(V_{ij},\gamma)+u)'\beta_x + Z_{ij}'\beta_z + B_{ij}'t\right]\right\} f_b(t;\theta)f_U(u;\alpha)$ w.r.t (ψ,γ) are bounded absolutely by the positive function $G(v,t,u)$ given in Assumption 5.2.2.

Assumption 5.2.6. The matrices

$$D_\psi = E\left[\frac{\partial\rho_i'(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_i(\psi_0)}{\partial\psi'}\right], \tag{5.11}$$

$$D_\gamma = E\left[\frac{\partial r_i'(\gamma_0)}{\partial\gamma}\frac{\partial r_i(\gamma_0)}{\partial\gamma'}\right] \tag{5.12}$$

are nonsingular.

Theorem 5.2.4. *As $N \to \infty$,*

1. *under assumptions 5.2.1-5.2.4, $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$;*

2. *under assumptions 5.2.1-5.2.6, $\sqrt{N}(\hat{\psi}_N - \psi_0) \xrightarrow{L} N(0, D_\psi^{-1} C D_\psi^{-1})$, where*

$$
\begin{aligned}
C &= E(C_1 C_1') && (5.13)\\
C_1 &= \frac{\partial \rho_i'(\psi_0)}{\partial \psi} A_i \rho_i(\psi_0) + D_{\psi\gamma} D_\gamma^{-1} \frac{\partial r_i'(\gamma)}{\partial \gamma} r_i(\gamma) && (5.14)\\
D_{\psi\gamma} &= E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} A_i \frac{\partial \rho_i(\psi_0)}{\partial \gamma'} \right] && (5.15)
\end{aligned}
$$

The second term in equation (5.14) is the correction term due to the first-step estimation of γ. If γ_0 is known or estimated using an independent sample from the main sample, then this term vanishes and the most efficient weight is given by $A_i^{opt} = E[\rho_i(\psi_0)\rho_i'(\psi_0)|V_i]^{-1}$ (Abarin and Wang 2006). In practice, direct calculation of A_i^{opt} is not feasible since it involves unknown parameters to be estimated. One possible solution is using a two-stage procedure. First, minimize $Q_N(\psi)$ using $A_i = I$ to obtain the first stage estimator $\hat{\psi}_{N1}$. Second, estimate A_i^{opt} by any nonparametric method or

$$
A^{opt} = \left(\frac{1}{N} \sum_{i=1}^{N} \rho_i(\hat{\psi}_{N1}) \rho_i'(\hat{\psi}_{N1}) \right)^{-1}, \tag{5.16}
$$

and minimizing $Q_N(\psi)$ again using A_i^{opt} to obtain the second stage estimator $\hat{\psi}_{N2}$. In practice, the calculation of A_i^{opt} may be difficult or inaccurate due to its high dimension, so one may consider using certain diagonal weight matrix. A detailed discussion on the choice of A_i^{opt} can be found in Li and Wang (2010).

In general, MME can be computed using Newton-Raphson algorithm as

$$\hat{\psi}^{(\tau+1)} = \hat{\psi}^{(\tau)} - \left(\frac{\partial^2 Q_N(\hat{\psi}^{(\tau)})}{\partial\psi\partial\psi'} \right)^{-1} \frac{\partial Q_N(\hat{\psi}^{(\tau)})}{\partial\psi},$$

where $\hat{\psi}^{(\tau)}$ denotes the estimate of ψ at the τ^{th} iteration,

$$\frac{\partial Q_N(\hat{\psi}^{(\tau)})}{\partial\psi} = 2\sum_{i=1}^{N} \frac{\partial\rho_i'(\hat{\psi}^{(\tau)})}{\partial\psi} A_i \rho_i(\hat{\psi}^{(\tau)}), \tag{5.17}$$

$$\frac{\partial^2 Q_N(\hat{\psi}^{(\tau)})}{\partial\psi\partial\psi'} = 2\sum_{i=1}^{N} \left[\frac{\partial\rho_i'(\hat{\psi}^{(\tau)})}{\partial\psi} A_i \frac{\partial\rho_i(\hat{\psi}^{(\tau)})}{\partial\psi'} + (\rho_i'(\hat{\psi}^{\tau}) A_i \otimes I) \frac{\partial \mathrm{vec}(\partial\rho_i'(\hat{\psi}^{(\tau)})/\partial\psi)}{\partial\psi'} \right]. \tag{5.18}$$

Since the second term in (5.18) has expectation zero, it can be ignored for computational simplicity.

When using the weight (5.16), the MME is able to safeguard against influential measurements. In particular, the influence function (IF) at a single contaminated data point v for subject l takes the form (Hampel et al., 1986)

$$\mathrm{IF}(v; \hat{\psi}_N, F) = -D_\psi(\hat{\psi}_N(F))^{-1} \frac{\partial\hat{\rho}_l'(v; \hat{\psi}_N(F))}{\partial\psi} A_l(v; \hat{\psi}_N(F)) \hat{\rho}_l(v; \hat{\psi}_N(F)), \tag{5.19}$$

where F is the underlying distribution and D_ψ is given in (5.11). If $\hat{\psi}_N$ is computed using the estimated weight (5.16), then analogous to the proof of Theorem 2.2.4 we can prove that $\left\| \mathrm{IF}(v; \hat{\psi}_N, F) \right\| \to 0$ as $\|v\| \to \infty$. Therefore, the influence function of $\hat{\psi}_N$ is bounded and $\hat{\psi}_N$ has a redescending property (Huber 2004), so it is robust to influential observations or outliers.

5.3 Simulation-based Estimator

The numerical computation of MME $\hat{\psi}_N$ is straightforward if the moments in (5.6)-(5.8) admit explicit forms. However, sometimes the integrals involved in these moments are intractable. In this case, we propose a simulation-based approach. The basic idea is to replace the integrals with their Monte Carlo simulators as follows. First, generate random points t_{is} and $u_{is}, i = 1, 2, \ldots, N;\ s = 1, 2, \ldots, 2S$ from known densities $l(t)$ and $h(u)$. Then use the first half of the points t_{is} and $u_{is}, s = 1, 2, \ldots, S$ to compute

$$\begin{aligned}
\kappa^1_{1,ij}(\psi) &= \frac{1}{S}\sum_{s=1}^{S} g\left[(m(V_{ij};\gamma)+u_{is})'\beta_x + Z'_{ij}\beta_z + B'_{ij}t_{is}\right]\frac{f_b(t_{is};\theta)f_U(u_{is};\alpha)}{l(t_{is})h(u_{is})} \\
\kappa^1_{2,ijk}(\psi) &= \frac{1}{S}\sum_{s=1}^{S} g\left[(m(V_{ij};\gamma)+u_{is})'\beta_x + Z'_{ij}\beta_z + B'_{ij}t_{is}\right] \\
&\quad g\left[(m(V_{ik};\gamma)+u_{is})'\beta_x + Z'_{ik}\beta_z + B'_{ik}t_{is}\right]\frac{f_b(t_{is};\theta)f_U(u_{is};\alpha)}{l(t_{is})h(u_{is})} \\
&\quad +\varphi_{jk}\phi\frac{1}{S}\sum_{s=1}^{S}\nu\left(g\left[(m(V_{ij};\gamma)+u_{is})'\beta_x + Z'_{ij}\beta_z + B'_{ij}t_{is}\right]\right)\frac{f_b(t_{is};\theta)f_U(u_{is};\alpha)}{l(t_{is})h(u_{is})} \\
\kappa^1_{3,ijk}(\psi) &= \frac{1}{S}\sum_{s=1}^{S}(m(V_{ik};\gamma)+u_{is})g\left[(m(V_{ij};\gamma)+u_{is})'\beta_x + Z'_{ij}\beta_z + B'_{ij}t_{is}\right]\frac{f_b(t_{is};\theta)f_U(u_{is};\alpha)}{l(t_{is})h(u_{is})}
\end{aligned}$$

and similarly use the second half of the points t_{is} and $u_{is}, s = S+1, S+2, \ldots, 2S$ to compute $\kappa^2_{1,ij}(\psi)$, $\kappa^2_{2,ijk}(\psi)$ and $\kappa^2_{3,ijk}(\psi)$. It is easy to see that $\kappa^\iota_{1,ij}(\psi)$, $\kappa^\iota_{2,ijk}(\psi)$ and $\kappa^\iota_{3,ijk}(\psi)$, $\iota = 1, 2$ are unbiased estimators for $\kappa_{1,ij}(\psi)$, $\kappa_{2,ijk}(\psi)$ and $\kappa_{3,ijk}(\psi)$ respectively. Finally, the simulation-based estimator

(SBE) for ψ is defined as

$$\hat{\psi}_{N,S} = \underset{\psi\in\Omega_\psi}{\operatorname{argmin}}\, Q_{N,S}(\psi) = \underset{\psi\in\Omega_\psi}{\operatorname{argmin}} \sum_{i=1}^{N} \hat{\rho}'_{i,1}(\psi) A_i \hat{\rho}_{i,2}(\psi), \tag{5.20}$$

where $\hat{\rho}_{i,\iota}(\psi) = (y_{ij} - \hat{\kappa}^{\iota}_{1,ij}(\psi), 1 \le j \le n_i, y_{ij}y_{ik} - \hat{\kappa}^{\iota}_{2,ijk}(\psi), y_{ij}W_{ik} - \hat{\kappa}^{\iota}_{3,ijk}(\psi), 1 \le j \le k \le n_i)'$. We refer this simulation technique as simulation-by-parts since $\hat{\rho}_{i,1}(\psi)$ and $\hat{\rho}_{i,2}(\psi)$ are constructed by using two independent sets of random points. The benefit of simulation by parts is that $\hat{\rho}_{i,1}(\psi)$ and $\hat{\rho}_{i,2}(\psi)$ are conditionally independent given $(Y_i, W_i, V_i, Z_i, B_i)$ so that $Q_{N,S}(\psi)$ is an unbiased simulator for $Q_N(\psi)$ for finite S. It is worth noting that the construction of simulated moments only requires b_i and U_{ij} to have certain known parametric forms (not necessary normal). For example, one can follow Davidian and Gallant (1993) and Zhang and Davidian (2001) to represent the density of b_i and U_{ij} by the standard seminonparametric density which includes normal, skewed, multi-modal, fat- or thin-tailed densities. One can also impose the Tukey(g, h) family distribution (Field and Genton 2006) for b_i as well which is generated by a single transformation of the standard normal and covers a variety of distributions.

Theorem 5.3.1. *Suppose that* $Supp(l) \supseteq Supp(f_b(\,\cdot\,;\theta))$ *for all* $\theta \in \Omega_{\theta 0}$, *and* $Supp(h) \supseteq Supp(f_U(\,\cdot\,;\alpha))$ *for all* $\alpha \in \Omega_{\alpha 0}$. *Then for any fixed* $S > 0$, *as* $N \to \infty$,

1. *under assumptions 5.2.1-5.2.4,* $\hat{\psi}_{N,S} \xrightarrow{a.s.} \psi_0$;

2. *under assumptions 5.2.1-5.2.6,* $\sqrt{N}(\hat{\psi}_{N,S} - \psi_0) \xrightarrow{L} N(0, D_\psi^{-1} C_S D_\psi^{-1})$,

where

$$C_S = E\left(C_{1S}C_{1S}'\right), \tag{5.21}$$

$$2C_{1S} = \frac{\partial \rho_{i,1}'(\psi_0)}{\partial \psi} A_i \rho_{i,2}(\psi_0) + \frac{\partial \rho_{i,2}'(\psi_0)}{\partial \psi} A_i \rho_{i,1}(\psi_0) + 2D_{\psi\gamma} D_{\gamma}^{-1} \frac{\partial r_i'(\gamma)}{\partial \gamma} r_i(\gamma) \tag{5.22}$$

Note that the above asymptotic results do not require the simulation size S tends to infinity because we use the simulation-by-parts technique to approximate moments. This is fundamentally different from other simulation-based methods in the literature which typically require S goes to infinity to obtain consistent estimators. However, due to the approximation of marginal moments, $\hat{\psi}_{N,S}$ is generally less efficient than $\hat{\psi}_N$. In general, analogous to the Corollary 4 in Wang (2004) we can show that the efficiency loss caused by simulation decreases at the rate $O(1/S)$.

5.4 Monte Carlo Simulation Studies

In this section, we evaluate the finite sample behavior of the proposed estimators, and compare them with the naive ML estimates. We carried out 500 Monte Carlo replications in each simulation study and reported the biases and the root mean square errors (RMSE). All computations are done in R and the naive ML estimates are obtained from `glmmPQL` package.

In the first simulation study, we considered the mixed Poisson model in Example 5.2.2. We simulated δ_{ij} from $N(0,1)$, set $z_i = 1$ for half the sample

and 0 for the remainder, and set $N = 100, 300$ and $n = 4$. In addition, we set $x_{ij} = 1.5 + 0.5v_{ij} + u_{ij}$, $v_{ij} \sim N(0, 1)$ and $u_{ij} \sim N(0, 0.25)$. Table 5.1 reports the simulation results. For the fixed effects associated with ME, β_0, β_x and β_{xz}, the MME is almost unbiased while the naive MLE is severely downward biased and attenuated towards zero. The MME is considerably more efficient than the naive MLE in terms of smaller RMSE. With the increase of sample size from $N = 100$ to 300, the RMSE and biases of MME are decreasing while the ones from the naive MLE stay almost the same. For exactly measured effect β_z, the MME still provides a better estimate in terms of biases and RMSE which may because z_i interacts with x_{ij}. The naive MLE for β_z is also biased towards zero. However, with the increase of sample sizes, the biases and RMSE reduces for the naive MLE as well as the MME. For the random effect σ_b^2, surprisingly both estimators provide quite satisfactory estimators with no apparent biases.

Table 5.1: Biases(RMSE) for the parameter estimates in the random intercept Poisson models

	$N = 100$		$N = 300$	
Parameter	Naive MLE	MME	Naive MLE	MME
$\beta_0 = 1.00$	1.37 (1.39)	-0.22 (0.27)	1.38 (1.38)	-0.20 (0.23)
$\beta_x = 1.00$	-0.79 (0.79)	-0.05 (0.08)	-0.79 (0.79)	-0.07 (0.08)
$\beta_z = -0.50$	0.38 (0.47)	0.06 (0.17)	0.36 (0.40)	0.05 (0.16)
$\beta_{xz} = 0.25$	-0.20 (0.22)	-0.01 (0.10)	-0.19 (0.20)	-0.02 (0.08)
σ_b^2=1.00	-0.03 (0.21)	-0.04 (0.19)	0.03 (0.28)	-0.05 (0.15)

In the second simulation study, we considered a logistic model for bi-

nary responses. In particular, we adopted the following model used in the simulation studies by Wang, Lin, Gutierrez, and Carroll (1998):

$$\text{logit}(Pr(y_{ij} = 1|b_i, x_{ij}, z_{ij})) = \beta_0 + \beta_x x_{ij} + \beta_z z_{ij} + b_i \tag{5.23}$$

where $b_i \sim N(0, 0.5)$, $z_{ij} \sim N(0, 1)$ and $\delta_{ij} \sim N(0, 1)$. In addition, we assumed an instrumental variable is observed that relates to x_{ij} though $x_{ij} = 1.5+0.5v_{ij}+u_{ij}$, $v_{ij} \sim N(0, 1)$ and $u_{ij} \sim N(0, 0.5)$. In the present simulation, we selected $N = 50, 100$ and $n = 3$. The closed form of the marginal moments are not available so we applied the SBE in this case. To compute the SBE, we chose the density of $N(0, 2)$ to be $h(u)$ and $l(t)$, and generated independent points u_{is} and t_{is}, $s = 1, ..., 2S$ using $S = 1000$. The simulation results are presented in Table 5.2. For the fixed effects associated with ME, β_0 and β_x, SBE is almost unbiased while the naive ML is severely downward biased and attenuated towards zero. With the increase of sample size from $N = 50$ to 100, the RMSE and biases of MME are decreasing while the ones from the naive ML stay almost the same. This is the same findings as the ones in the first simulation study. For exactly measured effect β_z, both estimates seem to be unbiased; however, the naive ML provides a better estimates in terms of smaller biases and RMSE. With the increase of sample size, the RMSE and biases from both methods are decreasing. For the random effect, the naive ML overestimates σ_b^2 with larger biases as well as RMSE. With the increase of sample size, both estimators lead to smaller biases and RMSE.

Table 5.2: Biases(RMSE) for the parameter estimates in the random intercept Logistic models

	$N = 50$		$N = 100$	
Parameter	Naive MLE	MME	Naive MLE	MME
$\beta_0 = 0.00$	1.65 (1.69)	0.02 (0.15)	1.61 (1.62)	0.01 (0.08)
$\beta_x = 2.00$	-1.31 (1.32)	0.07 (0.72)	-1.32 (1.32)	0.03 (0.12)
$\beta_z = 1.00$	-0.10 (0.22)	-0.05 (0.49)	-0.11 (0.16)	0.01 (0.22)
$\sigma_b^2 = 0.50$	0.64 (1.06)	0.11 (1.09)	0.51 (0.65)	0.05 (0.13)

Chapter 6

Summary and Future Work

Longitudinal data arise in many areas, such as medical and biological sciences, epidemiology, agriculture, social and environmental sciences. The distinct feature of longitudinal data is that individual subjects are measured repeatedly across time and these measurements are likely to be correlated within the same individual. Although there have been extensive methodological developments for the analysis of longitudinal data, there are still many emerging issues arising in practice. In particular, outlying data, missing data and measurement errors are very common in longitudinal studies, and many of these issues need to be addressed simultaneously in order to draw reliable conclusions from the data. Generalized linear mixed models have been widely used in the modeling of longitudinal data where the response is discrete. In statistical literature, the most popular estimation approach for the GLMM is the maximum likelihood method. However, it is usually difficult to obtain a closed-form expression for the likelihood function when the random

effects are multi-dimensional. Consequently, many methods have been proposed to approximate the integrals in the likelihood function. In addition, for computational convenience, these methods routinely require the normality assumption of random effects and within-subject error variances. Since the random effects are unobservable, it is not feasible to verify their distributional assumptions. It is thus natural to be concerned with these methods yield reliable results when the normal assumption is not appropriate.

This thesis consists of a few major contributions to the theory and method of GLMM inferences. In this thesis, we have proposed the second-order least squares estimation method for the GLMM. This approach does not require the parametric assumptions for the distributions of the unobserved random effects. This estimator can be easily computed if the two marginal moments admit an analytic form. The potential computational issue of deriving the moment equations with multiple integrals has been address by using the method of simulated moments. We have established the consistency and asymptotic normality of the proposed estimators under mild regularity conditions. The finite sample behavior of the proposed estimators have been examined and compared with maximum likelihood methods by simulation studies. The asymptotic confidence intervals and testing hypothesis for the parameters are not studied here but they can be a subject of future research.

Data contaminations or data outliers are common in longitudinal data.

It is known that likelihood-based methods are vulnerable to data outliers because they are based on the normal distribution. It is a concern that the second-order least squares estimators may lack robustness as the second moments used in the estimation may enlarge the outlier impact. We have studied the robustness property of the second-order least squares estimators by means of the influence function. We have proved that they have bounded influence functions under certain form of the estimated optimal weight, and hence, they are robust against data outliers. It is noticed in our simulation studies, there are some finite sample biases for the estimation of variance components by the second-order least squares estimators. These biases are downward-oriented and diminish with increasing sample sizes. We have investigated the source of this finite-sample bias and proposed a bias reduction technique by using independent weights. Simulation studies show that the bias reduction method works well in finite sample with small efficiency loss.

Incomplete longitudinal data are almost inevitable in longitudinal studies due to various reasons. For a valid analysis, a study of the missing mechanism is necessary. Based on the dependence of the missing data on the response process, Little and Rubin (2002) classified missing data mechanisms into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). We have shown that the second-order least squares estimators based on observed data are valid only under MCAR missing data mechanism. Therefore, we have adapted the inverse probability weight method and applied the multiple imputation

approach to accommodate MAR response data. Furthermore, we have suggested a few ways to compute the optimal weight matrix under the incomplete longitudinal data setting. A future research is to develop a strategy for using the second-order least squares in non-ignorable missing data problems.

In comparison with the likelihood-based methods, the second-order least squares approach produces exactly (rather than approximately) consistent estimates; and it requires less distributional assumptions since it allows random effects to have any parametric distribution (not necessarily normal). In comparison with the generalized estimating equation approaches and associated simulation-based methods, the proposed approach is computationally more attractive since it does not require the simulation size to go to infinity in order to produce exactly consistent estimates. Moreover, for computational convenience, generalized estimating equation methods routinely use the "working" correlation matrix which may yield inefficient estimates. In contrast, our approach does not necessarily require the "working" specification of the optimal weight matrix. Unlike the generalized estimating equation methods, the proposed estimators have a well defined objective function, which is useful for hypotheses testing and model selection. It is well known that in the presence of outliers the generalized estimating equation methods will fail to produce consistent estimators and lead to misleading conclusions. A further advantage of the proposed estimators is that they have a bounded influence function and they are robust against data outliers.

Measurement error or errors-in-variable is another challenging area in longitudinal data analysis. It is well-known that simply substituting a proxy variable for the unobserved covariate in the model will generally lead to biased and inconsistent estimates. This thesis has proposed the method of moments estimation for the generalized linear mixed model with measurement error using the instrumental variable approach. This method does not require parametric assumptions for the distributions of the unobserved covariates or of the measurement errors, and it allows random effects to have any parametric distributions (not necessarily normal). The methodology is illustrated through simulation studies. In our measurement error model formulation, we have restricted our attention to the case where only fixed effects are subject to measurement error. Although this is a common model used in the literature, it may not always be realistic. A possible extension of the proposed approach is for the estimation of a generalized linear mixed model with measurement errors in both fixed and random effects. Also, it would be worthwhile extending the proposed estimators for the situations in which discrete variables are measured with error. This problem is common referred to as misclassification (Carroll, Ruppert, Stefanski, and Crainiceanu 2006). Missing data and measurement error often arise simultaneously in a real world problem, so it would be valuable to develop the proposed methodology to cope with these situations.

Appendix A

Appendix: Technical Proofs

A.1 Proof of Theorem 2.2.1

First, for any $1 \leq i \leq N$, by assumptions 2.2.1 - 2.2.2 and Cauchy-Schwartz inequality, we have

$$\begin{aligned} E\left[\|W_i\| \sup_{\Omega} \sum_j (y_{ij} - x'_{ij}\beta)^2\right] &\leq 2\sum_j E\|W_i\| y_{ij}^2 + 2\sum_j E\|W_i\| \|x_{ij}\|^2 \sup_{\Omega} \|\beta\|^2 \\ &< \infty, \end{aligned}$$

and

$$
\begin{aligned}
& E\left[\|W_i\| \sup_{\Gamma} \sum_j \sum_k (y_{ij}y_{ik} - (x'_{ij}\beta x'_{ik}\beta + z'_{ij}Dz_{ik} + \delta_{jk}\sigma^2))^2\right] \\
& \leq 2\sum_j \sum_k E\|W_i\| y_{ij}^2 y_{ik}^2 + 2\sum_j \sum_k E\|W_i\| \sup_{\Gamma} (x'_{ij}\beta x'_{ik}\beta + z'_{ij}Dz_{ik} + \varphi\sigma^2)^2 \\
& \leq 2\sum_j \sum_k E\|W_i\| y_{ij}^2 y_{ik}^2 + 6\sum_j \sum_k E\|W_i\| \sup_{\Omega} \left\|x'_{ij}\beta x'_{ik}\beta\right\|^2 \\
& \quad +6\sum_j \sum_k E\|W_i\| \sup_{\Theta} \left\|z'_{ij}Dz_{ik}\right\|^2 + 6n_i \sup_{\Sigma} \sigma^4 E\|W_i\| \\
& \leq 2\sum_j \sum_k E\|W_i\| y_{ij}^2 y_{ik}^2 + 6\sum_j \sum_k E\|W_i\| \|x_{ij}\|^2 \|x_{ik}\|^2 \sup_{\Omega} \|\beta\|^2 \\
& \quad +6\sum_j \sum_k E\|W_i\| \|z_{ij}\|^2 \|z_{ik}\|^2 \sup_{\Theta} \|D\|^2 + 6n_i \sup_{\Sigma} \sigma^4 E\|W_i\| \\
& < \infty,
\end{aligned}
$$

which imply $E\sup_\Gamma \rho'_i(\psi)W_i\rho_i(\psi) \leq E\|W_i\|\sup_\Gamma \|\rho'_i(\psi)\|^2 < \infty$. Then, it follows from the uniform law of large numbers (ULLN, Jennrich 1969, Theorem 2), that $\frac{1}{N}Q_N(\psi)$ converges almost surely to $Q(\psi) = E\rho'_i(\psi)W_i\rho_i(\psi)$ uniformly for all ψ in Γ. Furthermore, we have

$$
\begin{aligned}
Q(\psi) &= Q(\psi_0) + 2E\rho'_i(\psi_0)W_i(\rho_i(\psi) - \rho_i(\psi_0)) + E(\rho_i(\psi) - \rho_i(\psi_0))'W_i(\rho_i(\psi) - \rho_i(\psi_0)) \\
&= Q(\psi_0) + E\left[(\rho_i(\psi) - \rho_i(\psi_0))'W_i(\rho_i(\psi) - \rho_i(\psi_0))\right]
\end{aligned}
$$

because $\rho_i(\psi) - \rho_i(\psi_0)$ does not depend on Y_i and hence

$$
E\left[\rho'_i(\psi_0)W_i(\rho_i(\psi) - \rho_i(\psi_0))\right] = E\left[E(\rho'_i(\psi_0)|X_i, Z_i)W_i(\rho_i(\psi) - \rho_i(\psi_0))\right] = 0.
$$

Therefore by assumption 2.2.3 $Q(\psi) \geq Q(\psi_0)$ and the equality holds if and only if $\psi = \psi_0$. Thus, all conditions of Lemma 3 in Amemiya (1973) are satisfied, so we have $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$, as $N \to \infty$.

A.2 Proof of Theorem 2.2.2

The first derivative $\partial Q_N(\psi)/\partial\psi$ exists and has the first-order Taylor expansion in Γ. Since $\partial Q_N(\hat{\psi}_N)/\partial\psi = 0$ and $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$, for sufficiently large N we have

$$\frac{\partial Q_N(\hat{\psi}_N)}{\partial\psi} = \frac{\partial Q_N(\psi_0)}{\partial\psi} + \frac{\partial^2 Q_N(\tilde{\psi}_N)}{\partial\psi\partial\psi'}(\hat{\psi}_N - \psi_0) = 0, \tag{A.1}$$

where $\left\|\tilde{\psi}_N - \psi_0\right\| \leq \left\|\hat{\psi}_N - \psi_0\right\|$. The first derivative of $Q_N(\psi)$ in (A.1) is given by

$$\frac{\partial Q_N(\psi)}{\partial\psi} = 2\sum_{i=1}^{N} \frac{\partial\rho_i'(\psi)}{\partial\psi} W_i\rho_i(\psi),$$

where

$$\begin{aligned}\frac{\partial\rho_i'(\psi)}{\partial\psi} &= (-(x_{ij},\ 0,\ 0)', 1 \leq j \leq n_i, \\ &\quad -((x_{ij}x_{ik}' + x_{ik}x_{ij}')\beta,\ (\partial\mathrm{vec}(D)/\partial\theta)\mathrm{vec}(z_{ij}z_{ik}'),\ \delta_{jk})', 1 \leq j \leq k \leq n_i\big).\end{aligned}$$

Moveover, since $\frac{\partial\rho_i'(\psi)}{\partial\psi}W_i\rho_i(\psi)$ are *i.i.d.*, it follows the Central Limit Theorem, as $N \to \infty$,

$$\frac{1}{\sqrt{N}}\frac{\partial Q_N(\psi_0)}{\partial\psi} \xrightarrow{L} N(0, 4C), \tag{A.2}$$

where C is as in (6).

The second derivative of $Q_N(\psi)$ in (A.1) is given by

$$\frac{\partial^2 Q_N(\psi)}{\partial\psi\partial\psi'} = 2\sum_{i=1}^{N}\left[\frac{\partial\rho_i'(\psi)}{\partial\psi}W_i\frac{\partial\rho_i(\psi)}{\partial\psi'} + (\rho_i'(\psi)W_i \otimes I)\frac{\partial\mathrm{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'}\right],$$

where I is the $2N(p+r+1)$ dimensional identity matrix, and

$$\frac{\partial\mathrm{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'} = -\left(\frac{\partial^2\mu_{ij}(\psi)}{\partial\psi\partial\psi'}, 1 \leq j \leq n_i, \frac{\partial^2\eta_{ijk}(\psi)}{\partial\psi\partial\psi'}, 1 \leq j \leq k \leq n_i\right)',$$

with

$$\frac{\partial^2 \mu_{ij}(\psi)}{\partial\psi\partial\psi'} = 0, \text{ and } \frac{\partial^2 \eta_{ijk}(\psi)}{\partial\psi\partial\psi'} = \begin{pmatrix} x_{ij}x'_{ik} + x_{ik}x'_{ij} & 0 \\ 0 & 0 \end{pmatrix}.$$

By assumptions 2.2.1-2.2.2 and Cauchy-Schwartz inequality

$$\begin{aligned} & E \sup_{\Gamma} \left\| \frac{\partial \rho'_i(\psi)}{\partial\psi} W_i \frac{\partial \rho_i(\psi)}{\partial\psi'} \right\| \leq E \left\| W_i \right\| \sup_{\Gamma} \left\| \frac{\partial \rho'_i(\psi)}{\partial\psi} \right\|^2 \\ \leq & \sum_j E \left\| W_i \right\| \left\| x_{ij} \right\|^2 + 2 \sum_j \sum_k E \left\| W_i \right\| \left\| x_{ij} \right\|^2 \left\| x_{ik} \right\|^2 \sup_{\Omega} \left\| \beta \right\|^2 \\ & + \sum_j \sum_k E \left\| W_i \right\| \sup_{\Theta} \left\| \frac{\partial \text{vec}(D)}{\partial\theta} \right\|^2 \left\| z_{ij} \right\|^2 \left\| z_{ik} \right\|^2 + n_i E \left\| W_i \right\| \\ < & \infty, \end{aligned}$$

and

$$\begin{aligned} & E \left(\left\| W_i \right\| \sup_{\Gamma} \left\| \frac{\partial \text{vec}(\partial \rho'_i(\psi)/\partial\psi)}{\partial\psi'} \right\|^2 \right) \leq E \left\| W_i \right\| \sup_{\Omega} \left(\left\| \frac{\partial^2 \eta_{ijk}(\psi)}{\partial\beta\partial\beta'} \right\|^2 \right) \\ \leq & 2 \sum_j \sum_k E \left\| W_i \right\| \left\| x_{ij} \right\|^2 \left\| x_{ik} \right\|^2 \\ < & \infty. \end{aligned}$$

Therefore,

$$\begin{aligned} & E \sup_{\Gamma} \left\| (\rho'_i(\psi) W_i \otimes I) \frac{\partial \text{vec}(\partial \rho'_i(\psi)/\partial\psi)}{\partial\psi'} \right\| \\ \leq & \sqrt{2N(p+r+1)} E \left\| W_i \right\| \sup_{\Gamma} \left\| \rho_i(\psi) \right\| \left\| \frac{\partial \text{vec}(\partial \rho'_i(\psi)/\partial\psi)}{\partial\psi'} \right\| \\ \leq & \sqrt{2N(p+r+1)} \left(E \left\| W_i \right\| \sup_{\Gamma} \left\| \rho_i(\psi) \right\|^2 \right)^{1/2} \left(E \left\| W_i \right\| \sup_{\Gamma} \left\| \frac{\partial \text{vec}(\partial \rho'_i(\psi)/\partial\psi)}{\partial\psi'} \right\|^2 \right)^{1/2} \\ \leq & \sqrt{2N(p+r+1)} \left(E \left\| W_i \right\| \sup_{\Gamma} \left\| \rho_i(\psi) \right\|^2 \right)^{1/2} \left(2 \sum_j \sum_k E \left\| W_i \right\| \left\| x_{ij} \right\|^2 \left\| x_{ik} \right\|^2 \right)^{1/2} \\ < & \infty. \end{aligned}$$

It follows from the ULLN, that $(1/N)\partial^2 Q_N(\psi)/\partial\psi\partial\psi' \xrightarrow{a.s.} \partial^2 Q(\psi)/\partial\psi\partial\psi'$ uniformly for all ψ in Γ, where

$$\frac{\partial^2 Q(\psi)}{\partial\psi\partial\psi'} = 2E\left[\frac{\partial\rho_i'(\psi)}{\partial\psi}W_i\frac{\partial\rho_i(\psi)}{\partial\psi'} + (\rho_i'(\psi)W_i\otimes I)\frac{\partial \text{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'}\right]$$

. Thus, it follows Lemma 4 of Amemiya (1973)

$$\frac{1}{N}\frac{\partial^2 Q_N(\tilde{\psi}_N)}{\partial\psi\partial\psi'} \xrightarrow{a.s.} \frac{\partial^2 Q(\psi_0)}{\partial\psi\partial\psi} = 2B,$$

which is due to the fact that

$$\begin{aligned} E\left[(\rho_i'(\psi_0)W_i\otimes I)\frac{\partial \text{vec}(\partial\rho_i'(\psi_0)/\partial\psi)}{\partial\psi'}\right] &= E\left[(E(\rho_i'(\psi_0)|X_i,Z_i)W_i\otimes I)\frac{\partial \text{vec}(\partial\rho_i'(\psi_0)/\partial\psi)}{\partial\psi'}\right] \\ &= 0. \end{aligned}$$

Since B is nonsingular, for sufficiently large N, we have

$$\sqrt{N}(\hat{\psi}_N - \psi_0) = -\left(\frac{1}{N}\frac{\partial^2 Q_N(\tilde{\psi}_N)}{\partial\psi\partial\psi'}\right)^{-1}\frac{1}{\sqrt{N}}\frac{\partial Q_N(\psi_0)}{\partial\psi}$$

Therefore, by, assumption 2.2.4 and Slutsky's theorem, we have $\sqrt{N}(\hat{\psi}_N - \psi_0) \xrightarrow{L} N(0, B^{-1}CB^{-1})$.

A.3 Proof of Theorem 2.2.4

The IF (2.15) is bounded if and only if $G(v;\hat{\psi}_N, F)$ is bounded. Write

$$\hat{U} = \frac{1}{N}\sum_{i=1}^{N}\rho_i\rho_i' = \frac{1}{N}(V_l + \rho_l\rho_l'),$$

where $V_l = \sum_{i\neq l}\rho_i\rho_i'$. Then by Sherman-Morrison-Woodbury formula, we have

$$\hat{U}^{-1} = N(V_l + \rho_l\rho_l')^{-1} = N\left(V_l^{-1} - \frac{V_l^{-1}\rho_l\rho_l'V_l^{-1}}{1+\rho_l'V_l^{-1}\rho_l}\right)$$

if V_l is nonsingular, V_l^{-1} and $\hat{U}^{-1}$ exist. Therefore,

$$U^{-1}\rho_l = N\left(V_l^{-1}\rho_l - \frac{V_l^{-1}\rho_l\rho_l'V_l^{-1}\rho_l}{1+\rho_l'V_l^{-1}\rho_l}\right) = N\left(\frac{V_l^{-1}\rho_l}{1+\rho_l'V_l^{-1}\rho_l}\right),$$

and accordingly,

$$\left\|\frac{\partial\rho_l'(\psi)}{\partial\psi}U_i^{-1}\rho_l\right\|^2 = N^2\left(\frac{\rho_l'V_l^{-1}\frac{\partial\rho_i(\psi)}{\partial\psi}\frac{\partial\rho_i'(\psi)}{\partial\psi}V_l^{-1}\rho_l}{1+\rho_l'V_l^{-1}\rho_l}\frac{1}{1+\rho_l'V_l^{-1}\rho_l}\right) \to 0$$

as $\|v\| \to \infty$.

A.4 Proof of Corollary 3.2.5.1

For any $1 \le i \le N$, by assumptions 1-3 and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\|\rho_i(\psi)\|^2 &\le 2\sum_j y_{ij}^2 + 2\sum_{j\le k} y_{ij}^2y_{ik}^2 + 2\sum_j\left(\int g(x_{ij}'\beta + z_{ij}'u)f_b(u;\theta)du\right)^2 \\
&\quad +4\sum_{j\le k}\left(\int g(x_{ij}'\beta + z_{ij}'u)g(x_{ik}'\beta + z_{ik}'u)f_b(u;\theta)du\right)^2 \\
&\quad +4\phi^2\sum_j\left(\int \nu(g(x_{ij}'\beta + z_{ij}'u))f_b(u;\theta)du\right)^2 \\
&\le 2\sum_j y_{ij}^2 + 2\sum_{j\le k} y_{ij}^2y_{ik}^2 + 2\sum_j\int g^2(x_{ij}'\beta + z_{ij}'u)f_b(u;\theta)du \\
&\quad +4\sum_{j\le k}\int g^2(x_{ij}'\beta + z_{ij}'u)f_b(u;\theta)du\int g^2(x_{ik}'\beta + z_{ik}'u)f_b(u;\theta)du \\
&\quad +4\phi^2\sum_j\left(\int \nu(g(x_{ij}'\beta + z_{ij}'u))f_b(u;\theta)du\right)^2
\end{aligned}$$

and therefore

$$\begin{aligned} E \sup_{\Gamma} \rho_i'(\psi) W_i \rho_i(\psi) &\leq E \left\| W_i \right\| \sup_{\Gamma} \left\| \rho_i'(\psi) \right\|^2 \\ &\leq 2n_i E \left\| W_i \right\| y_{ij}^2 + n_i(n_i+1) E \left\| W_i \right\| y_{ij}^2 y_{ik}^2 \\ &\quad + 2n_i E \left\| W_i \right\| \int G(X_i, Z_i, u) du \\ &\quad + 2n_i \left(n_i + 1 + 2 \sup_{\Sigma} \phi^2 \right) E \left\| W_i \right\| \left(\int G(X_i, Z_i, u) du \right)^2 \\ &< \infty. \end{aligned}$$

Hence by the ULLN, $\sup_{\psi \in \Gamma} \left| \frac{1}{N} Q_N(\psi) - Q(\psi) \right| \xrightarrow{a.s.} 0$, where $Q(\psi) = E[\rho_i'(\psi) W_i \rho_i(\psi)]$. Further, since $\rho_i(\psi) - \rho_i(\psi_0)$ does not depend on Y_i,

$$\begin{aligned} Q(\psi) &= E(\rho_i'(\psi) - \rho_i'(\psi_0) + \rho_i'(\psi_0)) W_i (\rho_i(\psi) - \rho_i(\psi_0) + \rho_i(\psi_0)) \\ &= Q(\psi_0) + E(\rho_i(\psi) - \rho_i(\psi_0))' W_i (\rho_i(\psi) - \rho_i(\psi_0)). \end{aligned}$$

It follows from assumption 3.2.4 that $Q(\psi) \geq Q(\psi_0)$ and the equality holds if and only if $\psi = \psi_0$. Thus, all conditions of Amemiya (1973) Lemma 3 are satisfied and therefore $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$, as $N \to \infty$.

A.5 Proof of Corollary 3.2.5.2

By assumption 5 and the dominated convergence theorem, the first derivative $\partial Q_N(\psi)/\partial \psi$ exists and has the first-order Taylor expansion in Γ. Since $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$, for sufficiently large N we have

$$\frac{\partial Q_N(\hat{\psi}_N)}{\partial \psi} = \frac{\partial Q_N(\psi_0)}{\partial \psi} + \frac{\partial^2 Q_N(\tilde{\psi}_N)}{\partial \psi \partial \psi'} (\hat{\psi}_N - \psi_0) = 0, \tag{A.3}$$

where $\left\|\tilde{\psi}_N - \psi_0\right\| \leq \left\|\hat{\psi}_N - \psi_0\right\|$. The first derivative of $Q_N(\psi)$ in (A.3) is given by

$$\frac{\partial Q_N(\psi)}{\partial \psi} = 2\sum_{i=1}^{N} \frac{\partial \rho_i'(\psi)}{\partial \psi} W_i \rho_i(\psi),$$

where

$$\frac{\partial \rho_i'(\psi)}{\partial \psi} = -\left(\frac{\partial \mu_{ij}(\psi)}{\partial \psi}, 1 \leq j \leq n_i, \frac{\partial \eta_{ijk}(\psi)}{\partial \psi}, 1 \leq j \leq k \leq n_i\right)$$

with nonzero first derivatives:

$$\begin{aligned}
\frac{\partial \mu_{ij}(\psi)}{\partial \beta} &= x_{ij} \int g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du, \\
\frac{\partial \mu_{ij}(\psi)}{\partial \theta} &= \int g(x_{ij}'\beta + z_{ij}'u) f_b^{(1)}(u;\theta) du, \\
\frac{\partial \eta_{ijk}(\psi)}{\partial \beta} &= x_{ij} \int g^{(1)}(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta) du \\
&\quad + x_{ik} \int g(x_{ij}'\beta + z_{ij}'u) g^{(1)}(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta) du \\
&\quad + \delta_{jk}\phi x_{ij} \int \nu^{(1)}(g(x_{ij}'\beta + z_{ij}'u)) g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du, \\
\frac{\partial \eta_{ijk}(\psi)}{\partial \theta} &= \int g(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b^{(1)}(u;\theta) du \\
&\quad + \delta_{jk}\phi \int \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b^{(1)}(u;\theta) du, \\
\frac{\partial \eta_{ijk}(\psi)}{\partial \phi} &= \delta_{jk} \int \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b(u;\theta) du.
\end{aligned}$$

Since $\frac{\partial \rho_i'(\psi)}{\partial \psi} W_i \rho_i(\psi)$ are *i.i.d.* with zero mean, it follows from the Central Limit Theorem that, as $N \to \infty$,

$$\frac{1}{\sqrt{N}} \frac{\partial Q_N(\psi_0)}{\partial \psi} \xrightarrow{L} N(0, 4C). \tag{A.4}$$

The second derivative of $Q_N(\psi)$ in (A.3) is given by

$$\frac{\partial^2 Q_N(\psi)}{\partial\psi\partial\psi'} = 2\sum_{i=1}^{N}\left[\frac{\partial\rho_i'(\psi)}{\partial\psi}W_i\frac{\partial\rho_i(\psi)}{\partial\psi'} + (\rho_i'(\psi)W_i\otimes I)\frac{\partial \mathrm{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'}\right],$$

where I is the $2N(p+r+1)$ dimensional identity matrix and

$$\frac{\partial \mathrm{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'} = -\left(\frac{\partial^2\mu_{ij}(\psi)}{\partial\psi\partial\psi'}, 1\le j\le n_i, \frac{\partial^2\nu_{ijk}(\psi)}{\partial\psi\partial\psi'}, 1\le j\le k\le n_i\right)'$$

with nonzero partial derivatives

$$\begin{aligned}
\frac{\partial^2\mu_{ij}(\psi)}{\partial\beta\partial\beta'} &= x_{ij}x_{ij}'\int g^{(2)}(x_{ij}'\beta + z_{ij}'u)f_b(u;\theta)du,\\
\frac{\partial^2\mu_{ij}(\psi)}{\partial\theta\partial\theta'} &= \int g(x_{ij}'\beta + z_{ij}'u)f_b^{(2)}(u;\theta)du,\\
\frac{\partial^2\mu_{ij}(\psi)}{\partial\beta\partial\theta'} &= x_{ij}\int g^{(1)}(x_{ij}'\beta + z_{ij}'u)f_b^{(1)}(u;\theta)du,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2\eta_{ijk}(\psi)}{\partial\beta\partial\beta'} &= x_{ij}x_{ij}'\int g^{(2)}(x_{ij}'\beta + z_{ij}'u)g(x_{ik}'\beta + z_{ik}'u)f_b(u;\theta)du\\
&\quad +2x_{ij}x_{ik}'\int g^{(1)}(x_{ij}'\beta + z_{ij}'u)g^{(1)}(x_{ik}'\beta + z_{ik}'u)f_b(u;\theta)du\\
&\quad +x_{ik}x_{ik}'\int g(x_{ij}'\beta + z_{ij}'u)g^{(2)}(x_{ik}'\beta + z_{ik}'u)f_b(u;\theta)du\\
&\quad +\delta_{jk}\phi x_{ij}x_{ij}'\int \nu^{(2)}(g(x_{ij}'\beta + z_{ij}'u))\left(g^{(1)}(x_{ij}'\beta + z_{ij}'u)\right)^2 f_b(u;\theta)du,\\
&\quad +\delta_{jk}\phi x_{ij}x_{ij}'\int \nu^{(1)}(g(x_{ij}'\beta + z_{ij}'u))g^{(2)}(x_{ij}'\beta + z_{ij}'u)f_b(u;\theta)du,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2\eta_{ijk}(\psi)}{\partial\theta\partial\theta'} &= \int g(x_{ij}'\beta + z_{ij}'u)g(x_{ik}'\beta + z_{ik}'u)f_b^{(2)}(u;\theta)du\\
&\quad +\delta_{jk}\phi\int \nu(g(x_{ij}'\beta + z_{ij}'u))f_b^{(2)}(u;\theta)du,
\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\beta\partial\theta'} &= x_{ij}\int g^{(1)}(x'_{ij}\beta + z'_{ij}u)g(x'_{ik}\beta + z'_{ik}u)f_b^{(1)}(u;\theta)du \\ &\quad + x_{ik}\int g(x'_{ij}\beta + z'_{ij}u)g^{(1)}(x'_{ik}\beta + z'_{ik}u)f_b^{(1)}(u;\theta)du \\ &\quad + \delta_{jk}\phi\int \nu^{(1)}(g(x'_{ij}\beta + z'_{ij}u))g^{(1)}(x'_{ij}\beta + z'_{ij}u)f_b^{(1)}(u;\theta)du,\end{aligned}$$

By assumption 3.2.1, 3.2.2, 3.2.3, 3.2.5 and Cauchy-Schwartz inequality,

we have

$$
\begin{aligned}
& E \sup_{\Psi} \left\| \frac{\partial \rho_i'(\psi)}{\partial \psi} W_i \frac{\partial \rho_i(\psi)}{\partial \psi'} \right\| \\
\leq \; & E \left\| W_i \right\| \sup_{\Psi} \left\| \frac{\partial \rho_i'(\psi)}{\partial \psi} \right\|^2 \\
\leq \; & E \left\| W_i \right\| \sup_{\Psi} \left(\sum_j \left\| \frac{\partial \mu_{ij}(\psi)}{\partial \beta} \right\|^2 + \sum_j \left\| \frac{\partial \mu_{ij}(\psi)}{\partial \theta} \right\|^2 \right. \\
& \left. + \sum_j \sum_k \left\| \frac{\partial \eta_{ijk}(\psi)}{\partial \beta} \right\|^2 + \sum_j \sum_k \left\| \frac{\partial \eta_{ijk}(\psi)}{\partial \theta} \right\|^2 + \sum_j \sum_k \left\| \frac{\partial \eta_{ijk}(\psi)}{\partial \phi} \right\|^2 \right) \\
\leq \; & \sum_j E \left\| W_i \right\| \left\| x_{ij} \right\|^2 \left\| \int \sup_{\Psi} g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du \right\|^2 \\
& + \sum_j E \left\| W_i \right\| \left\| \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) f_b^{(1)}(u;\theta) du \right\|^2 \\
& + 3 \sum_j \sum_k E \left\| W_i \right\| \left\| x_{ij} \right\|^2 \left\| \int \sup_{\Psi} g^{(1)}(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta) du \right\|^2 \\
& + 3 \sum_j \sum_k E \left\| W_i \right\| \left\| x_{ik} \right\|^2 \left\| \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) g^{(1)}(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta) du \right\|^2 \\
& + 3 n_i \sup_{\Sigma} \phi^2 E \left\| W_i \right\| \left\| x_{ij} \right\|^2 \left\| \int \sup_{\Psi} \nu^{(1)}(g(x_{ij}'\beta + z_{ij}'u)) g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du \right\|^2 \\
& + 2 \sum_j \sum_k E \left\| W_i \right\| \left\| \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b^{(1)}(u;\theta) du \right\|^2 \\
& + 2 n_i \sup_{\Sigma} \phi^2 E \left\| W_i \right\| \left\| \int \sup_{\Psi} \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b^{(1)}(u;\theta) du \right\|^2 \\
& + n_i E \left\| W_i \right\| \left\| \int \sup_{\Psi} \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b(u;\theta) du \right\|^2 \\
< \; & \infty,
\end{aligned}
$$

$$
\begin{aligned}
& E\left(\|W_i\| \sup_{\Psi}\left\|\frac{\partial \mathrm{vec}(\partial \rho_i'(\psi)/\partial\psi)}{\partial \psi'}\right\|^2\right) \\
\leq\ & E\|W_i\| \sup_{\Psi}\left(\left\|\frac{\partial^2 \mu_{ij}(\psi)}{\partial\beta\partial\beta'}\right\|^2 + \left\|\frac{\partial^2 \mu_{ij}(\psi)}{\partial\theta\partial\theta'}\right\|^2 + 2\left\|\frac{\partial^2 \mu_{ij}(\psi)}{\partial\beta\partial\theta'}\right\|^2\right) \\
& + E\|W_i\| \sup_{\Psi}\left(\left\|\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\beta\partial\beta'}\right\|^2 + \left\|\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\theta\partial\theta'}\right\|^2 + 2\left\|\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\beta\partial\theta'}\right\|^2\right) \\
\leq\ & \sum_j E\|W_i\| \left\| x_{ij}x_{ij}' \int \sup_{\Psi} g^{(2)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du\right\|^2 \\
& + \sum_j E\|W_i\| \left\| \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) f_b^{(2)}(u;\theta) du\right\|^2 \\
& + 2\sum_j E\|W_i\| \left\| x_{ij} \int \sup_{\Psi} g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b^{(1)}(u;\theta) du\right\|^2 \\
& + 5\sum_j \sum_k E\|W_i\| \left\| x_{ij}x_{ij}' \int \sup_{\Psi} g^{(2)}(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta) du\right\|^2 \\
& + 20\sum_j \sum_k E\|W_i\| \left\| x_{ij}'x_{ik}' \int \sup_{\Psi} g^{(1)}(x_{ij}'\beta + z_{ij}'u) g^{(1)}(x_{ik}'\beta + z_{ik}'u) du\right\|^2 \\
& + 5\sum_j \sum_k E\|W_i\| \left\| x_{ik}x_{ik}' \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) g^{(2)}(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta) du\right\|^2 \\
& + 5n_i \sup_{\Sigma} \phi^2 E\|W_i\| \left\| x_{ij}x_{ij}' \int \sup_{\Psi} \nu^{(2)}(g(x_{ij}'\beta + z_{ij}'u)) g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du\right\|^2 \\
& + 5n_i \sup_{\Sigma} \phi^2 E\|W_i\| \left\| x_{ij}x_{ij}' \int \sup_{\Psi} \nu^{(1)}(g(x_{ij}'\beta + z_{ij}'u)) g^{(2)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du\right\|^2 \\
& + 2\sum_j \sum_k E\|W_i\| \left\| \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b^{(2)}(u;\theta) du\right\|^2 \\
& + 2n_i \sup_{\Sigma} \phi^2 E\|W_i\| \left\| \int \sup_{\Psi} \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b^{(2)}(u;\theta) du\right\|^2 \\
& + 6\sum_j \sum_k E\|W_i\| \left\| x_{ij}' \int \sup_{\Psi} g^{(1)}(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b^{(1)}(u;\theta) du\right\|^2 \\
& + 6\sum_j \sum_k E\|W_i\| \left\| x_{ik}' \int \sup_{\Psi} g(x_{ij}'\beta + z_{ij}'u) g^{(1)}(x_{ik}'\beta + z_{ik}'u) f_b^{(1)}(u;\theta) du\right\|^2 \\
& + 6n_i \sup_{\Sigma} \phi^2 E\|W_i\| \left\| \int \sup_{\Psi} \nu^{(1)}(g(x_{ij}'\beta + z_{ij}'u)) g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b^{(1)}(u;\theta) du\right\|^2 \\
<\ & \infty.
\end{aligned}
$$

Therefore,

$$\begin{aligned}
& E \sup_{\Gamma} \left\| (\rho_i'(\psi) W_i \otimes I) \frac{\partial \text{vec}(\partial \rho_i'(\psi)/\partial \psi)}{\partial \psi'} \right\| \\
\leq \; & \sqrt{2N(p+r+1)} E \left\| W_i \right\| \sup_{\Gamma} \left\| \rho_i(\psi) \right\| \left\| \frac{\partial \text{vec}(\partial \rho_i'(\psi)/\partial \psi)}{\partial \psi'} \right\| \\
\leq \; & \sqrt{2N(p+r+1)} \left(E \left\| W_i \right\| \sup_{\Gamma} \left\| \rho_i(\psi) \right\|^2 \right)^{1/2} \left(E \left\| W_i \right\| \sup_{\Gamma} \left\| \frac{\partial \text{vec}(\partial \rho_i'(\psi)/\partial \psi)}{\partial \psi'} \right\|^2 \right)^{1/2} \\
< \; & \infty.
\end{aligned}$$

By the ULLN and Lemma 4 of Amemiya (1973), we have

$$\frac{1}{2N} \frac{\partial^2 Q_N(\psi)}{\partial \psi \partial \psi'} \xrightarrow{a.s.} E \left[\frac{\partial \rho_i'(\psi)}{\partial \psi} W_i \frac{\partial \rho_i(\psi)}{\partial \psi'} + (\rho_i'(\psi) W_i \otimes I) \frac{\partial \text{vec}(\partial \rho_i'(\psi)/\partial \psi)}{\partial \psi'} \right] = B \tag{A.5}$$

where the second equality holds because

$$E \left[(\rho_i'(\psi_0) W_i \otimes I) \frac{\partial \text{vec}(\partial \rho_i'(\psi_0)/\partial \psi)}{\partial \psi'} \right] = 0.$$

The result then follows from (A.3) - (A.5), assumption 3.2.6 and Slutsky's theorem.

A.6 Proof of Theorem 3.2.4.1

First, the conditional expectation satisfies

$$
\begin{aligned}
& E\left(\sup_{\Gamma}\|\rho_{i,1}(\psi)\|\,|Y_i,X_i,Z_i\right) \\
\leq\ & \sum_j |y_{ij}| + \sum_{j\leq k}|y_{ij}y_{ik}| + \frac{1}{S}\sum_j\sum_{s=1}^{S} E\left(\frac{\sup_\Gamma\left|g(x_{ij}'\beta+z_{ij}'u_{is})\right| f_b(u_{is};\theta)}{h(u_{is})}\middle| X_i,Z_i\right) \\
& + \frac{1}{S}\sum_{j\leq k}\sum_{s=1}^{S} E\left(\frac{\sup_\Gamma\left|g(x_{ij}'\beta+z_{ij}'u_{is})g(x_{ik}'\beta+z_{ik}'u_{is})\right| f_b(u_{is};\theta)}{h(u_{is})}\middle| X_i,Z_i\right) \\
& + \frac{\sup_\Sigma\phi}{S}\sum_j\sum_{s=1}^{S} E\left(\frac{\sup_\Gamma\left|\nu(g(x_{ij}'\beta+z_{ij}'u_{is}))\right| f_b(u_{is};\theta)}{h(u_{is})}\middle| X_i,Z_i\right) \\
\leq\ & \sum_j |y_{ij}| + \sum_{j\leq k}|y_{ij}y_{ik}| + \sum_j\left(\int\sup_\Gamma\left|g(x_{ij}'\beta+z_{ij}'u)\right| f_b(u;\theta)du\right) \\
& + \sum_{j\leq k}\left(\int\sup_\Gamma\left|g(x_{ij}'\beta+z_{ij}'u)g(x_{ik}'\beta+z_{ik}'u)\right| f_b(u;\theta)du\right) \\
& + \sup_\Sigma\phi\sum_j\left(\int\sup_\Gamma\left|\nu(g(x_{ij}'\beta+z_{ij}'u))\right| f_b(u;\theta)du\right).
\end{aligned}
$$

Similarly, the above upper bound applies to $E\left(\sup_\Gamma\|\rho_{i,2}(\psi)\|\,|Y_i,X_i,Z_i\right)$ as well. Further, since $\rho_{i,1}$ and $\rho_{i,2}$ are conditionally independent given (Y_i,X_i,Z_i), we have

$$
\begin{aligned}
& E\left(\sup_\Gamma|\rho_{i,1}(\psi)W_i\rho_{i,2}(\psi)|\right) \\
& \leq E\left[\|W_i\|\,E\left(\sup_\Gamma\|\rho_{i,1}(\psi)\|\,|Y_i,X_i,Z_i\right)E\left(\sup_\Gamma\|\rho_{i,2}(\psi)\|\,|Y_i,X_i,Z_i\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & E\left\|W_i\right\| \left(\sum_j |y_{ij}| + \sum_{j\leq k} |y_{ij}y_{ik}| + \sum_j \int \sup_{\Gamma} \left| g(x'_{ij}\beta + z'_{ij}u) \right| f_b(u;\theta)du \right. \\
& + \sum_{j\leq k} \int \sup_{\Gamma} \left| g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) \right| f_b(u;\theta)du \\
& \left. + \sup_{\Sigma} \phi \sum_j \int \sup_{\Gamma} \left| \nu(g(x'_{ij}\beta + z'_{ij}u)) \right| f_b(u;\theta)du \right)^2 .
\end{aligned}
$$

Analogous to the proof of Corollary 3.1.1, we have $E\left(\sup_{\Gamma} |\rho_{i,1}(\psi)W_i\rho_{i,2}(\psi)|\right) < \infty$, and therefore by the ULLN,

$$
\frac{1}{N} Q_{N,S}(\psi) \xrightarrow{a.s.} E\rho'_{i,1}(\psi)W_i\rho_{i,2}(\psi)
$$

uniformly in $\psi \in \Gamma$, where

$$
E\rho'_{i,1}(\psi)W_i\rho_{i,2}(\psi) = E[E(\rho'_{i,1}(\psi)|X_i, Z_i)W_i E(\rho'_{i,2}(\psi)|X_i, Z_i)] = Q(\psi).
$$

It has been proved previously that $Q(\psi)$ attains a unique minimum at $\psi_0 \in \Gamma$. Therefore, by Lemma 3 of Amemiya (1973), $\hat{\psi}_{N,S} \xrightarrow{a.s.} \psi_0$, as $N \xrightarrow{a.s.} \infty$.

A.7 Proof of Theorem 3.2.4.2

For sufficiently large N we have

$$
\frac{\partial Q_{N,S}(\psi_0)}{\partial \psi} + \frac{\partial^2 Q_{N,S}(\tilde{\psi}_{N,S})}{\partial \psi \partial \psi'} (\hat{\psi}_{N,S} - \psi_0) = 0, \tag{A.6}
$$

where $\left\| \tilde{\psi}_{N,S} - \psi_0 \right\| \leq \left\| \hat{\psi}_{N,S} - \psi_0 \right\|$ and the first derivative

$$
\frac{\partial Q_{N,S}(\psi)}{\partial \psi} = \sum_{i=1}^{N} \left(\frac{\partial \rho'_{i,1}(\psi)}{\partial \psi} W_i \rho_{i,2}(\psi) + \frac{\partial \rho'_{i,2}(\psi)}{\partial \psi} W_i \rho_{i,1}(\psi) \right)
$$

is a summation are i.i.d. terms with mean zero and common covariance matrix

$$\begin{aligned}
4C_S &= E\left[\frac{\partial\rho'_{i,1}(\psi_0)}{\partial\psi}W_i\rho_{i,2}(\psi_0)\rho'_{i,2}(\psi_0)W_i\frac{\partial\rho_{i,1}(\psi_0)}{\partial\psi'}\right] \\
&\quad +E\left[\frac{\partial\rho'_{i,1}(\psi_0)}{\partial\psi}W_i\rho_{i,2}(\psi_0)\rho'_{i,1}(\psi_0)W_i\frac{\partial\rho_{i,2}(\psi_0)}{\partial\psi'}\right] \\
&\quad +E\left[\frac{\partial\rho'_{i,2}(\psi_0)}{\partial\psi_0}W_i\rho_{i,1}(\psi_0)\rho'_{i,2}(\psi_0)W_i\frac{\partial\rho_{i,1}(\psi_0)}{\partial\psi'}\right] \\
&\quad +E\left[\frac{\partial\rho'_{i,2}(\psi_0)}{\partial\psi}W_i\rho_{i,1}(\psi)\rho'_{i,1}(\psi_0)W_i\frac{\partial\rho_{i,2}(\psi_0)}{\partial\psi'}\right].
\end{aligned}$$

Hence by the central limit theorem we have

$$\frac{1}{\sqrt{N}}\frac{\partial Q_{N,S}(\psi)}{\partial\psi} \xrightarrow{a.s.} N(0,4C_S). \tag{A.7}$$

Next, the second derivative is given by

$$\begin{aligned}
\frac{\partial^2 Q_{N,S}(\psi)}{\partial\psi\partial\psi'} &= \sum_{i=1}^{N}\left[\frac{\partial\rho'_{i,1}(\psi)}{\partial\psi}W_i\frac{\partial\rho_{i,2}(\psi)}{\partial\psi'}+(\rho'_{i,2}(\psi)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,1}(\psi)/\partial\psi)}{\partial\psi'}\right] \\
&\quad +\sum_{i=1}^{N}\left[\frac{\partial\rho'_{i,2}(\psi)}{\partial\psi}W_i\frac{\partial\rho_{i,1}(\psi)}{\partial\psi'}+(\rho'_{i,1}(\psi)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,2}(\psi)/\partial\psi)}{\partial\psi'}\right],
\end{aligned}$$

where I is the $2N(p+r+1)$ dimensional identity matrix. Similar to previous proofs, it can be shown that $\frac{1}{N}\frac{\partial^2 Q_{N,S}(\psi)}{\partial\psi\partial\psi'}$ converges to

$$\begin{aligned}
&E\left[\frac{\partial\rho'_{i,1}(\psi_0)}{\partial\psi}W_i\frac{\partial\rho_{i,2}(\psi_0)}{\partial\psi'}+(\rho'_{i,2}(\psi_0)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,1}(\psi_0)/\partial\psi)}{\partial\psi'}\right] \\
&+E\left[\frac{\partial\rho'_{i,2}(\psi_0)}{\partial\psi}W_i\frac{\partial\rho_{i,1}(\psi_0)}{\partial\psi'}+(\rho'_{i,1}(\psi_0)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,2}(\psi_0)/\partial\psi)}{\partial\psi'}\right],
\end{aligned}$$

uniformly for all $\psi\in\Gamma$. Since

$$E\left[\frac{\partial\rho'_{i,1}(\psi_0)}{\partial\psi}W_i\frac{\partial\rho_{i,2}(\psi_0)}{\partial\psi'}\right]=E\left[\frac{\partial\rho'_{i}(\psi_0)}{\partial\psi}W_i\frac{\partial\rho_{i}(\psi_0)}{\partial\psi'}\right]=B$$

and

$$E\left[(\rho'_{i,1}(\psi_0)W_i \otimes I)\frac{\partial \text{vec}(\partial \rho'_{i,2}(\psi_0)/\partial\psi)}{\partial \psi'}\right] = 0,$$

we have

$$\frac{1}{N}\frac{\partial^2 Q_{N,S}(\psi)}{\partial\psi\partial\psi'} \xrightarrow{a.s.} 2B. \tag{A.8}$$

Finally, the result follows from (A.6)-(A.8) and Slutsky's theorem.

A.8 Derivation of the Working Optimal Weight Matrix

A.8.1 Gaussian Assumption

Assume y_i is from a multivariate normal distribution, and we denote $\sigma_{ijk} = E(y_{ij} - u_{ij})(y_{ik} - u_{ik})$. The third moment of y_i, for all $j,\ k,\ l$ is

$$\begin{aligned}
\text{cov}(y_{ij}, y_{ik}y_{il}) &= E(y_{ij}y_{ik}y_{il}) - E(y_{ij})E(y_{ik}y_{il}) \\
&= E[(y_{ij} - \mu_{ij})(y_{ik}y_{il})] \\
&= E[(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})] + \mu_{il}\sigma_{ijk} + \mu_{ik}\sigma_{ijl} \\
&\quad +\mu_{ik}\mu_{il}E[(y_{ij} - \mu_{ij})] \\
&= \mu_{il}\sigma_{ijk} + \mu_{ik}\sigma_{ijl},
\end{aligned}$$

since $E[(y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})(y_{il} - \mu_{il})] = 0$ under normality assumption.

The fourth moment of y_i, for all $j,\ \ k,\ \ l,\ \ t$ is,

$$
\begin{aligned}
& \operatorname{cov}(y_{ij}y_{ik}, y_{il}y_{it}) \\
= \ & E(y_{ij}y_{ik}y_{il}y_{it}) - E(y_{ij}y_{ik})E(y_{il}y_{it}) \\
= \ & E[(y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{il}-\mu_{il})(y_{it}-\mu_{it})] \\
& +\mu_{ij}E[(y_{ik}-\mu_{ik})(y_{il}-\mu_{il})(y_{it}-\mu_{it})] + \mu_{ik}E[(y_{ij}-\mu_{ij})(y_{il}-\mu_{il})(y_{it}-\mu_{it})] \\
& +\mu_{il}E[(y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{it}-\mu_{it})] + \mu_{it}E[(y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{it}-\mu_{it})] \\
& +\mu_{ij}\mu_{ik}E[(y_{il}-\mu_{il})(y_{it}-\mu_{it})] + \mu_{ij}\mu_{il}E[(y_{ik}-\mu_{ik})(y_{it}-\mu_{it})] \\
& +\mu_{ij}\mu_{it}E[(y_{il}-\mu_{il})(y_{ik}-\mu_{ik})] + \mu_{ik}\mu_{il}E[(y_{ij}-\mu_{ij})(y_{it}-\mu_{it})] \\
& +\mu_{ik}\mu_{it}E[(y_{ij}-\mu_{ij})(y_{il}-\mu_{il})] + \mu_{il}\mu_{it}E[(y_{ij}-\mu_{ij})(y_{il}-\mu_{il})] \\
& +\mu_{ij}\mu_{ik}\mu_{il}\mu_{it} \\
& -(E[(y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})] + \mu_{ij}\mu_{ik})(E[(y_{il}-\mu_{il})(y_{it}-\mu_{it})] + \mu_{il}\mu_{it}) \\
= \ & \sigma_{ijl}\sigma_{ikt} + \sigma_{ijt}\sigma_{ikl} + \mu_{ik}\mu_{il}\sigma_{ijt} + \mu_{ij}\mu_{il}\sigma_{ikt} + \mu_{ik}\mu_{it}\sigma_{ijl} + \mu_{ij}\mu_{it}\sigma_{ikl}.
\end{aligned}
$$

since $E[(y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{il}-\mu_{il})(y_{it}-\mu_{it})] = \sigma_{ijk}\sigma_{ilt} + \sigma_{ijl}\sigma_{ikt} + \sigma_{ijt}\sigma_{ikl}$ under normality assumption.

A.8.2 Independence Assumption

Assume independence among the elements of y_i. The third moment is

$$\operatorname{cov}(y_{ij}, y_{ik}y_{il}) \ = \ E(y_{ij}y_{ik}y_{il}) - E(y_{ij})E(y_{ik}y_{il}).$$

(i). If $j = k = l$, we have

$$\begin{aligned}\text{cov}(y_{ij}, y_{ik}y_{il}) &= E(y_{ij}^3) - E(y_{ij})E(y_{ij}^2) \\ &= E[(y_{ij} - \mu_{ij})^3] + 3\mu_{ij}\sigma_{ijj} - 2\mu_{ij}^3 - \mu_{ij}\sigma_{ijj} \\ &= E[(y_{ij} - \mu_{ij})^3] + 2\mu_{ij}\sigma_{ijj} - 2\mu_{ij}^3.\end{aligned}$$

(ii). If $j = l \neq k$, since $E(y_{ij}y_{ik}y_{ij}) = E(y_{ij}^2)E(y_{ik})$ and $E(y_{ij}y_{ik}) = E(y_{ij})E(y_{ik})$ under independence assumption. Then it follows that

$$\begin{aligned}\text{cov}(y_{ij}, y_{ik}y_{il}) &= E(y_{ij}^2)E(y_{ik}) - (E(y_{ij}))^2E(y_{ik}) \\ &= \mu_{ij}\sigma_{ijj}.\end{aligned}$$

(iii). If $j = k \neq l$, similar to above, we have

$$\begin{aligned}\text{cov}(y_{ij}, y_{ik}y_{il}) &= E(y_{ij}^2)E(y_{il}) - (E(y_{ij}))^2E(y_{il}) \\ &= \mu_{il}\sigma_{ijj}.\end{aligned}$$

(iv). If $j \neq k \neq l$, $E(y_{ij}y_{ik}y_{il}) = E(y_{ij})E(y_{ik})E(y_{il})$ and $E(y_{ij})E(y_{ik}y_{il}) = E(y_{ij})E(y_{ik})E(y_{il})$ under independence assumption. We have obviously

$$\text{cov}(y_{ij}, y_{ik}y_{il}) = 0.$$

The fourth moment is

$$\text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) = E(y_{ij}y_{ik}y_{il}y_{it}) - E(y_{ij}y_{ik})E(y_{il}y_{it}).$$

(i). If $j = k = l = t$, we have

$$\begin{aligned} \text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) &= E(y_{ij}^4) - [E(y_{ij}^2)]^2 \\ &= E(y_{ij}^4) - \mu_{ij}^2 - \sigma_{ijj}. \end{aligned}$$

(ii). if $j = k = l \neq t$, $E(y_{ij}y_{ij}y_{ij}y_{it}) = E(y_{ij}^3)E(y_{it})$ and $E(y_{ij}y_{ij})E(y_{ij}y_{it}) = E(y_{ij}^2)E(y_{ij})E(y_{it})$ under independence assumption. Then it follows that

$$\begin{aligned} \text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) &= E(y_{ij}^3)\mu_{it} - E(y_{ij}^2)\mu_{ij}\mu_{it} \\ &= E[(y_{ij} - \mu_{ij})^3]\mu_{it} + \mu_{ij}^3\mu_{it} + 3\mu_{ij}\mu_{it}\sigma_{ijj} - (\mu_{ij}\mu_{it}\sigma_{ijj} + \mu_{ij}^3\mu_{it}) \\ &= E[(y_{ij} - u_{ij})^3]\mu_{it} + 2\mu_{ij}\mu_{it}\sigma_{ijj}. \end{aligned}$$

(iii). If $j = k = l \neq t$, similar to above, we have

$$\text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) = E[(y_{ij} - u_{ij})^3]\mu_{il} + 2\mu_{ij}\mu_{il}\sigma_{ijj}.$$

(iv). If $j \neq k \neq l \neq t$, under independence assumption, we have obviously,

$$\text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) = 0.$$

(v). If $(j = k) \neq (l = t)$, $E(y_{ij}y_{ij}y_{il}y_{il}) = E(y_{ij}^2)E(y_{il})^2$ under independence assumption. Then it follows that

$$\text{cov}(y_{ij}y_{ik}, y_{il}y_{it}) = E(y_{ij}^2)E(y_{il})^2 - E(y_{ij}^2)E(y_{il}^2) = 0.$$

A.9 Proof of Theorem 5.2.4.1

By assumption 5.2.1 and the Dominated Convergence Theorem (DCT), we have the first-order Taylor expansion about γ_0.

$$Q_N(\psi) = \sum_{i=1}^{N} \rho_i'(\psi) A_i \rho_i(\psi) + 2 \sum_{i=1}^{N} \rho_i'(\psi, \tilde{\gamma}) A_i \frac{\partial \rho_i(\psi, \tilde{\gamma})}{\partial \gamma'} (\hat{\gamma}_N - \gamma_0), \qquad \text{(A.9)}$$

where $\|\tilde{\gamma} - \gamma_0\| \leq \|\hat{\gamma}_N - \gamma_0\|$. Further, for any $1 \leq i \leq N$, by assumptions 5.2.1-5.2.3 and Cauchy-Schwartz inequality, we have

$$\begin{aligned}
& \|\rho_i(\psi)\|^2 \\
\leq\ & 2 \sum_j y_{ij}^2 + 2 \sum_{j \leq k} y_{ij}^2 y_{ik}^2 + 2 \sum_{j \leq k} \|y_{ij} W_{ik}\|^2 \\
& + 2 \sum_j \left(\int g \left[(m(V_{ij}; \gamma) + u)' \beta_x + Z_{ij}' \beta_z + B_{ij}' t \right] f_b(t; \theta) f_U(u; \alpha) dt du \right)^2 \\
& + 4 \sum_{j \leq k} \left(\int g \left[(m(V_{ij}; \gamma) + u)' \beta_x + Z_{ij}' \beta_z + B_{ij}' t \right] \right. \\
& \left. g \left[(m(V_{ik}; \gamma) + u)' \beta_x + Z_{ik}' \beta_z + B_{ik}' t \right] f_b(t; \theta) f_U(u; \alpha) dt du \right)^2 \\
& + 4\phi^2 \sum_j \left(\int \nu \{ g \left[(m(V_{ij}; \gamma) + u)' \beta_x + Z_{ij}' \beta_z + B_{ij}' t \right] \} f_b(t; \theta) f_U(u; \alpha) dt du \right)^2 \\
& + 2 \sum_{j \leq k} \left(\int \|m(V_{ik}; \gamma) + u\| \, g \left[(m(V_{ij}; \gamma) + u)' \beta_x + Z_{ij}' \beta_z + B_{ij}' t \right] f_b(t; \theta) f_U(u; \alpha) dt du \right)^2
\end{aligned}$$

and therefore

$$\begin{aligned}
E\sup_{\Omega_\psi}|\rho_i'(\psi)A_i\rho_i(\psi)| &\leq E\left\|A_i\right\|\sup_{\Omega_\psi}\left\|\rho_i(\psi)\right\|^2 \\
&\leq 2n_iE\left\|A_i\right\|y_{ij}^2+n_i(n_i+1)\left(E\left\|A_i\right\|y_{ij}^2y_{ik}^2+E\left\|A_i\right\|\left\|y_{ij}W_{ik}\right\|^2\right) \\
&\quad +2n_iE\left\|A_i\right\|\left(\int G(V_i,Z_i,t,u)dtdu\right) \\
&\quad +2n_i\left(n_i+1+2\sup_{\Omega_\phi}\phi^2\right)E\left\|A_i\right\|\left(\int G(V_i,Z_i,t,u)dtdu\right)^2 \\
&\quad +n_i(n_i+1)E\left\|A_i\right\|\left(\int G(V_i,Z_i,t,u)\left\|m(V;\gamma_0)+u\right\|dtdu\right)^2 \\
&< \infty.
\end{aligned}$$

Hence by the uniform law of large numbers (ULLN),

$$\sup_{\psi\in\Omega_\psi}\left|\frac{1}{N}\sum_{i=1}^N\rho_i'(\psi)A_i\rho_i(\psi)-Q(\psi)\right|\xrightarrow{a.s.}0, \tag{A.10}$$

where $Q(\psi)=E[\rho_i'(\psi)W_i\rho_i(\psi)]$. Similarly, by assumption 5.2.1-5.2.3 and 5.2.5 we can show

$$\left(E\sup_{\Omega_\psi,\Omega_\gamma}\left\|\rho_i'(\psi,\gamma)A_i\frac{\partial\rho_i(\psi,\gamma)}{\partial\gamma'}\right\|\right)^2\leq E\left\|A_i\right\|\sup_{\Omega_\psi,\Omega_\gamma}\left\|\rho_i'(\psi,\gamma)\right\|^2E\left\|A_i\right\|\left\|\frac{\partial\rho_i(\psi,\gamma)}{\partial\gamma'}\right\|^2<\infty,$$

then again by the ULLN,

$$\sup_{\Omega_\psi,\Omega_\gamma}\left\|\frac{1}{N}\sum_{i=1}^N\rho_i'(\psi,\gamma)A_i\frac{\partial\rho_i(\psi,\gamma)}{\partial\gamma'}\right\|=O(1)\ \ (a.s.)$$

Therefore,

$$\begin{aligned}
&\sup_{\Omega_\psi}\left\|\frac{1}{N}\sum_{i=1}^N\rho_i'(\psi,\tilde{\gamma})A_i\frac{\partial\rho_i(\psi,\tilde{\gamma})}{\partial\gamma'}(\hat{\gamma}_N-\gamma_0)\right\| \\
&\leq\sup_{\Omega_\psi,\Omega_\gamma}\left\|\frac{1}{N}\sum_{i=1}^N\rho_i'(\psi,\gamma)A_i\frac{\partial\rho_i(\psi,\gamma)}{\partial\gamma'}\right\|\left\|\hat{\gamma}_N-\gamma_0\right\|\xrightarrow{a.s.}0.
\end{aligned} \tag{A.11}$$

It follows (A.9) - (A.11) that

$$\sup_{\Omega_\gamma} \left| \frac{1}{N} Q_N(\psi) - Q(\psi) \right| \xrightarrow{a.s.} 0. \tag{A.12}$$

Furthermore, $Q(\psi) = Q(\psi_0) + E[\rho_i(\psi) - \rho_i(\psi_0))' A_i (\rho_i(\psi) - \rho_i(\psi_0)]$, then by assumption 5.2.4, $Q(\psi) \geq Q(\psi_0)$ and the equality holds if and only if $\psi = \psi_0$. Thus, all conditions of Amemiya (1973) Lemma 3 are satisfied and therefore $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$, as $N \to \infty$.

A.10 Proof of Theorem 5.2.4.2

By assumption 5.2.5 and the DCT, the first derivative $\partial Q_N(\psi)/\partial\psi$ exists and has the first-order Taylor expansion in the open neighborhood $\Omega_{\psi 0} \in \Omega_\psi$ of ψ_0. Since $\hat{\psi}_N \xrightarrow{a.s.} \psi_0$, for sufficiently large N we have

$$\frac{\partial Q_N(\hat{\psi}_N)}{\partial \psi} = \frac{\partial Q_N(\psi_0)}{\partial \psi} + \frac{\partial^2 Q_N(\tilde{\psi}_N)}{\partial \psi \partial \psi'} (\hat{\psi}_N - \psi_0) = 0, \tag{A.13}$$

where $\left\| \tilde{\psi}_N - \psi_0 \right\| \leq \left\| \hat{\psi}_N - \psi_0 \right\|$. The first and second derivative of $Q_N(\psi)$ in (A.13) are given in (5.17) and (5.18).

Analogous to the proof of Theorem 5.2.4.1, by assumption 5.2.1 - 5.2.5 and Cauchy-Schwartz inequality, we can verify that

$$E \sup_{\Omega_\psi} \left\| \frac{\partial \rho_i'(\psi)}{\partial \psi} A_i \frac{\partial \rho_i(\psi)}{\partial \psi'} \right\| \leq E \left\| A_i \right\| \sup_{\Omega_\psi} \left\| \frac{\partial \rho_i'(\psi)}{\partial \psi} \right\|^2 < \infty$$

and

$$E \sup_{\Omega_\psi} \left\| (\rho_i'(\psi) A_i \otimes I) \frac{\partial \text{vec}(\partial \rho_i'(\psi)/\partial \psi)}{\partial \psi'} \right\| < \infty.$$

Therefore by the ULLN and Lemma 4 of Amemiya (1973), we have

$$\frac{1}{2N}\frac{\partial^2 Q_N(\tilde{\psi})}{\partial\psi\partial\psi'} \xrightarrow{a.s.} E\left[\frac{\partial\rho_i'(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_i(\psi_0)}{\partial\psi'} + (\rho_i'(\psi_0)A_i\otimes I)\frac{\partial \text{vec}(\partial\rho_i'(\psi_0)/\partial\psi)}{\partial\psi'}\right] = D_\psi \tag{A.14}$$

where D_ψ is given in (5.11) and the second equality holds because

$$E\left[(\rho_i'(\psi_0)A_i\otimes I)\frac{\partial \text{vec}(\partial\rho_i'(\psi_0)/\partial\psi)}{\partial\psi'}\right] = 0.$$

Then by assumption 5.2.6 and (A.14), we rearrange (A.13) as

$$\sqrt{N}(\hat{\psi}_N - \psi_0) = (2D_\psi)^{-1}\left(-\frac{1}{\sqrt{N}}\frac{\partial Q_N(\psi_0)}{\partial\psi}\right) \tag{A.15}$$

For by assumption 5.2.5 and DCT, we have the first-order Taylor expansion of $\frac{\partial Q_N(\psi_0)}{\partial\psi}$ about γ_0:

$$\frac{\partial Q_N(\psi_0)}{\partial\psi} = 2\sum_{i=1}^N \frac{\partial\rho_i'(\psi_0)}{\partial\psi}A_i\rho_i(\psi_0) + \frac{\partial^2\tilde{Q}_N(\psi_0)}{\partial\psi\partial\gamma'}(\hat{\gamma}-\gamma_0), \tag{A.16}$$

where $\|\tilde{\gamma}-\gamma_0\| \le \|\hat{\gamma}-\gamma_0\|$ and

$$\frac{\partial^2\tilde{Q}_N(\psi_0)}{\partial\psi\partial\gamma'} = 2\sum_{i=1}^N\left[\frac{\partial\rho_i'(\psi_0,\tilde{\gamma})}{\partial\psi}A_i\frac{\partial\rho_i'(\psi_0,\tilde{\gamma})}{\partial\gamma'} + (\rho_i'(\psi_0,\tilde{\gamma})A_i\otimes I)\frac{\partial\text{vec}(\partial\hat{\rho}_i'(\psi_0,\tilde{\gamma})/\partial\psi)}{\partial\gamma'}\right].$$

Similarly to the derivation of (A.14), we can show

$$\frac{1}{2N}\frac{\partial^2\tilde{Q}_N(\psi_0)}{\partial\psi\partial\gamma'} \xrightarrow{a.s.} E\left[\frac{\partial\rho_i'(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_i(\psi_0)}{\partial\gamma'}\right] = D_{\psi\gamma}. \tag{A.17}$$

Then by (A.15)-(A.17), we have

$$\sqrt{N}(\hat{\psi}_N-\psi_0) = D_\psi^{-1}\left(-N^{-1/2}\sum_{i=1}^N\frac{\partial\rho_i'(\psi_0)}{\partial\psi}A_i\rho_i(\psi_0)\right) + D_\psi^{-1}D_{\psi\gamma}\sqrt{N}(\hat{\gamma}-\gamma_0) \tag{A.18}$$

Therefore, if $D_{\psi\gamma} = 0$ we can ignore the effect of $\hat{\gamma}$ and simply treated it as a known constant. If $D_{\psi\gamma} \neq 0$, we need to make some adjustments to the asymptotic variance of $\sqrt{N}(\hat{\psi}_N - \psi_0)$. Since $\hat{\gamma}_N - \gamma_0 = O_p(N^{-1/2})$, we have the first-order Taylor expansion in the open neighborhood $\Omega_{\gamma 0} \in \Omega_\gamma$ of γ_0

$$\frac{\partial \Psi(\hat{\gamma})}{\partial \gamma} = \frac{\partial \Psi(\gamma_0)}{\partial \gamma} + \frac{\partial^2 \Psi(\tilde{\gamma})}{\partial \gamma \partial \gamma'}(\hat{\gamma}_N - \gamma_0) = 0, \tag{A.19}$$

By assumption 5.2.6, we can have the following representation of $\hat{\gamma}_N$

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \;=\; D_\gamma^{-1}\left(-N^{1/2}\sum_{i=1}^{N}\frac{\partial r_i'(\gamma)}{\partial \gamma} r_i(\gamma)\right) = N^{-1/2}\sum_{i=1}^{N}\left(-D_\gamma^{-1}\frac{\partial r_i'(\gamma)}{\partial \gamma} r_i(\gamma)\right), \tag{A.20}$$

where D_γ is given in (5.12). Then plug it back into (A.18), we have

$$\sqrt{N}(\hat{\psi}_N - \psi_0) = -D_\psi^{-1} N^{-1/2}\sum_{i=1}^{N}\left(\frac{\partial \rho_i'(\psi_0)}{\partial \psi} A_i \rho_i(\psi_0) + D_{\psi\gamma} D_\gamma^{-1}\frac{\partial r_i'(\gamma)}{\partial \gamma} r_i(\gamma)\right) \tag{A.21}$$

Finally, the theorem follows from (A.13) - (A.21), CLT and Slutsky's Theorem.

A.11 Proof of Theorem 5.3.1.1

By assumption 5.2.1 and the DCT, $Q_{N,S}(\psi)$ has the first-order Taylor expansion about γ_0,

$$\begin{aligned} Q_{N,S}(\psi) &= \sum_{i=1}^{N} \rho'_{i,1}(\psi) A_i \rho_{i,2}(\psi) \\ &\quad + \sum_{i=1}^{N} \left[\rho'_{i,1}(\psi,\tilde{\gamma}) A_i \frac{\partial \rho_{i,2}(\psi,\tilde{\gamma})}{\partial \gamma'} + \rho'_{i,2}(\psi,\tilde{\gamma}) A_i \frac{\partial \rho_{i,1}(\psi,\tilde{\gamma})}{\partial \gamma'} \right] (\hat{\gamma}_N - \gamma_0), \end{aligned} \tag{A.22}$$

where $\|\tilde{\gamma}_N - \gamma_0\| \leq \left\| \hat{\psi}_N - \psi_0 \right\|$. Since $\rho_{i,1}$ and $\rho_{i,2}$ are conditionally independent given $(Y_i, W_i, V_i, Z_i, B_i)$, analogous to the proof of Theorem 5.2.4.1, by assumptions 5.2.1-5.2.3 and Cauchy-Schwartz inequality, we have

$$E\left(\sup_{\Gamma} |\rho'_{i,1}(\psi) A_i \rho_{i,2}(\psi)| \right) < \infty.$$

Hence by the ULLN, $\frac{1}{N}\sum_{i=1}^{N} \rho'_{i,1}(\psi) A_i \rho_{i,2}(\psi) \xrightarrow{a.s.} E\rho'_{i,1}(\psi) W_i \rho_{i,2}(\psi)$ uniformly in $\psi \in \Gamma$, where

$$E\rho'_{i,1}(\psi) W_i \rho_{i,2}(\psi) = E[E(\rho'_{i,1}(\psi)|Y_i, W_i, V_i, Z_i, B_i) W_i E(\rho'_{i,2}(\psi)|Y_i, W_i, V_i, Z_i, B_i)] = Q(\psi).$$

Similar to proof of Theorem 5.2.4.1, we can show that

$$\begin{aligned} &\sup_{\Gamma} \left\| \frac{1}{N} \sum_{i=1}^{N} \rho'_{i,2}(\psi,\tilde{\gamma}) A_i \frac{\partial \rho_{i,1}(\psi,\tilde{\gamma})}{\partial \gamma'} (\hat{\gamma}_N - \gamma_0) \right\| \\ &\leq \sup_{(\Gamma,\Upsilon)} \left\| \frac{1}{N} \sum_{i=1}^{N} \rho'_{i,2}(\psi,\tilde{\gamma}) A_i \frac{\partial \rho_{i,1}(\psi,\tilde{\gamma})}{\partial \gamma'} \right\| \|\hat{\gamma}_N - \gamma_0\| \xrightarrow{a.s.} 0. \end{aligned} \tag{A.23}$$

It then follows that

$$\sup_{\Gamma} \left| \frac{1}{N} Q_{N,S}(\psi) - Q(\psi) \right| \xrightarrow{a.s.} 0. \tag{A.24}$$

It has been proved previously that $Q(\psi)$ attains a unique minimum at $\psi_0 \in \Gamma$. Therefore, by Lemma 3 of Amemiya (1973), $\hat{\psi}_{N,S} \xrightarrow{a.s.} \psi_0$, as $N \to \infty$.

A.12 Proof of Theorem 5.3.1.2

For sufficiently large N, by assumption 5.2.5 we have the first-order Taylor expansion of $\partial Q_{N,S}(\psi)/\partial\psi$ about ψ_0:

$$\frac{\partial Q_{N,S}(\psi_0)}{\partial\psi} + \frac{\partial^2 Q_{N,S}(\tilde{\psi})}{\partial\psi\partial\psi'}(\hat{\psi}_{N,S} - \psi_0) = 0, \tag{A.25}$$

where $\left\|\tilde{\psi} - \psi_0\right\| \leq \left\|\hat{\psi}_{N,S} - \psi_0\right\|$ and the first and second derives are given by

$$\frac{\partial Q_{N,S}(\psi)}{\partial\psi} = \sum_{i=1}^{N}\left(\frac{\partial\hat{\rho}'_{i,1}(\psi)}{\partial\psi}A_i\hat{\rho}_{i,2}(\psi) + \frac{\partial\hat{\rho}'_{i,2}(\psi)}{\partial\psi}A_i\hat{\rho}_{i,1}(\psi)\right)$$

and

$$\begin{aligned}\frac{\partial^2 Q_{N,S}(\psi)}{\partial\psi\partial\psi'} &= \sum_{i=1}^{N}\left[\frac{\partial\hat{\rho}'_{i,1}(\psi)}{\partial\psi}A_i\frac{\partial\hat{\rho}_{i,2}(\psi)}{\partial\psi'} + (\hat{\rho}'_{i,2}(\psi)A_i\otimes I)\frac{\partial\mathrm{vec}(\partial\hat{\rho}'_{i,1}(\psi)/\partial\psi)}{\partial\psi'}\right] \\ &\quad + \sum_{i=1}^{N}\left[\frac{\partial\hat{\rho}'_{i,2}(\psi)}{\partial\psi}A_i\frac{\partial\hat{\rho}_{i,1}(\psi)}{\partial\psi'} + (\hat{\rho}'_{i,1}(\psi)A_i\otimes I)\frac{\partial\mathrm{vec}(\partial\hat{\rho}'_{i,2}(\psi)/\partial\psi)}{\partial\psi'}\right].\end{aligned}$$

Similar to the derivation of (A.14), we can show $\frac{1}{N}\frac{\partial^2 Q_{N,S}(\psi)}{\partial\psi\partial\psi'}$ converges to

$$\begin{aligned}&E\left[\frac{\partial\rho'_{i,1}(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_{i,2}(\psi_0)}{\partial\psi'} + (\rho'_{i,2}(\psi_0)A_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,1}(\psi_0)/\partial\psi)}{\partial\psi'}\right] \\ &+E\left[\frac{\partial\rho'_{i,2}(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_{i,1}(\psi_0)}{\partial\psi'} + (\rho'_{i,1}(\psi_0)A_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,2}(\psi_0)/\partial\psi)}{\partial\psi'}\right],\end{aligned}$$

uniformly for all $\psi \in \Gamma$. Since

$$E\left[\frac{\partial\rho'_{i,1}(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_{i,2}(\psi_0)}{\partial\psi'}\right] = E\left[\frac{\partial\rho'_i(\psi_0)}{\partial\psi}A_i\frac{\partial\rho_i(\psi_0)}{\partial\psi'}\right] = D_\psi$$

and

$$E\left[(\rho'_{i,2}(\psi_0)A_i \otimes I)\frac{\partial \text{vec}(\partial \rho'_{i,1}(\psi_0)/\partial\psi)}{\partial \psi'}\right] = 0,$$

we have

$$\frac{1}{N}\frac{\partial^2 Q_{N,S}(\psi)}{\partial\psi\partial\psi'} \xrightarrow{a.s.} 2D. \tag{A.26}$$

Again, $\partial Q_{N,S}(\psi_0)\partial\psi$ has the first-order Taylor expansion about γ_0:

$$\frac{\partial Q_{N,S}(\psi_0)}{\partial\psi} = \sum_{i=1}^{N}\left[\frac{\partial \rho'_{i,1}(\psi_0)}{\partial\psi}A_i\rho'_{i,2}(\psi_0) + \frac{\partial \rho'_{i,2}(\psi_0)}{\partial\psi}A_i\rho'_{i,1}(\psi_0)\right] + \frac{\partial^2 \tilde{Q}_{N,S}(\psi_0)}{\partial\psi\partial\gamma'}(\hat{\gamma}_N - \gamma_0). \tag{A.27}$$

Finally, Analogous to the proof of Theorem 5.2.4.2, the results follows from (A.20), (A.25)-(A.27), CLT and Slutsky's Theorem.

Bibliography

[1] Abarin, T. (2008). Second-order least squares estimation in regression models with application to measurement error problems. *PhD Dissertation*, University of Manitoba.

[2] Abarin, T., Li, H., Wang, L. and Briollais, L. (2010). Estimation in semi-parametric linear mixed effects models with measurement error on covariates and response. *Working Paper*, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada

[3] Abarin, T. and Wang, L. (2006). Comparison of GMM with second-order least squares estimator in nonlinear models. *Far East Journal of Theoretical Statistics*, 20, 179-196.

[4] Abarin, T. and Wang, L. (2010). Instrumental variable approach to covariate measurement error in generalized linear models. *Annals of the Institute of Statistical Mathematics*, DOI: 10.1007/s10463-010-0319-0.

[5] Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th print-

ing, New York: Dover.

[6] Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41, 997-1016.

[7] Altonji, J. G. and Segal, L.M. (1996). Small sample bias in GMM estimation of covariance structures. *Journal of Business and Economic Statistics*, 14, 353–366.

[8] Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics,* 61, 962-73.

[9] Bartlett, J.W., de Stavola, B.L. and Frost, C. (2009). Linear mixed models for replication data to efficiently allow for covariate measurement error. *Statistics in Medicine*, 28, 3158-3178.

[10] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistician Association*, 88, 9-25.

[11] Buonaccorsi, J.P., Demidenko, E. and Tosteson, T. (2000). Estimation in longitudinal random effects models with measurement error. *Statistica Sinica*, 10, 885-836.

[12] Buonaccorsia, J.P. and Lin, C. (2002). Berkson measurement error in designed repeated measures studies with random coefficients. *Journal of Statistical Planning and Inference*, 104, 53-72.

[13] Buzas, J.S. and Stefanski, L.A. (1996). Instrumental variable estimation in generalized measurement error models. *Journal of American Statistical Association*, 91, 999-1006.

[14] Carpenter, J.R., Kenward, M.G., Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, 169, 571-584.

[15] Carroll, R.J. and Stefanski, L.A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13, 1265-1282.

[16] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, Chapman & Hall, London.

[17] Christensen, R., Pearson, L.M. and Johnson, W. (1992). Case deletion diagnostics for mixed models. *Technometrics*, 34, 38-45.

[18] Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multi-phase sampling (with discussion). *Journal of the Royal Statistical Society, Series B*, 71-87.

[19] Davidian, M. and Gallant, A.R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80, 475-488.

[20] Dawber, T.R. (1980). *The Framingham Study. The Epidemiology of Atherosclerotic Disease*, Cambridge, MA: Harvard University Press.

[21] Dawber, T.R., Moore, F.E., Jr. and Mann, G.V. (1957). Coronary Heart Disease in the Framingham Study. *American Journal of Public Health*, 47,4-24.

[22] Demidenko, E. (2004). *Mixed Models: Theory and Applications*, New York: Wiley.

[23] Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-93.

[24] Diggle, P.J., Liang, K-J. and Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Oxford University Press Inc, New York.

[25] Durbin, J. and Koopman, S.J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84, 669-684.

[26] Field, C. A. and Genton, M. G. (2006). The multivariate g-and-h distribution. *Technometrics*, 48, 104-111.

[27] Fuller, W. (1987). *Measurement Error Models,* New York: John Wiley & Sons.

[28] Fitzmaurice, G.M., Davidian, M., Molenberghs, G. and Verbeke, G. (2008). *Longitudinal Data Analysis*, Boca Raton, Florida: Chapman & Hall/CRC.

[29] Fitzmaurice, G.M., Laird, N.M. and Zahner, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-108.

[30] Fitzmaurice, G.M., Molenberghs, G., and Lipsitz, S.R. (1995). Regression models for longitudinal binary responses with informative dropouts. *Journal of Royal Statistical Society, Series B*, 57, 691-704.

[31] Fuller, W. (1987). *Measurement Error Models,* New York: John Wiley & Sons.

[32] Gill, P.S. (2000). A robust mixed linear model analysis for longitudinal data. *Statistics in Medcine*, 19, 975-987.

[33] Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78, 45-51.

[34] Goldstein, H. (2003). *Multilevel Statistical Models,* 3rd edition, London: Edward Arnold

[35] Goldstein, H., and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 159, 505-513.

[36] Gouriéroux, C. and Monfort, A. (1997). *Simulation-Based Econometric Methods.* Oxford University Press.

[37] Greene, W. H.(2008). *Econometric analysis*, Prentice Hall, New Jersey.

[38] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

[39] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.

[40] Heagerty, P.J. and Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15, 1-26.

[41] Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, 40, 13-29.

[42] Horton, N. J., and Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55, 244-254.

[43] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

[44] Huber, P.J. (2004). *Robust Statistics*, New York: Wiley.

[45] Jacqmin-Gadda, H., Sibillot S., Proust, C., Molina J. and Thiebaut, R. (2006). Robustness of the linear mixed model to missspecified error distribution. *Computational Statistics & Data Analysis*, 51, 5142-5154.

[46] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, 40, 633-643.

[47] Jiang, J.M. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 720-729.

[48] Jiang, J.M. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, Berlin.

[49] Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 52 , 5066-5074.

[50] Laird, N.M. and Wair, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

[51] Lange, K., Little, R. and Taylor, J. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881-896.

[52] Lavori, P., Dawson, R. and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, 14, 1913-1925

[53] Li, H. (2005). A simulation study of the second-order least squares estimators for nonlinear mixed effects models. *Master's thesis, University of Manitoba.*

[54] Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal analysis using genaralized linear models. *Biometrika*, 73, 13-22.

[55] Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91, 1007-1016

[56] Lin, T.I. and Lee, J.C. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medcine*, 27, 1490-1507.

[57] Lindsey, J. K. (2000). Dropouts in longitudinal studies: definitions and models. *Journal of Biopharmaceutical Statistics*, 10, 503-525.

[58] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second edition*, John Wiley & Sons, Hoboken, N.J..

[59] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall: London.

[60] McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edition, New York: Wiley.

[61] Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester, UK: Wiley.

[62] Molenberghs, G., Kenward, M.G. and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84, 33-44.

[63] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer, New York.

[64] Nelder, J.A and Lee, Y. (2004). Conditional and marginal models: another view. *Statistical Science*, 19, 219-238.

[65] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.

[66] Noh, M. and Lee, Y. (2007). Robust modeling for inference from generalized linear model classes. *Journal of the American Statistical Association*, 102, 1059-1072.

[67] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57, 1027–1057.

[68] Pan, W., Zeng, D. and Lin, X. (2009). Estimation in semiparametric transition measurement error models for longitudinal data. *Biometrics*, 65, 728-736.

[69] Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.

[70] Pinheiro, J.C. and Chao, E.C. (2006) Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15, 58-81.

[71] Pinheiro, J.C., Liu C. and Wu, Y.N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10, 249-276.

[72] Preisser, J.S. and Qaqish, B.F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika*, 83, 551-562.

[73] Preisser, J.S. and Qaqish, B.F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics*, 55, 574-579.

[74] Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47, 825-839.

[75] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in FORTRAN*, Cambridge University Press.

[76] Qu, A. and Song, X.K. (2004). Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika*, 91, 447-459.

[77] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2, 1-21.

[78] Raine Study 2010, accessed 5 November 2010, <http://www.rainestudy.org.au/>.

[79] Richardson, A.M. (1997). Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association*, 92, 154-161.

[80] Rios, E.. Neuhauser, L., Margen, S. and Melnick, V. (1992). Accuracy of mothers' responses to questions about breast-feeding practices. *Food and Nutrition Bulletin*, 14, 115-118.

[81] 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.

[82] Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

[83] Rotnitzky, A. and Robins, J. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82, 805-820.

[84] Rotnitzky A., Robins J. M. and Scharfstein D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93, 1321-1339.

[85] Rubin, D.B. (1976). Infrence and missing data. *Biometrika*, 63, 581-592.

[86] Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 359, 538-543.

[87] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[88] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association,* 91, 473-489.

[89] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

[90] Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096-1120.

[91] Schenker, N. and Taylor, J.M.G. (1996).Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22, 425-446.

[92] Schennach, M. S. (2007). Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica*, 75, 201-239.

[93] Schneeweiss, H. and Augustin, T. (2006). Some recent advances in measurement error models and methods. *AStA Advances in Statistical Analysis*, 90, 183-197.

[94] Sinha, S.K. (2004). Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association*, 99, 451-460.

[95] Sinha, S.K. (2006). Robust inference in generalized linear mixed models for longitudinal data. *The Canadian Journal of Statistics*, 34, 1-18.

[96] Sutradhar, B.C. (2004). On exact quasilikelihood inference in generalized linear mixed models. *Sankhya : The Indian Journal of Statistics*, 66, 261-289.

[97] Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657-671.

[98] Tosteson, T., Buonaccorsi, J., and Demidenko, E. (1998). Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statistics in Medicine*, 17, 1959-1971.

[99] Tsiatis, T. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88, 447-458.

[100] Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23, 541-556.

[101] Vonesh, E. F., Wang, H., Nie L. and Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed effects models. *Journal of the American Statistical Association*, 97, 271-283.

[102] Wang, L. (2003). Estimation of nonlinear Berkson-type measurement error models. *Statistica Sinica*, 13, 1201-1210.

[103] Wang, L. (2004). Estimation of nonlinear models with Berkson measurement errors. *Annals of Statistics*, 32, 2559-2579.

[104] Wang, L. (2007). A unified approach to estimation of nonlinear mixed effects and Berkson measurement error models. *The Canadian Journal of Statistics*, 35, 233-248.

[105] Wang, L. and Hsiao, C. (1995). A simulated semiparametric estimation of nonlinear errors-in-variables models. *Working Paper*, Department of Economics, University of Southern California.

[106] Wang, L. and Hsiao, C. (2010). Method of moments estimation and identifiability of nonlinear semiparametric errors-in-variables models. *Journal of Econometrics*, in press.

[107] Wang, L. and Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics*, 60, 883-900.

[108] Wang, N. and Davidian, M. (1996). A note on covariate measurement error in nonlinear mixed effects models. *Biometrics*, 83, 801-812.

[109] Wang, N., Lin, X. and Guttierrez, R. G. (1999). A bias correction regression calibration approach in generalized linear mixed measurement error models. *Communication in Statistics - Theory and Methods*, 28, 217-233.

[110] 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.

[111] Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT. Press, Cambridge MA.

[112] Wu, C., Gumpertz, M.L. and Boos, D.D. (2001). Comparison of GEE, MINQUE, ML, and REML estimating equations for normally distributed data. *The American Statistician*, 55, 125-130.

[113] Yau, K.K.W. and Kuk, A.Y.C. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society: Series B*, 64, 101-117.

[114] 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.

[115] Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

[116] Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57, 795-802.

[117] Zhong, X.P., Fung, W.K. and Wei, B.C. (2002). Estimation in linear models with random effects and errors-invariables. *Annals of the Institute of Statistical Mathematics*, 54, 595-606.