2013-09-09

Linear Mixed Effects Models with Measurement Error: Bayesian Approach

Yuan, Zheng

UNIVERSITY OF CALGARY

Linear Mixed Effects Models with Measurement Error:

Bayesian Approach

by

Zheng Yuan

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

AUGUST, 2013

Abstract

Measurement error usually occurs in practice whenever we can not exactly observe the variables in a model. It has been long recognized that measurement error will bias the estimates if we do not correct it. Thus, it is significant for us to take into account measurement error in our analysis in order to obtain valuable results.

The Bayesian method is one of approaches for correcting measurement error in covariates in both linear models and linear mixed effects models. Bayesian approach became feasible and straightforward for many problems due to the availability of modern computers and computational tools such as the Markov chain Monte Carlo (MCMC) methods and `WinBUGS`.

In this paper, the first goal is to assess the effects of measurement error on naive analysis which ignore it in both linear models and linear mixed effects models. Then we focus on correcting measurement error through utilizing regression calibration methods and Bayesian methods, and comparing their performance in different situations. Estimating the regression coefficients using regression calibration methods and Bayesian methods in a linear mixed effects model with measurement error in time-varying covariates is mainly considered. We illustrate with real data analysis investigating the relationship between true dietary intake of beta-carotene and serum beta-carotene, and analyze the estimation results of naive methods, regression calibration methods and Bayesian methods.

Acknowledgements

First of all, I would like to express my deep gratitude to my supervisor Dr. Hyang Mi Kim for her unlimited encouragement and support in both academic area and my daily life over the past two years. She motivates my passion for seeking in a brand new statistics world and I feel really lucky to be one of her students.

Second, I have a lot thanks to all the faculty members at the Department of Mathematics and Statistics. I achieved a lot of knowledge from the courses taught by Dr. Gemai Chen, Dr. Xuewen Lu, Dr. John Collins, Dr. Murray Burke and Dr. Jingjing Wu. Their rigorous academic spirit and profound knowledge impressed me so much. I am grateful to my friends Chaoqun Ji, Shan Zhu, Sheng Li, Tasnima Abedin, Ce Bian and etc for their joyful companion and valuable suggestions.

Finally, I greatly acknowledge the Department of Mathematics and Statistics at the University of Calgary for providing me precious opportunity to study further.

Dedication

To My Dearest Mom and Dad

Table of Contents

List of Tables

List of Figures and Illustrations

Chapter 1

Introduction

Mixed effects models were developed to solve the problems of clustered data and have been a popular area in Statistics for the past years. We define the clustered data as the data in which the observations can be grouped into some disjoint classes, called clusters, according to some classification criterion. There are some examples of clustered data including repeated measures data and longitudinal data in which some observations can be made about the same individual. In a repeated measures study, multiple measurements of one or more variables are made on each individual. Longitudinal studies are vary similar with repeated measures study but the difference is the multiple measures on each individual are made over time. Data within the same cluster may be correlated, but data between different clusters are usually assumed to be independent in clustered data. This allows many statistical methods to analyze these correlated data.

Measurement error occurs frequently in practice whenever we can not exactly observe the variables in a model. It has been long recognized that measurement error will bias the estimates if we do not correct it. Further effects are unreliable coverage level of confidence intervals, then reduce the power of tests. So it is significant for us to consider measurement error in our analysis in order to obtain valuable results. Wang and Davidian (1996) are among the first researchers to research on the effect of measurement error on variance estimators. A detailed review of measurement error is Carroll et al. (2006).

In my thesis we will concentrate on the beta-carotene data which is from a "longitudinal validation study" conducted as part of a randomized clinical trial of beta-carotene dietary supplementation in prevention of recurrence of skin cancer (Tosteson, Buonaccorsi and Demidenko, 1998). The data contains serum beta-carotene measurements and measures

of beta-carotene intake which is based on a food frequency questionnaire. Both measures of serum beta-carotene and measures of beta-carotene intake were measured on 6 repeated days for 158 individuals, that we are able to treat as longitudinal data. Our target is to investigate the relationship between true dietary intakes of beta-carotene and serum beta-carotene. A main problem in statistical inference of the beta-carotene data is that the measures of beta-carotene intake are conducted based on a food frequency questionnaire, which leads to the measurement error problem. Some restrictions should be made in order to account for measurement error in the absence of validation data. We attempt to utilize a longitudinal model for the true dietary intakes of beta-carotene. The primary objective of my research is to explore the impact of measurement error by assessing the bias in naive estimators and to study approaches for correcting measurement error through implementing two approaches: regression calibration methods and Bayesian methods.

Simple linear models with measurement error and linear mixed effects models with measurement error will be the main subjects in this thesis. The thesis is organized as follows: In Chapter 2, some literature review about the topics in my thesis will be provided; In Chapter 3, some technical result concerning simple linear models with measurement error are considered. According to the simulation study, three different estimates (the naive estimate, the regression calibration estimate and the Bayesian estimate) are examined; The further research on the case of linear mixed effects models with measurement error will be done in Chapter 4. Simulation studies are conducted to assess the performance of the three different estimates. I make use of some existing `R` package for regression calibration methods and `WinBUGS` for Bayesian methods. `WinBUGS` is a programming language based software that is used to generate a random sample from the posterior distribution of the parameters of a Bayesian model; In the beta-carotene study that is designed to investigate the relationship between true dietary intakes of beta-carotene and serum beta-carotene, the true dietary intakes of beta-carotene are known to be measured with errors. The methods are applied to

beta-carotene data from an open web-site in Chapter 5; Chapter 6 draws the conclusion and discusses future study.

Chapter 2

Literature Review

In this chapter, the detailed background on the topics in my thesis will be considered. In section 2.1, the basic concept of measurement error is described. We will focus on the content of linear mixed effects models with measurement error in section 2.2. The details of Bayesian linear mixed effects models with measurement error are demonstrated in section 2.3.

2.1 Measurement error

The effect of measurement error on the independent variables in a regression model is a common problem in many scientific areas. There are substantial instances testifying that the implication of ignoring measurement error in inferential procedures may turn out in unreliable consequences. Refer to many reasons for the erroneous measurements, the most obvious ones being the inaccuracy of instruments and sampling error (Buonaccorsi, 2010). Some researchers do not consider measurement error since they are not aware of the measurement error, lack of softwares, or the information for correcting measurement error is not available for them. The high cost of exact measures, the subjective nature of some variable such as self-reported information and intrinsic biological variability are other reasons of the occurrence of measurement error. For instance, in epidemiologic studies, or during clinical trials, different measurements would be taken through different means and methods, some of which may be consistent with time, or vary with time. Some other times, researchers had run into deliberate measurement of wrong quantities due to substitution of a cheaper and more convenient method of measurement for the direct measurement.

There exist three typically components in the model with measurement error: the model for the true value called outcome model which can be any statistical model, the measurement

error model which specifies the relationship between the true values and observed values, and extra information which is utilized to carry out corrections for measurement error such as replicates, validation data in which both the true and mis-measured values are observed, and the instrumental variables. The simplest measurement error model is the classical measurement error model, which is an unbiased and additive measurement error model. An alternative model is the Berkson error model, which typically arises in laboratory studies and experimental situations in which the observed variable is controlled for. Usually, additional information is needed in order to guarantee the identifiability of the parameters. Additional data can be available in different forms such as the internal validation data set and replication data. Higgins, Davidian and Giltinan (1997) and Tosteson, Buonaccorsi, and Demidenko (1998) discovered that if a mis-measured covariate is observed longitudinally, then a structural model for the covariate with dimension less than the number of observations per subject allows all parameters to be identified (Carroll et al. 2006).

Different types of measurement error can arise in practice. An important distinction is made between differential and non-differential measurement errors. The error in the observed value W is non-differential if no additional information on Y is contained in W with respect to X. In this case, W is said to be a surrogate for X and the equivalent concept is conditional independence that is Y and W are independent given $X = x$. Otherwise, the error is said to be differential. Many different error sources can be encountered in applications, which implies that both non-differential and differential errors, with classical or Berkson components, can be defined. An accurate specification of the measurement error model is crucial due to the different impacts of the errors on the inferential results and the different available correction techniques. We can classify the correction techniques into two groups according to their interpretation of the unobserved variables. If the method makes no assumptions on the unobserved variables, that is, unknown non-random constants, then it is functional. On the contrary, the method is defined to be structural if it assumes the unobserved variables to be

random variables.

Methods for correcting measurement error

Many different measurement error correction techniques have been suggested in many literatures. They differ according to the assumptions about the distribution of the unobserved variable, the availability of additional data about the unobserved variable and the theoretical background of the approach, which may be parametric or nonparametric. We distinguish among different models relating the variable. Comprehensive reviews of covariate measurement error methods are provided in Fuller (1987), Gustafson (2004), and Carroll et al. (2006). Commonly used methods for covariate measurement errors in regression includes regression calibration methods, simulation extrapolation (SIMEX) methods, likelihood methods, approximation methods and Bayesian methods. Regression calibration methods and SIMEX methods make minimal assumptions on the distribution of the unobserved covariates. In contrast, likelihood methods and Bayesian methods make strong distribution assumptions on the unobserved covariates, so they are more efficient if the covariate distributions are correctly specified. In this thesis, we emphasize regression calibration methods and Bayesian methods for correcting measurement error.

Regression calibration method

The regression calibration method is a conceptually straightforward method to correct measurement error and has been successfully applicable to almost any regression models with measurement error on covariates. Compared with naive estimation, rather than using the observed mis-measured values W as the covariates in the regression model, regression calibration method attempts to model the distribution of the unobserved true values given the observed mis-measured values, and then substitute the unobserved true covariates by an estimated value of the conditional expectation $\hat{X} = E(X|W)$ in the regression model. After

the true covariate is approximated by an estimate, one perform a standard analysis as if there were no measurement error in covariates. Due to the simplicity of its application with existing softwares, the regression calibration method becomes a commonly adopted method to correct the measurement error on covariates in the regression models. However, it requires some prerequisite such as X and W must have a linear homoscedastic relationship between each other. This method will not be accurate if the requirement does not hold.

Bayesian method

The Bayesian method is found to be fashionable in science in the beginning of the 21 century. However, until the late 1980s, Bayesian statistics were considered only as an interesting alternative to the "classical" theory (Ntzoufras, 2009). Over the last two decades there has been an "MCMC revolution" in which the Bayesian method has become a highly popular and effective tool for the applied statistician (Ntzoufras, 2009). The main difference between classical statistical methods and Bayesian methods is that Bayesian methods consider parameters as random variables that are characterized by a prior distribution which is the assumed distributions for the parameters. Due to the availability of modern computers and computational tools such as the Markov chain Monte Carlo(MCMC) methods and `WinBUGS`, Bayesian methods for many problems become feasible and straightforward. The Bayesian inference is based on the posterior distribution, which is the conditional distribution of unobserved quantities, such as the parameters or unobserved data, given the observed data. The posterior distribution is our target and it summarizes all the information about the parameters.

The Bayesian inference is based on the rationale which is called Bayes Theorem. Assume two outcomes A and B, and $A = A_1 \cup \cdots \cup A_n$ for which $A_i \bigcap A_j = \emptyset$, $i \neq j$. Bayes Theorem

states that the conditional probability of A_i given B can be expressed as (Ntzoufras, 2009)

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} \tag{2.1}$$

In a general form, for any outcome A and B

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A) \tag{2.2}$$

This equation is also called $Bayes^{'}rule$, although it was originally found by Piere-Simon de Laplace (Hoffmann-Jorgensen, 1994).

We assume the unknown parameters $\boldsymbol{\theta}$ are random variables following a distribution with probability density function $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{\theta}_h)$ which is called prior distribution that expresses the information available to us before any data are involved in the statistical analysis. The parameters $\boldsymbol{\theta}_h$ in the prior distribution are called hyper-parameters and are often assumed to be known, which can be selected based on similar studies, expert opinions or even non-informative (Wu, 2010). Based on the $Bayes^{'}rule$, we can obtain the posterior distribution of parameters as

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \tag{2.3}$$

The Bayesian inference will depend on this posterior distribution which is the multiplication of the likelihood $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$ and the prior distribution $f(\boldsymbol{\theta})$. The prior distribution $f(\boldsymbol{\theta})$ has a significant effect on the Bayesian inference and we can test the sensitivity by selecting different prior distributions or different values of hyper-parameters. If we do not have any prior information, the non-informative prior $f(\boldsymbol{\theta}) \propto 1$ will be chosen, then we can obtain $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y})$. It means that the Bayesian inference is the same with the likelihood inference when we use the non-informative prior distribution.

The Bayesian method is another popular approach to correct measurement error. The Bayesian formulation of general measurement error problems has been developed (Clayton, 1992). Structural specifications entail the formulation of three sub-models: a response model

relating $\mathbf{X}$ and $\mathbf{Y}$, a measurement error model specifying the relationship between $\mathbf{W}$ and $\mathbf{X}$ and a prior model for the prior distribution of the unobserved true covariates $\mathbf{X}$. A graphical model with suitable conditional independence assumptions is used to link these sub-models. For a general linear model with the additive classical measurement error

$$f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1): \quad \text{response model with parameters} \quad \boldsymbol{\theta}_1 \tag{2.4}$$

$$f(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta}_2): \quad \text{measurement error model with parameters} \quad \boldsymbol{\theta}_2 \tag{2.5}$$

$$f(\mathbf{X}|\boldsymbol{\theta}_3): \quad \text{prior model with parameters} \quad \boldsymbol{\theta}_3 \tag{2.6}$$

An important assumption is that of non-differential measurement error

$$f(\mathbf{Y}|\mathbf{X}, \mathbf{W}; \boldsymbol{\theta}_1) = f(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta}_1) \tag{2.7}$$

Then the joint distribution can be written as

$$f(\mathbf{Y}, \mathbf{X}, \mathbf{W}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) = f(\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_2) f(\boldsymbol{\theta}_3) f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1) f(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta}_2) f(\mathbf{X}|\boldsymbol{\theta}_3) \tag{2.8}$$

In the Bayesian method, we attempt to make inference about unknown data $\mathbf{X}$ and unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ through deriving the posterior densities conditional on the observed data $(\mathbf{Y}, \mathbf{W})$. The joint posterior densities of the unknown values can be expressed as

$$\begin{aligned} f(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}, \mathbf{W}) &\propto f(\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_2) f(\boldsymbol{\theta}_3) f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1) f(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta}_2) f(\mathbf{X}|\boldsymbol{\theta}_3) \\ &= f(\boldsymbol{\theta}_1) f(\boldsymbol{\theta}_2) f(\boldsymbol{\theta}_3) \prod_{i=1}^{n} f(Y_i|X_i, \boldsymbol{\theta}_1) \prod_{i=1}^{n} f(W_i|X_i, \boldsymbol{\theta}_2) \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_3) \end{aligned} \tag{2.9}$$

It is very difficult for us to evaluate the joint or marginal densities based on either analytic approximation or numerical integration since the joint density function has so many unknown values. In order to avoid the intractable integrals, it is convenient for us to implement the Gibbs sampler generating dependent samples from joint and marginal posterior densities and make inference on the posterior distributions of unknowns.

The Gibbs sampler which is one of Markov chain Monte Carlo (MCMC), was introduced by Geman and Geman (1984). It has been widely used to compute approximate posterior densities in many statistical areas. The Gibbs sampler generates a Markov chain whose stationary distribution is the posterior distribution and its key feature is this chain can be simulated using only the joint densities of the parameters, the unobserved data and the observed data such as the product of the prior and the likelihood. One advantage of the Gibbs sampler is that, in each step, random values must be generated from uni-dimensional distribution for which a wide variety of computational tools can be implemented (Gilks, 1996). Suppose a set of k variables with joint distribution $f(\theta_1, \cdots, \theta_k)$ which is uniquely specified by the set of k full-conditional distributions

$$f(\theta_i|\boldsymbol{\theta}_{\setminus i}) \quad = \quad f(\theta_i|\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_k) \tag{2.10}$$

where $i = 1, \cdots, k$. First of all, we need to set the initial value $\boldsymbol{\theta}^{(0)}$. For each iteration of the algorithm, we will repeat the following procedure: 1. set $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t-1)}$; 2. update θ_i from $f(\theta_i|\theta_1^{(t)}, \cdots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \cdots, \theta_k^{(t-1)})$, $t = 1, \cdots, T$; 3. set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(current)}$ and save it as the general set of values at $t+1$ iteration of the algorithm. We can check the convergence of the Gibbs sampler through inspecting the sample summary statistics and density estimates.

2.2 Linear mixed effects models with measurement error

Mixed effects models provide a stage to model cluster dependence in which the response can be defined as a function of fixed effects, unobserved cluster specific random effects and an error term. The data within the same cluster are statistically dependent as they share common random effects. There are two types of parameters in a mixed effects model: fixed effects which associate with the average effects of predictors on the response; random effects which represent the effects of clusters on the repeated observations in corresponding clusters. Variance-covariance component which relates to the covariance structure of the random effects and the error term. In my thesis, I will restrict the random effects and the error term

to be based on normal distribution. Maximum likelihood method and restricted maximum likelihood method (Harville, 1974) have been generally adopted for analyzing linear mixed effects models (Longford, 1993). For nonlinear mixed effects models, statisticians are still debating on the estimation method although several methods proposed.

Linear mixed effects models are mixed effects models in which both the fixed effects and the random effects have a linear contribution to the response. For longitudinal data or clustered data, classical linear regression is inappropriate because the observations within each cluster may be correlated, which makes the independence assumption for classical model not work. To incorporate the correlation within clusters and the variation between clusters, we can obtain linear mixed effects models from classical linear regression models by adding random effects, and the magnitude of the random effects measures the variation between clusters. Assume Y_{ij} is the response for individual i at time t_{ij}, $i = 1, \cdots, G$, $j = 1, \cdots, n_i$. $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{in_i})'$ is the n_i repeated observations within individual i. Using the hierarchical notation of Laird and Ware (1982),

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad i = 1, \cdots, G \tag{2.11}$$

$$\mathbf{b}_i \sim N(0, \mathbf{D}) \tag{2.12}$$

$$\boldsymbol{\epsilon}_i \sim N(0, \sigma^2\mathbf{I}) \tag{2.13}$$

where $\mathbf{Y}_i$ is a $n_i \times 1$ response vector, $\mathbf{X}_i$ is a $n_i \times (p+1)$ design matrix for fixed effects containing covariates of individual i, $\boldsymbol{\beta}$ are $(p+1) \times 1$ regression coefficients for fixed effects, $\mathbf{Z}_i$ is a $n_i \times q$ design matrix for random effects, $\mathbf{b}_i$ is a $q \times 1$ matrix for random effects and $\boldsymbol{\epsilon}_i$ is a $n_i \times 1$ matrix for random errors of within individual measurements, which demonstrates the variability of the repeated measurements within each individual. $\sigma^2\mathbf{I}$ is a $n_i \times n_i$ variance-covariance matrix of within individual measurements. Wang and Heckman (2009) proved that linear mixed effects models are always identifiable if variance-covariance matrix of within individual measurements is $\sigma^2\mathbf{I}$. The matrix $\mathbf{D}$ is often unstructured but we can define it

as a diagonal matrix (Jenrich and Schluchter, 1986). The variance of the random effects $\mathbf{b_i}$ or the diagonal elements of $\mathbf{D}$ are sometimes called variance components, which measures the variability between individuals that are not explained by covariates. There are some special cases of model (2.11) such as variance components models (Searle, Casella and McCulloch, 1992), mixed effects ANOVA models (Miller, 1977), and linear models for longitudinal data (Laird and Ware, 1982). Maximum likelihood method (ML) and restricted maximum likelihood method (REML) are the typically estimation methods for the statistical inference of a linear mixed effects model (Laird and Ware, 1982; Lindstrom and Bates, 1988). We can only obtain the REML estimates by computer because the deriving procedure contains a rather complicated nonlinear optimization issue, resulting in no closed form expressions for the distribution of REML estimates. EM algorithm (Dempster, Laird and Rubin, 1977) and Newton-Raphson methods (Thisted, 1988) are the most common methods to solve the optimization, but the latter seems to be more efficient than the former (Lindstrom and Bates, 1988). As the independence assumption does not hold, the classical asymptotic theory for ML estimates (Lehmann, 1983) is not available for linear mixed effects models. Bayesian method is developed by using a hierarchical model approach. It is flexible in manipulating complicated situations like constrained parameters and non-Gaussian distributions for the random effects or error terms, but it also has drawbacks such as the selection of prior distribution for all the population parameters and the requirement of intensive computational effort.

The linear mixed models specially incorporate the variation within individuals and the variation between individuals. Therefore, it can be interpreted as a hierarchical two-stage model: the first stage specifies the within-individual variation and the second stage specifies the between-individual variation (Wu, 2010).

An additive classical measurement error can be written as

$$\mathbf{W}_i \;=\; \mathbf{X}_i + \mathbf{u}_i \quad i = 1, \cdots, G. \tag{2.14}$$

where $\mathbf{W}_i$ is the error-prone measure of $\mathbf{X}_i$, $E(\mathbf{u}_i) = \mathbf{0}$ and $\mathbf{u}_1, \cdots, \mathbf{u}_G$ are assumed independent with each other. To model the true covariates $\mathbf{X}_i$, we need to consider a covariate mixed effects model to incorporate between-individual variation and within-individual correlation.

For longitudinal data, a linear mixed effects model to address measuremtn error can be written as

$$\begin{aligned} \mathbf{W}_i &= \mathbf{X}_i + \mathbf{u}_i \\ &= \mathbf{M}_i\boldsymbol{\eta} + \mathbf{N}_i\boldsymbol{\delta}_i + \mathbf{u}_i \end{aligned} \tag{2.15}$$

where $\mathbf{W}_i$ is the observed measure of $\mathbf{X}_i$, $\mathbf{M}_i$ and $\mathbf{N}_i$ are known design matrices, $\boldsymbol{\eta}$ contains unknown fixed parameters, $\boldsymbol{\delta}_i$ are random effects and $\mathbf{u}_i$ are covariates measurement errors. Assume $\boldsymbol{\delta}_i$ are independent and identically distributed with $N(0, \boldsymbol{\Omega}_\delta)$, $\boldsymbol{\Omega}_\delta$ is an unknown covariance matrix, $\mathbf{u}_i$ are independent and identically distributed with $N(0, \sigma_u^2\mathbf{I})$. σ_u^2 presents the magnitude of the measurement error. Furthermore, we assume $\boldsymbol{\delta}_i$, $\mathbf{u}_i$, $\mathbf{b}_i$, $\boldsymbol{\epsilon}_i$ are independent with each other. In the measurement error model (2.15), the unobserved true covariates $\mathbf{X}_i$ can be written as

$$\mathbf{X}_i = \mathbf{M}_i\boldsymbol{\eta} + \mathbf{N}_i\boldsymbol{\delta}_i \tag{2.16}$$

The measurement error model can be fitted using standard methods for linear mixed effects models given $\mathbf{W}_i$. The lack of additional residual error in (2.16) allows for estimation of $\boldsymbol{\eta}$, σ_u^2, and $\boldsymbol{\Omega}_\delta$ from the $\mathbf{W}$ data.

2.3 Bayesian linear mixed effects models with measurement error

The general Bayesian approach can be applied to mixed effects models with measurement error. Bayesian estimation analytic expressions are often unavailable so Monte Carlo methods are often used, which can be computationally intensive. For linear models, some analytical expressions can be obtained. First, we consider Bayesian linear mixed model,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, i = 1, \cdots, G,$$

$$\begin{aligned} \mathbf{b}_i &\sim N(0, \mathbf{D}), \boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \mathbf{I}) \\ \boldsymbol{\beta} &\sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0); \text{ a prior distribution for the mean fixed parameter} \end{aligned} \tag{2.17}$$

The hyper-parameter $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ are known. We assume a non-informative prior for $\boldsymbol{\beta}$, i.e., $\boldsymbol{\Sigma}_0^{-1} = 0$ or $\boldsymbol{\beta} \sim \text{Uniform}(-\infty, \infty)$. For the convenience of presentation, we write the model in a more compact form as follow: Let $N = \sum_{i=1}^{G} n_i$, and let $\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_G)', \mathbf{b} = (\mathbf{b}_1, \cdots, \mathbf{b}_G)'$, and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_G)'$. Bayesian estimation of the mean parameter $\boldsymbol{\beta}$ can be based on the following posterior distribution of $\boldsymbol{\beta}$ given the observed data:

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{Y}) &= \frac{f(\mathbf{Y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})}{f(\mathbf{Y})} \\ &= \frac{\int f(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{b}) f(\boldsymbol{\beta}) f(\mathbf{b}) d\mathbf{b}}{\int \int f(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{b}) f(\boldsymbol{\beta}) f(\mathbf{b}) d\mathbf{b} d\boldsymbol{\beta}} \end{aligned} \tag{2.18}$$

Based on properties of multivariate normal distributions, it can be shown that the posterior distribution of $\boldsymbol{\beta}$ can be found (Searle et al, 1992; Davidian and Giltinan, 1995). Bayesian inference can be based on the Gibbs sampler along with rejection sampling method (Zeger and Karim, 1991; Gelman et al. 2003). A Gibbs sampler method to generate samples from the posterior distribution $f(\boldsymbol{\beta}, \mathbf{D}, \mathbf{b}|\mathbf{Y})$ is described as follow:

- generate $\boldsymbol{\beta}^{(t)}$ from $f(\boldsymbol{\beta}|\mathbf{D}^{(t-1)}, \mathbf{b}^{(t-1)}, \mathbf{Y})$
- generate $\mathbf{D}^{(t)}$ from $f(\mathbf{D}|\boldsymbol{\beta}^{(t)}, \mathbf{b}^{(t-1)}, \mathbf{Y})$
- generate $\mathbf{b}^{(t)}$ from $f(\mathbf{b}|\boldsymbol{\beta}^{(t)}, \mathbf{D}^{(t)}, \mathbf{Y})$ $\quad t = 1, \cdots, T.$

Once we generate many such samples, the posterior mean and posterior covariance can be approximated by the same mean and sample covariance based on the simulated data.

Now, with measurement error, Bayesian approach typically treats the unobserved true covariate $\mathbf{X}$ as missing data and imputes them many times by sampling from the conditional distribution of $\mathbf{X}$ given all other variables and observed data. Specifically by treating the unobserved true covariate $\mathbf{X}$ as missing data, we can write the "complete" likelihood as

follows:

$$f(\boldsymbol{\theta})f(\mathbf{Y}|\mathbf{X},\mathbf{b},\boldsymbol{\theta})f(\mathbf{b}|\boldsymbol{\theta})f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta})f(\mathbf{X}|\boldsymbol{\theta}) \tag{2.19}$$

where $\boldsymbol{\theta}$ is all unknown parameters. Bayesian inference is based on the following posterior distribution of $\boldsymbol{\theta}$ given the observed data:

$$f(\boldsymbol{\theta}|\mathbf{Y},\mathbf{W}) = \frac{\int\int f(\mathbf{Y},\mathbf{W},\mathbf{X},\mathbf{b}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\mathbf{x}d\mathbf{b}}{\int\int\int f(\mathbf{Y},\mathbf{W},\mathbf{X},\mathbf{b}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\mathbf{x}d\mathbf{b}d\boldsymbol{\theta}} \tag{2.20}$$

We again use the Gibbs sampler method, which generates many samples from the posterior distribution by iterating sampling from lower dimensional conditional distribution. The Gibbs sampler iterates between following steps in the i^{th} cluster:

- generate samples of the unobserved covariate $\mathbf{X}_i$ from its posterior distribution given the observed data and the last generated random effects $\mathbf{b}_i$ and parameters $\boldsymbol{\theta}$:

$$\mathbf{X}_i \sim f(\mathbf{X}_i|\mathbf{Y}_i,\mathbf{W}_i,\mathbf{b}_i,\boldsymbol{\theta}) \propto f(\mathbf{Y}_i|\mathbf{X}_i,\mathbf{b}_i,\boldsymbol{\theta})f(\mathbf{W}_i|\mathbf{X}_i,\boldsymbol{\theta})f(\mathbf{X}_i|\boldsymbol{\theta})$$

- generate samples of the random effect $\mathbf{b}_i$ from its posterior distribution given the observed data and the last generated unobserved values $\mathbf{X}_i$ and parameters $\boldsymbol{\theta}$:

$$\mathbf{b}_i \sim f(\mathbf{b}_i|\mathbf{Y}_i,\mathbf{W}_i,\mathbf{X}_i,\boldsymbol{\theta}) \propto f(\mathbf{Y}_i|\mathbf{X}_i,\mathbf{b}_i,\boldsymbol{\theta})f(\mathbf{b}_i|\boldsymbol{\theta})$$

- generate samples of the parameters $\boldsymbol{\theta}$ from its posterior distribution given the observed data and the last generated unobserved values $\mathbf{X}$ and random effects $\mathbf{b}$:

$$\boldsymbol{\theta} \sim f(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X},\mathbf{W},\mathbf{b}) \propto f(\boldsymbol{\theta})f(\mathbf{Y}|\mathbf{X},\mathbf{W},\mathbf{b},\boldsymbol{\theta})f(\mathbf{b}|\boldsymbol{\theta})f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta})f(\mathbf{X}|\boldsymbol{\theta})$$

Iterating the above three steps, the resulting sequence is a Markov chain which will converges to its stationary distribution, the target posterior distribution.

Throughout this thesis, we assume the classical measurement error is modeled as a structure with a constant variance and the measurement error is additive and non-differential. For simulation studies, the measurement error variance is assumed to be known. The repeated measurements will be used to estimate measurement error variance for longitudinal data in real data analysis.

Chapter 3

Simple Linear Models with Measurement Error

In this chapter, we focus on the estimation of the parameters in a simple linear model with measurement error in covariates. Simulation studies will be conducted to compare the performance of regression calibration methods and Bayesian methods. We assume the measurement error variance is known for identifiability problem and the measurement error is non-differential. In section 3.1, we construct a simple linear regression model with measurement error and assess the bias of estimator in naive estimation. The correcting methods consisting of moment-based corrections, regresson calibration methods and Bayesian methods are best described in section 3.2. Simulation results are demonstrated in section 3.3.

3.1 Simple linear models with measurement error

A classical simple linear model with parameters $\boldsymbol{\theta}_1 = (\beta_0, \beta_1, \sigma_\varepsilon^2)$ can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad i = 1, \cdots, n \tag{3.1}$$

where Y_i is the response, X_i is the predictor and the error in the equation $\epsilon_i s$ are assumed uncorrelated with mean 0 and constant variance σ_ϵ^2. The unobserved $X_i s$ are assumed to be independent and identically distributed with mean μ_x and variance σ_x^2. Let W_i denote the error-prone measurement for the predictor X_i, a classical additive measurement error model can be expressed as

$$W_i = X_i + u_i \quad i = 1, \cdots, n \tag{3.2}$$

where $u_i s$ are assumed to be independent and identically distributed with mean 0 and variance σ_u^2. Furthermore, We assume X_i, u_i and ϵ_i are mutually independent.

Assessing bias in naive estimators

The behavior of the naive estimation in a simple normal structural model with the normal additive measurement error is considered. Assume

$$(X_i, \epsilon_i, u_i)' \sim N_3\{(\mu_x, 0, 0)'; diag(\sigma_x^2, \sigma_\epsilon^2, \sigma_u^2)\}$$

Then

$$\begin{pmatrix} W_i \\ Y_i \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} \mu_x \\ \mu_y = \beta_0 + \beta_1 \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \sigma_{wy} \\ \sigma_{yw} & \sigma_y^2 \end{pmatrix} \right\} \tag{3.3}$$

where

$$\begin{pmatrix} \sigma_w^2 & \sigma_{wy} \\ \sigma_{yw} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 + \sigma_u^2 & \beta_1 \sigma_x^2 \\ \beta_1 \sigma_x^2 & \beta_1^2 \sigma_x^2 + \sigma_\epsilon^2 \end{pmatrix} \tag{3.4}$$

If the observed $(\mathbf{Y}, \mathbf{W})$ are jointly normal, the distribution of (Y_i, W_i) is characterized by the elements of mean vector and covariance matrix, that is, $(\mu_x, \mu_y, \sigma_w^2, \sigma_{wy}$ and $\sigma_y^2)$. Because the model contains six parameters $(\beta_0, \beta_1, \sigma_\epsilon^2, \mu_x, \sigma_x^2$ and $\sigma_u^2)$, there are many different configurations that lead to the same distribution of the observations. Therefore, the model is not identified. Therefore some additional information has to be provided, either a data structure or a model assumptions. σ_u is known or can be estimated and $\frac{\sigma_u}{\sigma_\epsilon}$ is known are the most common extra information. There are assumptions found in the literature, among other, σ_u is known or $\lambda = \frac{\sigma_u^2}{\sigma_w^2}$ is known makes the model identifiable.

Naive methods are conducted to utilize the observed mis-measured W_i as the true unobserved X_i in model (3.1)

$$E(Y_i|W_i) \quad = \quad \beta_{0naive} + \beta_{1naive} W_i \tag{3.5}$$

Taking the expectations conditional on W_i for (3.1)

$$E(Y_i|W_i) \quad = \quad E[E(Y_i|X_i, W_i)|W_i]$$

$$
\begin{aligned}
&= E[E(Y_i|X_i)|W_i] \quad \text{(non-differential measurement error assumption)} \\
&= E(\beta_0 + \beta_1 X_i|W_i) \\
&= \beta_0 + \beta_1 E(X_i|W_i) \\
&= \beta_0 + \Big(1 - \frac{\sigma_x^2}{\sigma_w^2}\Big)\beta_1\mu_x + \frac{\sigma_x^2}{\sigma_w^2}\beta_1 W_i \qquad (3.6)
\end{aligned}
$$

This means that

$$
\begin{aligned}
E(\hat{\beta}_{1naive}) &= \frac{\sigma_x^2}{\sigma_w^2}\beta_1 = \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right)\beta_1 = \lambda\beta_1 \\
E(\hat{\beta}_{0naive}) &= \beta_0 + \Big(1 - \frac{\sigma_x^2}{\sigma_w^2}\Big)\beta_1\mu_x \qquad (3.7)
\end{aligned}
$$

The slope is attenuated by $\lambda = \frac{\sigma_x^2}{\sigma_w^2} = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ and the intercept is shifted by $(1-\lambda)\beta_1\mu_x$. Where $\hat{\beta}_{1naive} = \sum_{i=1}^{n}[(W_i - \overline{W})^2]^{-1}\sum_{i=1}^{n}(W_i - \overline{W})(Y_i - \overline{Y})$ and $\hat{\beta}_{0naive} = \overline{Y} - \hat{\beta}_{1naive}\overline{W}$. If $\beta_1 \neq 0$ and $\sigma_u^2 > 0$, then $|\lambda\beta_1| < |\beta_1|$ leads to what is known as attenuation with the attenuation factor $\lambda = \frac{\sigma_x^2}{\sigma_w^2} = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} < 1$. As a result, the analysis will be involved in the attenuation problem as $\lambda < 1$ and $\hat{\beta}_{1naive}$ will be biased towards 0 as σ_u is large enough. The existence of measurement error in $\mathbf{X}$ gives rise to the estimation bias can not be reduced though increasing the sample size. With increase sample size, the impact of the attenuation is aggravated such that the estimates are tending towards becoming more precisely wrong.

3.2 Correcting Methods

3.2.1 Moment-based Correcting Method

Moment-based correcting method is one of the simplest methods that takes a linear transformation of the naive estimates of the coefficients in the models. The basic idea of this method is to correct for the bias in $\hat{\sigma}_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2$ as estimator of σ_x^2. The corrected

estimates can be expressed as

$$\begin{aligned}\hat{\beta}_1 &= \hat{\lambda}^{-1}\hat{\beta}_{1naive} = \frac{\hat{\sigma}_w^2}{\hat{\sigma}_x^2}\hat{\beta}_{1naive} \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1\overline{W} \end{aligned} \tag{3.8}$$

where $\hat{\lambda} = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_w^2} = \frac{\hat{\sigma}_w^2 - \sigma_u^2}{\hat{\sigma}_w^2}$ and σ_u^2 is known. (Fuller 1987; Rosner et al 1989; Thurston et al. 2003)

3.2.2 Regression Calibration Methods

Two primary steps for the regression calibration methods are performed in order to address the measurement errors in covariates: First, model and estimate the regression of X_i on W_i depending on parameters; Second, replace the unobserved X_i by its estimate and run a standard analysis to obtain parameter estimates. When $(\mathbf{X}, \mathbf{W})$ is approximately jointly normal, the regression of X_i on W_i is linear:

$$\begin{aligned}\hat{x_i} &= \hat{E}(X_i|W_i) = \hat{\mu}_x + \frac{\hat{\sigma}_x^2}{\hat{\sigma}_w^2}(W_i - \hat{\mu}_x) \\ &= \overline{W} + \frac{\hat{\sigma}_w^2 - \sigma_u^2}{\hat{\sigma}_w^2}(W_i - \overline{W}) \\ &= \Big(1 - \frac{\hat{\sigma}_w^2 - \sigma_u^2}{\hat{\sigma}_w^2}\Big)\overline{W} + \frac{\hat{\sigma}_w^2 - \sigma_u^2}{\hat{\sigma}_w^2}W_i \\ &= (1 - \hat{\lambda})\overline{W} + \hat{\lambda}W_i \end{aligned} \tag{3.9}$$

where σ_u^2 is assumed to be known or can be estimated and $\hat{\sigma}_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2$. It is important for us to realize that the regression calibration model is an approximate model for the observed data. Therefore it is not necessarily the same as the actual mean for the observed data while is only moderately different in many cases.

3.2.3 Bayesian Methods

For a simple linear model with the additive classical measurement error, the three sub-models can be expressed as

$$f(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta}_1): \quad \text{response model with parameters} \quad \boldsymbol{\theta}_1 = (\beta_0, \beta_1, \sigma_\epsilon^2) \tag{3.10}$$

$$f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta}_2): \quad \text{measurement error model with parameters} \quad \boldsymbol{\theta}_2 = (\sigma_u^2) \tag{3.11}$$

$$f(\mathbf{X}|\boldsymbol{\theta}_3): \quad \text{exposure model with parameters} \quad \boldsymbol{\theta}_3 = (\mu_x, \sigma_x^2) \tag{3.12}$$

An important assumption is that the measurement error is non-differential, then

$$f(\mathbf{Y}|\mathbf{X},\mathbf{W};\boldsymbol{\theta}_1) \;=\; f(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta}_1) \tag{3.13}$$

The joint distribution will be

$$f(\mathbf{Y},\mathbf{X},\mathbf{W},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3) = f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_3)\prod_{i=1}^{n} f(Y_i|X_i,\boldsymbol{\theta}_1)\prod_{i=1}^{n} f(W_i|X_i,\boldsymbol{\theta}_2)\prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_3) \tag{3.14}$$

The joint posterior densities of the unknown values can be written as

$$f(\mathbf{X},\boldsymbol{\theta}|\mathbf{Y},\mathbf{W}) \propto f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_3)\prod_{i=1}^{n} f(Y_i|X_i,\boldsymbol{\theta}_1)\prod_{i=1}^{n} f(W_i|X_i,\boldsymbol{\theta}_2)\prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_3) \tag{3.15}$$

We first need to derive the full conditional posterior distributions for all the unknown values in order to utilize the Gibbs sampler. The full conditional posterior distributions of unknown data $\mathbf{X}$ can be written as

$$\begin{aligned} f(X_i|Y_i,W_i,\boldsymbol{\theta}) \;&\propto\; f(Y_i|X_i,\boldsymbol{\theta}_1)f(W_i|X_i,\boldsymbol{\theta}_2)f(X_i|\boldsymbol{\theta}_3) \\ &=\; f(Y_i|X_i,\beta_0,\beta_1,\sigma_\epsilon^2)f(W_i|X_i,\sigma_u^2)f(X_i|\mu_x,\sigma_x^2) \end{aligned} \tag{3.16}$$

With normal assumption

$$\begin{aligned} f(X_i|Y_i,W_i,\boldsymbol{\theta}) \;&\propto\; \exp\left(-\frac{1}{2\sigma_\epsilon^2}(Y_i-\beta_0-\beta_1X_i)^2\right) \\ &\times\; \exp\left(-\frac{1}{2\sigma_u^2}(W_i-X_i)^2\right)\exp\left(-\frac{1}{2\sigma_x^2}(X_i-\mu_x)^2\right) \end{aligned} \tag{3.17}$$

The full conditional posterior distributions of unknown parameters $\boldsymbol{\theta}$ is

$$\begin{aligned} f(\beta_j|\beta_{\setminus j}, \sigma_\epsilon^2, \mathbf{Y}, \mathbf{X}) &\propto f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1) f(\beta_j|\mathbf{X}, \sigma_\epsilon^2) \\ &= \prod_{i=1}^{n} f(Y_i|X_i, \beta_0, \beta_1, \sigma_\epsilon^2) f(\beta_j) \end{aligned} \tag{3.18}$$

$$\begin{aligned} f(\sigma_\epsilon^2|\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) &\propto f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1) f(\sigma_\epsilon^2|\boldsymbol{\theta}_1, \mathbf{X}) \\ &= \prod_{i=1}^{n} f(Y_i|X_i, \beta_0, \beta_1, \sigma_\epsilon^2) f(\sigma_\epsilon^2) \end{aligned} \tag{3.19}$$

where we usually assume the unknown parameters are apriori independent. Based on the Gibbs sampler, given the initial values $\boldsymbol{\theta}^{(0)}$ and $\mathbf{X}^{(0)}$, we first draw the samples of the unobserved $\mathbf{X}$ from its full conditional posterior distribution (3.16). Then the samples of the unknown parameters $\boldsymbol{\theta}$ can be generated from (3.18) and (3.19). After repeating the above procedures many times for a burn-in period, finally we are able to obtain the desired samples from the posterior distribution (3.15).

3.3 Simulation Studies

A simulation study is conducted to evaluate and compare the performance of naive methods, regression calibration methods and Bayesian methods. Simulated data were generated: the sample size $n = 20$ and 100, the intercept of the regression $\beta_0 = 1$, the slope of the regression $\beta_1 = 0.5$, $\mu_x = 1$, $\sigma_x = 1$, $\sigma_\epsilon = 0.3$. We look at numerical summaries for parameters $\boldsymbol{\theta}$ with different measurement error variance ($\sigma_u^2 = 0.25, 0.49, 1$). In Bayesian analysis, the choice of prior distribution is important since it may affect the final results. We specify the parameters β_0, β_1 and μ_x have independent normal priors with mean 0 and precision 10^{-6}. It is quite common in Bayesian analysis to specify the normal distribution in terms of its precision instead of its variation. Thus, we specify τ_ϵ, τ_x, τ_u have independent gamma priors with parameters 3 and 1. $\sigma = \tau^{-\frac{1}{2}}, \lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} = \frac{\tau_u}{\tau_u+\tau_x}$. The number of iterations is $10,000$ with $1,000$ burn-in, 20 thin in each of the 3 chains. We simulated 500 sets of data for naive

methods, regression calibration methods and Bayesian methods, and present frequentist criteria, bias $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ and mean square error (MSE) $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (B(\hat{\theta}))^2$, to measure the performance of estimator accuracy and precision in the model.

Table 3.1: Bias and MSE (Simple linear models with measurement error)

Sample size of $n = 20$

parameter	σ_u^2	Naive	RC	Bayesian
$\beta_0 = 1$	0.25	0.0891(0.0235)	0.0386(0.0355)	0.0372(0.0196)
	0.49	0.1628(0.0450)	0.0665(0.1016)	0.0537(0.0203)
	1	0.2506(0.0801)	0.1341(0.4596)	0.0674(0.0298)
$\beta_1 = 0.5$	0.25	-0.0969(0.0183)	-0.0266(0.0231)	-0.0284(0.0099)
	0.49	-0.1608(0.0428)	-0.0760(0.0816)	-0.0393(0.0147)
	1	-0.2497(0.0696)	-0.1251(0.3570)	-0.0549(0.0161)

Sample size of $n = 100$

parameter	σ_u^2	Naive	RC	Bayesian
$\beta_0 = 1$	0.25	0.0987(0.0125)	0.0028(0.0042)	0.0473(0.0049)
	0.49	0.1674(0.0289)	0.0103(0.0061)	0.0792(0.0094)
	1	0.2510(0.0656)	0.0204(0.0187)	0.1101(0.0163)
$\beta_1 = 0.5$	0.25	-0.0984(0.0115)	-0.0009(0.0024)	-0.0493(0.0035)
	0.49	-0.1661(0.0283)	-0.0081(0.0040)	-0.0747(0.0079)
	1	-0.2515(0.0623)	-0.0183(0.0162)	-0.1012(0.0145)

From Table 3.1, we can realize that the naive estimates which do not account for measurement error are typically biased. The naive estimator of β_1 will be underestimated and as σ_u increases the attenuation of the slope estimator will get larger. For RC methods, the correction for measurement error is usually efficient based on the fact that estimates of β work well. As the sample size increases, both the bias and MSE of RC estimators are getting smaller, which indicates RC methods will accomplish a better adjustment for the bias when the size of sample is large enough. As to Bayesian methods, it seems like it performs better when the sample size is smaller. Compared with these three methods, both RC and Bayesian methods are better than naive methods which are struggle into the attenuation of slope. RC

methods achieve better estimates than Bayesian methods since RC estimators have relatively smaller bias and MSE in most cases. For example, when $n = 100$ and $\sigma_u^2 = 0.25$, the bias and MSE of $\hat{\beta}_1$ for RC methods $(-0.0009, 0.0024)$ are smaller than that of Bayesian methods $(-0.0493, 0.0035)$. However, the exception is that when the size of sample is small and σ_u is large, Bayesian methods may adjust the bias better than RC methods.

In general, naive estimates which do not account for measurement error are typically biased. In order to correct measurement error, we explore regression calibration estimates and Bayesian estimates. Although both methods achieve good performance on adjustment of the bias in simple linear models, the regression calibration method is preferred to be used due to its relatively small bias and MSE in most cases, and its simplicity. However, the Bayesian estimate will be an available alternative when the size of sample is small and the measurement error is large. Correcting for this bias entails what is usually referred to as a bias versus variance tradeoff. It means that the resulting corrected estimator will be more variable than the biased estimator. On the other hand, we need to adjust the standard error of the parameter estimates in the response model to reflect the uncertainty in the estimation of the covariate model using method such as the bootstrap methods or sandwich methods (Carroll et al, 2006).

Chapter 4

Linear Mixed Effects Models with Measurement Error

In this chapter, we are interested in the estimation of the parameters in a linear mixed model with measurement error in covariates. The general description of linear mixed effects model will be provided first. Both intercept-varying and intercept-slope varying models are considered. Simulation studies will be conducted to compare the performance of three approaches: naive methods, regression calibration methods and Bayesian methods. Similar with linear models with measurement error, the measurement error is non-differential and its variance is assumed to be known for identifiability problem. In section 4.1, both linear mixed effects (intercept-varying) model formation and linear mixed effects (intercept-slope varying) model formation are provided and the bias of naive estimators will be monitored. The methods for correcting measurement error including regression calibration methods and Bayesian methods are indicated in section 4.2. Simulation results and discoveries are presented in section 4.3.

4.1 Linear mixed effects models

A general linear mixed model implies that $\mathbf{Y}_i$ has a multi-normal distribution

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma^2\mathbf{I}) \quad i = 1, \cdots, G \tag{4.1}$$

which depends on the assumption that the random effects $\mathbf{b}_i$ and random errors $\boldsymbol{\epsilon}_i$ are linear while they are independent and normally distributed. Let $V(\boldsymbol{\gamma}) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma^2\mathbf{I}$ and $\boldsymbol{\gamma}$, called variance-covariance component parameters, denote the vector of all distinct parameters in the variance-covariance matrices $\mathbf{D}$ and $\sigma^2\mathbf{I}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ denote all parameters in the

linear mixed effects models. Then the likelihood of the observed data $\mathbf{Y}$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{Y}) &= \prod_{i=1}^{G} f(\mathbf{Y}_i|\boldsymbol{\theta}) = \prod_{i=1}^{G} f(\mathbf{Y}_i|\boldsymbol{\beta},\boldsymbol{\gamma}) \\ &= \prod_{i=1}^{G} \int f(\mathbf{Y}_i|\mathbf{b}_i,\boldsymbol{\beta},\sigma^2) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i \end{aligned} \tag{4.2}$$

where $f(\mathbf{Y}_i|\mathbf{b}_i,\boldsymbol{\beta},\sigma^2\mathbf{I}) = (2\pi)^{-\frac{n_i}{2}}|\sigma^2\mathbf{I}|^{-\frac{1}{2}}\exp[-(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i)'(\sigma^2\mathbf{I})^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i)]$ and $f(\mathbf{b}_i|\mathbf{D}) = (2\pi)^{-\frac{q}{2}}|\mathbf{D}|^{-\frac{1}{2}}\exp(-\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i)$. Given the variance-covariance parameters $\boldsymbol{\gamma}$, the value of $\boldsymbol{\beta}$ and σ^2 maximize (4.2) are

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{G}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{G}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i \tag{4.3}$$

$$\hat{\sigma}^2 = \frac{1}{G}\sum_{i=1}^{G}(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}})'\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{4.4}$$

Many estimation methods have been proposed over the years (Searle et al, 1992). Maximum likelihood method (ML) and restricted maximum likelihood method (REML) (Longford, 1993) are the most popular methods implemented to estimate the parameters. Compared with ML estimates, REML estimates take into account the estimation of the fixed effect while calculating the degrees of freedom associated with the variance-covariance component estimates. Thus, we prefer to utilize REML to estimate the variance-covariance component parameters. The estimation is usually achieved through maximizing the ML or REML based on numerical optimization. We are able to utilize a Bayesian perspective to estimate the random effects which reflect how much the subject-specific profiles deviate from the overall average profile. The posterior distribution of $\mathbf{b}_i$ given the data $\mathbf{Y}_i$ can be expressed as

$$f(\mathbf{b}_i|\mathbf{Y}_i) = \frac{f(\mathbf{Y}_i|\mathbf{b}_i)f(\mathbf{b}_i)}{\int f(\mathbf{Y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i} \tag{4.5}$$

We can utilize the posterior mean to estimate $\mathbf{b}_i$ as its posterior distribution is a multivariate normal distribution.

$$E(\mathbf{b}_i|\mathbf{Y}_i) = \int \mathbf{b}_i f(\mathbf{b}_i|\mathbf{Y}_i)d\mathbf{b}_i = \mathbf{D}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}) \tag{4.6}$$

After the unknown parameters $\boldsymbol{\theta}$ are substituted by their ML or REML estimates, the resulting estimates for the random effects $\mathbf{b}_i$ can be written as

$$\hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{D}}\mathbf{Z}_i'\hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{4.7}$$

4.1.1 Linear mixed effects (intercept-varying) models with measurement error

A linear mixed effects (intercept-varying) model with the additive classical measurement error can be expressed as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + a_i + \epsilon_{ij} \tag{4.8}$$

$$W_{ij} = X_{ij} + u_{ij} \tag{4.9}$$

where $i = 1, \cdots, G, j = 1, \cdots, n_i, n = \sum_{i=1}^{G} n_i$.

In matrix form, $\mathbf{Y}_i = \mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{Z}_i^*\mathbf{b}_i^* + \boldsymbol{\epsilon}_i$, where $\mathbf{Y}_i$ is $n_i \times 1$ random vector of outcomes, $\mathbf{X}_i^* = (\mathbf{1}, \mathbf{X}_i)$, $\mathbf{X}_i$ is $n_i \times 1$ vector of covariates subject to measurement error, $\boldsymbol{\beta}^* = (\beta_0, \beta_1)'$, $\mathbf{Z}_i^* = \mathbf{1}$ is $n_i \times 1$ constant vector, $\mathbf{b}_i^* = (a_i)$, and $\boldsymbol{\epsilon}_i$ is $n_i \times 1$ random vector with mean $\mathbf{0}$ and covariance $\sigma_\epsilon^2\mathbf{I}$.

Assume

$$(x_{ij}, \epsilon_{ij}, a_i, u_{ij})' \sim N_4\{(\mu_x, 0, 0, 0)'; diag(\sigma_x^2, \sigma_\epsilon^2, \sigma_a^2, \sigma_u^2)\}$$

Then

$$\begin{pmatrix} W_{ij} \\ Y_{ij} \end{pmatrix} \sim N_2\left\{\begin{pmatrix} \mu_x \\ \mu_y = \beta_0 + \beta_1\mu_x \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \sigma_{wy} \\ \sigma_{yw} & \sigma_y^2 \end{pmatrix}\right\} \tag{4.10}$$

where

$$\begin{pmatrix} \sigma_w^2 & \sigma_{wy} \\ \sigma_{yw} & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 + \sigma_u^2 & \beta_1\sigma_x^2 \\ \beta_1\sigma_x^2 & \beta_1^2\sigma_x^2 + \sigma_a^2 + \sigma_\epsilon^2 \end{pmatrix} \tag{4.11}$$

An induced model for $Y_{ij}|W_{ij}$ can be written as

$$Y_{ij}|W_{ij} \;=\; \beta_0 + \beta_1 E(X_{ij}|W_{ij}) + \epsilon_{ij}^* = \beta_0 + (1-\lambda)\beta_1\mu_x + \lambda\beta_1 W_{ij} + \epsilon_{ij}^* \tag{4.12}$$

where $\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$, $E(X_{ij}|W_{ij}) = (1-\lambda)\mu_x + \lambda W_{ij}$, ϵ^* has covariance $\sigma_\epsilon^2 + \sigma_a^2 + \beta_1^2\sigma_x^2(1-\lambda)$. It is obvious that the naive estimators of either β_1 or the variance parameters are usually biased. However, the bias can not be identified immediately since neither the fixed effects or the covariance has the same structure with the original model. As the fixed effects have the same form with the original model, $\beta_1 E(X_{ij}|W_{ij})$ needs to be written as $\beta_1^* W_i$, which is not always true. As the covariance structure is preserved, the bias in the naive estimators of the variance-covariance parameters can be identified. In general, the asymptotic biases for any of the naive estimators can also be examined through the estimating equations (Wang et al).

4.1.2 Linear mixed effects (intercept-slope varying) models with measurement error

A linear mixed effects (intercept-slope varying) model can be written as

$$\begin{aligned} Y_{ij} &= \alpha_i + \beta_i Z_{ij} + \beta_x X_{ij} + \epsilon_{ij} \\ &= \alpha_0 + a_i + (\beta_0 + b_i) Z_{ij} + \beta_x X_{ij} + \epsilon_{ij} \\ &= \alpha_0 + \beta_0 Z_{ij} + \beta_x X_{ij} + a_i + b_i Z_{ij} + \epsilon_{ij} \end{aligned} \tag{4.13}$$

where $i = 1, \cdots, G$, $j = 1, \cdots, n_i$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, X_{ij} is a predictor for individual i, $X_{ij} \sim N(\mu_x, \sigma_x^2)$, which is measured with error. Z_{ij} is another predictor without measurement error. α_0 is the fixed effect intercept term, $\alpha_i = \alpha_0 + a_i$ is the intercept for the ith individual, β_0 is the fixed effect slope term, $\beta_i = \beta_0 + b_i$ is the slope for the ith individual, and β_x is a fixed slope for the error-prone predictor $\mathbf{X}$. Consider the correlation between the varying-intercept and varying-slope exists, then

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \right\}$$

In matrix form, $\mathbf{Y}_i = \mathbf{X}_i^* \boldsymbol{\beta}^* + \mathbf{Z}_i^* \mathbf{b}_i^* + \boldsymbol{\epsilon}_i$, where $\mathbf{Y}_i$ is $n_i \times 1$ random vector of outcomes, $\mathbf{X}_i^* = (\mathbf{1}, \mathbf{Z}_i, \mathbf{X}_i)$, $\mathbf{X}_i$ is $n_i \times 1$ vector of covariates subject to measurement error, $\boldsymbol{\beta}^* =$

$(\alpha_0, \beta_0, \beta_x)'$, $\mathbf{Z}_i^* = (\mathbf{1}, \mathbf{Z}_i)$, $\mathbf{b}_i^* = (a_i, b_i)'$, and $\boldsymbol{\epsilon}_i$ is $n_i \times 1$ random vector with mean $\mathbf{0}$ and covariance $\sigma_\epsilon^2 \mathbf{I}$.

4.2 Correcting Methods

4.2.1 Regression Calibration Methods

Three main steps for the regression calibration methods are conducted in the linear mixed effects models, in order to address the measurement error in covariates: First, model and estimate the regression of $\mathbf{x}_i$ on $\mathbf{w}_i$; Second, replace $\mathbf{x}_i$ in the response model by its estimate $E(\mathbf{X}_i|\mathbf{W}_i)$ and perform a standard analysis on the appropriate response model; Third, adjust the resulting standard errors to account for the estimation of parameters in the first step based on the bootstrap or sandwich method. With the additive classical measurement error in covariates $W_{ij} = X_{ij} + u_{ij}$, u_{ij}s are measurement error in covariates and assumed identically independently distributed with mean 0 and variance σ_u^2, an estimate of the unobserved true covariate can be derived as

$$\begin{aligned} \hat{x}_{ij} &= \hat{E}(X_{ij}|W_{ij}) \\ &= \overline{W} + \frac{\hat{\sigma}_x^2}{\hat{\sigma}_w^2}(W_{ij} - \overline{W}) \\ &= (1 - \hat{\lambda})\overline{W} + \hat{\lambda} W_{ij} \end{aligned} \tag{4.14}$$

where $\hat{\lambda} = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_w^2} = \frac{\hat{\sigma}_w^2 - \sigma_u^2}{\hat{\sigma}_w^2}$, σ_u^2 is assumed to be known or can be estimated, $\overline{W} = \frac{1}{n}\sum_{i=1}^{G}\sum_{j=1}^{n_i} W_{ij}$. Then we substitute x_{ij} by $\hat{x}_{ij}$ in the linear mixed effects models and perform an analysis on the approximate response model.

4.2.2 Bayesian Methods

Intercept-varying model

For a linear mixed effects (intercept-varying) model with the additive classical measurement error

$$f(\mathbf{Y}|\mathbf{X},\mathbf{a},\boldsymbol{\theta}_1): \quad \text{response model with parameters} \quad \boldsymbol{\theta}_1=(\beta_0,\beta_1,\sigma_a^2,\sigma_\epsilon^2) \tag{4.15}$$

$$f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta}_2): \quad \text{measurement error model with parameters} \quad \boldsymbol{\theta}_2=(\sigma_u^2) \tag{4.16}$$

$$f(\mathbf{X}|\boldsymbol{\theta}_3): \quad \text{exposure model with parameters} \quad \boldsymbol{\theta}_3=(\mu_x,\sigma_x^2) \tag{4.17}$$

A key assumption is that the measurement error is non-differential

$$f(\mathbf{Y}|\mathbf{X},\mathbf{a},\mathbf{W};\boldsymbol{\theta}_1) \;=\; f(\mathbf{Y}|\mathbf{X},\mathbf{a};\boldsymbol{\theta}_1) \tag{4.18}$$

The joint distribution can be written as

$$\begin{aligned} f(\mathbf{Y},\mathbf{X},\mathbf{W},\mathbf{a},\boldsymbol{\theta}) &= f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_3) \\ &\times \prod_{i=1}^{G}\prod_{j=1}^{n_i} f(Y_{ij}|X_{ij},a_i,\boldsymbol{\theta}_1)\prod_{i=1}^{G}\prod_{j=1}^{n_i} f(W_{ij}|X_{ij},\boldsymbol{\theta}_2)\prod_{i=1}^{G}\prod_{j=1}^{n_i} f(X_{ij}|\boldsymbol{\theta}_3) \end{aligned} \tag{4.19}$$

We are interested in performing inference about unknown data $\mathbf{X}$ and unknown parameters $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3)$ by the posterior densities conditional on the observed data $(\mathbf{Y},\mathbf{W})$. The joint posterior densities of the unknown values in the linear mixed models can be written as

$$\begin{aligned} f(\mathbf{X},\mathbf{a},\boldsymbol{\theta}|\mathbf{Y},\mathbf{W}) &\propto f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_3) \\ &\times \prod_{i=1}^{G}\prod_{j=1}^{n_i} f(Y_{ij}|X_{ij},a_i,\boldsymbol{\theta}_1)\prod_{i=1}^{G}\prod_{j=1}^{n_i} f(W_{ij}|X_{ij},\boldsymbol{\theta}_2)\prod_{i=1}^{G}\prod_{j=1}^{n_i} f(X_{ij}|\boldsymbol{\theta}_3) \end{aligned} \tag{4.20}$$

The full conditional posterior distributions of unknown data $\mathbf{X}$ will be

$$\begin{aligned} f(X_{ij}|Y_{ij},W_{ij},\boldsymbol{\theta},a_i) &\propto f(Y_{ij}|X_{ij},a_i,\boldsymbol{\theta}_1)f(W_{ij}|X_{ij},\boldsymbol{\theta}_2)f(X_{ij}|\boldsymbol{\theta}_3) \\ &= f(Y_{ij}|X_{ij},\beta_0,\beta_1,a_i,\sigma_\epsilon^2)f(W_{ij}|X_{ij},\sigma_u^2)f(X_{ij}|\mu_x,\sigma_x^2) \end{aligned} \tag{4.21}$$

With normal assumption

$$\begin{aligned} f(X_{ij}|Y_{ij}, W_{ij}, \boldsymbol{\theta}, a_i) &\propto \exp\left(-\frac{1}{2\sigma_\epsilon^2}(Y_{ij} - \beta_0 - \beta_1 X_{ij} - a_i)^2\right) \\ &\times \exp\left(-\frac{1}{2\sigma_u^2}(W_{ij} - X_{ij})^2\right) \exp\left(-\frac{1}{2\sigma_x^2}(X_{ij} - \mu_x)^2\right) \end{aligned} \tag{4.22}$$

The full conditional posterior distributions of random effects a_i

$$f(a_i|\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i, \boldsymbol{\theta}) \propto f(\mathbf{Y}_i|\mathbf{X}_i, a_i, \boldsymbol{\theta}_1) f(a_i|\sigma_a^2) \tag{4.23}$$

With normal assumption

$$f(a_i|\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i, \boldsymbol{\theta}) \propto \prod_{j=1}^{n_i} \exp\left(-\frac{1}{2\sigma_\epsilon^2}(Y_{ij} - \beta_0 - \beta_1 X_{ij} - a_i)^2\right) \exp\left(-\frac{1}{2\sigma_a^2} a_i^2\right) \tag{4.24}$$

The full conditional posterior distribution of the unknown parameters $\boldsymbol{\theta}$

$$f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{a}) \propto f(\mathbf{Y}|\mathbf{X}, \mathbf{a}, \boldsymbol{\theta}_1) f(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta}_2) f(\mathbf{X}|\boldsymbol{\theta}_3) f(\boldsymbol{\theta}) \tag{4.25}$$

Similar with linear models with measurement error, we try to use the Gibbs sampler to generate samples from the posterior distribution to escape high dimensional integration. Given the initial values $\boldsymbol{\theta}^{(0)}$,$\mathbf{X}^{(0)}$,$\mathbf{a}^{(0)}$, the samples of the unobserved $\mathbf{X}$ can be generated from its full conditional posterior distribution given the observed data, $\boldsymbol{\theta}^{(current)}$, and $\mathbf{a}^{(current)}$. Then we draw the samples of the random effects $\mathbf{a}$ from its full conditional posterior distribution given the observed data, $\mathbf{X}^{(current)}$ generated from the last step, and $\boldsymbol{\theta}^{(current)}$. Finally, we produce the samples of $\boldsymbol{\theta}$ from its full conditional posterior distribution given the observed data, $\mathbf{X}^{(current)}$ and $\mathbf{a}^{(current)}$. Repeating the above steps many times, the resulting sequence converges to the target posterior distribution $f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{W})$.

Intercept-slope varying model

For a linear mixed effects model (intercept-slope varying) with the additive classical measurement error

$f(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}_1)$: response model with parameters $\boldsymbol{\theta}_1 = (\alpha_0, \beta_0, \beta_x, \sigma_\epsilon^2, \sigma_a^2, \sigma_b^2, \rho)$

(4.26)

$$f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta}_2): \quad \text{measurement error model with parameters} \quad \boldsymbol{\theta}_2 = (\sigma_u^2) \tag{4.27}$$

$$f(\mathbf{X}|\boldsymbol{\theta}_3): \quad \text{exposure model with parameters} \quad \boldsymbol{\theta}_3 = (\mu_x, \sigma_x^2) \tag{4.28}$$

The non-differential measurement error is also hold, which leads to

$$f(\mathbf{Y}|\mathbf{X},\mathbf{Z},\mathbf{a},\mathbf{b},\mathbf{W};\boldsymbol{\theta}_1) \;=\; f(\mathbf{Y}|\mathbf{X},\mathbf{Z},\mathbf{a},\mathbf{b};\boldsymbol{\theta}_1) \tag{4.29}$$

Then the joint distribution can be written as

$$\begin{aligned} f(\mathbf{Y},\mathbf{X},\mathbf{Z},\mathbf{W},\mathbf{a},\mathbf{b},\boldsymbol{\theta}) &= f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_3) \\ &\times f(\mathbf{Y}|\mathbf{X},\mathbf{Z},\mathbf{a},\mathbf{b},\boldsymbol{\theta}_1)f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta}_2)f(\mathbf{X}|\boldsymbol{\theta}_3) \end{aligned} \tag{4.30}$$

Similar with the materials in chapter 3, we focus on analyzing about unknown data $\mathbf{X}$ and unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ by the posterior densities conditional on the observed data $(\mathbf{Y}, \mathbf{W})$. Then the joint posterior densities of the unknown values in the linear mixed effects models can be expressed as

$$\begin{aligned} f(\mathbf{X},\mathbf{a},\mathbf{b},\boldsymbol{\theta}|\mathbf{Y},\mathbf{Z},\mathbf{W}) &\propto f(\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_3) \\ &\times f(\mathbf{Y}|\mathbf{X},\mathbf{Z},\mathbf{a},\mathbf{b},\boldsymbol{\theta}_1)f(\mathbf{W}|\mathbf{X},\boldsymbol{\theta}_2)f(\mathbf{X}|\boldsymbol{\theta}_3) \end{aligned} \tag{4.31}$$

The full conditional posterior distributions of unknown data $\mathbf{X}$ will be

$$\begin{aligned} f(X_{ij}|Y_{ij},Z_{ij},W_{ij},\boldsymbol{\theta},a_i,b_i) &\propto f(Y_{ij}|X_{ij},Z_{ij},a_i,b_i,\boldsymbol{\theta}_1)f(W_{ij}|X_{ij},\boldsymbol{\theta}_2)f(X_{ij}|\boldsymbol{\theta}_3) \\ &= f(Y_{ij}|X_{ij},Z_{ij},\alpha_0,\beta_0,\beta_x,a_i,b_i,\sigma_\epsilon^2)f(W_{ij}|X_{ij},\sigma_u^2) \\ &\times f(X_{ij}|\mu_x,\sigma_x^2) \end{aligned} \tag{4.32}$$

With normal assumption

$$\begin{aligned} f(X_{ij}|Y_{ij},W_{ij},Z_{ij},\boldsymbol{\theta},a_i,b_i) &\propto \exp\left(-\frac{1}{2\sigma_\epsilon^2}(Y_{ij}-\alpha_0-\beta_0 Z_{ij}-\beta_x X_{ij}-a_i-b_i Z_{ij})^2\right) \\ &\times \exp\left(-\frac{1}{2\sigma_u^2}(W_{ij}-X_{ij})^2\right)\exp\left(-\frac{1}{2\sigma_x^2}(X_{ij}-\mu_x)^2\right) \end{aligned} \tag{4.33}$$

The full conditional posterior distributions of random effects a_i, b_i will be

$$\begin{aligned} f(a_i, b_i|\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}) &\propto f(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, a_i, b_i, \boldsymbol{\theta}_1) f(a_i, b_i|\sigma_a^2, \sigma_b^2, \rho) \\ &= \prod_{j=1}^{n_i} \exp\left(-\frac{1}{2\sigma_\epsilon^2}(Y_{ij} - \alpha_0 - \beta_0 Z_{ij} - \beta_x X_{ij} - a_i - b_i Z_{ij})^2\right) \\ &\times \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{a_i^2}{\sigma_a^2} - 2\rho\left(\frac{a_i}{\sigma_a}\right)\left(\frac{b_i}{\sigma_b}\right) + \frac{b_i^2}{\sigma_b^2}\right]\right) \end{aligned} \tag{4.34}$$

The full conditional posterior distributions of the unknown parameters $\boldsymbol{\theta}$ can be written as

$$f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{a}, \mathbf{b}) \propto f(\mathbf{Y}|\mathbf{X}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}_1) f(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta}_2) f(\mathbf{X}|\boldsymbol{\theta}_3) f(\boldsymbol{\theta}) \tag{4.35}$$

Then we use the Gibbs samplers to draw samples from the posterior distributions to avoid the intractable integrals. Given the initial values $\boldsymbol{\theta}^{(0)}$, $\mathbf{X}^{(0)}$, $\mathbf{a}^{(0)}$, $\mathbf{b}^{(0)}$, we can generate the samples of the unobserved $\mathbf{X}$ from its full conditional posterior distribution given the observed data, $\boldsymbol{\theta}^{(current)}$, $\mathbf{a}^{(current)}$, and $\mathbf{b}^{(current)}$. Then the samples of the random effects $\mathbf{a}$ and $\mathbf{b}$ can be produced from their full conditional posterior distribution given the observed data, $\mathbf{X}^{(current)}$ generated from the last step and $\boldsymbol{\theta}^{(current)}$. The last step is generating the samples of $\boldsymbol{\theta}$ from their full conditional posterior distributions given the observed data, $\mathbf{X}^{(current)}$, $\mathbf{a}^{(current)}$, $\mathbf{b}^{(current)}$. We obtain the resulting sequence after repeating the above steps many times.

4.3 Simulation Studies

In this section, we conduct simulation studies for a linear mixed effects (intercept-varying) model with measurement error and a linear mixed effects (intercept-slope varying) model with measurement error. The performance of naive estimations, regression calibration estimations and Bayesian estimations are contrasted depending on the simulation results.

4.3.1 Intercept-varying model

The total number of measurements is 42 and 84 in $G = 7$ and 14 individuals with 6 replicates. As mixed effects models typically have so many parameters and it is not feasible to examine all estimates, we focus on numerical summaries for $\boldsymbol{\theta} = (\beta_0, \beta_1)$ with different measurement error variance ($\sigma_u^2 = 0.25, 0.49, 0.81$). The number of iterations is $20,000$ with $5,000$ burn-in, 5 thin in each of the 3 chains. We simulated 100 sets of data for naive methods, regression calibration methods and Bayesian methods.

Table 4.1: Bias and MSE (Linear mixed effects (varying-intercept) models with measurement error)

$G = 7$ and $n_i = 6$

parameter	σ_u^2	Naive	RC	Bayesian
$\beta_0 = 5$	0.25	0.1996(0.1018)	0.0096(0.0547)	0.0184(0.0492)
	0.49	0.3136(0.1629)	0.0219(0.1129)	0.0587(0.0608)
	0.81	0.4327(0.2598)	0.0753(0.1755)	0.1139(0.0756)
$\beta_1 = 1$	0.25	-0.1978(0.0517)	-0.0127(0.0154)	-0.0287(0.0094)
	0.49	-0.3308(0.1262)	-0.0242(0.0528)	-0.0469(0.0136)
	0.81	-0.4390(0.1986)	-0.0695(0.1211)	-0.0889(0.0243)

$G = 14$ and $n_i = 6$

parameter	σ_u^2	Naive	RC	Bayesian
$\beta_0 = 5$	0.25	0.2108(0.0582)	0.0034(0.0203)	0.0431(0.0135)
	0.49	0.3289(0.1395)	0.0131(0.0478)	0.0751(0.0293)
	0.81	0.4472(0.2464)	0.0289(0.0760)	0.1408(0.0599)
$\beta_1 = 1$	0.25	-0.1983(0.0467)	-0.0043(0.0045)	-0.0323(0.0034)
	0.49	-0.3436(0.1138)	-0.0107(0.0142)	-0.0517(0.0081)
	0.81	-0.4525(0.1904)	-0.0323(0.0331)	-0.1247(0.0222)

We present the bias and MSE in Table 4.1 for the linear mixed effects (intercept-varying) models. Evidently, the results indicate that substantial bias is incurred if measurement errors are not properly treated. Specially, the naive estimator of β_1 is underestimated and as σ_u increases the attenuation of the slope estimator will get larger. On the other hand, we can find that the bias of the slope estimates will be larger as the size of group increases under

the condition of the fixed σ_u. For regression calibration methods, the adjustment of bias is usually available in most cases. Especially with the same number of observations in each group, when the number of groups increases, both the bias and MSE of regression calibration estimators change to smaller. Which indicates that regression calibration methods achieve good performance on the correction of the measurement error when the group size is large enough under the condition that the number of observations in each group is fixed. Bayesian methods also present significant improvement on the estimation especially when the group size is small (with the same number of observations in each group) and σ_u is large. Compare with these three methods for a linear mixed effects (intercept varying) model, both regression calibration methods and Bayesian methods work better than naive methods which ignores the measurement error. Regression calibration methods work better than Bayesian methods.

4.3.2 Intercept-slope varying model

We conduct a simulation study to evaluate and compare the performance of naive methods, regression calibration methods and Bayesian methods in linear mixed effects (intercept-slope varying) models. In order to measure the performance of estimator accuracy and precision, we will show the frequentist criteria, bias and MSE. The total number of observations is 42 and 84 in $G = 7$ and 14 individuals with 6 replicates. It is difficult for us to monitor all estimate as a lot of parameters exist in a linear mixed effects model. We are interested in numerical summaries for some parameters $\boldsymbol{\theta} = (\alpha_0, \beta_0, \beta_x)$ with different measurement error variance ($\sigma_u^2 = 0.25, 0.49, 0.81$). In regression model, we choose a multivariate normal distribution as a prior distribution for the mean parameters $\boldsymbol{\beta}$, that is, typically $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_0$ are hyper-parameters. For the non-informative priors, $\boldsymbol{\beta} \sim \text{Uniform}(-\infty, \infty)$ or $\boldsymbol{\Sigma}_0^{-1} = 0$ will be chosen. As to the variance-covariance matrix for the intercepts and slopes, the scaled Wishart distribution is selected as prior distribution. The number of iterations is $50,000$ with $10,000$ burn-in, 5 thin in each of the 3 chains. We simulated 100 sets of data for naive methods, regression calibration methods and Bayesian

methods.

Table 4.2: Bias and MSE (Linear mixed effects (intercept-slope varying) models with measurement error)

$G = 7$ and $n_i = 6$

parameter	σ_u^2	Naive	RC	Bayesian
$\alpha_0 = 5$	0.25	0.1121(0.0676)	0.0091(0.0382)	0.0194(0.0319)
	0.49	0.1573(0.0841)	0.0189(0.0522)	0.0321(0.0513)
	0.81	0.2037(0.1031)	0.0401(0.0819)	0.0611(0.0627)
$\beta_0 = 1$	0.25	-0.0143(0.0117)	-0.0143(0.0117)	-0.0132(0.0112)
	0.49	-0.0199(0.0108)	-0.0199(0.0108)	-0.0196(0.0107)
	0.81	-0.0373(0.0086)	-0.0373(0.0086)	-0.0359(0.0087)
$\beta_x = 0.5$	0.25	-0.1038(0.0219)	-0.0089(0.0107)	-0.0167(0.0065)
	0.49	-0.1807(0.0583)	-0.0172(0.0263)	-0.0294(0.0082)
	0.81	-0.2081(0.0731)	-0.0385(0.0516)	-0.0450(0.0121)

$G = 14$ and $n_i = 6$

parameter	σ_u^2	Naive	RC	Bayesian
$\alpha_0 = 5$	0.25	0.0926(0.0381)	0.0045(0.0240)	0.0212(0.0093)
	0.49	0.1660(0.0574)	0.0147(0.0271)	0.0423(0.0185)
	0.81	0.2061(0.0719)	0.0374(0.0308)	0.0769(0.0243)
$\beta_0 = 1$	0.25	-0.0099(0.0081)	-0.0099(0.0081)	-0.0106(0.0081)
	0.49	-0.0101(0.0053)	-0.0101(0.0053)	-0.0109(0.0052)
	0.81	-0.0120(0.0041)	-0.0120(0.0041)	-0.0127(0.0041)
$\beta_x = 0.5$	0.25	-0.1015(0.0177)	-0.0019(0.0014)	-0.0181(0.0037)
	0.49	-0.1797(0.0443)	-0.0093(0.0032)	-0.0376(0.0060)
	0.81	-0.2126(0.0675)	-0.0128(0.0042)	-0.0631(0.0113)

We provides the bias and MSE in Table 4.2 for the linear mixed effects (intercept-slope varying) models. Similar with the results in the linear mixed effects (intercept varying) models, the ignorance of measurement error has a significant effect on the estimations. For naive methods, the estimator of β_x is underestimated and the bias of the estimates will get larger as σ_u increases. Regression calibration methods still play well in the linear mixed effects (intercept-slope varying) model especially when the size of group is large (with the same number of observations in each group). Bayesian methods also have good performance on the correction of measurement error especially when the group size is small (with the

same number of observations in each group) and σ_u is large enough. Besides, we are able to realize that these three methods obtain similar estimations on β_0 which is the fixed coefficient of the covariates without measurement error. Finally, both regression calibration methods and Bayesian methods accomplish better adjustment for the bias in the linear mixed effects models (intercept-slope varying) than naive methods. Usually regression calibration methods are better than Bayesian methods except the case that the group size is small and σ_u is large, in terms of both bias and MSE.

4.3.3 Conclusion

The model used in this chapter is different from what we have in Chapter 3. We added some random effect into the model which accommodates the variation that existed between the groups (individuals). In fact, what we observe in this chapter is almost the same with Chapter 3 to the attributes demonstrated by both regression calibration methods and Bayesian methods. When dealing with linear mixed effects models with measurement error in covariates, we would recommend regression calibration methods over Bayesian methods and naive methods. The regression calibration method has been described by many literatures as a less complicated and more intuitive method for estimating the parameters in a linear mixed effects model with measurement error, which we realized in this chapter.

Chapter 5

Data Analysis

In this chapter, we first provide a brief background description of the beta-carotene data which obtained from an open web-site:

`http://www.math.umass.edu/~johnpb/`

and a more detailed description of this data set can be found in Demidenko, Tosteson and Buonaccorsi (2000). Our target is to investigate the relationship between true dietary intakes of beta-carotene and serum beta-carotene. A major challenge in statistical analysis of the beta-carotene data is that the measures of beta-carotene intake are conducted based on a food frequency questionnaire (FFQ), which leads to the measurement error problems. We attempt to fit the data allowing measurement error with a linear mixed effects model and apply two correcting methods (regression calibration methods and Bayesian methods) we discussed in chapter 4 as an example. More reliable and valid data analysis can be found in other statistical articles (Demidenko, Tosteson and Buonaccorsi, 2000). In section 5.1, a simple description of beta-carotene data is given. The construction of the model for beta-carotene data will be discussed in section 5.2. We try to utilize both regression calibration methods and Bayesian methods to correct measurement error in section 5.3. The results and conclusion will be summarized in section 5.4.

5.1 Data

The beta-carotene data contains seven variables and we are interested in four of them: individual id, days of measurements, serum beta-carotene measurements and measures of beta-carotene intake based on a food frequency questionnaire. Both measures of serum beta-carotene and measures of beta-carotene intake from a food frequency questionnaire

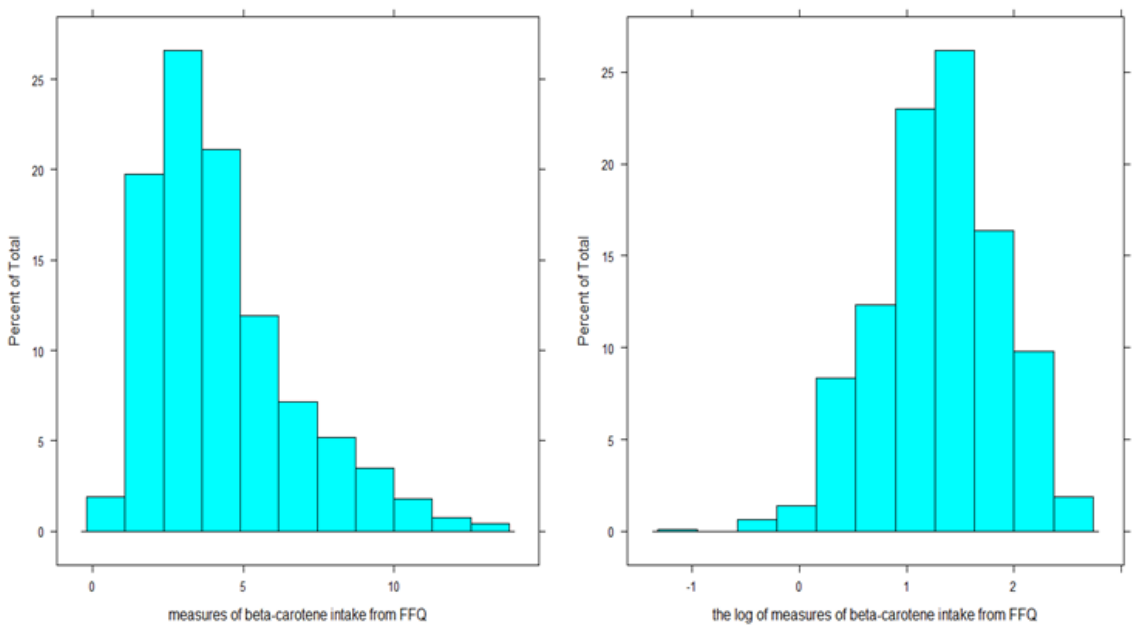

Figure 5.1: beta-carotene intake from FFQ

were included on days, 1, 2, 3, 4, 5, 6. There are 158 individuals, each having six serum measures of beta-carotene and six measures of beta-carotene intake based on a food frequency questionnaire (FFQ).

Figure 5.1 presents a histogram of measures of beta-carotene intake from FFQ which shows that the data is right-skewed distributed. In order to satisfy the normal assumption, we will do a log transformation for the measures of beta-carotene intake from FFQ. The figure shows that the log of measures of beta-carotene intake from FFQ is fairly normally distributed. Similar with the measures of beta-carotene intake from FFQ, the serum measures of beta-carotene is also right-skewed distributed in Figure 5.2. The figure demonstrates that the log of serum measures of beta-carotene is normally distributed.

Figure 5.3 presents a simple linear regression plot of the log of the serum measures of beta-carotene versus the log of measures of beta-carotene intake from FFQ for each individual and day. This plot is used to check the relationship between the two variables based on individuals and shows that the data is approximately linearly related for most of the individuals. We consider a linear mixed effects model to incorporate between-individual variation

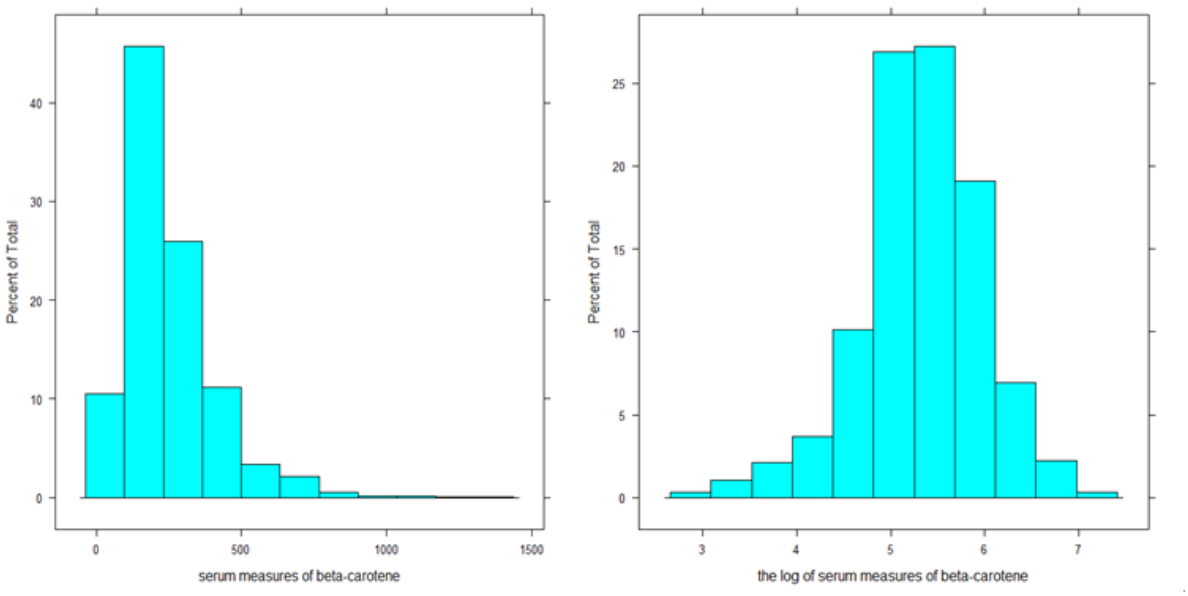

Figure 5.2: serum beta-carotene

and within-individual variation, assuming the covariate values change smoothly over time. The individual plot is provided to display the relationship or the likely differences between individuals and it is suspected that both individual intercept and slope are different, which motivates us to consider a varying coefficient. We proceed to test which model is appropriate out of intercept varying or intercept-slope varying and the test result shows that intercept-slope varying model is the proper option resulting from the fact that it has a relatively small AIC (1009.034 vs. 1016.910). Therefore, we decide to work with the linear mixed effects (intercept-slope varying) model. Figure 5.4 is the plot of the standardized residual against the fitted values and individual id, which indicates that the model fits the data reasonably well.

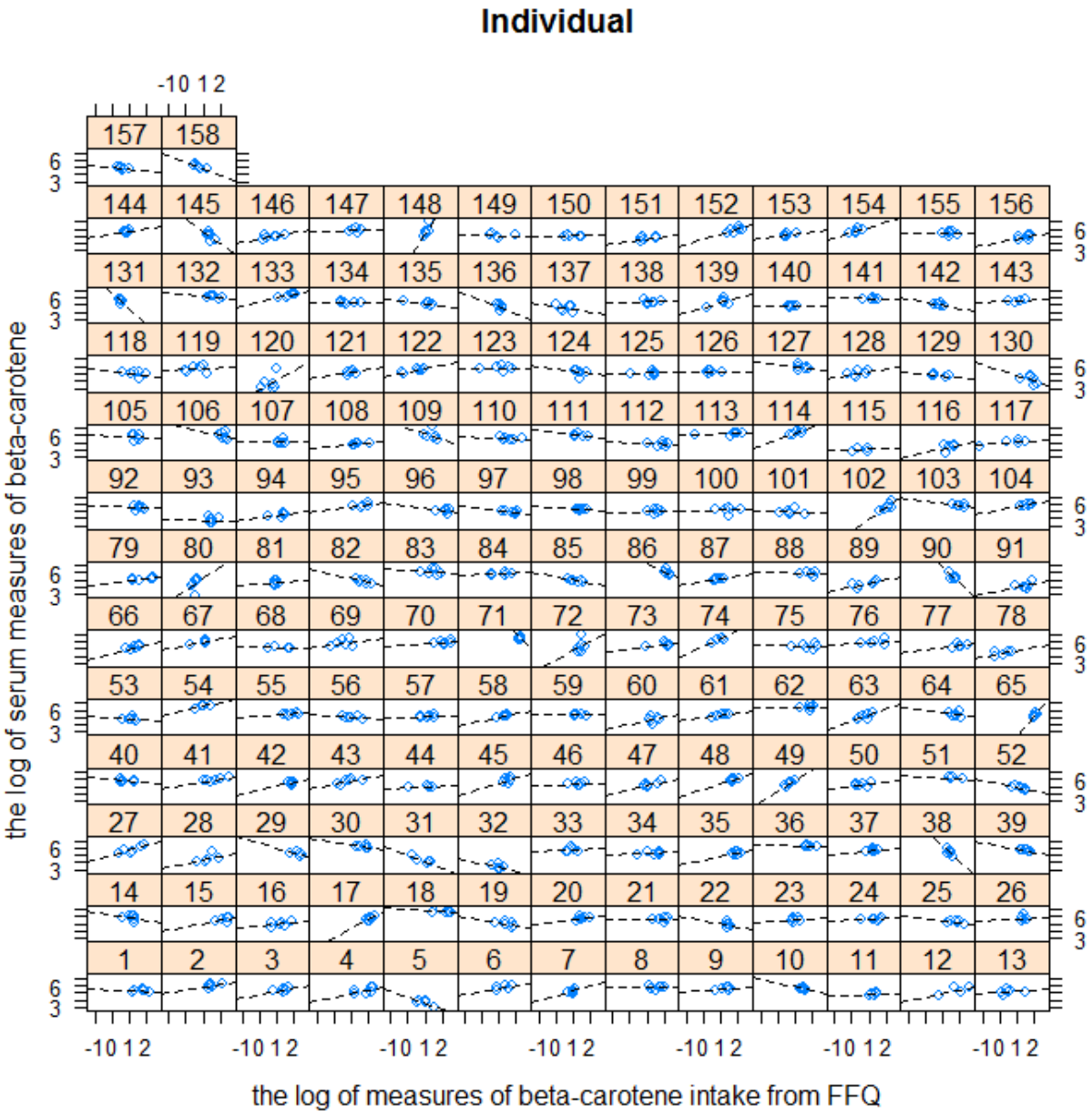

Figure 5.3: log of serum beta-carotene vs log of beta-carotene from FFQ

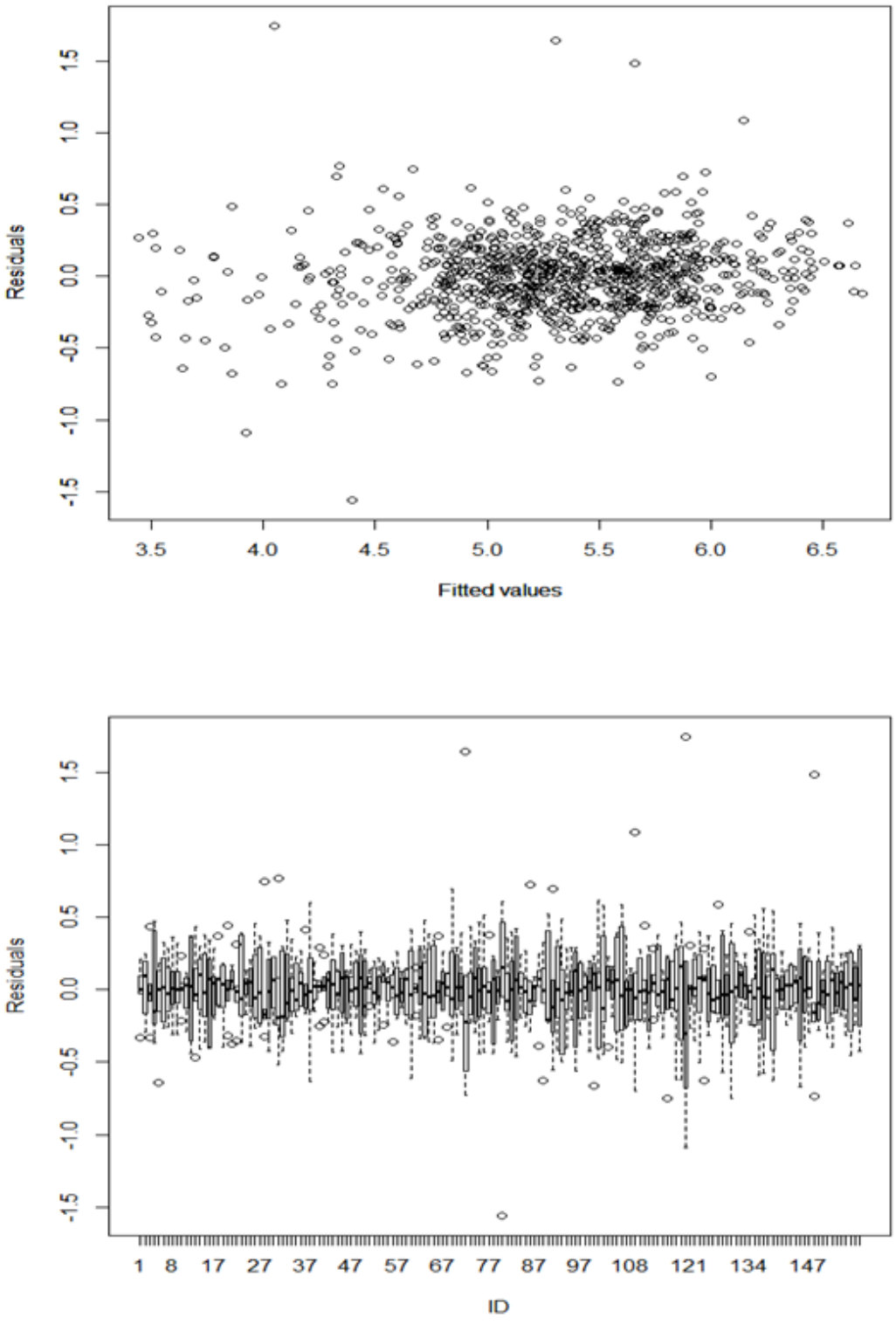

Figure 5.4: plot of standardized residual

5.2 Model

Consider a intercept-slope varying linear mixed effects model for the beta-carotene data

$$Y_{ij} = \alpha_0 + a_i + (\beta_0 + b_i)Z_{ij} + \beta_x X_{ij} + \epsilon_{ij} \quad i = 1, \cdots, 158, j = 1, \cdots, 6. \tag{5.1}$$

where Y_{ij} is the log of the serum measures of beta-carotene, X_{ij} is the log of measures of true diet intake of beta-carotene. $Z_{ij} = j - 1$ is a time variable (from 0 to 5), a_i represents a random subject effect while b_i allows a random time trend in the serum level after conditioning on dietary intake. ϵ_{ij}s are assumed to have mean 0 and variance σ^2_ϵ. $\boldsymbol{\epsilon}_i$ are assumed to be independent with a_i and b_i, $Cov(a_i, b_i) = \boldsymbol{\Sigma}$. The true intake $\mathbf{X}$ is unobserved and there is no replication in a year. An additive classical measurement error model can be written as $\mathbf{W}_i = \mathbf{X}_i + \mathbf{u}_i$, where $\mathbf{W}_i$ is the measured beta-carotene intakes based on a food frequency questionnaire. We can treat the repeated measurements of beta-carotene intake from FFQ over time as "replicates" and fit an empirical covariates mixed effects model to observe measures of true beta-carotene intake in the longitudinal data model by setting $\mathbf{M}_i = \mathbf{N}_i = \mathbf{R}$ in (2.16)

$$\mathbf{X}_i = \mathbf{R}\boldsymbol{\eta} + \mathbf{R}\boldsymbol{\delta}_i \tag{5.2}$$

$$\mathbf{R}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \end{pmatrix} \tag{5.3}$$

which is constructed and proved by Tosterson et al. (1998) and Buonaccorsi et al. (2000). Tosteson et al. (1998) present some justification for this being a reasonable approximation in this example. Where $\boldsymbol{\eta}$ consisting of unknown fixed parameters, $\boldsymbol{\delta}_i$ are random effects with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Omega}_\delta$, $\boldsymbol{\Sigma}_X = \mathbf{R}\boldsymbol{\Omega}_\delta\mathbf{R}'$. $\mathbf{R}\boldsymbol{\eta}$ presents the fixed part of the model while $\mathbf{R}\boldsymbol{\delta}_i$ captures the random effects part. Then the additive classical measurement error can be written as

$$\mathbf{W}_i = \mathbf{R}\boldsymbol{\eta} + \mathbf{R}\boldsymbol{\delta}_i + \mathbf{u}_i \tag{5.4}$$

5.3 Correcting methods for measurement error

Estimation based on regression calibration methods

The regression calibration methods are conducted to fit $Y_{ij} = \alpha_0 + \beta_0 Z_{ij} + \beta_x \hat{x}_{ij} + \epsilon^*_{ij}$ as a linear mixed effects model. First, we attempt to fit the covariates linear mixed effects models to obtain an estimate of the unobserved true covariates $\hat{x}_{ij} = \hat{E}(X_{ij}|\hat{\eta}, \hat{\mathbf{\Omega}}_\delta, \hat{\sigma}^2_u)$. Second, we replace the x_{ij} in the linear mixed effects models by $\hat{x}_{ij}$ and perform a standard analysis on the approximate response model. After fitting the covariates linear mixed effects models, we obtain

$$\hat{\boldsymbol{\eta}} = \begin{pmatrix} 1.2589 \\ 0.0146 \end{pmatrix} \quad \hat{\mathbf{\Omega}}_\delta = \begin{pmatrix} 0.2251 & -0.0115 \\ -0.0115 & 0.0041 \end{pmatrix} \quad \hat{\sigma}^2_u = 0.1191 \tag{5.5}$$

Figure 5.5 shows a simple linear regression plot of the log of serum measures of beta-carotene versus the estimated true beta-carotene intake for each individual and day based on regression calibration methods.

Estimation based on Bayesian methods

Based on the materials of linear mixed effects (intercept-slope varying) models with measurement error using Bayesian methods in section 4.2.2, the three sub-models can be written as

$$f(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{a}, \mathbf{b}, \boldsymbol{\theta}_1): \quad \text{response model with parameters} \quad \boldsymbol{\theta}_1 = (\alpha_0, \beta_0, \beta_x, \sigma^2_\epsilon, \sigma^2_a, \sigma^2_b, \rho) \tag{5.6}$$

$$f(\mathbf{W}|\mathbf{X}, \boldsymbol{\theta}_2): \quad \text{measurement error model with parameters} \quad \boldsymbol{\theta}_2 = (\sigma^2_u) \tag{5.7}$$

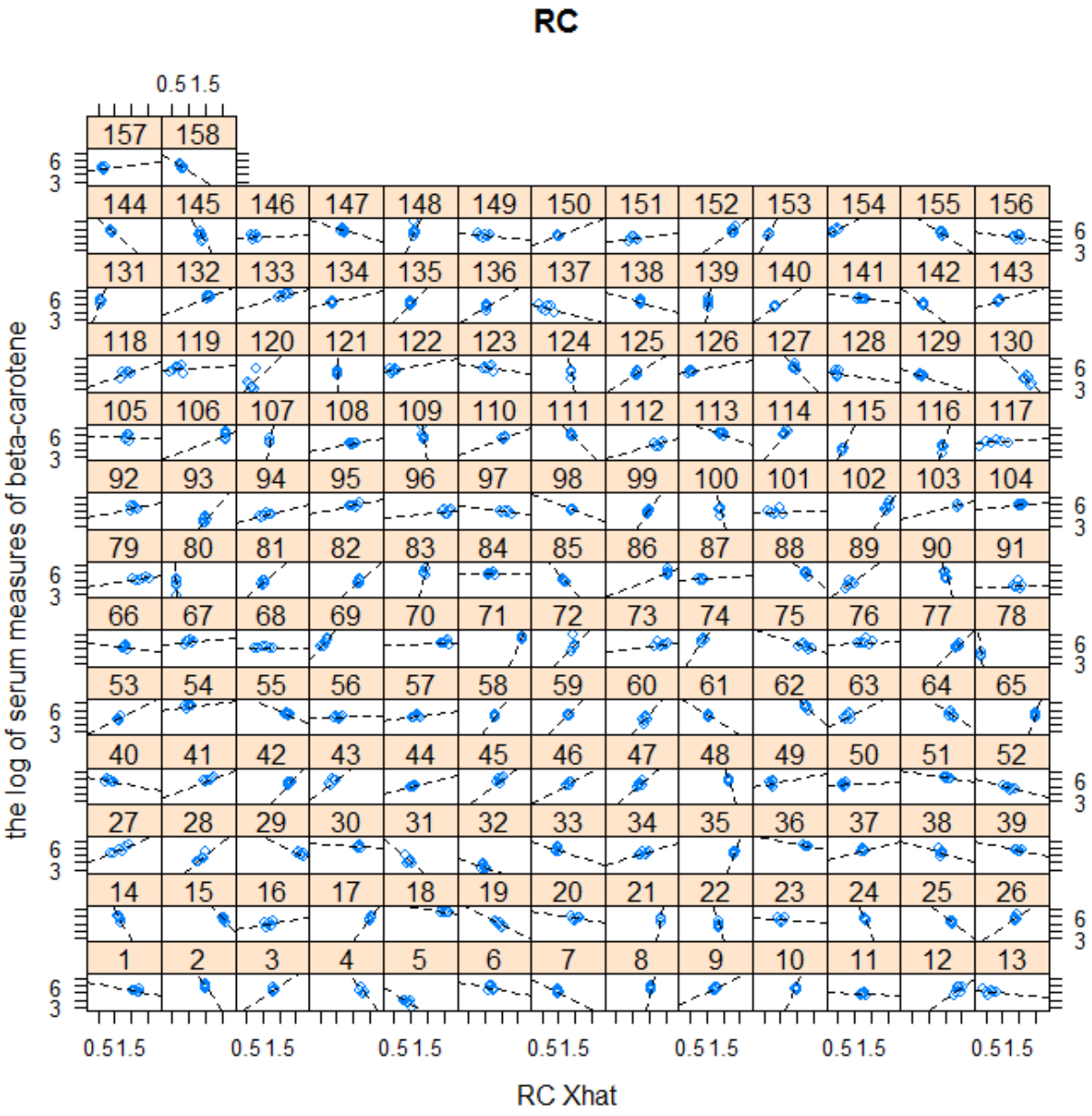

Figure 5.5: log of serum beta-carotene vs RC estimate of true beta-carotene intake

$$f(\mathbf{X}|\boldsymbol{\theta}_3): \quad \text{exposure model with parameters} \quad \boldsymbol{\theta}_3 = (\mu_x, \sigma_x^2) \tag{5.8}$$

A multivariate non-informative normal distribution is chosen as a prior distribution for the mean parameters and the scaled Wishart distribution is selected as the prior distribution for the variance-covariance matrix of the intercepts and slopes. The number of iterations is $50,000$ with $20,000$ burn-in, 5 thin in each of the 3 chains. $\hat{\sigma}_u^2 = 0.1191$ obtained from regression calibration methods is used in the program.

5.4 Results and Conclusion

In order to study the impact of measurement error in covariates on the relationship between the measures of true beta-carotene intake and the serum measures of beta-carotene in the beta-carotene trial, we attempt to compare the performance of regression calibration methods (RC) and Bayesian methods in a linear mixed effects (intercept-slope varying) model. Table 5.1 shows the naive, RC and Bayesian estimates of $\hat{\alpha}_0$, $\hat{\beta}_0$, and $\hat{\beta}_x$ with standard errors. We conclude that the most dramatic effect of the correction for the measurement error is about the estimate of β_x which changes from 0.1502 in the naive estimation to 0.4398 by using regression calibration methods which correct the measurement error. Our simulation study indicates that the estimation by the regression calibration methods is more reliable if the data has a large number of groups (individuals). Therefore, we suggest to utilize $\hat{\beta}_x = 0.4398$ as the best estimate of β_x in this example. On the other hand, we realize that the Bayesian method also accomplishes good performance on the correction for measurement error although it is not as good as the regression calibration method.

Table 5.1: Estimates with standard error

	Naive	RC	Bayesian
$\hat{\alpha}_0$	5.1043(0.0611)	4.7398(0.1232)	4.8713(0.0803)
$\hat{\beta}_0$	0.0088(0.0071)	0.0045(0.0072)	0.0086(0.0071)
$\hat{\beta}_x$	0.1502(0.0297)	0.4398(0.0704)	0.3763(0.0415)

Chapter 6

Conclusion and Future Research

Measurement errors in variables are common problems in practice, where it is usually difficult for us to measure variables accurately. If the observed data is measured with errors but treated as true values, that is, the measurement errors are not taken into account, statistical inference will be biased and misleading. Therefore, it has significance for us to address measurement errors in order to obtain valid statistical analysis. The primary goal of this thesis is to utilize regression calibration methods and Bayesian methods to correct measurement errors in covariates in both linear models and linear mixed effects models. Based on our study, we can realize that regression calibration methods tend to be more powerful for regression models in most cases and produce approximately unbiased estimates of the main parameters in the response model. However, its major drawback is that it fails to incorporate the uncertainty in the estimation of the true covariates in the first step, so that the standard error of the main parameter estimates may be under-estimated (Wu, 2010). The Bayesian methods also accomplish good adjustment for the bias, especially when the sample size is small and the measurement error is relatively large, comparing with regression calibration methods. So it can be chosen as an alternative method in some cases. But the major challenge for us to use Bayesian methods is the computer programming problem and we are not able to prove the results by mathematical derivations.

In general, we can conclude that: First, it will be very misleading if we ignore measurement error and analyze the data as if the values were all correctly measured; Second, we can not reduce biases caused by measurement error through increasing sample size; Third, implementation of measurement error correcting methods requires computer programming; Finally, the regression calibration method is more powerful and simpler than the Bayesian

method for correcting measurement error in the covariates of linear models and linear mixed effects models.

There are still a lot of work related with this thesis need to be done in the future. For example, we attempt to test the effect of the between-group variability on the estimation of main parameters in the linear mixed effects models, and explore the reason why the estimation of variance components is biased in measurement error models. On the other hand, the common feature of methods examined is applicable to problems where only a single covariate is measured with error and the other covariates are considered as error-free values. As the dimension of $\mathbf{X}$ increases, the extension of most of the procedures will not be straightforward and their applications may be less attractive. The research for extension of the existing methods to higher dimensions of unobserved covariates is required so as to make them appropriate for more realistic issues.

In conclusion, the occurrence of measurement error has significant effects on the estimation. Thus, it is important for us to be aware of the existence of measurement error and their potential impacts in practice.

Bibliography

[1] Buonaccorsi, J., Demidenko, E. and Tosteson, T. (2000). Estimation in longitudinal random effects models with measurement error. *Statistica Sinica*, 10, 885 – 903.

[2] Buonaccorsi, J. P. (2010). *Measurement Errors: Models, Methods and Applications*. Chapman and Hall.

[3] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Craniceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. New York: Chapman and Hall.

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22 – 37). *Journal of the Royal Statistical Society, Ser. B*, 39, 1 – 22.

[5] Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

[6] Geman, S. and Geman, D. (1984). Stochatic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721 – 741.

[7] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC.

[8] Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman and Hall/CRC Press.

[9] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383 – 385.

[10] Hoffmann-Jorgensen, J. (1994). *Probability with a View Towards Statistics*, Vol.1, Probability Series. New York: Chapman and Hall.

[11] Jennrich, R. I. and Schuchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805 − 820.

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

[13] Lehmann, E. L. (1983). *Theory of Point Estimaiton*. New York: Wiley.

[14] Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014 − 1022.

[15] Longford, N. T. (1993).*Random Coefficient Models*. New York: Oxford Univerisity Press.

[16] Miller, J. J. (1977). Asympototic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann. of Statistics*, 5, 746 − 762.

[17] Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New York: Wiley.

[18] Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.

[19] Thisted, R. A. (1988). *Elements of Statistical Computing*. London: Chapman and Hall.

[20] Tosteson, T., Buonaccorsi, J. P. 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.

[21] Wang, N., and Davidian, M. (1996). A Note on Covariate Measurement Error in Nonlinear Mixed Effects Models. *Biometrika*, 83, 801 − 812.

[22] Wang, W. and Heckman, N. (2009). Identifiability in linear mixed models. *Technical report*, Department of Statistics, Univerisity of British Columbia.

[23] Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman and Hall/CRC Press.