Model and Inference Issues Related to Exposure-Disease Relationships

by

Li Xing

MSc, Simon Fraser University, 2007
MSc, University of British Columbia, 2005
BSc, Hebei University of Technology, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University Of British Columbia
(Vancouver)

August 2014

© Li Xing, 2014
Abstract

The goal of my thesis is to make contributions on some statistical issues related to epidemiological investigations of exposure-disease relationships.

Firstly, when the exposure data contain missing values and measurement errors, we build a Bayesian hierarchical model for relating disease to a potentially harmful exposure while accommodating these flaws. The traditional imputation method, called the group-based exposure assessment method, uses the group exposure mean to impute the individual exposure in that group, where the group indicator indicates that the exposure levels tend to vary more across groups and less within groups. We compare our method with the traditional method through simulation studies, a real data application, and theoretical calculation. We focus on cohort studies where a logistic disease model is appropriate and where group exposure means can be treated as fixed effects. The results show a variety of advantages of the fully Bayesian approach, and provide recommendations on situations where the traditional method may not be suitable to use.

Secondly, we investigate a number of issues surrounding inference and the shape of the exposure-disease relationship. Presuming that the relationship can be expressed in terms of regression coefficients and a shape parameter, we investigate how well the shape can be inferred in settings which might typify epidemiologic investigations and risk assessment. We also consider a suitable definition of the average effect of exposure, and investigate how precisely this can be inferred. We also examine the extent to which exposure measurement error distorts inference about the shape of the exposure-disease relationship. All these investigations require a family of exposure-disease relationships indexed by a shape parameter. For this purpose, we employ a family based on the Box-Cox transformation.
Thirdly, matching is commonly used to reduce confounding due to lack of randomization in the experimental design. However, ignoring measurement errors in matching variables will introduce systematically biased matching results. Therefore, we recommend to fit a trajectory model to the observed covariate and then use the estimated true values from the model to do the matching. In this way, we can improve the quality of matching in most cases.
Preface

This thesis was completed under the supervision of Professor Paul Gustafson.

Chapter 2 of this thesis is based on the paper “A Comparison of Bayesian Hierarchical Modeling with Group-based Exposure Assessment in Occupational Epidemiology.” *Statistics in Medicine*, Xing L, Burstyn I, Richardson DB, Gustafson P, 2013 [56]. This research problem was identified and designed by my thesis advisor Professor Paul Gustafson and by our collaborator Professor Igor Burstyn. As the first author, I proposed the model, developed the algorithm, conducted the mathematical calculation, and performed the empirical analysis and simulation studies. I drafted the manuscript; my supervisor and other coauthors helped me to revise it. Professor David Richardson provided the data from the Savannah River Site (SRS) Study. The use of the SRS data has been approved by UBC Research Ethics Board, under the project title, SRSBM, and the Certificate Number of the Ethics Certificate obtained, H09-03190. And Professor Paul Gustafson is the principal investigator of this project.

Chapter 3 of this thesis is based on the paper “On the Shape of an Exposure-Disease Relationship, the Average Effect of Exposure, and the Impact of Exposure Measurement Error.” *under review for Statistics in Medicine*, Xing L, Burstyn I, Gustafson P, 2014 [57]. This research problem was identified and designed by my thesis advisor Professor Paul Gustafson. As the first author, I worked on various conjectures, conducted numerical analysis and mathematical calculation, and performed data analysis and simulation studies. I drafted the manuscript, my supervisor and other coauthor helped me to revise it.
Table of Contents

Abstract ................................................................. ii
Preface ................................................................. iv
Table of Contents .................................................... v
List of Tables ........................................................ vii
List of Figures ........................................................ ix
Acknowledgments ....................................................... xii

1 Introduction .......................................................... 1

2 A Comparison of Bayesian Hierarchical Modeling with Group-based Exposure Assessment in Occupational Epidemiology ........ 4
  2.1 Introduction ..................................................... 4
  2.2 Bayesian Hierarchical Modeling ................................ 7
  2.3 The Algorithm ................................................... 9
  2.4 Simulation Study ................................................ 11
  2.5 Bias of the Group-based Exposure Assessment Method .......... 17
  2.6 An Empirical Example .......................................... 24
  2.7 Discussion ....................................................... 28

3 On the Shape of an Exposure-Disease Relationship, the Average Effect of Exposure, and the Impact of Exposure Measurement Error . 36
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>36</td>
</tr>
<tr>
<td>3.2</td>
<td>Logistic Box-Cox Model</td>
<td>39</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Parameter Estimates in the Model</td>
<td>40</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Factors in Epidemiological Studies</td>
<td>45</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Average Predictive Effect</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>Measurement Errors in the Model</td>
<td>59</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>Two-step Approach in Matching Process for Longitudinal Matching Covariates in the Presence of Measurement Error</td>
<td>72</td>
</tr>
<tr>
<td>4.1</td>
<td>Background</td>
<td>72</td>
</tr>
<tr>
<td>4.2</td>
<td>Baltimore Experience Corps Study</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>Statistical Modeling</td>
<td>76</td>
</tr>
<tr>
<td>4.4</td>
<td>Matching Algorithm</td>
<td>78</td>
</tr>
<tr>
<td>4.5</td>
<td>Treatment Effect</td>
<td>79</td>
</tr>
<tr>
<td>4.6</td>
<td>Simulation Study</td>
<td>80</td>
</tr>
<tr>
<td>4.7</td>
<td>Matching Quality and Measurement Error</td>
<td>86</td>
</tr>
<tr>
<td>4.8</td>
<td>Discussion</td>
<td>91</td>
</tr>
<tr>
<td>5</td>
<td>Discussion</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>95</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 100 samples observed in each group</td>
<td>12</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 1000 samples observed in each group</td>
<td>13</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.4$ and 1000 samples observed in each group</td>
<td>14</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.4$ and 100 samples observed in each group</td>
<td>15</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>The SSE and the ASE of the estimated $\beta_1$ when $\beta_1 = 0.6$ and 100 samples are observed in each group</td>
<td>16</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 100 samples observed in each group under the average difference between group exposure means equal to 0.3</td>
<td>23</td>
</tr>
<tr>
<td>Table 2.7</td>
<td>Estimates from logistic regression using true exposure intensity to predict leukemia mortality</td>
<td>24</td>
</tr>
<tr>
<td>Table 2.8</td>
<td>Reduced occupations of the SRS male workers and their summaries statistics</td>
<td>26</td>
</tr>
</tbody>
</table>
Table 2.9 | Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 20 samples observed in each group | 29

Table 2.10 | Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.4$ and 20 samples observed in each group | 30

Table 2.11 | Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 100 samples observed in each group under different priors for the scale parameters | 32

Table 2.12 | Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 20 samples observed in each group under different priors for the scale parameters | 33
**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The standardized biases versus $\sigma_B$ under $\sigma_W = 0.5$ and $\beta_1 = 0.6$. The vertical bars are 95% confidence intervals due to simulation variability.</td>
<td>18</td>
</tr>
<tr>
<td>2.2</td>
<td>Plot of $(1 - \rho_g)(1 - 2\rho_g)$ as a function of the group exposure mean, $\mu_g$, based on $\beta_0 = -4$ and $\beta_1 = 0.4$.</td>
<td>22</td>
</tr>
<tr>
<td>3.1</td>
<td>The histogram of the level of acid phosphatase in King-Armstrong units.</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Distributions of the exposure variable.</td>
<td>46</td>
</tr>
<tr>
<td>3.3</td>
<td>The risks of disease as a function of exposure in the selected settings.</td>
<td>48</td>
</tr>
<tr>
<td>3.4</td>
<td>Asymptotic standard deviation of $\hat{\lambda}$ over different settings.</td>
<td>50</td>
</tr>
<tr>
<td>3.5</td>
<td>Asymptotic standard deviation of $\hat{\lambda}$ over different settings.</td>
<td>51</td>
</tr>
<tr>
<td>3.6</td>
<td>The asymptotic standard deviation of the estimated average predictive effect, $\hat{\Delta}$, and the asymptotic standard deviation of $\hat{\lambda}$ under $P_1 = 0.02$ and different values of factor A, B, and D.</td>
<td>52</td>
</tr>
<tr>
<td>3.7</td>
<td>The asymptotic standard deviation of the estimated average predictive effect, $\hat{\Delta}$, and the asymptotic standard deviation of $\hat{\lambda}$ under $P_1 = 0.02$ and different values of factor A, B, and D.</td>
<td>53</td>
</tr>
<tr>
<td>3.8</td>
<td>Coefficient of variation of $\hat{\lambda}$ over different settings.</td>
<td>54</td>
</tr>
<tr>
<td>3.9</td>
<td>The average predictive effect, $\Delta$, and the large-sample limiting coefficient, $\gamma_1^*$, from the simple logistic model under different settings.</td>
<td>58</td>
</tr>
</tbody>
</table>
Figure 3.10 The asymptotic standard deviation of $\hat{y}$ from the misspecified model and the asymptotic standard deviation of $\hat{\Delta}$ under different settings. ......................................................... 59

Figure 3.11 The difference between the shape parameter, $\lambda$, in the true model (3.1) and the large-sample limit, $h^*$, in the misspecified model (3.19) in the presence of the multiplicative measurement error with $\sigma_e = 0.1$ across different settings. .................. 62

Figure 3.12 The difference between the shape parameter, $\lambda$, in the true model (3.1) and the large sample limit, $h^*$, in the model (3.19) in the presence of the multiplicative measurement error $\sigma_e = 0.01$ across different settings. .................. 63

Figure 3.13 The difference between the shape parameter, $\lambda$, in the true model (3.1) and the large sample limit, $h^*$, in the model (3.19) across different settings under $\sigma_e = 0.1$. .................. 64

Figure 3.14 The difference between the shape parameter, $\lambda$, in the true model (3.1) and the large sample limit, $h^*$, in the model (3.19) across different settings under $\sigma_e = 0.01$. .................. 65

Figure 3.15 The difference between the true probability and the fitted one from the misspecified model (3.19). .................. 66

Figure 3.16 The estimated probability of the nodal involvement in the prostate cancer patients via the fitted logistic Box-Cox model. .................. 67

Figure 4.1 Boxplot of the global distance from different conditions. .......... 82
Figure 4.2 Boxplot of the global distance from different conditions. .......... 83
Figure 4.3 Boxplot of the global distance from different conditions. .......... 84
Figure 4.4 Boxplot of the global distance from different conditions. .......... 85
Figure 4.5 Boxplot of the estimated treatment effects from different conditions. ................................................................. 87
Figure 4.6 Boxplot of the estimated treatment effects from different conditions. ................................................................. 88
Figure 4.7 Boxplot of the estimated treatment effects from different conditions. ................................................................. 89
Figure 4.8  Boxplot of the estimated treatment effects from different conditions. ................................. 90
Acknowledgments

I would like to express my special appreciation and thanks to my PhD supervisor, Professor Paul Gustafson, you have been a tremendous supervisor for me. Thanks for your priceless advice on my research and for letting the research work become a delightful and enjoyable journey. I also would like to express my thanks to our collaborator, Professor Igor Burstyn at the Department of Environmental and Occupational Health of Drexel University, who provided valuable and critical advice in epidemiology. Also thanks to Professors Lang Wu and Rollin Brant for serving as my committee members along the way, and thanks to our other collaborators, Professor David Richardson at the Department of Epidemiology, University of North Carolina, who provided the data from the Savannah River Site Study and gave valuable advice in epidemiology. And also thanks to Professors Qianli Xue, Jeanine M Parisi, and George Rebok at the Johns Hopkins Center on Aging and Health for hosting me and working together with us on the matching problems.

I feel so lucky to have been able to study at the UBC Statistics Department, where I had the good fortune to have been mentored by Professors Harry Joe, William J. Welch, Ruben H. Zamar, Matías Salibián-Barrera, and others. I learned so much from their insight, advice and tremendous knowledge of statistics. Also I wish also to acknowledge my Master degrees’ supervisors, Professor Charmaine Dean of the Department of Statistical and Actuarial Science, University Western Ontario and Professor Kee Yuen Lam of the Mathematics Department, UBC. Thanks also to Peggy Ng, Andrea Sollberger, and Zin Kurji in the UBC Statistics department office, who always gave efficient assistance with sincere care rather than merely carrying out duties.

A special thanks to my beloved husband, Xuekui Zhang, who takes lot of fam-
ily responsibility to support my studies. Thanks to my friends and colleagues for their encouragement and generous help in my life.

A final acknowledgement goes to NSERC and University 4-year Fellowships of UBC.
Chapter 1

Introduction

Investigation of relationships between exposures and diseases is a central topic in epidemiologic studies. The exposure variable could be workers’ exposure to potentially hazardous chemicals and/or ionizing radiation in the workplace in the occupational epidemiologic settings, could be the level of biomarkers in the human body in medical research, or could be the level of air pollution and/or heavy metal exposure in environmental health studies. And the disease outcome is typically binary with 1 indicating the presence of the disease and 0 indicating the absence of the disease. The goal of my thesis is to develop statistical methods to assist in finding associations between exposures and diseases under the consideration of some typical study concerns, which include the imperfectly measured exposure data, various potential shapes of relationship between exposure and disease, and the lack of randomization in the study design.

As for concerns in such studies, first of all, it is often the case that the exposure measurements are from readings in devices which are prone to measurement errors. If ignored in the analysis, these would introduce biased inference. Also, particularly in occupational epidemiologic studies, it is very costly to measure everyone’s exposure in the workplace. So the investigators randomly sample some workers to have their exposure measured, which introduces missing data in exposure measurements. However, the disease outcome data are typically relatively complete since they are obtained from the hospital records. Therefore, we cannot simply ignore the incomplete exposure data nor just use the naive traditional imputation method,
which may introduce biased results. Secondly, to model the shape of the relationship between exposure and disease, the linear model and the logit-linear model are overused in epidemiology due to convenience, but nonlinear relationships are quite possible and, therefore, require more attention. The final concern is that a lot of epidemiologic studies on exposure-disease relationship are observational cohort studies. Due to the lack of randomization in the experimental design, there are potential confounders in estimating the relationship between exposure and disease. Therefore, it is challenging to appropriately reduce confounding when there is also measurement error in the confounders. I address these issues separately in the following chapters.

In Chapter 2, we build a Bayesian hierarchical model for relating disease to a potentially harmful exposure, which can accommodate the missingness and measurement errors in the exposure measurement. The traditional imputation method, called the group-based exposure assessment method, uses the group exposure mean to impute the individual exposure in that group, where the group indicator indicates that the exposure levels tend to vary more across groups and less within groups. We compare our method with the traditional method through simulation studies, a real data application, and theoretical calculation. We focus on cohort studies where a logistic disease model is appropriate and where group exposure means can be modeled as fixed effects. The results show a variety of advantages of the fully Bayesian approach, and lead to recommendations on situations where the traditional group-based exposure assessment method may not be suitable to use.

In Chapter 3, we explore some conjectures concerning the nonlinear shapes of the exposure-disease relationship. Presuming that the relationship can be expressed in terms of regression coefficients and a shape parameter, we investigate how well the shape can be inferred in settings which might typify epidemiologic investigations and risk assessment. We also consider a suitable definition of the average effect of exposure, and investigate how precisely this can be inferred. This is done both in the case of using a model acknowledging uncertainty about the shape parameter and in the case of using a simple model ignoring this uncertainty. We also examine the extent to which exposure measurement error distorts inference about the shape of the exposure-disease relationship. All these investigations require a family of exposure-disease relationships indexed by a shape parameter. For this
purpose, we employ a family based on the Box-Cox transformation.

In Chapter 4, we look into how to reduce confounding of the observed covariates in estimating the relationship between exposure and disease since in the presence of measurement error in the confounders, the traditional matching method can provide systematically biased matching results. We propose to use the mixed-effects model to fit the repeated measurements or the longitudinal measurements of the matching covariates. Then we can adjust for the measurement errors by using matching on variables from the model. Simulation studies show us the this pre-matching process improve the matching quality and, as a consequence, treatment effect can be better estimated.

Discussion on our current investigations and future research topics on exposure-disease relationships are mentioned in Chapter 5.
Chapter 2

A Comparison of Bayesian Hierarchical Modeling with Group-based Exposure Assessment in Occupational Epidemiology

2.1 Introduction

In occupational cohort studies, interest lies in the association between potentially harmful occupational exposures occurring in the workplace and diseases outcomes. A binary outcome variable is often available for each cohort member. However, exposure quantities are usually measured for only a sample of members due to cost, or may be historically available only on a sample. For the occupational exposure, we often have repeated measurements, which are used to characterize day-to-day variations and measurement error. In this chapter we focus on a setting in which we wish to examine the association between a person’s average occupational exposure intensity, $x$, and a binary outcome, $y$. As for the measurement errors, there are two common types. One is the classical measurement error model, which states that
the observed exposure, $x^*$, is randomly distributed around the truth, $x$,

$$x^* = x + \varepsilon, \quad \varepsilon \sim N(0, \sigma_W^2), \quad \text{and} \quad x \perp \varepsilon.$$ 

Please note that $x^*$ is typically on the logarithmic scale of the observed exposure since the distribution of the observed exposure on the original scale is often right-skewed. The other is called the Berkson error model. In a Berkson error model, it is assumed that we have a predefined targeted value. The underlying true value, $x$, is equal to the targeted value, $\tilde{x}$, plus a random error. That is,

$$x = \tilde{x} + \varepsilon, \quad \varepsilon \sim N(0, \sigma_{Be}^2), \quad \tilde{x} \perp \varepsilon.$$ 

It is typical in a controlled experiment, where we know a predefined level such as a temperature or radiation exposure level, and we wish to bring the experimental quantity to this value, but the underlying true value may be away from this targeted value by an error. In an occupational epidemiology study, the exposure scores are sometimes assigned based on work areas, jobs, or other factors, and exposure levels are classified based on these work groups. In such cases, group exposure means are estimated with uncertainty and retain features of classical measurement error structure as in Berkson’s “modified controlled experiment” [4].

There are two methods to assign average exposure intensity scores to cohort members by drawing on available exposure measurements for a subsample of the cohort. One is called group-based exposure assessment; the other is called individual-based exposure assessment. Assume $x_{gij}^*$ is the log transformed exposure measurement for the $j$th measurement of the $i$th subject in the $g$th group and it satisfies the following model.

$$x_{gij}^* = \log(\text{exposure}) = \mu_g + \gamma_{gil} + \epsilon_{gij} = x_{gil} + \epsilon_{gij},$$

where $\mu_g$ is the fixed group exposure mean, $\gamma_{gil}$ is the random effect for the $i$th subject in the $g$th group, $x_{gil}$ is the true exposure based on log scale for the $g$th group and $i$th individual, and $\epsilon_{gij}$ is the measurement error with $i = 1, \ldots, n_g$ and $g = 1, \ldots, G$.

For the group-based exposure assessment, the group exposure mean estimated from the sample of individuals from the $g$th group, $\bar{X}_{g}^* = n_g^{-1} \sum_{i=1}^{n_g} \sum_{j=1}^{J} x_{gij}^*$, is used to
approximate the true exposure, \( x_{gi} \), for all people in that group. For the individual-based exposure assessment, if there are measurements for that individual, the individual exposure mean, \( \bar{X}_{gi} = J^{-1} \sum_{j=1}^{J} x_{gi,j} \), is used to approximate the true exposure, \( x_{gi} \), for that individual. For the classical exposure model, Tielemans et al. [53] showed that, when regressing continuous disease outcomes on the exposure measurements, the group-based exposure assessment method provided less biased estimates of the association parameter, and the individual-based exposure assessment method provided more biased, but more precise estimates of the association parameter. Kim et al. [29] confirmed this result. However, when encountering missing their exposure measurements, the individual-based exposure assessment method uses the complete-case analysis strategy. That is, the subjects without any observed exposure will be ignored. This strategy can cause loss of information [35], and, therefore, is not recommended in general. The group-based exposure assessment method works as a simple imputation method. That is, \( \bar{X}_{g} \) is imputed as the exposure measurement for all people in that group, and then a model for disease status given imputed exposure and covariates is fitted and interpreted as if it were a model for disease status given actual exposure and covariates. This can improve the utility of the data, especially when there is a large proportion of missing exposure data. Thus occupational health researchers are more interested in the group-based exposure assessment method, and there is more work on it in the literature, which we describe in detail as follows.

When the outcome variable is binary, Kim et al. [28] showed that the group-based exposure assessment method corresponds to an approximate Berkson error model when the group exposure mean is well estimated, such as when it is based on large sample size. Then \( E(X_{gi}| \bar{X}_{g}^*) = \bar{X}_{g}^* + (m - 1)(\bar{X}_{g}^* - \mu_g) \approx \bar{X}_{g}^* \), where \( m = \text{Cov}(X_{gi}, \bar{X}_{g}^*) / \text{Var}(\bar{X}_{g}^*) \). Then they postulated an approximate Berkson error model as \( X_{gi} = \bar{X}_{g}^* + \epsilon_{gi} \), where \( \epsilon_{gi} \) is normally distributed and \( \text{Cov}(\bar{X}_{g}^*, \epsilon_{gi}) = 0 \). Based on this Berkson error model, Kim and Burstyn [27] showed that Bayesian methods can provide less biased estimates of the association parameter compared to the group-based exposure assessment method, when the group exposure means are far apart and there are moderate between-subject variations. Kim et al. [29] further relaxed the requirement on the independence between the group exposure mean and the measurement error, and introduced a quasi-Berkson measurement
error model, which allows for correlation between the group exposure mean and
the random error when the group exposure mean is a random effect. For this quasi-
Berkson type model, they compared the estimates of the association parameters
based on assuming random group effects and assuming fixed group effects. They
also stated that if there is a fixed set of studied occupational groups in the study,
we should treat the group as fixed, while if we would like to draw inferences on all
possible occupational groups, it is more natural to assume that groups are created
through a random draw of all possible groupings. Under the assumption of either
fixed or random group exposure means, the estimates of regression coefficients are
less attenuated with a large sample size used to estimate group exposure means,
when between-subject variability is small and the spread between group exposure
means is large. Attenuation bias is virtually absent when a large number of mea-
surements is used to estimate fixed-effect but not random-effect group exposure
means (due to covariance of group exposure means and random error).

In this chapter, we develop a fully Bayesian method, which should in principle
give certain advantages. As for dealing with missing data, the group-based expo-
sure assessment method ignores the between-worker variation within a group by
simply replacing workers’ exposure levels by the sample group exposure mean,
which distorts the marginal distributions and the measures of covariations of the
exposure data and, therefore, may lead to invalid inference [35]. Our Bayesian
method is based on a likelihood approach, which can accommodate all the uncer-
tainty from the data, and also provide the posterior variance to describe the uncer-
tainty about the parameters and the exposure measurements [22]. Secondly, given
the groups are known, we use the classical type measurement error model to adjust
for measurement errors for each worker based on his repeated exposure measure-
ments. We concentrate our efforts on exploring properties of the fully Bayesian
approach in a cohort study with the group exposure means as the fixed effects in
the analysis.

2.2 Bayesian Hierarchical Modeling

We propose to build a hierarchical model, which contains three sub-models: a
response model to link disease status and true exposure, an exposure model to
model true exposure based on occupational categories, and a measurement error model to adjust for measurement errors. The Bayesian method is our top candidate to jointly fit these models, and it has a natural way to deal with missing data, which can be automatically imputed by simulated draws via Markov chain Monte Carlo based on the joint distribution assumptions.

Let \( y_{gi} \) be the binary disease outcome for the \( i \)-th individual in the \( g \)-th group, \( s_{gi} \) be the original true exposure level, with \( i = 1, \cdots, n_g \), and \( g = 1, \cdots, G \). Our response model can be described as

\[
H(\mathbb{E}(y_{gi}|s_{gi})) = \beta_0 + \beta_1 f(s_{gi}) \tag{2.1}
\]

where \( H(\cdot) \) is the link function, particularly here we use logit link function, and \( f(\cdot) \) is some transformation function. We focus on the logistic link function in this chapter. For simplicity we do not write out the covariate effects, though they could be easily added in (2.1). And we consider the log-transformed exposure data with the assumption that \( s_{gi} \) follows a log-normal distribution, which is often a realistic assumption for occupational exposure data. The measured exposure, \( s^*_{gi} \), has multiplicative measurement error, which can be written as \( s^*_{gi} = s_{gi}e_{gi} \), where \( e_{gi} \) also follows a log-normal distribution. Denote \( x_{gi} = \log(s_{gi}) \), and \( x^*_{gi} = \log(s^*_{gi}) \). Thus our measurement error model can be written as

\[
x^*_{gi} | x_{gi} \sim N(x_{gi}, \sigma_W^2),
\]

where \( x^*_{gi} \) can be treated as the \( j \)-th repeated measurement of the exposure intensity with \( j = 1, 2, \cdots, J \), and \( \sigma_W \) is the within-worker standard deviation. We assume that measurement errors are nondifferential, that is, these \( x^*_{gi} \)'s are not dependent on the response, \( y_{gi} \) given \( x_{gi} \). Our exposure model states the distribution of \( x_{gi} \) as follows:

\[
x_{gi} \sim N(\mu_g, \sigma_B^2), \tag{2.2}
\]

where \( \mu_g \) is the exposure mean in the \( g \)th group, and \( \sigma_B \) is the between-worker standard deviation within a group. In our study, the group is treated as a fixed effect. For priors, we assume \( \beta_0 \sim N(0, \sigma_{\beta_0}^2), \beta_1 \sim N(0, \sigma_{\beta_1}^2), \mu_g \sim N(0, \sigma_{\mu}^2) \) for \( g = 1, 2, \cdots, G \), \( \sigma_W^2 \sim \text{Inv-Gamma}(\alpha_W, \gamma_W) \), and \( \sigma_B^2 \sim \text{Inv-Gamma}(\alpha_B, \gamma_B) \).
2.3 The Algorithm

Based on the models proposed in the previous section, we calculate the full conditional posterior distribution for all the parameters \((\beta, \sigma^2_W, \sigma^2_B, \mu_1, \cdots, \mu_G)\) and the latent variables \((x_{gi}’s)\). For \(\sigma^2_W, \sigma^2_B, \) and \(\mu_g’\)’s, we can write out the closed forms of their full conditional posterior distributions. For the remaining parameters and latent variables, their full conditional posterior distributions are only available up to a proportionality constant. So we propose to use a MCMC algorithm, Metropolis within Gibbs [7]. Specifically, in each iteration, for \(\sigma^2_W, \sigma^2_B, \) and \(\mu_g’\)’s, we use Gibbs sampling to sample candidates, which are all accepted for updating, while for \(\beta\) and \(x_{gi}’\)’s, we use random walk Metropolis. For the random walk Metropolis, in each \(t\)-th iteration of the Metropolis sampling, we propose their candidate values based on the normal distributions centered at their sampling values in the \((t-1)\)-th iteration and with some pre-tuned standard deviations. Those candidates may be accepted or rejected based on the acceptance ratio, \(r\). The definition of this acceptance ratio can be found in [7]. If the acceptance ratio, \(r\), is larger than 1, we use the candidate value. If \(r < 1\), we use the candidate value by probability \(r\) and stay at the \((t-1)\)-th sampled value by probability \(1-r\). Note that for efficient convergence, the standard deviations of the normal distributions in the random walks are determined by adaptively tuning the overall acceptance rates such that the rates fall in the range between 20% and 50% [7]. Our algorithm is coded in R 2.12.2.

Usually there are missing values in the exposure measurements and these are an essential feature of the group-based exposure assessment approach where, by definition, exposure measurements are not available for all subjects for whom health outcome was determined. It is infeasible to make measurements on the whole cohort. So, ideally, the investigators randomly choose some members of each group to measure their exposure, and in this ideal situation the missing mechanism is missing completely at random (MCAR) within each group, which is assumed by the group-based exposure assessment method. For the fully Bayesian method, we extend the missing mechanism to missing at random (MAR) within each group. Please also note that MAR is based on the random sampling conditional on the observed data, which may not always the case due to complex real situations. For example, some workers may have missing records due to loss of compliance from
more exposure. In this case, the missingness is based on unobserved data. In real analysis, the statement of the exceptions for the missing data is encouraged [32]. In our analysis, in addition, we assume if one measurement is missing, the other measurements for that subject are missing as well. Denote the exposure measurements as

\[ X^* = \left( X^*_{gij} \right)_{g=1}^{G} \]

where \( X_{gij} \) is the observed exposure measurement and \( X_{gij}^m \) is the missing exposure measurement, the latent true exposure as \( X \), and the outcomes as \( Y \). Let the missing indicator for the \( i \)th worker in the \( g \)th group be

\[
m_{gi} = \begin{cases} 
1 & \text{if } x_{gij}^* \text{ is missing for } j = 1, \cdots, J. \\
0 & \text{if } x_{gij}^* \text{ is observed for } j = 1, \cdots, J.
\end{cases}
\]

and \( M = (m_{gi}) \) with \( i = 1, 2, \cdots, n_g \) and \( g = 1, 2, \cdots, G \). We know that this missing mechanism does not depend on the missing data. Let \( \theta \) be the vector of parameters of data, \( \phi \) be the vector of parameters of the missing-data process, and \( \theta \) and \( \phi \) are assumed to be distinct. The full likelihood function can be written as

\[
L(\theta, X, \phi | Y, M, X^*) = f(Y, M, X^* | \theta, X, \phi)
\]

We further assume that the missingness indicator \( m_{gi} \) is conditionally independent of \( x_{gi}^* \) given \( y_{gi} \) and \( g \) for all \( g = 1, \cdots, G \) since the true exposure is a latent quantity and the exposure measurements are not observed for all workers. Then our likelihood function would be extended to

\[
L(\theta, X, \phi | Y, M, X^*) = \prod_{g=1}^{G} \prod_{i=1}^{n_g} f(y_{gi} | \theta, x_{gi}) f(x_{gi}^* | \theta, x_{gi}) f(m_{gi} | y_{gi}, g, x_{gi}, \phi)
\]

Also if we add the prior information, \( p(\theta) \), of \( \theta \), we can see the posterior distri-
bution of $\theta$ is proportional to $f(Y, X^* | \theta, X)p(\theta)$. Then we do not need to model $M$, and we can say that the missing mechanism is ignorable for Bayesian inference [35].

### 2.4 Simulation Study

To compare the performance of the fully Bayesian method and the group-based exposure assessment method, we design the following simulation studies corresponding to some occupational epidemiological settings [27, 28, 30, 31]. We simulate data with 5000 workers, who are categorized into five groups, with the true group exposure means $\mu_g = 1.1, 2.1, 3.1, 4.1, \text{and } 5.1$. In each group, there are 1000 workers. For each worker, one binary disease outcome is measured. We know that the exposure may only be measured for a sample of workers per group in a cohort study and there are two repeated measurements for those workers who had measured exposure. We choose the number of the observed samples per group varying as 1000 (complete data), 100 (90% subjects with missing exposure data), or 20 (98% subjects with missing exposure data). We randomly select these observed workers within each group. We set $b_0 = -4$, $b_1 = 0.4$ or 0.6, $\sigma_W$ varying as 0.1, 0.5, and 1.0, and $\sigma_B$ varying as 0.1, 0.5, 1.0, 1.5 and 2.0. Thus the typical risk per group is varying from 2.8% to 12.3% given $b_1 = 0.4$, or from 3.4% to 28.1% given $b_1 = 0.6$, under the mean exposure for each group.

For each setting we simulate 100 data sets, and for each data set we do the parameter estimation using one MCMC chain with 10,000 iterations and storing every 10-th iteration. The chosen priors are $b_0, b_1 \sim N(0, 100)$, $\sigma_W, \sigma_B \sim \text{Inv-Gamma}(0.0001, 0.0001)$, and $\mu_g \sim N(0, 10000)$ with $g = 1, 2, \cdots, G$. The first half of the chain is discarded as burn-in, and the remaining half is kept for inference. The estimated posterior mean from the chain is considered as the estimate of the corresponding parameter. And the estimated posterior standard deviation is considered as the standard error of the corresponding parameter.

We compare the precision, variances, and coverage probabilities of the 95% credible intervals of the estimated association parameter $b_1$ from the two methods.
Table 2.1: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 100 samples observed in each group

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>FB</td>
<td>0.0069</td>
<td>0.0013</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0059</td>
<td>0.0013</td>
<td>0.96</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0052</td>
<td>0.0014</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0061</td>
<td>0.0014</td>
<td>0.95</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>-0.0016</td>
<td>0.0015</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0060</td>
<td>0.0016</td>
<td>0.93</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0055</td>
<td>0.0015</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0023</td>
<td>0.0014</td>
<td>0.95</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0031</td>
<td>0.0015</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0022</td>
<td>0.0014</td>
<td>0.94</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>FB</td>
<td>-0.0051</td>
<td>0.0017</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0024</td>
<td>0.0016</td>
<td>0.90</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0065</td>
<td>0.0018</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0225</td>
<td>0.0019</td>
<td>0.85</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0042</td>
<td>0.0019</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0225</td>
<td>0.0020</td>
<td>0.85</td>
</tr>
<tr>
<td>1.5</td>
<td>0.5</td>
<td>FB</td>
<td>-0.0075</td>
<td>0.0020</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0227</td>
<td>0.0021</td>
<td>0.80</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0064</td>
<td>0.0024</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0570</td>
<td>0.0049</td>
<td>0.51</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5</td>
<td>FB</td>
<td>0.0015</td>
<td>0.0022</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0570</td>
<td>0.0049</td>
<td>0.53</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0143</td>
<td>0.0024</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0573</td>
<td>0.0050</td>
<td>0.50</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>FB</td>
<td>0.0071</td>
<td>0.0032</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.1013</td>
<td>0.0121</td>
<td>0.17</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>FB</td>
<td>0.0016</td>
<td>0.0031</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.1013</td>
<td>0.0120</td>
<td>0.14</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0199</td>
<td>0.0032</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.1017</td>
<td>0.0122</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Table 2.2: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_i = 0.6$ and 1000 samples observed in each group.

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>FB</td>
<td>0.0066</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0060</td>
<td>0.0013</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0063</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0058</td>
<td>0.0013</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0065</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0056</td>
<td>0.0013</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0052</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0015</td>
<td>0.0012</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0056</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0017</td>
<td>0.0012</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0059</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0019</td>
<td>0.0012</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0044</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0203</td>
<td>0.0014</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0054</td>
<td>0.0009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0205</td>
<td>0.0014</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0063</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0207</td>
<td>0.0014</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0011</td>
<td>0.0006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0532</td>
<td>0.0037</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0017</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0534</td>
<td>0.0037</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0029</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0536</td>
<td>0.0037</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0957</td>
<td>0.0099</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0007</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0958</td>
<td>0.0099</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0015</td>
<td>0.0006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0960</td>
<td>0.0100</td>
</tr>
</tbody>
</table>
Table 2.3: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.4$ and 1000 samples observed in each group

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>FB</td>
<td>0.0093</td>
<td>0.0023</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0095</td>
<td>0.0022</td>
<td>0.95</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0095</td>
<td>0.0023</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0094</td>
<td>0.0022</td>
<td>0.95</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0094</td>
<td>0.0023</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0092</td>
<td>0.0022</td>
<td>0.95</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0061</td>
<td>0.0019</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0066</td>
<td>0.0019</td>
<td>0.97</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0064</td>
<td>0.0020</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0065</td>
<td>0.0019</td>
<td>0.97</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>FB</td>
<td>0.0066</td>
<td>0.0020</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0063</td>
<td>0.0019</td>
<td>0.97</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0051</td>
<td>0.0014</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0048</td>
<td>0.0016</td>
<td>0.97</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0053</td>
<td>0.0014</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0047</td>
<td>0.0016</td>
<td>0.97</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>FB</td>
<td>0.0059</td>
<td>0.0016</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0046</td>
<td>0.0015</td>
<td>0.97</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0023</td>
<td>0.0010</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0057</td>
<td>0.0015</td>
<td>0.98</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0022</td>
<td>0.0010</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0058</td>
<td>0.0015</td>
<td>0.98</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>FB</td>
<td>0.0024</td>
<td>0.0011</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0059</td>
<td>0.0015</td>
<td>0.98</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0007</td>
<td>0.0006</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0176</td>
<td>0.0014</td>
<td>0.98</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0177</td>
<td>0.0014</td>
<td>0.98</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>FB</td>
<td>0.0012</td>
<td>0.0006</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0179</td>
<td>0.0014</td>
<td>0.99</td>
</tr>
</tbody>
</table>
**Table 2.4:** Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.4$ and 100 samples observed in each group

(FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method)

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>FB</td>
<td>0.0098</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0095</td>
<td>0.0022</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0084</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0094</td>
<td>0.0022</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>FB</td>
<td>0.0011</td>
<td>0.0021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0092</td>
<td>0.0022</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0077</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0061</td>
<td>0.0020</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0053</td>
<td>0.0019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0059</td>
<td>0.0019</td>
</tr>
<tr>
<td>0.1</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0019</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0056</td>
<td>0.0019</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0090</td>
<td>0.0021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0033</td>
<td>0.0019</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0063</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0032</td>
<td>0.0018</td>
</tr>
<tr>
<td>0.1</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0011</td>
<td>0.0019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0029</td>
<td>0.0018</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0054</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0082</td>
<td>0.0021</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0025</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0083</td>
<td>0.0021</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0064</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0085</td>
<td>0.0021</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0041</td>
<td>0.0025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0217</td>
<td>0.0024</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5</td>
<td>FB</td>
<td>0.0016</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0218</td>
<td>0.0023</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0074</td>
<td>0.0022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0221</td>
<td>0.0024</td>
</tr>
</tbody>
</table>
Table 2.1 shows us when $\beta_1 = 0.6$, $\sigma_B$ is moderate or large, and only 100 workers are observed per group, the estimated $\beta_1$ from the fully Bayesian method tends to have smaller bias, smaller mean squared error (MSE), and more accurate coverage rates than for the group-based exposure assessment method. For instance, when $\sigma_B = 2$ and $\sigma_W = 0.1$, the absolute bias of the group-based exposure assessment method is 14 fold that of the fully Bayesian method, the MSE of the group-based exposure assessment is around 4 times that of the fully Bayesian method, and the coverage of the group-based exposure assessment is only 17% compared with 89% from the fully Bayesian method. When $\sigma_B$ becomes larger, the advantages of the fully Bayesian method are more prominent, which is consistent with the fact that since the group-based exposure assessment method ignores the between-worker variance within each group, more between-worker variance can cause more bias. On the other hand, when $\sigma_W$ becomes large, the measurement becomes poor, which makes both methods less accurate. However, we did not find prominent differences between the two methods when $\beta_1$ is set as 0.4, as shown in Tables 2.3 and 2.4.

Table 2.5: The SSE and the ASE of the estimated $\beta_1$ when $\beta_1 = 0.6$ and 100 samples are observed in each group

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>GB</th>
<th>1.0</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>2.0</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_W$</td>
<td>SSE</td>
<td>0.038</td>
<td>0.039</td>
<td>0.040</td>
<td>0.041</td>
<td>0.041</td>
<td>0.042</td>
<td>0.043</td>
<td>0.042</td>
<td>0.043</td>
<td>0.030</td>
<td>0.030</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>0.034</td>
<td>0.034</td>
<td>0.034</td>
<td>0.032</td>
<td>0.032</td>
<td>0.032</td>
<td>0.030</td>
<td>0.030</td>
<td>0.030</td>
<td>0.056</td>
<td>0.056</td>
<td>0.053</td>
</tr>
<tr>
<td>FB</td>
<td>SSE</td>
<td>0.042</td>
<td>0.043</td>
<td>0.045</td>
<td>0.049</td>
<td>0.047</td>
<td>0.047</td>
<td>0.056</td>
<td>0.056</td>
<td>0.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ASE</td>
<td>0.041</td>
<td>0.041</td>
<td>0.042</td>
<td>0.046</td>
<td>0.047</td>
<td>0.046</td>
<td>0.051</td>
<td>0.050</td>
<td>0.048</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We calculate two kinds of summary standard errors of the estimated $\beta_1$. One is the simulated standard error (SSE), which is defined as the standard deviation of the estimated $\beta_1$'s from all simulated data under the same condition. The other is the averaged standard error (ASE), which is calculated as the mean of the estimated standard errors of $\beta_1$ from each simulation data set under the same condition. The SSE and the ASE of the fully Bayesian method agree well in all of our simulated
conditions. However, for the group-based exposure assessment method, when there is a large amount of missing data, Table 2.5 shows the SSE is much larger than the ASE. This confirms that the group-based exposure assessment method underestimates the standard error of the association parameter.

As for the biases, we inspect the raw bias, the standardized bias, and the percentage bias. Table 2.1 and Figure 2.1 show that the biases of the estimated $\beta_1$ from the fully Bayesian method are much smaller than those from the group-based exposure assessment method when $\beta_1 = 0.6$ and $\sigma_B$ is moderate or large. For the fully Bayesian method, the biases are within 5% of the value of $\beta_1$, and the amount of missing data has an effect on the magnitude of the biases. For example, when there are only 20 workers observed, the biases are much greater than those with 100 workers observed and complete data. However, as $\sigma_B$ increases, the differences between these three types of data tend to become less, and the magnitude of bias is also becoming smaller. For the group-based exposure assessment method, the magnitude of the bias sharply increases when $\sigma_B$ increases, regardless of the amount of missing data. Also the gaps of biases between data with different amount of missing data become larger and larger as $\sigma_B$ increases.

### 2.5 Bias of the Group-based Exposure Assessment Method

In addition to the simulation study, we can theoretically explore the cause of the bias of the group-based exposure assessment method for binary outcomes, if we assume that the group exposure means are known precisely instead of estimated from a sample as it is done in practice. The response model under consideration is

$$
\Pr(y_{gi} = 1|x_{gi}, \beta_0, \beta_1) = \logit^{-1}(\beta_0 + \beta_1 x_{gi}).
$$

We can obtain the relationship between $y_{gi}$ and $\mu_g$ by integrating out the $x_{gi}$’s. After plugging in the exposure model (2.2), we have

$$
\Pr(y_{gi} = 1|\mu_g, \epsilon_{gi}, \beta_0, \beta_1) = \logit^{-1}(\beta_0 + \beta_1 \mu_g + \beta_1 \epsilon_{gi}).
$$
Figure 2.1: The standardized biases versus $\sigma_B$ under $\sigma_W = 0.5$ and $\beta_1 = 0.6$. The vertical bars are 95% confidence intervals due to simulation variability.
To get the link between $y_{gi}$ and $\mu_g$, we need to integrate out the error term:

$$
\Pr(y_{gi} = 1 | \mu_g, \beta_0, \beta_1) = \int_{-\infty}^{+\infty} \logit^{-1}(\beta_0 + \beta_1 \mu_g + \beta_1 \varepsilon_{gi}) \frac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left(-\frac{\varepsilon_{gi}^2}{2\sigma_B^2}\right) d\varepsilon_{gi}
$$

$$
= \int_{-\infty}^{+\infty} \frac{1}{1 + \frac{\exp(-\beta_1 \varepsilon_{gi})}{\exp(B_0 + \beta_1 \mu_g)} \sqrt{2\pi\sigma_B^2}} \exp\left(-\frac{\varepsilon_{gi}^2}{2\sigma_B^2}\right) d\varepsilon_{gi}, \quad (2.3)
$$

where $\varepsilon_{gi} \sim N(0, \sigma_B^2)$. To calculate the integral (2.3), we do Taylor expansion of $\left(1 + \frac{\exp(-\beta_1 \varepsilon_{gi})}{\exp(B_0 + \beta_1 \mu_g)}\right)^{-1}$ on $-\beta_1 \varepsilon_{gi}$ around 0. Let $A = \exp(B_0 + \beta_1 \mu_g)$ and $z = -\beta_1 \varepsilon_{gi}$. Consider the function

$$
g(z) = \left(1 + \frac{e^z}{A}\right)^{-1}.
$$

We have

$$
g'(z) = -\frac{Ae^z}{(A + e^z)^2},
$$

$$
g^{(2)}(z) = \frac{Ae^z(-A + e^z)}{(A + e^z)^3},
$$

$$
g^{(3)}(z) = -\frac{Ae^z(A^2 - 4Ae^z + e^z)}{(A + e^z)^4},
$$

$$
g^{(4)}(z) = \frac{Ae^z(-A + e^z)(A^2 - 10Ae^z + e^z)}{(A + e^z)^5}.
$$

Then the first three terms of the expansion of $g$ at $z = 0$ is

$$
g = g(0) + g'(0)z + g^{(2)}(0)\frac{z^2}{2!} + g^{(3)}(0)\frac{z^3}{3!} + g^{(4)}(0)\frac{\eta^4}{4!} = \frac{A}{1 + A} - \frac{A}{(A + 1)^2}z + \frac{A(1 - A)}{(A + 1)^3}\frac{z^2}{2!} - \frac{A(1 - A)(A^2 - 10A + 1)}{(A + 1)^4}\frac{z^3}{3!} + \frac{A(1 - A)(A^2 - 10A + 1)\eta^4}{(A + 1)^5}\frac{\eta^4}{4!}, \quad (2.4)
$$
where η lies between z and 0. Plugging (2.4) in to integration (2.3), we have

\[
\Pr(y_{gi} = 1 | \mu_g, \beta_0, \beta_1) = \frac{A}{1 + A} + \frac{A(1 - A)}{2!(A + 1)^3} \beta_1^2 \sigma_B^2 \\
+ \frac{A(1 - A)(A^2 - 10A + 1)}{4!(A + 1)^5} \int_{-\infty}^{+\infty} \frac{\eta^4}{\sqrt{2\pi \sigma_B^2}} \exp \left( -\frac{\varepsilon_{gi}^2}{2\sigma_B^2} \right) \, d\varepsilon_{gi},
\]

where \( A = \exp(\beta_0 + \beta_1 \mu_g) \). The residual term is

\[
\frac{A(1 - A)(A^2 - 10A + 1)}{4!(A + 1)^5} \int_{-\infty}^{+\infty} \frac{\eta^4}{\sqrt{2\pi \sigma_B^2}} \exp \left( -\frac{\varepsilon_{gi}^2}{2\sigma_B^2} \right) \, d\varepsilon_{gi}.
\]

What is more, we have

\[
\frac{A(1 - A)(A^2 - 10A + 1)}{4!(A + 1)^5} \int_{-\infty}^{+\infty} \frac{\eta^4}{\sqrt{2\pi \sigma_B^2}} \exp \left( -\frac{\varepsilon_{gi}^2}{2\sigma_B^2} \right) \, d\varepsilon_{gi} \leq
\]

\[
\frac{A(1 - A)(A^2 - 10A + 1)}{4!(A + 1)^5} \int_{-\infty}^{+\infty} \frac{\eta^4}{\sqrt{2\pi \sigma_B^2}} \exp \left( -\frac{\varepsilon_{gi}^2}{2\sigma_B^2} \right) \, d\varepsilon_{gi} \\
= \frac{A(1 - A)(A^2 - 10A + 1)}{4!(A + 1)^5} \beta_1^4 \mathcal{E}(\varepsilon_{gi}) \\
= \frac{A(1 - A)(A^2 - 10A + 1)}{8(A + 1)^5} \beta_1^4 \sigma_B^2.
\]

After we ignore the residual term, we get the approximation of the relationship between \( y_{gi} \) and \( \mu_{gi} \).

\[
\Pr(y_{gi} = 1 | \mu_g, \beta_0, \beta_1) \approx \frac{A}{1 + A} + \frac{A(1 - A) \beta_1^2 \sigma_B^2}{(A + 1)^3 \cdot 2!},
\]
where $A = \exp(\beta_0 + \beta_1 \mu_g)$. Denote $\beta = (\beta_0, \beta_1)'$, and let

$$
\rho_g(\beta) = \frac{\exp(\beta_0 + \beta_1 \mu_g)}{1 + \exp(\beta_0 + \beta_1 \mu_g)}.
$$

Then this probability equation of the true model can be written as

$$
\Pr(y_{gi} = 1|\mu_g, \beta_0, \beta_1) \approx \rho_g \left[ 1 + (1 - \rho_g)(1 - 2\rho_g) \frac{\beta_1^2 \sigma_B^2}{2} \right].
$$

(2.5)

In comparison, when we assume that the group exposure means are known, the relationship between $y_{gi}$ and $\mu_g$ is assumed as

$$
\Pr(y_{gi} = 1|\mu_g, \beta_0, \beta_1) = \rho_g.
$$

Also since they are not available in the real studies, $\mu_g's$ are approximated by their corresponding sample group exposure means. Therefore we can see that, when the sample group exposure mean is very close to the true group exposure mean, which is often true under large sample sizes, the group-based exposure assessment method may not work well under any of the following conditions: (1), the association parameter is big, i.e., the disease risk increases dramatically when there is one-unit increase in the exposure; (2), the between-worker variation is relatively large; and (3), the prevalence of disease is rare, which makes $\rho_g$ far from 1 and $1/2$.

There is another conclusion from our theoretical calculation, which can be supported by the simulation results mentioned in [27]. Assume there is the same increment between the adjacent group exposure means. Kim and Burstyn [27] inspected the bias and MSE of the estimated $\beta_1$ when $\beta_0 = -4, \beta_1 = 0.4, \mu_1 = 0.1$, and the increment varies as $0.3, 0.5, 1, \text{and} 1.414$ in their simulation studies. They concluded that when the increment between groups is large and the between-subject variance is relatively small, the group-based exposure assessment method gives unbiased estimates, while it fails to adjust for the measurement errors properly when the increment between groups is small. Their conclusions agree with our theoretical calculation. From equation (2.5), we can see that, given $\beta_0 = -4$ and $\beta_1 = 0.4$, the closer the function $(1 - \rho_g)(1 - 2\rho_g)$ is to 0, the better the group-based exposure assessment method performs. Figure 2.2 is a plot of $(1 - \rho_g)(1 - 2\rho_g)$ as a func-
Figure 2.2: Plot of \((1 - \rho_{\theta})(1 - 2\rho_{\theta})\) as a function of the group exposure mean, \(\mu_\theta\), based on \(\beta_0 = -4\) and \(\beta_1 = 0.4\).
tion of the $g$th group exposure mean, $\mu_g$, with $\beta_0 = -4$ and $\beta_1 = 0.4$. It shows that $(1 - \rho_g) (1 - 2\rho_g)$ is decreasing to 0 as $\mu_g$ increases in the range of our consideration. Also given $\mu_1$, the smallest group exposure mean, the larger the increment of groups, the larger means the other groups have. If the group exposure means are more distributed to the right side of the curve in Figure 2.2, the group-based exposure assessment method will have less bias. Then we can conclude that under these given conditions, the further apart the group exposure means are, the better performance we can expect from the group-based exposure assessment method.

**Table 2.6:** Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 100 samples observed in each group under the average difference between group exposure means equal to 0.3

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_W$</th>
<th>$\sigma_B$</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0927</td>
<td>0.0263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0054</td>
<td>0.0257</td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td>FB</td>
<td>-0.0844</td>
<td>0.0207</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0542</td>
<td>0.0333</td>
</tr>
<tr>
<td>2.0</td>
<td></td>
<td>FB</td>
<td>-0.0595</td>
<td>0.0144</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.1111</td>
<td>0.0422</td>
</tr>
</tbody>
</table>

To confirm our aforementioned conclusion and also to compare the performance of the group-based exposure assessment and the fully Bayesian methods under smaller average difference between group exposure means, we run supplementary simulation studies under the average difference between group exposure means equal to 0.3 (with group exposure means equal to 0.1, 0.4, · · · , 1.3), $\beta_0 = -4$, $\beta_1 = 0.6$, $\sigma_W = 1$, and $\sigma_B = 1.0, 1.5, 2$, which are selected to match the simulation settings in [27]. The simulation results are shown in Table 2.6. We can see that under fixed $\sigma_W$, as $\sigma_B$ increases the performance of the group-based exposure assessment method worsens according to the three listed criteria. This is consistent with our theoretical finding. As $\sigma_B$ increases, there are also some trends
for the fully Bayesian method, i.e., slightly increased bias, decreased MSE, and reduced coverage rate. Those indicate that increasing between group variance will also make the fully Bayesian method perform worse. However, compared with that of the group-based exposure assessment method, the performance of the fully Bayesian method is relatively stable and consistent with the simulation results under the average difference between group exposure means equal to 1. In addition, the fully Bayesian method is less biased, has smaller MSE and similar coverage rate when $\sigma_B = 2$.

### 2.6 An Empirical Example

To illustrate the application of the fully Bayesian method, we use a subset of the data from a study of exposure to radiation and mortality from leukemia conducted at the Savannah River Site (SRS) (South Carolina) [40, 41]. The study includes a total of 18,883 workers hired between 1950 and 1986 who were followed through 2002 to ascertain causes of death. We analyze data for the subcohort of 15,264 male workers using the fully Bayesian method, and also using the group-based exposure assessment method, and comment on the analyses resulting from both methods.

#### Table 2.7: Estimates from logistic regression using true exposure intensity to predict leukemia mortality

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-6.406</td>
<td>1.297</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>True intensity exposure (log-scale)</td>
<td>0.268</td>
<td>0.108</td>
<td>0.013</td>
</tr>
<tr>
<td>Total exposure days (log-scale)</td>
<td>0.264</td>
<td>0.152</td>
<td>0.081</td>
</tr>
<tr>
<td>Black</td>
<td>-0.657</td>
<td>0.476</td>
<td>0.167</td>
</tr>
<tr>
<td>Pay monthly</td>
<td>0.021</td>
<td>0.327</td>
<td>0.948</td>
</tr>
<tr>
<td>Pay weekly</td>
<td>-0.265</td>
<td>0.343</td>
<td>0.440</td>
</tr>
<tr>
<td>Attained Age</td>
<td>-0.008</td>
<td>0.010</td>
<td>0.448</td>
</tr>
</tbody>
</table>

The outcome variable, $Y$, is leukemia mortality. The predictor of interest in the current analysis is the exposure intensity to ionizing radiation. The annual exposure measurements and the annual total days of employment were recorded. As
suggested by [40], there is an induction and latency period between exposure to
ionizing radiation and an observed change in risk of leukemia mortality. Since ex-
posure data were available for the period 1950 - 1999, while follow-up spanned
the period 1950-2002, a 3-year lag was the minimal lag assumption possible to
evaluate in these analyses. Therefore, for the exposure measurements and days of
exposure, we only consider those measured or counted 3 years before the leukemia
outcome or before the exit from the study. Also we can get a worker’s exposure in-
tensity estimate for a year using his annual measurement divided by his annual total
employment days. For purposes of evaluating the methodologies, we let $x_i$ be the
average of the $i$th worker’s annual intensities from study entry to the lag year. The
average number of such years is 12.9 (SE = 9.3). Conceptually we will treat $x_i$ as
the true exposure intensity, in order to evaluate the methodologies when less infor-
mation about exposure is available. To estimate the association between leukemia
mortality and radiation exposure intensity, we use a logistic regression with the
true exposure intensity, $x_i$, as a predictor. To reduce confounding, we adjust our
analysis using the following confounders: total days of exposure; the attained age,
which is defined as age of death if the worker is deceased during the study, or age
of exiting the study otherwise; sex; race (Black vs. other); and pay code (used to
control for socioeconomic differences in mortality and classified on the basis of
the worker’s pay schedule when hired as paid monthly, weekly, or hourly) [40].
Table 2.7 shows us the estimated coefficients, associated standard errors, and their
p-values. There is a significant association between exposure intensity and the
leukemia outcome. We also test the difference between the coefficient of the true
exposure intensity (log-scale) and that of the total exposure days (log-scale), and
find that we cannot reject the null hypothesis that there is no difference between
these two coefficients. Therefore, there is some indication that it is cumulative ex-
posure which is influencing the outcome. This would be an interesting topic for
further investigation.

To conduct group-based exposure assessment, we define exposure groups by
combining similar occupations that were recorded for the workers in the original
study and term them as “reduced occupations.” Though some workers changed
their occupations during the study, for simplicity we choose their longest occupa-
Table 2.8: Reduced occupations of the SRS male workers and their summaries statistics

<table>
<thead>
<tr>
<th>Reduced Occupations</th>
<th>Num of Leukemia</th>
<th>Crude Risk of Disease</th>
<th>Geometric Mean of radiation exposure in excess relative rate/10mSv (Geometric SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clerical and kindred non-manual workers</td>
<td>9 (898)</td>
<td>1%</td>
<td>0.158 (3.376)</td>
</tr>
<tr>
<td>Technicians, Analysts, and Assistants</td>
<td>4 (970)</td>
<td>0.41%</td>
<td>0.429 (3.085)</td>
</tr>
<tr>
<td>Scientific and medical services</td>
<td>2 (508)</td>
<td>0.39%</td>
<td>0.166 (2.671)</td>
</tr>
<tr>
<td>Engineering jobs</td>
<td>9 (2018)</td>
<td>0.45%</td>
<td>0.116 (2.622)</td>
</tr>
<tr>
<td>Supervisors and managers</td>
<td>7 (1244)</td>
<td>0.56%</td>
<td>0.185 (3.073)</td>
</tr>
<tr>
<td>General Service Operator</td>
<td>6 (1711)</td>
<td>0.35%</td>
<td>0.155 (2.840)</td>
</tr>
<tr>
<td>Operators</td>
<td>16 (2062)</td>
<td>0.78%</td>
<td>0.437 (3.440)</td>
</tr>
<tr>
<td>Skilled trades</td>
<td>13 (1542)</td>
<td>0.84%</td>
<td>0.327 (2.965)</td>
</tr>
<tr>
<td>Other semi-skilled workers</td>
<td>3 (670)</td>
<td>0.45%</td>
<td>0.076 (1.789)</td>
</tr>
</tbody>
</table>

tions during the study as their unique occupations during the study. We assume in each occupation, the exposure intensity is following a log-normal distribution. Table 2.8 shows us a summary of the data in each reduced occupation. To imagine a study with only two repeated measurements for each subject, we randomly select two exposure intensity measurements for the $i$th worker, $x_{i1}$ and $x_{i2}$, and treat them as two repeated measurements of the true exposure, $x_i$. We use the fully Bayesian and the group-based exposure assessment methods separately to analyze the data. We iterate this approach 100 times and obtain 100 model estimates from each method. The average of the estimated association parameter from the fully Bayesian method is 0.284 (ASE = 0.142). The average of the estimated association parameter from the group-based exposure assessment method is 0.228 (ASE = 0.271). The average of estimated within-worker variance is 1.226 and that of the between-worker variance is 0.988. Relative to the association based on observing the true exposure intensities shown in Table 2.7, we can see that the estimates of the association parameter from the fully Bayesian method are less biased for the true association between exposure and leukemia. There is some attenuation in the coefficient estimate from the group-based exposure assessment method, which may be due to lack of adjustment for measurement errors, and there is also a greater uncertainty from the group-based exposure assessment method, which is shown from the ASE of the group-based exposure assessment method doubling that from the fully Bayesian method. This is most likely because the group-based exposure assessment method does not make full use of the data. On the other side, for both methods we calculate the power to detect the significant association indi-
cated from the full data. The results show us based on the significance level set as 0.05, the fully Bayesian method detects a significant signal for 55 of 100 replications, while the group-based exposure assessment method does not detect for any of them. Therefore, we would like to conclude that the fully Bayesian method can better recover the signal from the data compared with the group-based exposure assessment method in this data example.

To examine the behavior of the fully Bayesian method under a more realistic prior for the association parameter, $\beta_1$, we have tried the prior distribution of $\beta_1$ as $N(0, \log_2 2)$. We repeat the above analysis. It turns out that the average of the estimated association parameter from the fully Bayesian method is 0.268 (ASE = 0.140), the average of estimated within-worker variance is 1.226 and that of the between-worker variance is 0.988, and a significant signal is detected for 47 of 100 replications. We can see that this informative prior gives a big improvement for the accuracy of the association parameter.

To examine both methods in the presence of a large proportion of missing data, we randomly select 90% of the subjects within each occupation group and treat them as missing their exposure measurements. After that we randomly select two repeated measurement as we did before and fit the models using both methods. We repeat this procedure 100 times. The average of the estimated association parameter from the fully Bayesian method is 0.299 (ASE = 0.256). The average of the estimated association parameter from the group-based exposure assessment method is 0.243 (ASE = 0.257). The average of estimated within-worker variance is 1.403 and that of the between-worker variance is 0.516. Compared to the full data results in Table 2.7 then, the two methods seem to be performing similarly. However, the fully Bayesian method gives 95%, 90%, and 80% credible intervals for the exposure intensity coefficient which excludes zero in 0, 9, and 36 of the 100 replications respectively. The corresponding numbers of significant findings for the group-based method are 0, 0 and 2 out of the 100 replications. So we conclude that even under the presence of missing data, the fully Bayesian method is more powerful in this data example.

Although a conditional logistic regression model would be a standard approach to analysis of matched case-control data [40], in order to illustrate our method, we use unconditional likelihood with adjustment for the matching factors. Our
analysis shows the significant association between exposure and mortality based on the unmatched data as was shown in [40].

2.7 Discussion

From our mathematical derivation and our simulation studies, we can conclude that there are a few conditions under which the group-based exposure assessment method may not work well. One of them is the existence of big variation between workers within a group. The bias of the estimated $\beta_1$ from the group-based exposure assessment method is almost proportional to the magnitude of the between-worker variance under certain settings, as shown by [28]. The second condition which may cause the group-based exposure assessment method to perform poorly is low prevalence of the disease outcome. Kim et al. [29] have run simulations under different disease prevalences for the group-based exposure assessment method. Based on their simulation results, we can see that under lower prevalence the group-based exposure assessment method tends to have much larger MSE. And the third condition is a big increase in risk given one-unit increase in exposure. Notice that some of these conditions commonly exist in occupational epidemiological studies, although it is comforting to note that for weak associations that are believed to dominate epidemiology today, all methods considered work well [13]. So we recommend the fully Bayesian method, which shows its good performance over different parameter settings from our testing. In addition, the fully Bayesian method can provide estimates of uncertainty arising from the data, such as the variance of the measurement errors and the variance between workers within a group, and estimates of other quantities including the association parameters and latent variables simultaneously through iterative simulation process. This should be more accurate than separated steps to get those estimates, since the variance structure is well considered [8]. This will further benefit epidemiologists and occupational hygienists to obtain estimates related to compliance and risk in assessing occupational exposure and overexposure [54].

We need to mention that there are other factors, which can also affect the behavior of the two methods. One such factor is the sample size. For the group-based
Table 2.9: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 20 samples observed in each group

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>FB</th>
<th>GB</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.0066</td>
<td>0.0014</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0060</td>
<td>0.0014</td>
<td>0.95</td>
</tr>
<tr>
<td>0.5</td>
<td>FB</td>
<td>0.0104</td>
<td>0.0019</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0053</td>
<td>0.0018</td>
<td>0.92</td>
</tr>
<tr>
<td>1.0</td>
<td>FB</td>
<td>0.0162</td>
<td>0.0033</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0034</td>
<td>0.0028</td>
<td>0.82</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>0.0073</td>
<td>0.0021</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0027</td>
<td>0.0019</td>
<td>0.87</td>
</tr>
<tr>
<td>0.5</td>
<td>FB</td>
<td>0.0097</td>
<td>0.0024</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0035</td>
<td>0.0022</td>
<td>0.87</td>
</tr>
<tr>
<td>1.0</td>
<td>FB</td>
<td>0.0130</td>
<td>0.0034</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0056</td>
<td>0.0031</td>
<td>0.77</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>0.0152</td>
<td>0.0048</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0258</td>
<td>0.0038</td>
<td>0.73</td>
</tr>
<tr>
<td>0.5</td>
<td>FB</td>
<td>0.0268</td>
<td>0.0199</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0265</td>
<td>0.0039</td>
<td>0.68</td>
</tr>
<tr>
<td>1.0</td>
<td>FB</td>
<td>0.0181</td>
<td>0.0066</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0285</td>
<td>0.0047</td>
<td>0.65</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>0.0204</td>
<td>0.0113</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0653</td>
<td>0.0089</td>
<td>0.42</td>
</tr>
<tr>
<td>0.5</td>
<td>FB</td>
<td>0.0222</td>
<td>0.0167</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0658</td>
<td>0.0090</td>
<td>0.44</td>
</tr>
<tr>
<td>1.0</td>
<td>FB</td>
<td>0.0167</td>
<td>0.0158</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.0677</td>
<td>0.0096</td>
<td>0.43</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>0.0203</td>
<td>0.0174</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.1169</td>
<td>0.0193</td>
<td>0.16</td>
</tr>
<tr>
<td>0.5</td>
<td>FB</td>
<td>0.0176</td>
<td>0.0182</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.1173</td>
<td>0.0193</td>
<td>0.15</td>
</tr>
<tr>
<td>1.0</td>
<td>FB</td>
<td>0.0031</td>
<td>0.0193</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.1190</td>
<td>0.0199</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Table 2.10: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.4$ and 20 samples observed in each group

(FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method)

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>FB</td>
<td>0.0098</td>
<td>0.0023</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0096</td>
<td>0.0023</td>
<td>0.93</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>-0.0213</td>
<td>0.0931</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0090</td>
<td>0.0023</td>
<td>0.92</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>-0.0188</td>
<td>0.0792</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0075</td>
<td>0.0027</td>
<td>0.88</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0069</td>
<td>0.0025</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0059</td>
<td>0.0023</td>
<td>0.95</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>FB</td>
<td>-0.0417</td>
<td>0.1340</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0053</td>
<td>0.0023</td>
<td>0.95</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>FB</td>
<td>-0.0354</td>
<td>0.1042</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0036</td>
<td>0.0025</td>
<td>0.92</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0128</td>
<td>0.0037</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0010</td>
<td>0.0028</td>
<td>0.89</td>
</tr>
<tr>
<td>1.0</td>
<td>0.5</td>
<td>FB</td>
<td>-0.0295</td>
<td>0.0894</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>0.0004</td>
<td>0.0028</td>
<td>0.89</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>FB</td>
<td>-0.0036</td>
<td>0.0311</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0012</td>
<td>0.0030</td>
<td>0.87</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0094</td>
<td>0.0061</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0146</td>
<td>0.0038</td>
<td>0.82</td>
</tr>
<tr>
<td>1.5</td>
<td>0.5</td>
<td>FB</td>
<td>0.0099</td>
<td>0.0061</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0151</td>
<td>0.0038</td>
<td>0.81</td>
</tr>
<tr>
<td>2.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0163</td>
<td>0.0104</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0333</td>
<td>0.0055</td>
<td>0.61</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5</td>
<td>FB</td>
<td>0.0247</td>
<td>0.0278</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0337</td>
<td>0.0054</td>
<td>0.63</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>FB</td>
<td>0.0196</td>
<td>0.0283</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0351</td>
<td>0.0056</td>
<td>0.64</td>
</tr>
</tbody>
</table>
exposure assessment method, if the sample size is very small, the sample mean can be far away from the true mean. The conditional mean imputation will cause very poor estimation of $\beta_1$. For the fully Bayesian method, if the sample size is small, the inference will be most likely affected by the priors, and since less information can be used in updating the posterior distribution, more biased estimates could be provided given the same length of MC chain. For small sample size, two observations from our simulation studies are noteworthy. When the sample size is very small, e.g. 100 workers per group and only 10 exposure measurements from each group, both methods fail for some simulated data sets. Therefore, in practice, more caution is required when the sample size is small. Secondly, we observe that the two methods behave differently at small sample size. For example, when there are 1000 workers per group, but only 20 exposure measurements from each group, Table 2.9 and Table 2.10 show us that the group-based exposure assessment method is less biased under the small between-group standard deviation given comparable coverage rates. When $\sigma_B$ is large, the fully Bayesian method performs better since the coverage of the group-based exposure assessment method deteriorates.

We also look at the sensitivity of the fully Bayesian method based to the selection of priors. In the simulation studies, we used diffuse priors for all the parameters. However, Gelman [17] mentioned some concern on inverse gamma distribution of the scale parameters and suggested using a uniform prior on the standard deviation, and a half-t family when the number of groups is small. So we did a small simulation study to investigate how our posterior estimates can change if we use different hyper-parameters of the inverse gamma distributions in our current settings. Instead of Inv-Gamma(0.0001, 0.0001), we let $\sigma_W, \sigma_R \sim$ Inv-Gamma(0.01, 0.01), and we also vary the number of observed workers per group as 20 and 100. Tables 2.11 and 2.12 shows us the model estimates, which are very close to those from Inv-Gamma(0.0001, 0.0001) priors shown in Tables 2.1 and 2.9, separately. This reflects the robustness of our modeling to some degree. However, we suggest readers to use more robust priors such as the uniform priors written in our attached WinBUGs code and when the number of groups is less than 5, consider the half-t family mentioned in Gelman [17]. Also the Cauchy
Table 2.11: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 100 samples observed in each group under different priors for the scale parameters

(FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method)

<table>
<thead>
<tr>
<th>$\sigma_B = 0.5$</th>
<th>$\sigma_W = 0.1$</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>0.0042</td>
<td>0.0015</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>GB</td>
<td>-0.0023</td>
<td>0.0014</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>$\sigma_W = 0.5$</td>
<td>FB</td>
<td>0.0038</td>
<td>0.0015</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0022</td>
<td>0.0014</td>
<td>0.94</td>
</tr>
<tr>
<td>$\sigma_W = 1.0$</td>
<td>FB</td>
<td>-0.0043</td>
<td>0.0016</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0024</td>
<td>0.0016</td>
<td>0.90</td>
</tr>
<tr>
<td>$\sigma_B = 1.0$</td>
<td>$\sigma_W = 0.1$</td>
<td>FB</td>
<td>0.0067</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0225</td>
<td>0.0019</td>
<td>0.86</td>
</tr>
<tr>
<td>$\sigma_W = 0.5$</td>
<td>FB</td>
<td>0.0043</td>
<td>0.0019</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0225</td>
<td>0.0020</td>
<td>0.85</td>
</tr>
<tr>
<td>$\sigma_W = 1.0$</td>
<td>FB</td>
<td>-0.0063</td>
<td>0.0019</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0227</td>
<td>0.0021</td>
<td>0.80</td>
</tr>
<tr>
<td>$\sigma_B = 1.5$</td>
<td>$\sigma_W = 0.1$</td>
<td>FB</td>
<td>0.0060</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0570</td>
<td>0.0049</td>
<td>0.51</td>
</tr>
<tr>
<td>$\sigma_W = 0.5$</td>
<td>FB</td>
<td>0.0024</td>
<td>0.0022</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0570</td>
<td>0.0049</td>
<td>0.53</td>
</tr>
<tr>
<td>$\sigma_W = 1.0$</td>
<td>FB</td>
<td>-0.0149</td>
<td>0.0023</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>-0.0573</td>
<td>0.0050</td>
<td>0.50</td>
</tr>
</tbody>
</table>

class of prior distributions for logistic regression parameters may have advantages in handling separation and sparsity issues [19].

In the literature there is another Bayesian approach for the occupational exposure data. Kim et al. [28] showed that when estimated group exposure mean is a good approximation of the true group exposure mean, the estimated group exposure mean is related to true exposure for each worker by an approximate Berkson error model. Consequently, Kim and Burstyn [27] assume a Berkson exposure error structure in a Bayesian framework. Also they further explored a setting where the standard deviation between workers with a group is known. Based on these assumptions, they developed a Bayesian method, called Berkson-Bayesian group-
Table 2.12: Bias, MSE and coverage rates of the fully Bayesian and the group-based exposure assessment methods when $\beta_1 = 0.6$ and 20 samples observed in each group under different priors for the scale parameters

( FB is for the fully Bayesian method, and GB is for the group-based exposure assessment method )

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\sigma_W$</th>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Coverage Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0070</td>
<td>0.0021</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0027</td>
<td>0.0019</td>
<td>0.87</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>FB</td>
<td>0.0093</td>
<td>0.0024</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0035</td>
<td>0.0022</td>
<td>0.87</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0140</td>
<td>0.0035</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0056</td>
<td>0.0031</td>
<td>0.78</td>
</tr>
<tr>
<td>1.0</td>
<td>0.1</td>
<td>FB</td>
<td>0.0144</td>
<td>0.0046</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0258</td>
<td>0.0038</td>
<td>0.73</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>FB</td>
<td>0.0253</td>
<td>0.0212</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0265</td>
<td>0.0039</td>
<td>0.68</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>FB</td>
<td>0.0249</td>
<td>0.0141</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0285</td>
<td>0.0047</td>
<td>0.65</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
<td>FB</td>
<td>0.0168</td>
<td>0.0090</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0653</td>
<td>0.0089</td>
<td>0.44</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>FB</td>
<td>0.0165</td>
<td>0.0124</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0658</td>
<td>0.0090</td>
<td>0.45</td>
</tr>
<tr>
<td>1.0</td>
<td></td>
<td>FB</td>
<td>0.0159</td>
<td>0.0218</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GB</td>
<td>-0.0677</td>
<td>0.0096</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Based on similar simulation settings, the fully Bayesian method provides less biased estimates in most cases and reduction by a factor of 4 in mean squared error when $\sigma_B = 2$. Also the fully Bayesian method avoids plugging in an estimate of $\sigma_B$ as if it were known and it assumes the classical measurement error type, which is closer to reality.
for occupational epidemiological studies. Both the Berkson-Bayesian group-based method and the fully Bayesian method can be trivially implemented in WinBUGS. The WinBUGS code for the fully Bayesian method is given below:

```r
model{
  for (i in 1 : M){ # M is the total sample size
    y[i] ˜ dbern(p[i]) # response model
    logit(p[i]) <- beta0 + beta1 * x[i] # true likelihood
    xx1[i] ˜ dnorm(x[i],tau.W) # measurement error model
    xx2[i] ˜ dnorm(x[i],tau.W) # measurement error model
    x[i] ˜ dnorm(mu.g[Group[i]],tau.B) # true exposure distribution
  }

  for (j in 1:G){ # G is the number of groups
    mu.g[j] ˜ dnorm(0, 0.01) # prior of mean of each group
  }

  tau.W <- pow(sigma.W, -2) # tau.W=1/sigma.W^2

  beta0 ˜ dnorm(0, 1.0E-2) # prior of beta0
  beta1 ˜ dnorm(0, 1.0E-2) # prior of beta1

  sigma.W ˜ dunif(0.001, 100) # prior of sigma.W
  sigma.B ˜ dunif(0.001, 100) # prior of sigma.B
}
```

When researchers plan to use the fully Bayesian method, simulations on relevant settings are recommended to see how much missingness and what sample sizes are optimal. In the future, we would modify the fully Bayesian method to fit a variety of studies. For example, we would modify the measurement error model if those errors are not classical. We would change the response model for the time to event outcome data. We would also accommodate different exposure distributions. On the other hand, the fully Bayesian method could also be extended for situations where exposure and outcome data are available on disjoint sets of individuals. This commonly occurs in practice when health status of persons whose exposure was monitored is unknown and is only linked to health effects through knowledge of group membership [32]. Another possible extension of the fully Bayesian approach is to use random group effects since this may be more natural to some epidemiologic problems. Generally, we hope the availability of a more
appropriate method of analysis encourages epidemiologists to use the group-based study design for estimating exposure-disease associations in occupational settings.
Chapter 3

On the Shape of an Exposure-Disease Relationship, the Average Effect of Exposure, and the Impact of Exposure Measurement Error

3.1 Introduction

The exposure-disease relationship analysis, also called dose-response analysis, plays a central role in risk assessment in epidemiologic studies, especially when the exposure is a continuous quantity and the disease is indicated by a binary outcome. For example, in occupational epidemiology, workers exposed to some potential hazard in the workplace may later develop the disease of interest. Therefore, we would like to assess the risk based on the level of the exposure through modeling the shape of the exposure-disease relationship. A popular approach to the dose-response analysis is categorical analysis, which breaks exposure into categories, usually based on the percentile method, and then looks for trend or relative risks. However, as mentioned in [20], one potential pitfall is that when exposure effects
are limited to extreme ends of the exposure scale, the estimated hazard from data can be attenuated if the high exposure level is combined with some lower level into a single category in the data analysis. Instead, Greenland [20] suggested using fractional polynomial regression and spline regression, which seems under-used in epidemiologic research. Steenland and Deddens [48] mentioned the importance of the spline method, which can provide a close fit of the dose-response curve, and also suggest using parametric models for easier interpretation. In this chapter, we focus on a parametric model, particularly though a Box-Cox type model, which is flexible and indexed by a shape parameter.

The general model for the association between a binary disease outcome, $Y$, and an exposure variable, $X$, could be written as follows. For the $i$-th subject, we have

$$
\log \left( \frac{p_i}{1-p_i} \right) = g(x_i),
$$

(3.1)

where $p_i = E(Y_i|X_i = x_i)$, function $g$ describes the shape of the exposure-disease relationship, and here we let $g(x) = \beta_0 + \beta_1 x^{(\lambda)}$. The function $x^{(\lambda)}$ is the Box-Cox transformation, sometimes also called the power transformation [5], defined as

$$
x^{(\lambda)} = \begin{cases} 
\frac{x^{\lambda} - 1}{\lambda} & \lambda > 0 \\
\log(x) & \lambda = 0
\end{cases}.
$$

(3.2)

This function has a nice continuity property at $\lambda = 0$ as we can see that $\lim_{\lambda \to 0^+} x^{(\lambda)} = \log(x)$, and this continuity extends to any degree of derivatives. Clearly the shape parameter, $\lambda$, can fit different patterns with the change of the predictor, $X$. More specifically $\lambda = 1$ indicates a linear pattern, $\lambda > 1$ indicates an increasing slope pattern, and $\lambda < 1$ indicates a decreasing slope pattern. This model is a special case of the fractional polynomial logistic model, which is discussed in [44].

As the nonlinearity of the Box-Cox transformation may introduce some difficulty in the model interpretation, and in some situations it may be hard to precisely estimate the shape parameter, we can consider the average predictive effect, as an alternative summary of the effect of a predictor. Gelman and Pardoe [18] suggested averaging the effect of a predictor over the population distribution of predictors. In a linear regression with predictors only included as linear main effects, the regres-
sion coefficients coincide with the predictive effects of their corresponding predictors. In a model containing nonlinear and/or interaction terms of a predictor, the average predictive effect of the predictor is defined as the expectation of its slope based on given distribution of the predictor. So we can think the average predictive effect as an averaged slope across its entire distribution. Examples are shown by Liu and Gustafson [36] in linear regression models and by Gustafson [23] in the survival analysis context. We adapt their definitions to the logistic regression context and define the average predictive effect as follows:

\[ \Delta = E \left[ \frac{d \{ \logit(E(Y|X)) \}}{dX} \right], \]

presuming this expectation exists.

When predictor \( X \) is poorly measured, we cannot ignore the measurement error. There are two common types of errors, the additive measurement error and the multiplicative measurement error [9]. For the classic additive measurement error model, the measured predictor, \( X^* \), is equal to the true value, \( X \), plus a random error, \( U \). That is,

\[ X^* = X + U, \]

where \( X \) and \( U \) are independent, and \( U \) follows the normal distribution centered at 0. As for the multiplicative measurement error model, which can be appropriate for a positive, skewed distribution of exposure that often arises in epidemiological applications, a common model is

\[ X^* = XW, \]

(3.3)

where \( X \) and \( W \) are independent, and \( W \) follows the log-normal(0, \( \sigma_w^2 \)) distribution.

Measurement errors can distort the relationship between the response and predictors. In the linear regression, there is attenuation in the \( \beta \) coefficient for the non-differential additive measurement errors [9]. There is a similar attenuation in the shape parameter when we fit a power regression model and the predictor follows a log-normal distribution for multiplicative non-differential measurement errors [12]. Küchenhoff and Carroll [33] showed that due to measurement errors an original threshold relationship between true predictor and response becomes a
smooth relationship between the measured predictor and the response. When we consider the generalized linear regression, the effect of measurement error is more complex. Stefanski and Carroll [50] found that in the simple logistic regression, the measurement error tends to attenuate the predicted probabilities towards 0.5. There does not appear to be any discussion in literature specifically about the effect of measurement error in the Box-Cox family applied to logistic regression.

In section 3.2, we start with a data example on the application of the logistic Box-Cox model, and then introduce the settings for all the issues we are going to discuss. Due to the difficulty of direct interpretation of the parameters in the model and/or the requirement of large sample sizes in order to precisely estimate the shape parameter under some conditions, we investigate using the average predictive effect as an alternative way to summarize the relationship between the outcome and the predictor. We also discuss using the estimated slope coefficient from the potentially misspecified simple logistic regression as an estimate of the average predictive effect. In section 3.3, we look at the distortion of the parameter estimates in the logistic Box-Cox family model under different degrees of measurement error. The chapter closes with a discussion in section 3.4.

3.2 Logistic Box-Cox Model

We start our discussion with precisely measured data. In this case, the maximum likelihood estimates (MLEs) from the fitted model based on the observed data are not contaminated by the measurement errors, and, therefore, can be treated as appropriate estimates of the parameters in the true model, assuming no model misspecification. Given data \((Y, X)\), we can obtain MLEs of the parameters in the logistic Box-Cox model (3.1), based on the profile likelihood method. The log likelihood function of model (3.1) is

\[
I(\beta_0, \beta_1; Y, X) = \sum_{i=1}^{n} Y_i \left[ \beta_0 + \beta_1 \left( \frac{x_1^i - 1}{\lambda} \right) \right] - \log \left[ 1 + \exp \left( \beta_0 + \beta_1 \left( \frac{x_1^i - 1}{\lambda} \right) \right) \right],
\]
where \( Y = (y_1, y_2, \cdots, y_n)^T \) and \( X = (x_1, x_2, \cdots, x_n)^T \). The MLEs, \((\hat{\beta}_0, \hat{\beta}_1, \hat{\lambda})\), can be obtained by solving the equations below,

\[
\nabla l(\hat{\beta}_0, \hat{\beta}_1, \hat{\lambda} | Y, X) \bigg|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1, \lambda = \hat{\lambda}} = 0.
\]

The above equations include the score functions relating to \( \beta_0 \) and \( \beta_1 \),

\[
\sum_{i=1}^{n} \left[ \left( \frac{1}{\hat{\lambda}^2 - 1} \right) \left( y_i - \frac{\exp \left( \hat{\beta}_0 + \hat{\beta}_1 \frac{x_i^\hat{\lambda} - 1}{\hat{\lambda}} \right)}{1 + \exp \left( \hat{\beta}_0 + \hat{\beta}_1 \frac{x_i^\hat{\lambda} - 1}{\hat{\lambda}} \right)} \right) \right] = 0. \tag{3.4}
\]

Defining a new variable \( V = (x^\hat{\lambda} - 1)/\hat{\lambda} \), we can rewrite equation (3.4) as

\[
\sum_{i=1}^{n} \left[ \left( \frac{1}{v_i} \right) \left( y_i - \frac{\exp \left( \hat{\beta}_0 + \hat{\beta}_1 v_i \right)}{1 + \exp \left( \hat{\beta}_0 + \hat{\beta}_1 v_i \right)} \right) \right] = 0. \tag{3.5}
\]

We can see that given \( \hat{\lambda} \), the above equation (3.5) could be solved using standard logistic regression software, if we assume \((Y, V)\) are our data. Therefore, we can use the profile likelihood method to get the MLE. First of all, we need to set an upper bound \( B_u \) of the estimate of \( \lambda \), since a very large value of \( \lambda \) is meaningless and implausible in the disease-exposure model. Then we do a grid search in the interval \([0, B_u]\) with 1000\([B_u]\) equally spaced values of \( \lambda \), fit the model at each value, and select the particular value which maximizes the likelihood function.

### 3.2.1 Parameter Estimates in the Model

The prostate cancer data from [6], which were analyzed in detail in [11], provide a good example to illustrate the logistic Box-Cox family model. The study involved 53 prostatic cancer patients who had a laparotomy to see if the cancer had spread to the surrounding lymph nodes. The outcome variable, \( Y \), is binary indicating presence (1) or absence (0) of nodal involvement. One of the risk factors is the level of acid phosphatase in serum (in King-Armstrong units), \( X \). The phosphatase is a measure of the degree of tissue damage. Figure 3.1 shows that the distribution of the phosphatase variable is positively skewed. Under the assumption of a log-
Figure 3.1: The histogram of the level of acid phosphatase in King-Armstrong units. The solid curve is an estimated density curve, and the dashed curve is the density curve of log-normal($-0.42, 0.32$).

normal($\mu, \sigma^2$) distribution for the phosphatase, we get $\hat{\mu} = -0.42$ and $\hat{\sigma} = 0.32$, which indicates some mild skewness. We fit the logistic Box-Cox model to the data using the profile likelihood mentioned above. The final fitted model is

$$\text{logit} \left[ \mathbb{E}(Y | X) \right] = 0.40 + 2.24\log(X).$$

We can see that phosphatase is positively associated with the nodal involvement, and one unit increase in phosphatase in log-scale is associated with 9.39 times increase in the odds of nodal involvement. Since $\hat{\lambda} = 0$ is at the boundary of the parameter space, we bootstrap 500 samples to check the variability of the parameter estimates. It turns out that for 91% of these samples $\hat{\lambda}$ is at the lower bound of 0, 3.4% yield $\hat{\lambda} = 5$, which is the imposed upper bound, and the remaining 5.6% yield $\hat{\lambda}$ between 0 and 5. Furthermore, the 95% CI for $\hat{\beta}_0$ is $[-0.75, 2.35]$, and the 95% CI for $\hat{\beta}_1$ is $[0.33, 6.51]$.

We can see that $\lambda$ plays an important role in determining the shape of exposure-disease relationship, with a one unit difference in $\lambda$ corresponding to a big difference in the shape. As an example, $\lambda = 0$ corresponds to a log-transformation in the
logistic model, but $\lambda = 1$ represents a simple logistic model. Therefore, we start with evaluating the precision of the estimator of $\lambda$. We calculate the Fisher information matrix to obtain the asymptotic variance of the MLE of $\lambda$. One conclusion is that given fixed values of other parameters, when $\beta_1$ goes to 0, the asymptotic variance of $\hat{\lambda}$ goes to infinity. Denote $Z$ as a random variable following the standard normal distribution. Then we can write $X = \exp(\mu + Z\sigma)$. The likelihood function for a single observation is

$$L(\beta_0, \beta_1, \lambda | X = x, Y = y) = \left[ \frac{\exp\left( \beta_0 + \beta_1 \frac{x^\lambda - 1}{\lambda} \right)}{1 + \exp\left( \beta_0 + \beta_1 \frac{x^\lambda - 1}{\lambda} \right)} \right]^y \left[ \frac{1}{1 + \exp\left( \beta_0 + \beta_1 \frac{x^\lambda - 1}{\lambda} \right)} \right]^{1-y} f(x),$$

where $f(x)$ is the probability density function of $X$. Then the log likelihood function for a single observation is

$$l(\beta_0, \beta_1, \lambda | X = x, Y = y) = y \left( \beta_0 + \beta_1 \frac{x^\lambda - 1}{\lambda} \right) - \log \left[ 1 + \exp\left( \beta_0 + \beta_1 \frac{x^\lambda - 1}{\lambda} \right) \right] + \log[f(x)]. \quad (3.6)$$

Based on the log likelihood function (3.6), we can calculate the Fisher information.
the inverse of the Fisher information matrix shown in (3.7). That is,

\[ I_1(\theta) \]

\[
= -E \begin{bmatrix}
\frac{\partial^2 I(X, Y)}{\partial q \partial q} & \frac{\partial^2 I(X, Y)}{\partial q \partial p} & \frac{\partial^2 I(X, Y)}{\partial q \partial \lambda} \\
\frac{\partial^2 I(X, Y)}{\partial p \partial q} & \frac{\partial^2 I(X, Y)}{\partial p \partial p} & \frac{\partial^2 I(X, Y)}{\partial p \partial \lambda} \\
\frac{\partial^2 I(X, Y)}{\partial \lambda \partial q} & \frac{\partial^2 I(X, Y)}{\partial \lambda \partial p} & \frac{\partial^2 I(X, Y)}{\partial \lambda \partial \lambda}
\end{bmatrix}
\]

\[
= -E \begin{bmatrix}
E \left( \frac{\partial^2 I(X, Y)}{\partial q \partial q} | X \right) & E \left( \frac{\partial^2 I(X, Y)}{\partial q \partial p} | X \right) & E \left( \frac{\partial^2 I(X, Y)}{\partial q \partial \lambda} | X \right) \\
E \left( \frac{\partial^2 I(X, Y)}{\partial p \partial q} | X \right) & E \left( \frac{\partial^2 I(X, Y)}{\partial p \partial p} | X \right) & E \left( \frac{\partial^2 I(X, Y)}{\partial p \partial \lambda} | X \right) \\
E \left( \frac{\partial^2 I(X, Y)}{\partial \lambda \partial q} | X \right) & E \left( \frac{\partial^2 I(X, Y)}{\partial \lambda \partial p} | X \right) & E \left( \frac{\partial^2 I(X, Y)}{\partial \lambda \partial \lambda} | X \right)
\end{bmatrix}
\]

\[
= E \begin{bmatrix}
P(1-P) & P(1-P)V & P(1-P)\beta_1 \frac{\phi(z)}{\phi(z)} \\
P(1-P)V & P(1-P)V^2 & P(1-P)\beta_1 V \frac{\phi(z)}{\phi(z)} \\
\frac{1}{\lambda} & \frac{1}{\lambda} & \frac{1}{\lambda} + \left( \frac{1}{\lambda} \right)^2
\end{bmatrix}
\]

\[
\phi(z) = \frac{1}{\lambda} \left( \frac{1}{\lambda} \right)^2
\]

where \( \theta = (\beta_0, \beta_1, \lambda)^T \), \( P = E(Y|X) \), \( \phi(z) \) is the probability density function of the standard normal distribution, and

\[
V = \frac{X^2 - 1}{\lambda}.
\]

To calculate the asymptotic variance of \( \lambda \), we need to calculate the (3,3) entry of the inverse of the Fisher information matrix shown in (3.7). That is,

\[
\text{Avar}(\hat{\lambda}) = \frac{1}{\det[I_1(\theta)]} C_{33},
\]

where \( C_{33} \) is the matrix cofactor of \( I_1(\theta) \) with

\[
C_{33} = (-1)^{3+3} \det \begin{bmatrix}
\int_{-\infty}^{+\infty} p(1-p)\phi(z)dz & \int_{-\infty}^{+\infty} p(1-p)v\phi(z)dz \\
\int_{-\infty}^{+\infty} p(1-p)v\phi(z)dz & \int_{-\infty}^{+\infty} p(1-p)v^2\phi(z)dz
\end{bmatrix}
\]

\[
= \left( \int_{-\infty}^{+\infty} p(1-p)\phi(z)dz \right) \left( \int_{-\infty}^{+\infty} p(1-p)v^2\phi(z)dz \right) \\
- \left( \int_{-\infty}^{+\infty} p(1-p)v\phi(z)dz \right) \left( \int_{-\infty}^{+\infty} p(1-p)v\phi(z)dz \right),
\]

43
and
\[
\det(I_l(\theta)) = \beta_1^2 \left( \int_{-\infty}^{\infty} p(1-p)\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v^2\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p) \left( \frac{\partial v}{\partial \lambda} \right)^2 \phi(z)dz \right) \\
- \left( \int_{-\infty}^{\infty} p(1-p)\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v\frac{\partial v}{\partial \lambda} \phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p) \left( \frac{\partial v}{\partial \lambda} \right)^2 \phi(z)dz \right) \\
- \left( \int_{-\infty}^{\infty} p(1-p)v\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v\frac{\partial v}{\partial \lambda} \phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p) \left( \frac{\partial v}{\partial \lambda} \right)^2 \phi(z)dz \right) \\
+ \left( \int_{-\infty}^{\infty} p(1-p)\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v^2\frac{\partial v}{\partial \lambda} \phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p) \frac{\partial v}{\partial \lambda} \phi(z)dz \right) \\
+ \left( \int_{-\infty}^{\infty} p(1-p)v\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v^2\frac{\partial v}{\partial \lambda} \phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p) \frac{\partial v}{\partial \lambda} \phi(z)dz \right) \\
- \left( \int_{-\infty}^{\infty} p(1-p)\phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v\frac{\partial v}{\partial \lambda} \phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v \frac{\partial v}{\partial \lambda} \phi(z)dz \right) \left( \int_{-\infty}^{\infty} p(1-p)v^2 \phi(z)dz \right). 
\]

In general, we cannot get closed-form expressions for the integrals shown in the above equation (3.8). However, numerical evaluation of the integrals is possible by using the Gaussian-Hermite Quadrature (GHQ). In GHQ, we use the following approximation,
\[
\int_{-\infty}^{\infty} e^{-t^2} f(t) dt \approx \sum_{i=1}^{m} w_i f(t_i)
\]
where \(m\) is the number of sample points used, and \(\{t_i\}_{i=1,...,m}\) are the roots of the (physicists’ version of the) Hermite polynomial \(H_m(t)\) \((i = 1, 2, ..., m)\). The associated weights \(w_i\) are given in [1] as
\[
w_i = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 [H_{m-1}(x_i)]^2}.
\]

In the logistic Box-Cox model, the value of \(\beta_1\) reflects the association between the outcome and the predictor. If we only have a weak association between \(Y\) and \(X\), it may be hard to get a precise estimate of the relationship. Therefore, for small values of \(\beta_1\), we may obtain a large variance for the estimated \(\lambda\). As we can see when \(\beta_1 = 0\), all values of \(\lambda\) yield the same distribution of \((Y|X)\). So next we look at this mathematically. Since
\[
\lim_{\beta_1 \to 0} p(x) = \exp(\beta_0) \frac{\exp(\beta_0)}{1 + \exp(\beta_0)},
\]

based on Lebesgue’s Dominated Convergence Theorem [3], we have

\[
\lim_{b_1 \to 0} C_{33} = \left[ \frac{\exp(b_0)}{(1 + \exp(b_0))^2} \right]^2 \left[ \int_{-\infty}^{+\infty} \phi(z) dz \int_{-\infty}^{+\infty} v^2 \phi(z) dz - \left( \int_{-\infty}^{+\infty} v \phi(z) dz \right)^2 \right],
\]

and

\[
\lim_{b_1 \to 0} \beta_1^2 \det(I_1(\theta)) = \left[ \frac{\exp(b_0)}{(1 + \exp(b_0))^2} \right]^3 \left[ \int_{-\infty}^{+\infty} \phi(z) dz \int_{-\infty}^{+\infty} v^2 \phi(z) dz \int_{-\infty}^{+\infty} \left( \frac{\partial v}{\partial \lambda} \right)^2 \phi(z) dz \\
- \int_{-\infty}^{+\infty} \phi(z) dz \int_{-\infty}^{+\infty} v \frac{\partial v}{\partial \lambda} \phi(z) dz \int_{-\infty}^{+\infty} v \frac{\partial v}{\partial \lambda} \phi(z) dz \\
- \int_{-\infty}^{+\infty} v \phi(z) dz \int_{-\infty}^{+\infty} v \phi(z) dz \int_{-\infty}^{+\infty} \left( \frac{\partial v}{\partial \lambda} \right)^2 \phi(z) dz \\
+ \int_{-\infty}^{+\infty} v \phi(z) dz \int_{-\infty}^{+\infty} \frac{\partial v}{\partial \lambda} \phi(z) dz \int_{-\infty}^{+\infty} v \frac{\partial v}{\partial \lambda} \phi(z) dz \\
+ \int_{-\infty}^{+\infty} \frac{\partial v}{\partial \lambda} \phi(z) dz \int_{-\infty}^{+\infty} v \phi(z) dz \int_{-\infty}^{+\infty} v \frac{\partial v}{\partial \lambda} \phi(z) dz \\
- \int_{-\infty}^{+\infty} \frac{\partial v}{\partial \lambda} \phi(z) dz \int_{-\infty}^{+\infty} \frac{\partial v}{\partial \lambda} \phi(z) dz \int_{-\infty}^{+\infty} v^2 \phi(z) dz \right].
\]

Therefore, we can get that \( \text{Avar}(\hat{\lambda}) = O(\beta_1^{-2}) \). And when the association between outcome and predictor is not strong, a relatively large sample size may be required for reasonable estimation of \( \lambda \). Due to the complexity of the integrations in the Fisher information matrix, we cannot get closed-form expressions in general, but we can carry out the numerical calculation across a comprehensive range of settings.

### 3.2.2 Factors in Epidemiological Studies

In a typical epidemiological study, there are four main factors to consider,

1. factor A: the shape of exposure distribution;
2. factor B: the shape of exposure-disease relationship;

3. factor C: the disease rarity;

4. factor D: the strength of exposure-disease association.

We vary the levels of the four factors factorially in order to evaluate estimator performance systematically and comprehensively. Let the exposure variable, $X$,

![Figure 3.2: Distributions of the exposure variable.](image)

follow the log-normal($\mu, \sigma^2$) distribution. Without loss of generality, we fix its 95-th percentile as 1, and let $\sigma$ vary as 0.5, 1 and 2 to represent different levels of skewness. Figure 3.2 shows the selected distributions of $X$, which range from mild skewness to severe skewness. As for B, we let the shape parameter, $\lambda$, in the Box-Cox transformation vary as 0, 0.5, 1 and 2, which corresponds to log, square root, linear and square functions respectively. We use the probability of disease at the 5-th percentile of the exposure to indicate the disease rarity, denoting this as $P_1$. For the strength of exposure-disease association, we consider the ratio of the probability of the disease at 95-th percentile of the exposure to the probability of
the disease at 5-th percentile, which is denoted as

\[ R = \frac{\Pr(Y = 1 | X \text{ is at 95-th percentile})}{\Pr(Y = 1 | X \text{ is at 5-th percentile})}. \]  

(3.9)

We let \( P_1 \) vary as 0.02 and 0.1 to represent low and moderate disease prevalence respectively, and let \( R \) vary as 1.1, 2 and 5 to represent weak, medium and strong associations respectively.

Figure 3.3 shows the disease risk as a function of the exposure in the above-described settings. In each panel, the distribution of exposure and disease rarity is fixed. The risk functions vary due to the difference of the shape parameters in the model and the risk ratios indicating the strength of the association. We can see that given the distribution of exposure, as the exposure-disease association becomes stronger, the risk differences between different shape parameters at the same exposure level become larger. Also, given the association, as the distribution becomes more skewed, the risk differences between different shape parameters at the same exposure level become larger. These indicate that the skewness and the strength of association may be related to the precision in estimating the shape parameter.

### 3.2.3 Average Predictive Effect

In the logistic Box-Cox model, the shape parameter, \( \lambda \), and the association parameter, \( \beta_1 \), jointly determine the slope of log odds at \( X = x \) as we have

\[ \left. \frac{d(\text{logit}(P))}{dX} \right|_{X=x} = \beta_1 x^{\lambda-1}, \]  

(3.10)

where \( P = \Pr(Y|X) \). Due to the difficulty in direct interpretation of this slope and the shape parameter, \( \lambda \), we are going to discuss another quantity with an easier interpretation. For our model (3.1), assume that \( X \) follows a log-normal(\( \mu, \sigma^2 \))
Figure 3.3: The risks of disease as a function of exposure in the selected settings. The solid curves in each panel correspond to $\lambda = 0$, the dashed curves are for $\lambda = 0.5$, the dotted curves are for $\lambda = 1$, and the dashed-dotted curves are for $\lambda = 2$. The distribution of the exposure for each panel is shown at the bottom of each panel by segmented gray lines.
distribution, so that the average predictive effect can be calculated as

$$
\Delta = E \left[ \frac{d(\logit(P))}{dX} \right] = \beta_1 \int_{-\infty}^{+\infty} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\logit(P) - \mu)^2}{2\sigma^2}} x^{\lambda-1} dx = \beta_1 e^{\frac{\lambda-1}{2}[(\lambda-1)\sigma^2 + 2\mu]}. \tag{3.11}
$$

We can see that the average predictive effect represents average slope in a nonlinear model. When the true model is linear, we have \( \lambda = 1 \), and, based on the above equation (3.11), we have \( \Delta = \beta_1 \). Generally \( \Delta \) is determined by \( \beta_1, \lambda, \mu, \) and \( \sigma \). Under given \( \mu \) and \( \sigma^2 \), an estimate of \( \Delta \) could be written as

$$
\hat{\Delta} = \hat{\beta}_1 e^{\frac{1}{2}[(\lambda-1)\sigma^2 + 2\mu]} \tag{3.12}
$$

When \( \mu \) and \( \sigma^2 \) are unknown, we can estimate them based on the empirical distribution of the predictor, which is independent from the model fitting. As for the asymptotic variance of \( \hat{\Delta} \), based on the multivariate delta method, we have

$$
\text{Avar}(\hat{\Delta}) \approx \begin{pmatrix}
\frac{\partial \Delta}{\partial \beta_1}, & \frac{\partial \Delta}{\partial \lambda}
\end{pmatrix} \Sigma \begin{pmatrix}
\frac{\partial \Delta}{\partial \beta_1}, & \frac{\partial \Delta}{\partial \lambda}
\end{pmatrix}^T = \Delta^2 \left( \frac{\text{Avar}(\hat{\beta}_1)}{\beta_1^2} + \frac{2[(\lambda-1)\sigma^2 + \mu]}{\beta_1} \text{Acov}(\hat{\beta}_1, \hat{\lambda}) + \left[(\lambda-1)\sigma^2 + \mu\right]^2 \text{Avar}(\hat{\lambda}) \right), \tag{3.13}
$$

where

$$
\begin{pmatrix}
\frac{\partial \Delta}{\partial \beta_1}, & \frac{\partial \Delta}{\partial \lambda}
\end{pmatrix} = \begin{pmatrix}
e^{\frac{1}{\lambda}[(\lambda-1)\sigma^2 + 2\mu]}, & e^{\frac{1}{\lambda}[(\lambda-1)\sigma^2 + 2\mu]} \beta_1 e^{\frac{1}{2}[(\lambda-1)\sigma^2 + 2\mu]} \\
[(\lambda-1)\sigma^2 + \mu] \beta_1 e^{\frac{1}{2}[(\lambda-1)\sigma^2 + 2\mu]}, & [(\lambda-1)\sigma^2 + \mu] e^{\frac{1}{2}[(\lambda-1)\sigma^2 + 2\mu]} \end{pmatrix},
$$

and

$$
\Sigma = \begin{pmatrix}
\text{Avar}(\hat{\beta}_1), & \text{Acov}(\hat{\beta}_1, \hat{\lambda}) \\
\text{Acov}(\hat{\beta}_1, \hat{\lambda}), & \text{Avar}(\hat{\lambda})
\end{pmatrix}
$$
is the asymptotic variance-covariance matrix of \( \hat{\beta}_1 \) and \( \hat{\lambda} \). And if our sample size \( n \) is sufficiently large \([10]\), we have

\[
\text{Var}(\hat{\lambda}) \approx \frac{1}{n} \text{Avar}(\hat{\lambda}).
\] (3.14)

Based on equation (3.13), given \( \mu \) and \( \sigma^2 \), we calculate the asymptotic variances of \( \hat{\lambda} \) and that of \( \hat{D} \) in all the settings mentioned above, to evaluate estimator precision under different conditions.

**Figure 3.4:** Asymptotic standard deviation of \( \hat{\lambda} \) over different settings. R1 is for \( R = 1.1 \), R2 is for \( R = 2 \), and R3 is for \( R = 5 \).

Figure 3.4 shows that the asymptotic variance of \( \hat{\lambda} \) tends to decrease as the exposure-disease association becomes stronger, as the skewness of distribution of exposure becomes more severe, as the disease becomes more common, and as the shape parameter increases. As for the asymptotic variance of \( \hat{\lambda} \) in Figure 3.5, we observe the same trend for the disease rarity and for the shape parameter, but the opposite trend for more skewed distribution. As an example, in the rare disease case (\( P_1 = 0.02 \)) and lower risk ratio case (\( R = 1.1 \)), ASD(\( \hat{\lambda} \)) doubles when the distribution of \( X \) changes from log-normal\((-0.82, 0.25)\) to log-normal\((-1.64, 1)\) and increases around 20 times when the distribution of \( X \) changes from log-normal\((-1.64, 1)\) to log-normal\((-3.29, 4)\). As for the relationship of ASD(\( \hat{\lambda} \)) with the strength of the association, for mildly-skewed distribution of \( X \), ASD(\( \hat{\lambda} \)) decrease as the associa-
Figure 3.5: Asymptotic standard deviation of $\hat{\Delta}$ over different settings. R1 is for $R = 1.1$, R2 is for $R = 2$, and R3 is for $R = 5$. The lines with solid dots are for $P_1 = 0.02$; the lines with open circles are for $P_1 = 0.1$; the solid lines are for $\lambda = 0$; the dashed lines are for $\lambda = 0.5$; the dotted lines are for $\lambda = 1$; and the dashed-dotted lines are for $\lambda = 2$. From left to right in the third panel, figure is only shown within the range of $[0, 200]$ in vertical axis for better comparison with its left two figures and we plot it again in the fourth panel in its complete range.

...tion becomes stronger. However, for a very skewed distribution, it seems ASD($\hat{\Delta}$) tends to increase for stronger associations for most of the cases.

Figures 3.6 and 3.7 show us the comparison between ASD($\hat{\lambda}$) and ASD($\hat{\Delta}$). We can see that under the same distribution of exposure, as the shape parameter $\lambda$ increases, both ASD($\hat{\lambda}$) and ASD($\hat{\Delta}$) decrease. Under the mildly-skewed distribution of exposure, ASD($\hat{\lambda}$) seems always be greater than ASD($\hat{\Delta}$), especially when the exposure-disease association is weak. As an example, under the case with rare disease ($P_1 = 0.02$), low risk ratio ($R = 1.1$) and mild skewness of exposure, no matter what the shape parameter is, ASD($\hat{\lambda}$) is around 10 times more than ASD($\hat{\Delta}$). This indicates that to achieve the similar precision in the parameter estimates, we need 100 times more observations for estimating $\lambda$ than for $\Delta$. However, under very skewed distribution of exposure, ASD($\hat{\lambda}$) is much less than ASD($\hat{\Delta}$) when $\lambda$ is less than 1. Especially under the case with rare disease ($P_1 = 0.02$), low risk...
Figure 3.6: The asymptotic standard deviation of the estimated average predictive effect, $\hat{\Delta}$, and the asymptotic standard deviation of $\hat{\lambda}$ under $P_1 = 0.02$ and different values of factor A, B, and D. R1 is for $R = 1.1$, R2 is for $R = 2$, and R3 is for $R = 5$. The lines with solid dots represent ASD($\hat{\lambda}$) and the lines with open circle represent ASD($\hat{\Delta}$). The range in the vertical axis is limited from 0 to 660, except for the left one in the bottom row due to its large values.

ratio ($R = 1.1$) and $\lambda = 0$, ASD($\hat{\Delta}$) is around 20 times more than ASD($\hat{\lambda}$). This exploration tells us that when the exposure distribution is mildly-skewed, inference on $\Delta$ requires less data and affords a more direct interpretation. However, on the other side, when exposure distribution is very skewed, it may not always be cost effective to do inference on the average predictive effect instead of $\lambda$, due to the inflation of the variance of the estimate of $\Delta$.

Another quantity of interest is the coefficient of variation (CV) of $\hat{\Delta}$, which is
Figure 3.7: The asymptotic standard deviation of the estimated average predictive effect, $\hat{\Delta}$, and the asymptotic standard deviation of $\hat{\lambda}$ under $P_1 = 0.02$ and different values of factor A, B, and D. R1 is for $R = 1.1$, R2 is for $R = 2$, and R3 is for $R = 5$. The lines with solid dots represent ASD($\hat{\lambda}$) and the lines with open circle represent ASD($\hat{\Delta}$). The range in the vertical axis is limited from 0 to 300, except for the left one in the bottom row due to its large values.

Figure 3.8 shows the coefficient of variation of $\hat{\Delta}$ at different levels of the four factors. We can see that CV of $\hat{\Delta}$ tends to decrease as the disease becomes more common, as the exposure-disease association becomes stronger, and as $\lambda$ increases. However, when the distribution of the exposure becomes more skewed, the CV of $\hat{\Delta}$ increases considerably. Therefore, for very skewed exposure distributions, it is

$$H(\hat{\Delta}) = \frac{\sqrt{\text{Avar}(\hat{\Delta})}}{\hat{\Delta}} =$$

$$= \sqrt{\frac{\text{Avar}(\hat{\beta}_1)}{\beta_1^2} + \frac{2[(\lambda - 1) \sigma^2 + \mu]}{\beta_1} \text{Acov}(\hat{\beta}_1, \hat{\lambda}) + [(\lambda - 1) \sigma^2 + \mu]^2 \text{Avar}(\hat{\lambda})}.$$
Figure 3.8: Coefficient of variation of $\hat{\Delta}$ over different settings. R1 is for $R = 1.1$, R2 is for $R = 2$, and R3 is for $R = 5$.

harder to determine the true signal of $\Delta$ given the same sample size and other parameter values. This doubly confirms that the preferable inference condition for $\Delta$ is under mildly-skewed exposure distribution when exposure is perfectly measured.

We know that $\Delta$ represents the average slope of the effect of the exposure. If a logit-linear model only contains the main effect of the predictor, the average predictive effect is equal to the regression coefficient in the simple logistic regression. So we conjecture that when we fit a logit-linear model to a logit-nonlinear relationship, though it is a misspecified model, the estimated slope parameter may still be a good estimate of the average predictive effect. The simple logistic model can be written as

$$\text{logit} \left( \Pr (Y = 1 | X = x) \right) = \gamma_0 + \gamma_1 x.$$  

We would like to determine if the large-sample limiting slope, $\gamma_1^*$, from fitting the simple logistic regression is close to the average predictive effect and to compare the asymptotic variance of the estimated average predictive effect with that of the estimate of $\gamma_1$. Since we assume independent and identical distributions of the observations, we focus our discussion on just one observation. The misspecified
log likelihood function for a single observation is
\[ I_1(\gamma_0, \gamma_1 | X, Y) = Y(\gamma_0 + \gamma_1 X) - \log(1 + \exp(\gamma_0 + \gamma_1 X)). \]

To get the large-sample limiting coefficients, \( \gamma_0^* \) and \( \gamma_1^* \), we need to find the maximizers for the expectation of the log likelihood function. That is,
\[
(\gamma_0^*, \gamma_1^*) = \arg\max_{(\gamma_0, \gamma_1) \in \mathbb{R}^2} \left\{ E(I_1(\gamma_0, \gamma_1)) \right\}.
\]

We can obtain the maximizers based on setting the first partial derivative equations equal to 0.
\[
\frac{\partial E(I_1)}{\partial \gamma_i} \bigg|_{\gamma_0 = \gamma_0^*, \gamma_1 = \gamma_1^*} = E \left( \frac{\partial I_1}{\partial \gamma_i} \right) \bigg|_{\gamma_0 = \gamma_0^*, \gamma_1 = \gamma_1^*} = 0,
\]
where \( i = 1, 2 \). That is, we have
\[
E \left[ \begin{pmatrix} 1 \\ X \end{pmatrix} \left( Y - \frac{\exp(\gamma_0^* + \gamma_1^* X)}{1 + \exp(\gamma_0^* + \gamma_1^* X)} \right) \right] = 0.
\]

And, after using the double expectation theorem and taking the expectation on \( Y \) firstly, we have
\[
E \left[ \begin{pmatrix} 1 \\ X \end{pmatrix} \left( P - \frac{\exp(\gamma_0^* + \gamma_1^* X)}{1 + \exp(\gamma_0^* + \gamma_1^* X)} \right) \right] = 0, \tag{3.15}
\]
where
\[
P = E(Y|X) = \frac{\exp \left( \beta_0 + \beta_1 \frac{x_i^4 - 1}{x} \right)}{1 + \exp \left( \beta_0 + \beta_1 \frac{x_i^4 - 1}{x} \right)}.
\]

Due to the misspecified likelihood, the inverse of the Fisher Information matrix can no longer provide the asymptotic variances of the parameters. Instead we use the sandwich estimation idea \([15, 55]\), whereby
\[
\text{Avar}(\hat{\gamma}) = J_1^{-1}(\gamma^*) V_1(\gamma^*) J_1^{-1}(\gamma^*), \tag{3.16}
\]
where \( J_1 = E(H(I_1)), H(I_1) \) is the Hessian matrix, \( V_1 = \text{Var}(\nabla I_1), \) and \( \gamma^* = (\gamma_0^*, \gamma_1^*)^T \).

We calculate this asymptotic variance as follows. First the gradient of the log-
likelihood is
\[ \nabla l_1(\gamma) = \left( \frac{\partial l}{\partial \gamma_0}, \frac{\partial l}{\partial \gamma_1} \right) = \left( \begin{array}{c} Y - P^* \\ (Y - P^*)X \end{array} \right), \]
where \( P^* = \exp(\gamma_0 + \gamma_1 X) / (1 + \exp(\gamma_0 + \gamma_1 X)) \). Then its variance is
\[ V_1(\gamma) = \Var(\nabla l_1(\gamma)) \]
\[ = \mathbb{E} \left( \nabla l_1(\gamma) \right)^T \nabla l_1(\gamma) \]
\[ = \begin{pmatrix} \mathbb{E} [(Y - P^*)^2] & \mathbb{E} [(Y - P^*)^2 X] \\ \mathbb{E} [(Y - P^*)^2 X] & \mathbb{E} [(Y - P^*)^2 X^2] \end{pmatrix} \]
\[ = \begin{pmatrix} \mathbb{E} [(P - 2PP^* + P^2)] & \mathbb{E} [(P - 2PP^* + P^2)X] \\ \mathbb{E} [(P - 2PP^* + P^2)X] & \mathbb{E} [(P - 2PP^* + P^2)X^2] \end{pmatrix}. \]

Secondly, the Hessian matrix of the likelihood is
\[ H(l_1) = \begin{pmatrix} \frac{\partial^2 l}{\partial \gamma_0^2} & \frac{\partial^2 l}{\partial \gamma_0 \partial \gamma_1} \\ \frac{\partial^2 l}{\partial \gamma_1 \partial \gamma_0} & \frac{\partial^2 l}{\partial \gamma_1^2} \end{pmatrix} = \begin{pmatrix} -P^* (1 - P^*) & -P^* (1 - P^*)X \\ -P^* (1 - P^*)X & -P^* (1 - P^*)X^2 \end{pmatrix}. \]

Then its expectation is
\[ J_1(\gamma) = -\mathbb{E}(H(l_1)) = \begin{pmatrix} \mathbb{E}(P^* (1 - P^*)) & \mathbb{E}(P^* (1 - P^*)X) \\ \mathbb{E}(P^* (1 - P^*)X) & \mathbb{E}(P^* (1 - P^*)X^2) \end{pmatrix}. \]

We need to solve equation (3.15) to get the large-sample limiting coefficient estimate of \( \gamma \) and then to calculate equation (3.16) to get the asymptotic variance of the MLE of \( \gamma \). Since we cannot get the closed-form solution for both of them, we do both based on numerical calculation. First of all, given \( \mu \) and \( \sigma^2 \), we simulate \( N \) samples from the distribution of \( X \), and use the sample mean to approximate the expectation. That is, for equation (3.15), we solve
\[ \frac{1}{N} \sum_{i=1}^{N} \left[ \begin{pmatrix} 1 \\ x_i \end{pmatrix} \right] (p_i - p_i^*) = 0, \quad (3.17) \]
where
\[ p_i = E(Y | X = x_i) = \frac{\exp \left( \beta_0 + \beta_1 \frac{x_i}{\lambda} \right)}{1 + \exp \left( \beta_0 + \beta_1 \frac{x_i}{\lambda} \right)}, \]

and
\[ p_i^\ast = \frac{\exp (\gamma_0^\ast + \gamma_1^\ast x_i)}{1 + \exp (\gamma_0^\ast + \gamma_1^\ast x_i)}. \]

If we treat \((x_i, p_i)_{i=1, \ldots, N}\) as our data, then the equation (3.17) can be solved by standard logistic regression software, such as the `glm()` function in the R software [39], which allows us to pretend that the outcome isn’t binary. We call the standard errors of the coefficients reported from the `glm()` function as the simulation errors since these errors are introduced by the limitation of a finite simulated sample size.

And for equation (3.16), we get
\[
\text{Avar}(\hat{\gamma}) \approx \hat{f}_1^{-1}(\gamma') \tilde{V}_1(\gamma') \hat{f}_1^{-1}(\gamma'), \tag{3.18}
\]

where
\[
\hat{f}_1(\gamma') = \left( \frac{1}{N} \sum_{i=1}^{N} [p_i^\ast (1 - p_i^\ast)], \frac{1}{N} \sum_{i=1}^{N} [p_i^\ast (1 - p_i^\ast)x_i] \right),
\]

and
\[
\tilde{V}_1(\gamma') = \left( \frac{1}{N} \sum_{i=1}^{N} [(p_i - 2p_ip_i^\ast + p_i^\ast)^2] \right), \frac{1}{N} \sum_{i=1}^{N} [(p_i - 2p_ip_i^\ast + p_i^\ast x_i)] \right), \]

In our example, we choose \(N = 100,000\). The bias between the average predictive effect, \(\Delta\), and the large-sample limiting coefficient, \(\gamma_1^\ast\), has a clear pattern shown in Figure 3.9. Given the exposure distribution, we see that \(\gamma_1^\ast\) is less than \(\Delta\) when \(\lambda < 1\), \(\gamma_1^\ast\) is greater than \(\Delta\) when \(\lambda\) is above 1, and they exactly agree with each other when \(\lambda = 1\). Also the bias is smaller when \(\lambda\) is closer to 1 and when the disease is more common. However, as the distribution of the exposure becomes more skewed, the bias increases, especially under \(\lambda = 0\). As an example, under the severely skewed distribution of exposure, the bias can be larger than 60 when
Figure 3.9: The average predictive effect, $\Delta$, and the large-sample limiting coefficient, $\gamma_1^\ast$, from the simple logistic model under different settings. The lines with solid dots are for $P_1 = 0.02$; the lines with open circles are for $P_1 = 0.1$; the solid lines are for $\lambda = 0$; the dashed lines are for $\lambda = 0.5$; the dotted lines are for $\lambda = 1$; and the dashed-dotted lines are for $\lambda = 2$; the gray line is the 45 degree line. From left to right, in the third panel, the vertical axis of the figure is restricted to $[0,6]$ for better comparison with the left two figures. We plot it again in the fourth panel over its complete range.

$P_1 = 0.1$ and $\lambda = 0$. Therefore, the conditions producing small biases should be that (1), the exposure variable is not very skewed, and (2), the shape parameter is close to 1. Going beyond the analysis of the bias, Figure 3.10 shows the comparison of asymptotic variances between $\hat{\Delta}$ and $\hat{\gamma}_1$. In most settings, the asymptotic variance of $\hat{\gamma}_1$ is less than the asymptotic variance of $\hat{\Delta}$. Also, $A_{\text{var}}(\hat{\gamma}_1)$ is much more stable across different distribution of exposure, disease rarity, risk ratio and the change of shape parameters. Therefore, under mildly-skewed exposure distributions and shape parameter close to 1, we could consider using $\hat{\gamma}_1$ to replace $\hat{\Delta}$, since the variance is substantially reduced even though we may incur some bias.

For the prostate cancer data from [6], we estimate the average predictive effect of the phosphatase as 3.59 with 95% CI $[0.52, 9.78]$ from the 500 bootstrap samples. This tells us that one King-Armstrong unit increase in the level of the acid phosphatase may introduce more cases of nodal involvement in the prostate cancer patients. At the same time, the estimated slope coefficient from the simple logistic regression model is 2.04 with 95% CI $[-0.11, 8.16]$ from the bootstrap, indicating
explore the biases produced under different degrees of measurement errors across different settings. We may get biased parameter estimates in the logistic Box-Cox model. For non-precisely measured data, the measurement errors in the predictor could distort the true relationship between the outcome, \( Y \), and the predictor, \( X \). Therefore, we may get biased parameter estimates in the logistic Box-Cox model. We explore the biases produced under different degrees of measurement errors across different models.

**Figure 3.10:** The asymptotic standard deviation of \( \hat{\gamma}_1 \) from the misspecified model and the asymptotic standard deviation of \( \hat{\lambda} \) under different settings. R1 is for \( R = 1.1 \), R2 is for \( R = 2 \), and R3 is for \( R = 5 \). The lines with solid dots are for \( P_1 = 0.02 \) and the lines with open circles are for \( P_1 = 0.1 \). The solid lines are for ASD(\( \hat{\lambda} \)), and the dashed lines are for ASD(\( \hat{\gamma}_1 \)). The range of the horizontal axis is limited from 0 to 120, except for the left two panels in the bottom row.

some evidence of an association.

### 3.3 Measurement Errors in the Model

For non-precisely measured data, the measurement errors in the predictor could distort the true relationship between the outcome, \( Y \), and the predictor, \( X \). Therefore, we may get biased parameter estimates in the logistic Box-Cox model. We explore the biases produced under different degrees of measurement errors across different models.
all of the proposed settings. Denote $X^\ast$ as the measured predictor, which contains the multiplicative measurement error (3.3). Given $X \sim \log$-normal$(\mu, \sigma^2)$, we have $X^\ast \sim \log$-normal$(\mu, \sigma^*_x)$, with $\sigma^2 = \sigma^2 + \sigma^2_e$. We can describe the Box-Cox model with the binary outcome, $Y$, and the predictor measurement, $X^\ast$, as

$$\text{logit}[E(Y = 1 | X^\ast = x^\ast)] = \rho_0 + \rho_1 x^\ast(h).$$

(3.19)

We assume that $(\rho_0, \rho_1, h)$ is in the same parameter space with $(\beta_0, \beta_1, \lambda)$. Presuming the model (3.1) describes the underlying relationship between the outcome, $Y$, and the predictor, $X$, the above model (3.19) is then a misspecified model. Estimators that maximize the misspecified likelihood will tend to their large-sample limits as the sample size goes to infinity [55]. Therefore, we focus our discussion on these large-sample limits. Denote $(\rho_0^\ast, \rho_1^\ast, h^\ast)$ as the large-sample limits of the estimates of $(\rho_0, \rho_1, h)$. To look into the distortion of the parameter estimates from measurement errors, we numerically evaluate the bias in the parameter estimation and also the bias in the disease probability estimation. We can evaluate these by solving the following optimization equation,

$$(\rho_0^\ast, \rho_1^\ast, h^\ast) = \arg\max_{(\rho_0, \rho_1, h) \in \Theta} \{E(l(\rho_0, \rho_1, h|Y, X^\ast))\},$$

(3.20)

where the likelihood function for a single observation is

$$l(\rho_0, \rho_1, h|Y, X^\ast) = Y \left[ \rho_0 + \rho_1 \left( \frac{X^{\ast h} - 1}{h} \right) \right] - \log \left[ 1 + \exp \left( \rho_0 + \rho_1 \left( \frac{X^{\ast h} - 1}{h} \right) \right) \right].$$

We can obtain the maximizer, $(\rho_0^\ast, \rho_1^\ast, h^\ast)$, in equation (3.20) based on the gradient at $(\rho_0^\ast, \rho_1^\ast, h^\ast)$ being a zero vector. That is,

$$\nabla E(l(\rho_0, \rho_1, h|Y, X^\ast))_{\rho = \rho^\ast, h = h^\ast} = 0.$$

Based on the equations corresponding to $\rho_0$ and $\rho_1$, we have

$$E \left[ \left( \frac{1}{X^{\ast h^\ast} - 1} \right) \left( Y - \frac{\exp \left( \rho_0^\ast + \rho_1^\ast \frac{X^{\ast h^\ast} - 1}{h^\ast} \right)}{1 + \exp \left( \rho_0^\ast + \rho_1^\ast \frac{X^{\ast h^\ast} - 1}{h^\ast} \right)} \right) \right] = 0.$$
Denote \( P = \mathbb{E}(Y|X) \). Then we have

\[
\mathbb{E} \left[ \left( \frac{1}{X^{*h^*} - 1} \right) \left( P - \frac{\exp\left( \rho_0^* + \rho_1^* \frac{X^{*h^*} - 1}{h^*} \right)}{1 + \exp\left( \rho_0^* + \rho_1^* \frac{X^{*h^*} - 1}{h^*} \right)} \right) \right] = 0. \tag{3.21}
\]

Define a new variable \( V = \frac{(X^{*h^*} - 1)}{h^*} \). We can rewrite equation (3.21) as

\[
\mathbb{E} \left[ \left( \frac{1}{V} \right) \left( P - \frac{\exp\left( \rho_0^* + \rho_1^* V \right)}{1 + \exp\left( \rho_0^* + \rho_1^* V \right)} \right) \right] = 0. \tag{3.22}
\]

For each \( (\beta_0, \beta_1, \lambda) \), we use the following algorithm to calculate the theoretical values of \( \rho_0^*, \rho_1^* \) and \( h^* \) given different degrees of measurement error. Set an upper bound on \( \lambda \), denoted as \( B_u \). Then we do a grid search in the interval \([0, B_u]\) with 1000\([B_u]\) equally spaced values, fit the model at \( \lambda \) equal to each value, and select the particular value which can maximize the likelihood function. We let \( \sigma_e \) vary as 0.1 and 0.01 to represent two different degrees of measurement error. When \( \sigma_e = 0.1 \), the measured values are most likely in a range of 82% – 122% of the true values due to multiplicative measurement errors, and we treat this situation as a representative of rough measurements. We use \( \sigma_e = 0.01 \) to typify fairly precise measurements since the range shrinks to 98% – 102% of the true values.

Figure 3.11 shows that under the rough measurements of the exposure, the large-sample limit of the shape parameter, \( h^* \), in the misspecified model (3.19) is always smaller than the shape parameter, \( \lambda \), in the true model (3.1). And the extent of the attenuation is associated with the true value of the shape parameter, \( \lambda \), the strength of the exposure-disease association, and the skewness of the exposure distribution. Within these three factors, the value of \( \lambda \) seems to play the most prominent role in the extent of attenuation, while the effects of the strength of the exposure-disease association and the skewness of the exposure distribution can only be seen clearly at a relatively large value of \( \lambda \). As an illustration, when \( \lambda \) is less than 1, the attenuation is less than 0.025 regardless of the other conditions. As \( \lambda \) increases, the attenuation becomes more severe. When \( \lambda = 2 \), the attenuation ranges from 0 to 0.6. Also in this situation, we can see a clear pattern that as the strength of exposure-disease association increases, the attenuation increases, and as
Figure 3.11: The difference between the shape parameter, $\lambda$, in the true model (3.1) and the large-sample limit, $h^*$, in the misspecified model (3.19) in the presence of the multiplicative measurement error with $\sigma_e = 0.1$ across different settings. The solid line is under $\sigma = 0.5$ in the exposure distribution, the dashed line is under $\sigma = 1$ in the exposure distribution, and the dashed dotted line is under $\sigma = 2$ in the exposure distribution. The lines with solid dots are for $P_1 = 0.02$, and the lines with open circles are for $P_1 = 0.1$. 
Figure 3.12: The difference between the shape parameter, \( \lambda \), in the true model (3.1) and the large sample limit, \( h^* \), in the model (3.19) in the presence of the multiplicative measurement error \( \sigma_e = 0.01 \) across different settings. The solid line is under \( \sigma = 0.5 \) in the exposure distribution, the dashed line is under \( \sigma = 1 \) in the exposure distribution, and the dashed dotted line is under \( \sigma = 2 \) in the exposure distribution. The lines with solid dots are for \( P_1 = 0.02 \) and the lines with open circles are for \( P_1 = 0.1 \).
Figure 3.13: The difference between the true probability and the estimated probability from the misspecified model (3.19) across different settings under $\sigma_e = 0.1$. The solid curves are for $R = 1.1$, the dashed curves are for $R = 2$, and the dashed-dotted curves are for $R = 5$. The curves with triangles are for $P_1 = 0.02$, and the curves with open circles are for $P_1 = 0.1$. 
Figure 3.14: The difference between the true probability and the estimated probability from the misspecified model (3.19) across different settings under $\sigma_x = 0.01$. The solid curves are for $R = 1.1$, the dashed curves are for $R = 2$, and the dashed-dotted curves are for $R = 5$. The curves with triangles are for $P_l = 0.02$, and the curves with open circles are for $P_l = 0.1$. The figures in all the panels except the right bottom panel are plotted in the range of $(-0.010, 0.010)$ in vertical axis for better comparison. The figure in the right bottom panel is plotted in a wider range in vertical axis due to its special large values on vertical axis.
the exposure distribution becomes more skewed, the attenuation increases as well. Under $\sigma_e = 0.01$, we have similar results except, unsurprisingly, the attenuation is milder under each setting, as shown in Figure 3.12. If we look at the bias in estimating $\Pr(Y = 1|X)$ via the fitted $(Y|X^*)$ relationship, we can see that the fitted probability is most biased when the distribution of exposure is very skewed and the shape parameter $\lambda$ is 2, which is shown in Figure 3.14. In this scenario, under $\sigma_e = 0.1$, we could estimate 5 fewer cases per 100 samples than the truth when the measured exposure is around 0.75, and could estimate around 8 more cases per 100 samples than the truth when the measured exposure is close to 1.5, which is far in the right tail. Also it seems that the fitted probability is always inflated at small values of exposure but then attenuated at large values of the exposure. We

![Figure 3.15](image_url)

**Figure 3.15:** The difference between the true probability and the fitted one from the misspecified model (3.19).

know that in simple logistic regression, measurement errors in the predictor tend to attenuate fitted probabilities towards 0.5 [50]. Figure 3.15 indicates that we may have the same attenuation pattern for this nonlinear logistic regression model.

As for a data example, we continue to use the prostate cancer data [6]. Firstly, we simulate 100 sets of measurement errors with the same sample size of the prostate cancer data from log-normal $(0, \sigma_e^2)$ distribution with $\sigma_e = 0.1$ and 0.01
Figure 3.16: The estimated probability of the nodal involvement in the prostate cancer patients via the fitted logistic Box-Cox model. In the left panel, the solid curve comes from the original data, the dashed curves come from three randomly selected simulated data sets containing the multiplicative measurement error with $\sigma_e = 0.1$ separately, and the dotted curves come from three randomly selected simulated data sets containing the multiplicative measurement error with $\sigma_e = 0.01$ separately. In the right panel, the solid curve comes from the original data, the dashed curve is the average of the fitted probability from 100 simulated data sets containing the multiplicative measurement error with $\sigma_e = 0.1$, and the dotted curve is the average of the fitted probability from 100 simulated data sets containing the multiplicative measurement error with $\sigma_e = 0.01$. Note in both panel the dotted curves are right under the solid curve.

respectively. Secondly, we multiply the acid phosphatase predictor in the prostate cancer data with each set of simulated errors. In this way, we have simulated data sets with different degrees of multiplicative measurement errors, while the original prostate cancer data play the role of data without measurement errors. For each simulated-error data set, we fit both the logistic Box-Cox model and the simple logistic model, and summarize the results based on the degree of the measurement errors. Figure 3.16 shows us the estimated probability of the nodal involvement given the level of acid phosphatase, using the fitted logistic Box-Cox model. We
can see that when the standard deviation of the measurement error is 0.01, it is hard to see the difference between the estimated probability based on data with and without measurement errors from the three randomly selected simulated data sets and from the average of the 100 simulated data sets. When the standard deviation of the measurement error distribution increases to 0.1, the difference in the estimated probability given the phosphatase level is apparent, but is small especially when the phosphatase is below its 95th percentile, 1.12. The change pattern of the difference exactly agree with what we find from numerical analysis. Also based on the numerical results, since the phosphatase distribution is not very skewed and \( \hat{\lambda} = 0 \), we expect the bias could be reduced if there were more samples in the original data. However, there is a lot of variation across the simulated datasets mainly due to Monte Carlo errors. For the three randomly selected simulated data sets, there can be attenuation or inflation at the large value of the acid phosphatase. As an example, when the measurement of the level of acid phosphatase is 1.0 King-Armstrong units, we predict 5 fewer cases of nodal involvement in every 100 prostate cancer patients than the estimated number from the original data in one scenario, 10 fewer cases in another scenario, and 10 more cases in a third scenario. And when the measurement of the level of acid phosphatase is 1.5 King-Armstrong units, the difference in the number of nodal involvement in the prostate cancer patients increases to around 10 for every 100 prostate cancer patients for the three sets of simulated data.

3.4 Discussion

In this chapter, we assume that the predictor, \( X \), follows the log-normal(\( \mu, \sigma^2 \)). The log-scale parameter, \( \mu \), controls the plausible range of \( X \) given \( \sigma \). The shape parameter, \( \sigma \), plays the role of a fraction of the association parameter. To explain more on this, let us first consider the following transformation. Denote \( \eta = \sigma \lambda \) and \( W = \exp(\mu / \sigma + Z) \). Then we have

\[
X^{(\lambda)} = \frac{X^\lambda - 1}{\lambda} = \sigma \frac{\exp\left(\frac{\mu}{\sigma} \eta + Z \eta\right) - 1}{\eta} = \sigma \frac{W^\eta - 1}{\eta} = \sigma W^{(\eta)}.
\]
And the model (3.1) becomes

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \sigma w_i(\eta).$$

We can see that no matter what $\sigma$ is, after transformation, we can always construct a new predictor with shape parameter equal to 1 for which $\beta_1 \sigma$ plays the role of the association parameter. Based on this, we have one direct conclusion. That is, given $\beta_1$, when $\sigma$ goes to 0, $\text{Avar}(\hat{\lambda})$ goes to infinity. This is because we have the $\text{Avar}(\hat{\eta})$ goes to infinity when $\sigma$ goes to 0 based on the calculation in the Appendix, and $\text{Avar}(\hat{\eta}) = \text{Avar}(\sigma \hat{\lambda}) = \sigma^2 \text{Avar}(\hat{\lambda})$. So we know that given the association parameter, if there is only a small variation of the predictor in log-scale, we may not get precise estimation of the shape parameter.

Our exploration of the Box-Cox type model could be naturally extended to the fractional polynomial model. A single covariate fractional polynomial of degree $m$ could be expressed as

$$\phi_m(X|\xi, p) = \xi_0 + \sum_{j=1}^{m} \xi_j X^{(p_j)},$$  \hspace{1cm} (3.23)

where $m$ is a positive integer, $p = (p_1, \cdots, p_m)$ is a real-valued vector of powers with $p_1 < \cdots < p_m$ and $\xi = (\xi_0, \xi_1, \cdots, \xi_m)$ are real-valued coefficients [44]. Also the round bracket notation represents the Box-Tidwell transformation,

$$X^{(p_j)} = \begin{cases} X^{p_j} & \text{if } p_j \neq 0 \\ \ln X & \text{if } p_j = 0. \end{cases}$$

We can see that when $m = 1$, it is equivalent to the Box-Cox family after we linearly transform the coefficients. As for fitting the fractional polynomial model, there is an R contributed package, mph [39], which can work for multiple covariates given the degree for each covariate. As the fractional polynomials can provide quite flexible shapes to fit for nonlinear pattern data, Keogh et al. [26] showed the effect of measurement errors in a proportional hazard model in some specified nonlinear models when fitting fractional polynomial models though simulation studies. This is different from our discussion about the effect of measurement errors in this chap-
ter, since our discussion is within the logistic Box-Cox family model, and our focus is to look at the distortion of the parameter estimates under different degree of random errors. Similar discussion could also be made for the fractional polynomial family models. As there are more complex situations in epidemiologic research, [49] mentioned that splines with knots are good candidate for risk assessment, especially when response attenuates at high exposure level and epidemiologists also prefer a liner exposure-disease shape in the low dosage. Therefore, the spline model is also in our further exploration.

We have explored the average predictive effect and its relationship with the coefficient in the misspecified linear model. As an average slope between logit[E(Y|X)] and X, if the distribution of exposure is not very skewed, the average predictive effect requires much less data to achieve the similar estimator precision compared with the shape parameter λ. So, sometimes it is more worthwhile to use the average predictive effect as an inferential target rather than the original parameters in the model. Another conjecture is that the fitted coefficient from the simple logistic regression may estimate well the average predictive effect from a more complex nonlinear model. In this chapter, using the logistic Box-Cox model as an example, we see that when the distribution of exposure is not very skewed and the shape parameter is close to 1, the large-sample limit of the coefficient is close to the average predictive effect, and the asymptotic variance of the estimated coefficient can be much smaller than that of the average predictive effect. Therefore, we are facing a bias-variance trade off in the model selection. In some conditions, we may prefer using the estimated coefficient from the simple model as an alternative to the estimated average predictive effect from a more complex model.

From our investigation, the measurement errors tend to overestimate the risk at low exposure and underestimate the risk at high exposure. In practice, the risk at low exposure is used to determine threshold for acceptable risk. Therefore, over-estimating risk in the low range would lead to unjustifiably low exposure thresholds/regulatory limits. So for non-precisely measured data, appropriate adjustment is recommended. A convenient way to do this is though a Bayesian hierarchical model. Gustafson [22] and Carroll et al. [9] showed detailed steps on implementing this in nonlinear models using Bayesian methods. As a byproduct, the average predictive effect could be calculated based on the fitted parameters from the Bayesian
hierarchical model. In this situation we may be less interested in the coefficient in the simple logistic model due to its bias caused by measurement error. Also if we fit a misspecified Bayesian hierarchical model with the simple logistic regression as one level, the computational complexity may be very similar to using the nonlinear model as one level.
Chapter 4

Two-step Approach in Matching Process for Longitudinal Matching Covariates in the Presence of Measurement Error

4.1 Background

Epidemiologic studies of exposure-disease relationships are often observational. We would like to compare the risks of diseases between the exposed and non-exposed groups, but due to lack of randomization in the treatment/exposure assignment, there are potential biases in estimating the causal effect, due to the imbalance of the covariates distributions between the treatment and control groups. Matched sampling is a common technique to achieve balance of the observed covariates between the two groups. Matching has been discussed extensively in the literature, especially when it is time-independent. A time-independent matching happens if the treatment units receive the treatment around the same time, and, therefore, we only consider two time points in the design and analysis procedure: (1), the baseline when the matching covariates are measured before any treatment initiation; and (2), after the treatment when the outcome is evaluated. For time-independent
matching, Stuart [52] provided “a structure for thinking about the matching methods and guidance on their use, coalescing the existing research (both old and new) and providing a summary of where the literature on matching methods is now and where it should be headed.”

In the meantime, in some observational studies, the study units receive the treatment at different time points after baseline. In this scenario, a time-dependent matching may be preferred, since if we select a control to match to a treatment unit, we would like to use the value of covariates measured right before the treatment initiation rather than values measured at baseline. The risk-set matching is an example of the time-dependent matching [34, 37]. In a risk-set matching, a study unit who receives treatment at time \( t \) will be matched by another study unit, selected from the set of study units who have not received treatment by time \( t \), but may receive treatment later. In this chapter, we propose a different approach for the time-dependent matching. We would like to model the trajectory from the repeated or longitudinal measurements of the matching variables before the initiation of the treatment, then the matched controls are selected from the units in the control group through the distance measure estimated based on the trajectory. There are two major benefits using the trajectory model before matching: (1), the measurement errors in the observed matching variables can be automatically adjusted for if we use estimated true values from the trajectory model to do the matching; and (2), the distance can be more accurately calculated, since the matching quantities plugged into the distance are more accurate and also we can have a better estimate of the variance-covariance matrix estimated from the trajectory model, which can especially improve the accuracy of the Mahalanobis distance, since the variance-covariance matrix is central to this distance measure. More details regarding the two benefits are described below.

First of all, regarding measurement errors in matched sampling, Steiner et al. [51] demonstrated that “unreliability of measurement can degrade the ability of propensity scores to reduce bias.” Althauser and Rubin [2], McShane et al. [38] provided ways to adjust for measurement errors in the linear regression and conditional logistic regression for matched sampling. As for the time-dependent matching, denote the measured matching variables for the \( i \)th subject as \( \mathbf{Z}_i^\ast \) and the true
values as $\mathbf{Z}_i$. A common trajectory model can be written as

$$
\mathbf{Z}_i(t) = \mathbf{f}(t) + \mathbf{Q}_i \mathbf{b}_i(t) + \mathbf{e}_i(t),
$$

(4.1)

where $\mathbf{Z}_i(t) = \mathbf{f}(t) + \mathbf{Q}_i \mathbf{b}_i(t)$, $\mathbf{f}(t)$ is a function to describe the population average over time, $\mathbf{Q}_i$ is a design matrix for the random effects for the $i$-th subject, $\mathbf{b}_i(t)$ is the vector of random effects for the $i$-th subject, and $\mathbf{e}_i(t)$ is the vector of random errors for the $i$-th subject. The estimated true variables, $\hat{\mathbf{Z}}_i(t)$, from the model will be used for matching. In this way, we can automatically adjust for the measurement errors.

Secondly, as for the distance measures in the time-dependent matching, Lu [37] used the propensity scores to measure the closeness between two units, which are defined as the hazard component based on the Cox proportional model when the event is defined as time to treatment. Other than time-dependent propensity score matching, Li et al. [34] showed us a way to define the Mahalanobis distance for the time-dependent matching, using an observational study of interstitial cystitis as an example. For the interstitial cystitis data, patients were evaluated at baseline and at intervals of approximately every 3 months. Pain, urgency, and nocturnal frequency of voiding were measured at baseline and every three months afterwards. After baseline, some patients receive treatment at different time points. They categorize the measurements as baseline measurement, and later measurements for each patient and pool the data together. That is, they have six quantities used to calculate the variance-covariance matrix, which contains baseline and current pain, urgency, and frequency measurements. We can see that for a control unit, all measurements after the baseline are treated as independent records in calculation of the variance-covariance matrix while the correlation between measurements for a patient working as control has been ignored. This may introduce biased estimates of the distances since it allows replicated counts of samples in estimation of variance-covariance matrices. In our proposed method, the trajectory model can provide better estimates of the variance-covariance matrix for the Mahalanobis distance. Based on (4.1), we have $\text{Var}(\mathbf{Z}_i(t)) = \mathbf{Q}_i \text{Var}(\mathbf{b}_i(t)) \mathbf{Q}_i^T$. Then the Mahalanobis dis-
distance is defined as
\[
d(Z_i(t), Z_j(t)) = (Z_i(t) - Z_j(t))^T \Sigma_i(t)^{-1} (Z_i(t) - Z_j(t)),
\]
where \(\Sigma_i(t) = \text{Var}(Z_i(t))\), the \(i\)-th unit receives treatment at time \(t\) and the \(j\)-th unit is in the control group.

Based on the above benefits, we propose to fit the trajectory model before matching and call this approach the two-step matching approach. In section 4.2, we introduce the Baltimore Experience Corps study, in which the two-step method could be applied. In section 4.3, we describe the trajectory model designed for this study. The matching algorithm is given in section 4.4. In section 4.5, we define the treatment effect for the time-dependent matching. In section 4.6, we conduct simulation studies for different trajectory models and compare the matching quality and the bias of treatment effect estimates between the two-step matching approach and the traditional matching approach. The mathematical calculation on the bias introduced by measurement error in the distance matching is described in section 4.7. And we close our discussion in section 4.8.

### 4.2 Baltimore Experience Corps Study

The Baltimore Experience Corps study (EC study), USA [16] concerns a senior volunteer program, which places people 60 and older in meaningful roles in the public elementary schools, with the roles designed to have high impact on children’s outcomes, as well as the health and functional status of the older adults. Along with this program, an observational study, the Experience Corps project 2, was conducted to determine if this new program improves the educational outcomes of children. During the course of the study from Fall 2006 to Spring 2011, a total of 27 Baltimore city public schools participated in the EC program, including 16 that received treatment since the 2006/2007 academic year, 4 that received treatment since the 2007/2008 academic year, 2 that received treatment since the 2009/2010 academic year, and 5 that received treatment in the 2010/2011 academic year. EC schools that meet the certain criteria were selected by a committee of investigators and GHCC in consultation with the city school system [16]. For an EC school, the senior volunteers started in September of the year of treatment.
initiation, to help students in their academic learning. In following March, students took the tests which are used to evaluate their academic performance. The Stanford Achievement Tests Series Tenth Edition (Stanford 10) are used for children under grade 3, and the Maryland School Assessment (MSA) tests are used for Grade 3 and above, in reading and math, to meet the testing requirements of the federal No Child Left Behind Act. The outcome variables of interest are the school-level scores at grade 3, and then the school scores at grade 2 in previous year are treated as matching variables. Since we assume that most students are promoted to the next grade and most of them maintain their academic performance, we can balance school-level student academic performance level between the treatment and the control groups based on their performance in the previous year.

In the educational data setting, the measurement errors can be interpreted as all kinds of deviations from their underlying truth, which is different from the traditional sense of errors made in the process of measurement using some instruments. For example, the school-level test scores can be estimated from the average of individual student scores. But the student individual scores would be affected by student retention, transferring, and/or teacher switching. So we assume the estimated school-level scores based on the average of individual scores are the underlying true school-level scores plus some random error.

### 4.3 Statistical Modeling

In randomized experiments, before treatment initiation ideally the joint distribution of covariates in the treatment group and the joint distribution of the covariates in the control group are close to identical due to the random assignment. However, in observational studies, these distributions may be different. So we denote quantities differently in the treatment group and control group. Let $P_1$ be the distribution of covariates in the treatment schools and $P_2$ be the distribution of covariates in the control schools. We have $N_1$ randomly selected treatment schools from $P_1$ and $N_2$ randomly selected control schools from $P_2$ with $N_2 > N_1$. The matching vector $Z = (z_1, z_2, \ldots, z_K)^T$ contains testing scores for $K$ topics. As an example, the topics could be mathematics, reading, writing, and comprehension. Assume that given the group, the random-effects vector of each school in that group shares the same
variance-covariance matrix of the matching covariates at a given time. And we can define the distance between the treatment school $i$ and the control school $j$ as

$$d_{i,j} = (Z_i(t) - Z_j(t))^T \Sigma_1^{-1}(t) (Z_i(t) - Z_j(t)),$$ (4.2)

where $\Sigma_1(t) = \text{Var}(Z_i(t))$, and the treatment school $i$ receives the treatment right after year $t$. The trajectory model for treatment schools can be described as

$$z_{ik}(t) = \alpha_1(t) + a_{1k}(t) + b_{1i}(t),$$ (4.3)

where $\alpha_1(t)$ is the overall mean for treatment schools, $a_{1k}(t)$ is the random topic effect for the $k$th topic at year $t$, $b_{1i}(t)$ is the school random effect for the $i$th school at year $t$, $i = 1, 2, \cdots, N_1, k = 1, 2, \cdots, K$. Note that the model is only applied to the treatment schools before the treatment initiation. For control schools, we have

$$z_{jk}(t) = \alpha_2(t) + a_{2k}(t) + b_{2j}(t),$$ (4.4)

where $\alpha_2(t)$ is the overall mean for control schools, $a_{2k}(t)$ is the random topic effect for the $k$th topic at year $t$, $b_{2j}(t)$ is the school random effect for the $j$th school at year $t$, $j = 1, 2, \cdots, N_2, k = 1, 2, \cdots, K$. For the control schools, we can model them during the entire study. Assume the random effects follow normal distributions centered at 0 with $\text{Var}(a_{1k}(t)) = \sigma_{a1}^2(t)$, $\text{Var}(a_{2k}(t)) = \sigma_{a2}^2(t)$, $\text{Var}(b_{1i}(t)) = \sigma_{b1}^2(t)$, and $\text{Var}(b_{2j}(t)) = \sigma_{b2}^2(t)$. Assuming the classical type random error, we have

$$z^\ast_{ik}(t) = z_{ik}(t) + e_{ik}(t).$$ (4.5)

We assume that within each group, the random effects and the random errors are mutually independent, and the random errors follow normal distributions centered at 0 with $\text{Var}(e_{ik}(t)) = \text{Var}(e_{jk}(t)) = \sigma_e^2(t)$. Based on our statistical model, we propose a two-step approach for the matching. That is, fit a mixed effects model to all of covariates before treatment initiation. That is, we use the covariates during the entire study for control schools and the covariates before treatment initiation for treatment schools. And then use estimated true school-level scores to do the matching. We will compare this approach with direct matching using the raw scores.
4.4 Matching Algorithm

Due to the longitudinal feature of time-dependent matching, sequential matching is feasible. For sequential matching, we start matching treatment units who receive treatment at earliest time point and after that we match treatment units who receive treatment at the second earliest time points while the previously matched controls are removed from the control pool. In this way, we sequentially match until the last time point of the study. This chronological matching procedure is called sequential matching. However, since we conduct matching without replacement, it may introduce poorly matched pairs in later years, as the treatment schools that received treatment at earlier times have the priority in choosing better matched controls. So we prefer the simultaneous matching approach, which aims to consider all the treatment schools simultaneously. In the simultaneous matching, there are also two options: the “greedy” nearest-available matching or the optimal matching method. In the nearest-available matching, we randomly order the treatment schools, and choose for each treatment school in turn the nearest-available control school available as a match [14, 45]. For this matching strategy, the matching quality may depend on the order of the treatment units to match for. As an example, when there are not many controls available, the competition of getting a well-matched control for a treatment unit is extensive [21]. The treatment unit can get better matched controls if it is matched earlier. A complete matching is defined as each treatment school being matched with a control school. In this work, we particularly assume the complete matching is formed based on matching without replacement. For the optimal matching, we look at all of the possible complete matchings, and choose the one which minimize the global distance. The optimal matching is guaranteed to find the global optimum through finding the path in a minimum cost flow in the network, which was discussed in [21, 24, 42] and is implemented in R package optimatch [39].

In this chapter, we conduct the optimal pair-matching rather than the nearest-available matching. In constructing a complete matching, once a control school is selected to match for a treatment school at a particular year, it will not be considered as a candidate for the remaining unmatched treatment schools. The global
distance is defined as
\[ D = \sum_{i=1}^{N_1} d_{i,j}, \tag{4.6} \]
where \( d_{i,j} \) is defined in equation (4.2). We choose the complete matching which minimizes the global distance.

### 4.5 Treatment Effect

The causal treatment effect can be defined based on the well known counterfactual model [47]. Denote \( Y(1)(t) \) as the response vector of a school after receiving the treatment at year \( t \), and \( Y(2)(t) \) as the response of the same school if it were assigned to the control group at the same year. Let \( T \) represents the treatment assignment with \( T = 1 \) indicating the treatment assignment and \( T = 2 \) indicating the control assignment. We would like to match for \( Z(t) \), which are measured right before the treatment. Assume the treatment assignment is strongly ignorable [43]. That is, the treatment assignment is independent of the potential outcomes conditional on the matching variables, which can be denoted below.

\[
(Y(1)(t), Y(2)(t)) \perp T \mid (Z(t-1) = z(t-1)), \text{ for all } z(t-1), \tag{4.7}
\]

and \( 0 < \Pr(T = 1 | Z(t-1) = z(t-1)) < 1 \), for all \( z(t-1) \). Denote \( E(Y(1)(t) | Z(t-1)) = R_1(Z(t-1)) \) and \( E(Y(2)(t) | Z(t-1)) = R_2(Z(t-1)) \), where \( R_1 \) is the expected treatment surface as a function of \( Z(t-1) \) and \( R_2 \) is the expected control surface as a function of \( Z(t-1) \). Then the treatment effect conditional on \( Z(t-1) \) can be written as \( R_1(Z(t-1)) - R_2(Z(t-1)) \). Assume the parallel response surfaces with the same distance at all years. Then we can further write

\[
R_q(Z(t-1)) = \tau_q + V(Z(t-1)),
\]

where \( V(Z(t-1)) \) is a function to represent the surface parallel to the treatment and control surfaces of the matching vector, and \( q = 1, 2 \). Then the treatment effect can be defined as

\[
\tau = E_1 [R_1(Z(t-1)) - R_2(Z(t-1))],
\]
where $E_1(.)$ means expectation based on the distribution in the treatment group. The estimator of $\tau$ we are going to use is

$$\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N_1} [Y_i(1)(t) - Y_{j_i}(2)(t)],$$

(4.8)

where $t$ is the one year before the treatment time for the treatment school $i$, the control school $j_i$ is the matched to the treatment school $i$, and the complete matching minimizes the global distance.

### 4.6 Simulation Study

To compare the matching quality and the accuracy of the treatment effect estimation between the two-step approach and the traditional matching approach, we carry out the following simulation studies.

There are 200 schools enrolled in a five-year study, and 24 schools are pre-selected as treatment schools. Without loss of generality, school 1 to 6 receive treatment at year 1; school 7 to 12 receive treatment at year 2; school 13 to 18 receive treatment at year 3; and school 19 to 24 receive treatment at year 4. The remaining 176 control schools are never treated during the entire study. There are three test scores recorded from tests in reading, vocabulary, and writing at grade 2. The trajectory models of the treatment schools and the control schools for the school-level scores at grade 2 are described separately below. The trajectory model for the treatment schools is

$$z_{ik}^*(t) = \alpha_{1k}(t) + b_{1i}(t) + e_{1ik}(t),$$

(4.9)

where $i = 1, 2, \cdots, 24$, $k = 1, 2, 3$, $\text{Var}(b_{1i}(t)) = \sigma_b^2(t)$, and $\text{Var}(e_{1ik}(t)) = \sigma_e^2(t)$; the trajectory model for the control schools is

$$z_{jk}^*(t) = \alpha_{2k}(t) + b_{2j}(t) + e_{2jk}(t),$$

(4.10)

where $j = 25, 26, \cdots, 200$, $k = 1, 2$, and 3, $\text{Var}(b_{2j}(t)) = \sigma_b^2(t)$, and $\text{Var}(e_{2jk}(t)) = \sigma_e^2(t)$. In this model, we assume no topic random effects. We can see that $\alpha_{1k}(t)$ and $\alpha_{2k}(t)$ represent the longitudinal fixed effects of the $k$-th topic for treatment
schools and control schools separately, $b_{1i}(t)$ and $b_{2j}(t)$ represent the longitudinal random effects of the $i$-th treatment school and of the $j$-th control school separately, $e_{1ik}(t)$ and $e_{1jk}(t)$ are random errors. The response is just the addition of all the topic scores after receiving the treatment for a treatment school or the addition calculated separately for every year for a control school.

In our simulation settings, we let the fixed effects be constant over years with $\alpha_{1k}(t) = \alpha_k + \eta$ and $\alpha_{2k}(t) = \alpha_k - \eta$, where $\alpha_1 = \alpha_2 = \alpha_3 = 78$ and $\eta$ vary as 0, 0.5, 1 and 2. We also let the random effects be independent of time with $\sigma_{b_1}^2 = 1$, $\sigma_{b_2}^2$ varies as 0.5, 1, and 2, and $\sigma_e^2$ varies as 1/9, 1 and 4. The choices of $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$ are based on some discussion of the ratio of the variances between treatment and control distributions [46].

We simulate 1000 data sets in each scenario. For each data set, we fit only one trajectory model with school intercept random effect instead of fitting two models for treatment group and control group separately since we would like to keep the trajectory model simple and also the sample size for treatment school is usually small. After that we conduct 1–1 optimal matching based on the estimated true school scores. We also conduct 1–1 optimal matching directly from the original school scores. We compare the global distances of the final complete matching from the two methods and also the estimated treatment effects from both methods.

Figures 4.1, 4.2, 4.3, and 4.4 show us that in all of the considered settings the global distances from the two-step approach are smaller than those from the direct matching, which indicate the two-step matching can improve the matching quality in the presence of measurement errors. We say a setting is matching favored if the final selected complete matching is close to perfect matching, which has the global distance equal to 0. For matching favored settings, we can see apparent improvement on the global distances. For example, when $\eta = 0$, it would be easier to find a better matched control for treatment schools since the values of their matching variables are close to each other. In these setting, we can see that no matter what the degree of the measurement errors is, the global distances can be decreased by using the two-step approach. While when $\eta = 2$, the values of matching variables between the treatment and control groups are far away from each other given their standard deviations. Under those conditions, when the degree of measurement er-
Figure 4.1: Boxplot of the global distance from different conditions. A: boxplot of the optimal true global distances; B₁: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_c^2 = \frac{1}{9}$; B₂: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_c^2 = \frac{1}{9}$; C₁: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_c^2 = 1$; C₂: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_c^2 = 1$; D₁: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_c^2 = 4$; and D₂: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_c^2 = 4$. 
Figure 4.2: Boxplot of the global distance from different conditions. A: boxplot of the optimal true global distances; B1: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = \frac{1}{9}$; B2: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = \frac{1}{9}$; C1: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 1$; C2: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 1$; D1: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 4$; and D2: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 4$. 

\[ \sigma_{\text{opt}}^2 / \sigma_{\text{true}}^2 = 1/10, \eta = 0.5 \] 

\[ \sigma_{\text{opt}}^2 / \sigma_{\text{true}}^2 = 1/2, \eta = 0.5 \] 

\[ \sigma_{\text{opt}}^2 / \sigma_{\text{true}}^2 = 1, \eta = 0.5 \]
Figure 4.3: Boxplot of the global distance from different conditions. A: boxplot of the optimal true global distances; B_1 : boxplot of the optimal global distances based on two-step approach under measurement errors with \( \sigma_e^2 = \frac{1}{5} \); B_2 : boxplot of the optimal global distances based on direct matching under measurement errors with \( \sigma_e^2 = \frac{1}{5} \); C_1 : boxplot of the optimal global distances based on two-step approach under measurement errors with \( \sigma_e^2 = 1 \); C_2 : boxplot of the optimal global distances based on direct matching under measurement errors with \( \sigma_e^2 = 1 \); D_1 : boxplot of the optimal global distances based on two-step approach under measurement errors with \( \sigma_e^2 = 4 \); and D_2 : boxplot of the optimal global distances based on direct matching under measurement errors with \( \sigma_e^2 = 4 \).
Figure 4.4: Boxplot of the global distance from different conditions. A: boxplot of the optimal true global distances; B1: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = \frac{1}{5}$; B2: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = \frac{1}{5}$; C1: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 1$; C2: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 1$; D1: boxplot of the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 4$; and D2: boxplot of the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 4$. 
rors is weak, we can only see a little bit of improvement. And the improvement becomes stronger when the degree of measurement errors increases. Other factors which can reflect whether the matching condition is favored or not are the standard deviations of the school random effect. Given fixed \( \eta \) and fixed variance of the control school, if the ratio of the variance of the treatment schools to that of the control schools is smaller, it is relatively easier to get a better matching result. Therefore, we can see that when the ratio of \( \sigma_{b1}^2 \) to \( \sigma_{b2}^2 \) decreases, we can see more apparent improvement regarding the global distance.

Figures 4.5, 4.6, 4.7, and 4.8 show us that under matching favored conditions the estimated treatment effect is unbiased and the two-step approach can provide estimators with less variance. However, when the matching is under unfavored conditions, we can get biased estimates even from matching based on true values of the matching variables.

4.7 Matching Quality and Measurement Error

Using Mahalanobis distance matching methods, if the matching covariates contain measurement errors, we can get systematically biased results. For the matching based on the measured matching covariates, the distance of a matched pair can be written as

\[
d^* = (Z_i - Z_{ji})^T \Sigma_i^{-1} (Z_i - Z_{ji}),
\]

where \( \Sigma_i^* \) is the variance-covariance matrix of the distribution of the measured matching covariates in the treatment group. If the matching is perfect, we have \( d^* = 0 \), and equivalently we have \( Z_i^* = Z_{ji}^* \). Under the assumption of classical measurement errors (4.5), we have

\[
Z_i^* = Z_i + E_i, \quad \text{and} \quad Z_{ji}^* = Z_{ji} + E_{ji}.
\]

(4.12)

Therefore, the true distance of this matched pair is

\[
d = (Z_i - Z_{ji})^T \Sigma_i^{-1} (Z_i - Z_{ji}) = (E_i - E_{ji})^T \Sigma_i^{-1} (E_i - E_{ji}),
\]

(4.13)
Figure 4.5: Boxplot of the estimated treatment effects from different conditions. A is from the optimal true global distances; B1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = \frac{1}{9}$; B2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = \frac{1}{9}$; C1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 1$; C2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 1$; D1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 4$; and D2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 4$. 
Figure 4.6: Boxplot of the estimated treatment effects from different conditions. A is from the optimal true global distances; B1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = \frac{1}{9}$; B2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = \frac{1}{9}$; C1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 1$; C2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 1$; D1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 4$; and D2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 4$. 

\[ \frac{\sigma_{\theta_1}}{\sigma_{\theta_2}} = 1/10 \quad \eta = 0.5 \]

\[ \frac{\sigma_{\theta_1}}{\sigma_{\theta_2}} = 1/2 \quad \eta = 0.5 \]

\[ \frac{\sigma_{\theta_1}}{\sigma_{\theta_2}} = 1 \quad \eta = 0.5 \]
Figure 4.7: Boxplot of the estimated treatment effects from different conditions. A is from the optimal true global distances; B\(_1\) is from the optimal global distances based on two-step approach under measurement errors with \(\sigma_e^2 = \frac{1}{10}\); B\(_2\) is from the optimal global distances based on direct matching under measurement errors with \(\sigma_e^2 = \frac{1}{4}\); C\(_1\) is from the optimal global distances based on two-step approach under measurement errors with \(\sigma_e^2 = 1\); C\(_2\) is from the optimal global distances based on direct matching under measurement errors with \(\sigma_e^2 = 1\); D\(_1\) is from the optimal global distances based on two-step approach under measurement errors with \(\sigma_e^2 = 4\); and D\(_2\) is from the optimal global distances based on direct matching under measurement errors with \(\sigma_e^2 = 4\).
Figure 4.8: Boxplot of the estimated treatment effects from different conditions. A is from the optimal true global distances; B1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = \frac{1}{9}$; B2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = \frac{1}{9}$; C1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 1$; C2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 1$; D1 is from the optimal global distances based on two-step approach under measurement errors with $\sigma_e^2 = 4$; and D2 is from the optimal global distances based on direct matching under measurement errors with $\sigma_e^2 = 4$. 
where $\Sigma_1$ is the variance-covariance matrix of the distribution of the true matching covariates in the treatment group, and $E = (e_1, e_2, \cdots, e_k)$ is the vector of random errors. Based on the expected value of matrix products shown in [25], for the perfectly matched pair under measurement errors, we have that the expectation of the distance becomes

$$
E(d|Z_i^* = Z_{ji}^*) = E \left[ (E_{ji} - E_i)^T \Sigma_1^{-1} (E_i - E_{ji}) | Z_i^* = Z_{ji}^* \right]
$$

$$
= E \left[ (E_{ji} - E_i) | Z_i^* = Z_{ji}^* \right]^T \Sigma_1^{-1} E \left[ (E_i - E_{ji}) | Z_i^* = Z_{ji}^* \right]
$$

$$
+ \text{tr}(2\text{Var}(E_i|Z_i^* = Z_{ji}^*) \Sigma_1^{-1})
$$

(4.14)

where $\text{tr}(.)$ is the trace function, which calculates the sum of the diagonal of a matrix. If the random errors in both groups are all independent with each other, we have

$$
\text{Var}(E_i|Z_i^* = Z_{ji}^*) = \sigma_e^2 I,
$$

where $I$ is the identity matrix. Therefore, based on equation (4.14) we have

$$
E(d|Z_i^* = Z_{ji}^*) = 2\sigma_e^2 \text{tr}(\Sigma_1^{-1}).
$$

(4.15)

As for a perfect complete matching based on the measured matching covariates, given independent random errors between treatment and control groups, the bias of the global distance is $2N_1\sigma_e^2 \text{tr}(\Sigma_1^{-1})$, which vanishes only when $\sigma_e^2 = 0$, i.e., there are no measurement errors. If the matching conditions are not preferable, the bias may be much larger than the bias from the perfect matching.

### 4.8 Discussion

Based on the simulation results and mathematical calculation, we can see that compared with the traditional matching approach the two-step approach can improve the matching quality in the cross-sectional setting and/or when there is no longitudinal trend in the data, especially under matching favored conditions. We propose to conduct more simulation studies in the future, focused on the settings where there are the various longitudinal trends in the matching covariates data, and/or various types of variance-covariance structure in the matching covariates. Also
more theoretical work is under way for longitudinal data.

As for the estimate of the treatment effect, we use the average of the difference of outcome within matched pair at specified time point. For time-independent matching, Rubin [46] compared five different estimates of the treatment effect including the average of difference of outcome across matched pair and four other regression adjusted estimates. For time-dependent matching, we would also like to consider model adjusted estimates while the linear regression model may be fitted at each time point or we would model the matched pair using a mixed-effects model and then do the adjustment based on the model output.

Finally we would like to apply the two-step method to the EC study data or other real data with similar structure in the future.
Chapter 5

Discussion

My work on issues in the studies of exposure-disease relationships can go further and deeper.

First of all, in my thesis, focus has only been given to binary outcomes. However, in many studies, time-to-disease is also available. Therefore, I would like to modify my models to accommodate time-to-event outcomes as well. As examples, in Chapter 2, I can modify the logistic regression sub-model in the Bayesian hierarchical model into a Cox proportional hazard model, if the disease outcome is a time-to-event variable. And in Chapter 3, it would be interesting to discuss dose-response relationships in the Cox regression setting.

Secondly, I can work more on statistical computation. In Chapter 2, for the Bayesian hierarchical model I built, I developed the specified MCMC algorithm for this type of model. This is much faster than using WinBUGs. To share the algorithm with other researchers, I would like to formalize the algorithm into a R package. This could benefit the researchers, who would like to run simulations and/or do analysis with extra large sample sizes, as WinBUGs may take a very long time to finish. For Chapter 3, for the model fitting, I used profile likelihood, which is commonly used for nonlinear models. I can try other numerical analysis methods to speed the program, though currently running time is not an issue.

Finally, in the future, more attention should be given to the study design. Chapters 2 and 3 mainly focus on prospective cohort studies. I address some problems in observational studies in Chapter 4. There is much work left for observational
studies if we use other matching strategies and consider various longitudinal patterns.
Bibliography


[29] H.-M. Kim, D. Richardson, D. Loomis, M. V. Tongeren, and I. Burstyn. Bias in the estimation of exposure effects with individual- or group-based...


