ABSTRACT

BERNHARDT, PAUL WILLIAM. Statistical Modeling with Covariates Subject to Detection Limits. (Under the direction of Huixia Judy Wang and Daowen Zhang.)

Censored observations are increasingly common in biomedical datasets. Though censoring is often associated with time-to-event data, censored observations also arise due to detection limits. When it is of interest to include censored variables in a model, care must be taken when deriving parameter and standard error estimates. While a great deal of research has been conducted on statistical methods for handling censored responses, relatively little research has focused on modeling with censored covariates. Traditionally used methods for handling covariates subject to detection limits, such as replacing censored observations with a function of the detection limit or only using complete cases, perform poorly when there is a moderate or high level of censoring.

In this thesis, we explore statistical methodology in various settings when covariates are subject to censoring at known detection limits. Specifically, we study how to handle censored covariates in generalized linear, survival, and joint modeling scenarios. For generalized linear and survival models, we develop a new “improper” multiple imputation procedure which has the advantage of being straightforward to employ and very computationally efficient. The consistency and asymptotic normality of the estimator are established and an easy-to-calculate approximate variance estimator is recommended. We also propose using an iterative version of this multiple imputation method which approximates the expectation-maximization algorithm for maximizing the likelihood but does not require iterating to convergence to obtain consistent parameter estimates. We additionally suggest several other methods for handling censored covariates in generalized linear models that are extensions from missing data literature, though we show both theoretically and with simulations that these competing methods are biased, more variable, or more computationally intensive than the proposed improper multiple imputation methods.
We also consider modeling data for which the covariates subject to censoring are longitudinal in nature. Specifically, we consider modeling a binary outcome when several covariates are observed repeatedly over a period of time and are subject to detection limits. We employ joint modeling techniques to incorporate the longitudinal aspect of the covariates. Traditionally, joint models involve describing the longitudinal covariates using mixed models and then including the random effects from the mixed models in the model for the response. We take this general approach but adapt the methodology to allow for censoring on the longitudinal covariates. We also propose a normal approximation in the expectation step of an expectation-maximization algorithm for obtaining maximum likelihood estimates that reduces the extensive computational time required to fit joint models in practice.

In this thesis, we take a parametric approach to handling censored covariates. For this reason, we briefly discuss methods for flexibly modeling the covariate data as well as ways to check the distributional assumptions on the covariates. We suggest a variety of techniques that have not been discussed in the context of censored covariate data.

The research in this thesis was motivated by the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. The GenIMS study was designed specifically to discover the relationship between three biological assays and outcomes such as 90-day survival and long-term survival time. These three assays, or biomarkers, were all subject to censoring with a total of 64% of individuals in the study having at least one censored biomarker at baseline. These biomarkers were also measured daily over a period of up to 30 days, with almost 50% total censoring. We analyze the GenIMS dataset using the methods described in this thesis in the context of logistic regression models, accelerated failure time survival models, and joint models of a binary outcome and multiple longitudinal covariates.
Statistical Modeling with Covariates Subject to Detection Limits

by
Paul William Bernhardt

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics
Raleigh, North Carolina
2013

APPROVED BY:

________________________________________________________________________
Marie Davidian                                             Anastasios Tsiatis

________________________________________________________________________
Huixia Judy Wang                                             Daowen Zhang
Co-chair of Advisory Committee                          Co-chair of Advisory Committee
DEDICATION

To my wife, who has been the foundation of my support over the last ten years, and to my parents, who have been there for me every step of the way.
BIOGRAPHY

The author was born in the city of Lancaster, Pennsylvania on September 6th, 1985, and grew up in the nearby town of Millersville, PA. He went to elementary, middle, and high school in the Penn Manor School District. He received his bachelor’s degree in 2008 in mathematics with secondary teaching certification and a minor in statistics from Messiah College in Grantham, Pennsylvania. He earned a master’s degree in statistics from North Carolina State University, Raleigh, North Carolina, in 2010 and has spent the last three years working toward his doctoral degree, also at North Carolina State University.
ACKNOWLEDGEMENTS

First, and foremost, I would like to thank my advisors Dr. Huixia Judy Wang and Dr. Daowen Zhang for their constant support and advice. When I chose to work with them three years ago, I did not necessarily know how to “do” research, but they guided me and helped me along the way as I needed it. I would especially like to thank Dr. Wang for the project idea and for extra assistance in writing and publishing along the way and Dr. Zhang for his invaluable advice throughout the research process. Additionally, I would like to thank Dr. Marie Davidian and Dr. Anastastios Tsiatis, whom, despite their very busy schedules agreed to be on my committee and provided excellent advice for improving my project. I thank Dr. Marlin Eby, my undergraduate mentor and statistics professor at Messiah College, for recognizing my ability in statistics, encouraging me to seek further education at a high-quality institution like North Carolina State University, and promoting me to others far more than I deserve. Finally, I would like to thank my wife, Tawny, and my entire family for their love and support.
# TABLE OF CONTENTS

LIST OF TABLES ................................................. viii

LIST OF FIGURES .............................................. xii

Chapter 1 Introduction .................................. 1
  1.1 Motivation and Review .............................. 1
  1.2 Dissertation Outline ............................... 3

Chapter 2 Generalized Linear Models with Covariates Subject to Detection Limits ........................................ 6
  2.1 Introduction ......................................... 6
  2.2 Model Set-up ....................................... 9
  2.3 Conventional Analysis Methods .................. 11
    2.3.1 Complete Case Estimation .................... 11
    2.3.2 Simple Imputation Methods ................. 12
    2.3.3 Maximum Likelihood Estimation ............ 14
    2.3.4 Bayesian Methods ............................ 15
  2.4 Improper Multiple Imputation Method .......... 17
    2.4.1 Estimation of $\theta$ ...................... 18
    2.4.2 Estimation of $\text{Var}(\hat{\theta})$ ......... 21
  2.5 Simulation Study .................................. 22
    2.5.1 Estimation of $\beta$ ....................... 23
    2.5.2 Variance Estimation of $\hat{\beta}_{IMI}$ .... 27
  2.6 Application to the GenIMS Study .............. 28
  2.7 Discussion ....................................... 31

Chapter 3 Survival Models with Covariates Subject to Detection Limits ........................................ 33
  3.1 Introduction ....................................... 33
  3.2 Seminonparametric Accelerated Failure Time Model ........................................ 35
    3.2.1 Accelerated Failure Time Model with Seminonparametric Error 35
    3.2.2 Fitting a Seminonparametric Accelerated Failure Time Model .......... 40
  3.3 Conventional Analysis Methods .................. 43
    3.3.1 Complete Case Estimation .................... 43
    3.3.2 Simple Imputation Methods ................. 44
    3.3.3 Maximum Likelihood Method ................. 44
    3.3.4 Bayesian Methods ............................ 45
  3.4 Improper Multiple Imputation Method .......... 47
  3.5 Iterated Improper Multiple Imputation Method .. 51
  3.6 Simulation Study .................................. 53
  3.7 Application to the GenIMS Study .............. 57
  3.8 Discussion ....................................... 60
Chapter 4 Joint Modeling of a Binary Response and Multiple Longitudinal Covariates Subject to Detection Limits

4.1 Introduction ................................................. 62
4.2 Joint Modeling of a Binary Response and Multiple Longitudinal Covariates .... 66
  4.2.1 Models and Notation .................................. 67
  4.2.2 Maximization via the EM Algorithm ....................... 68
  4.2.3 Approximate EM Algorithm ................................ 71
4.3 Joint Modeling of a Binary Response and Multiple Covariates Subject to Detection Limits ........................................... 76
  4.3.1 Longitudinal Model with Censoring ......................... 77
  4.3.2 Approximate EM Algorithm with Multiple Covariates Subject to Detection Limits ........................................... 77
4.4 Simulation Study ............................................. 82
  4.4.1 Joint Model with One Longitudinal Covariate .......... 83
  4.4.2 Joint Model with Three Longitudinal Covariates ......... 85
  4.4.3 Computational Advantages .................................. 88
4.5 Application to the GenIMS Study ................................ 88
4.6 Discussion .................................................. 93

Chapter 5 Flexible Models for the Conditional Distribution of the Covariates Subject to Detection Limits

5.1 Introduction .................................................. 95
5.2 Flexible Modeling Methods ...................................... 97
  5.2.1 Multivariate Normal ...................................... 97
  5.2.2 Box-Cox Transformations .................................. 99
  5.2.3 Seminonparametric Distribution .............................. 101
  5.2.4 Series of Univariate Regressions ......................... 103
  5.2.5 Other Methods ........................................... 104
5.3 Model Checking ............................................... 105
  5.3.1 Distributional Tests ...................................... 105
  5.3.2 Information Criterion Methods ......................... 106
  5.3.3 Truncated vs. Censored Percentages ...................... 106
  5.3.4 Graphical Checks ......................................... 107
  5.3.5 Expected Value of the Response .......................... 111
5.4 Discussion .................................................. 111

REFERENCES .................................................. 113

APPENDICES .................................................... 121
Appendix A Chapter 2 Proofs and Technical Details ......................... 122
  A.1 Regularity Conditions ...................................... 122
  A.2 Proof of Proposition 1 ....................................... 123
  A.3 Proof of Theorem 1 .......................................... 123
  A.4 Proof of Theorem 2 .......................................... 127
Appendix A.5 Proof Sketch of Theorem 3
Appendix B Additional Simulations for Generalized Linear Models with Covariates Subject to Detection Limits
Appendix C Chapter 3 Proofs and Technical Details
  C.1 Regularity Conditions
  C.2 Proof of Proposition 2
  C.3 Proof Sketch of Theorem 4
  C.4 Proof of Theorem 5
Appendix D Technical Details of the Approximate EM Algorithm Procedure
Appendix E An Overview of Adaptive Gaussian-Hermite Quadrature Methods
**LIST OF TABLES**

Table 2.1  Simulation results for a logistic regression on 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Time (s): average elapsed computational time in seconds for a single simulated dataset; Omni: omniscient method; CC: complete case analysis; SUB: substitution by DL/$\sqrt{2}$; Mean1: conditional mean imputation method conditioning only on the DL; Mean2: conditional mean imputation method conditioning on the covariates only; Mean3: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper (Bayesian) multiple imputation method; IMI: improper (frequentist) multiple imputation method. .......................... 25

Table 2.2  A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.001 and 0.003. SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}_{IMI}$; SE$_C$: average standard error estimated using (2.8); SE$_A$: average standard error estimated using (2.9); SE$_R$: average standard error estimated using (2.6); SE$_B$: average standard error estimated via a bootstrap procedure. .......................... 27

Table 2.3  Parameter and standard error estimates for coefficients of log-cytokine and demographic covariates in the logistic regression model for 90-day survival, based on the GenIMS dataset. Par.: parameter estimate; SE: standard error estimate; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method. .......................... 30

Table 3.1  Common distributions for AFT models .......................... 36

Table 3.2  Simulation results for an AFT regression model with Weibull errors, left, and a mixture of Weibull and lognormal errors, right, for 1000 datasets. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CC: complete case analysis; SUB: Substitution by DL/$\sqrt{2}$; IMI: improper multiple imputation method; IIMI: iterated improper multiple imputation method .......................... 56
Table 3.3  Parameter and standard errors estimates for coefficients of log-cytokine and demographic covariates in the AFT model for survival time, based on the GenIMS dataset; first set of estimates is based on using fixed 90-day censoring times, second set of estimates is based on using censoring times from last check-up time. Par.: parameter estimate; SE: standard error estimate; CC: complete case analysis; IMI: improper multiple imputation method; IIMI: iterated improper multiple imputation method. ................................. 59

Table 4.1  Simulation results for the joint model of a binary response and a single longitudinal covariate not subject to censoring. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability. ................................. 84

Table 4.2  Simulation results for the joint model of a binary response and a single longitudinal covariate subject to 20%, 40%, and 60% censoring. Par.: Parameter; SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability. ................................. 85

Table 4.3  Simulation results for the joint model of a binary response and three longitudinal covariates not subject to censoring. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability. ................................. 86

Table 4.4  Simulation results for the joint model of a binary response and three longitudinal covariates subject to 25% and 50% total censoring. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability. ................................. 87

Table 4.5  Parameter and standard error estimates for the coefficients of log-cytokine and demographic covariates in the joint model for 90-day survival, based on the GenIMS dataset. Each pair of estimates for the log-cytokines represents the baseline effect of the covariate as well as the effect corresponding to the longitudinal trajectory. ................................. 92

Table 5.1  Simulation results for a logistic regression on 2000 datasets with 200 individuals and an incorrect distributional assumption on the covariates. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of \( \hat{\beta} \); SE: average estimated standard error; CP: empirical 95% coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method. ................................. 98
Table 5.2  Simulation results for a logistic regression on 2000 datasets with 500 individuals and an incorrect distributional assumption on the covariates. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

Table 5.3  Simulation results for a logistic regression on 2000 datasets with 200 individuals and a Box-Cox transformation applied to the covariates. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

Table B.1  Simulation results for a logistic regression on 2000 datasets with 500 observations and two covariates of correlation $\rho = 0.5$. Biases which are statistically significant at the $\alpha = 0.99$ level are shown in bold. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Time (s): average elapsed computational time in seconds for a single simulated dataset; Omni: omniscient method; CC: complete case analysis; SUB: substitution by DL/$\sqrt{2}$; Mean1: conditional mean imputation method conditioning only on the DL; Mean2: conditional mean imputation method conditioning on the covariates only; Mean3: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

Table B.2  A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 datasets with 500 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.0005 and 0.0015. SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}_{IMI}$; $SE_C$: average standard error estimated using (2.8); $SE_A$: average standard error estimated using (2.9); $SE_R$: average standard error estimated using (2.6); $SE_B$: average standard error estimated via a bootstrap procedure.
Table B.3 Simulation results for a logistic regression on 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Biases which are statistically significant at the $\alpha = 0.99$ level are shown in bold. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Time (s): average elapsed computational time in seconds for a single simulated dataset; Omni: omniscient method; CC: complete case analysis; SUB: substitution by DL/$\sqrt{2}$; Mean1: conditional mean imputation method conditioning only on the DL; Mean2: conditional mean imputation method conditioning on the covariates only; Mean3: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

Table B.4 A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.00009 and 0.0006. SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}_{IMI}$; $SE_C$: average standard error estimated using (2.8); $SE_A$: average standard error estimated using (2.9); $SE_R$: average standard error estimated using (2.6); $SE_B$: average standard error estimated via a bootstrap procedure.
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Normal Q-Q plots for censored cytokines: log(TNF), log(IL-6), and log(IL-10). The Y-axis corresponds to the sample quantiles above the $p_0^{th}$ quantile of the cytokine biomarkers, where $p_0$ is the sample censoring proportion. The X-axis corresponds to the quantiles of a $N(0,1)$.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Distribution of the random effects conditional on the data in a joint model with a single longitudinal covariate described by a linear mixed model with a random intercept, $b_0$, and a random slope, $b_1$, using simulated data and various degrees of censoring.</td>
<td>79</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Distribution of the random effects conditional on the data in a joint model with three longitudinal covariates described by linear mixed models with a random intercept and slope term, using simulated data and various degrees of censoring. The intercepts and slope random effects for the three covariates are represented by $(b_0, b_2, b_4)$ and $(b_1, b_3, b_5)$, respectively.</td>
<td>80</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Longitudinal plots of log(TNF), log(IL-6), and log(IL-10) for a random sample of patients in the GenIMS study.</td>
<td>90</td>
</tr>
<tr>
<td>Figure 5.1</td>
<td>Normal Q-Q plots for log($X_1$) and log($X_2$) using data from a single simulated dataset.</td>
<td>99</td>
</tr>
<tr>
<td>Figure 5.2</td>
<td>Histogram of a sample with 50% censoring below the DL = $-1.5$. The curve of a normal distribution fitted to the censored data by maximum likelihood is overlayed on the histogram, with probability rather than frequency plotted on the Y-axis.</td>
<td>108</td>
</tr>
<tr>
<td>Figure 5.3</td>
<td>Q-Q plots of the data observed above the DL = $-1.5$ versus quantiles of an assumed distribution. The left Q-Q plot incorrectly plots the quantiles of a truncated distribution on the X-axis while the right Q-Q plot correctly plots the upper $1-p_0$ quantiles of an untruncated distribution.</td>
<td>109</td>
</tr>
<tr>
<td>Figure 5.4</td>
<td>Q-Q plot of the data observed above the DL = $-1.5$ versus the upper $1-p_0$ quantiles of a fitted normal distribution. The data should follow the 45 degree reference line through the origin if the normal distribution accurately describes the data.</td>
<td>110</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Motivation and Review

Biomedical datasets frequently contain variables that are subject to censoring. Though censoring is commonly associated with time-to-event data, censored data also arise due to detection limits (DLs). In practice, censoring due to DLs often occurs when medical instruments are unable to measure a biological factor below a certain known value. Two well-known examples of left-censoring due to DLs in biomedical datasets include viral RNA measurements in individuals with the human immunodeficiency virus (Paxton et al., 1997; Hughes, 1999; Lyles et al., 2000; Thiebaut et al., 2005; Vock et al., 2012) and antibody concentration in response to vaccines (Moulton and Halsey, 1995). In the motivating Genetic and Inflammatory Markers of Sepsis (GenIMS) study, multiple variables are subject to DLs, including the three biomarkers of primary interest in the study.

Statistical methodology for handling censored variables has been explored in several contexts. Helsel (2012), for example, summarized many strategies for computing summary statistics and group comparisons from datasets with one or more censored variables. Also, a great deal of research has been devoted to survival modeling, where the interest lies in modeling a censored response (see Klein and Moeschberger, 2003 for an introduction). Research in statistical mod-
eling with data subject to DLs has mainly focused on the situation where the censored data are longitudinal responses (Pettitt, 1986; Lyles et al., 2000; Thiebaut et al., 2005; Wang and Fygenson, 2009; Lee and Kong, 2011; Vock et al., 2012; Chen et al., 2013). Only recently have researchers considered treating variables subject to DLs as predictors, or covariates, in a model.

We can view covariates subject to DLs as a special type of measurement error (see, for example, Richardson and Ciampi, 2003). In this context, rather than observing the true covariate value $X$, we actually observe $X^* = \max(X, d)$ for a lower DL $d$ and $X^* = \min(X, d)$ for an upper DL $d$. This type of measurement error is unusual in the sense that it depends completely on the value of the variable subject to censoring and is strictly one-directional – greater than or equal to zero for lower DLs and less than or equal to zero for upper DLs. Methods for dealing with measurement error generally assume that the measurement error is both random and normally distributed, neither of which apply for censoring due to DLs.

Perhaps more appropriately, we can consider censoring on the covariates to be a type of missing data. Several modeling methods have been developed for analyzing missing covariate data (Lipsitz et al., 1999; Ibrahim et al., 2002, 2005; Herring and Ibrahim, 2001; Herring et al., 2004; Nan et al., 2009). However, these methods do not generally apply to data censored due to DLs.

Traditionally, three common approaches have been used to handle censored covariate data. First, a complete case analysis simply discards data from individuals with censored covariates. We show in Sections 2.3 and 3.3 that the complete case method leads to consistent estimates in generalized linear models (GLMs) and survival models when covariates are censored due to DLs, but efficiency is lost due to the discarding of data. A second common approach replaces censored observations with a fixed value such as the DL, $DL/2$, or $DL/\sqrt{2}$ (Hornung and Reed, 1990; Sutton-Tyrrell et al., 2010; Stein et al., 2009, among many others). Using these naive substitution methods leads to biased parameter estimates and incorrect inference (Lynn, 2001; Austin and Brunner, 2003; Lubin et al., 2004; Rigobon and Stoker, 2007; D’Angelo and Weissfeld, 2008; Rigobon and Stoker, 2009; Helsel, 2012). Another conventional approach is conditional
mean imputation, in which censored values are replaced with their conditional expectations, possibly given the other observed variables (Lynn, 2001; Austin and Hoch, 2004; Giovanini, 2008; Nie et al., 2010; Arunajadai and Rauh, 2012). Though conditional mean imputation is more sophisticated than the substitution approach, we prove in Section 2.3 that this method generally leads to biased estimates for GLMs. Additionally, as with substitution methods, it is generally unclear how to estimate standard errors.

These traditional methods all have serious flaws, especially when the degree of censoring on the covariates is high. While a few methods have been developed in recent years for handling covariates subject to DLs in generalized linear and survival models, these methods are either very limited in their approach or computationally intensive. Our main objective in this dissertation is to develop practical and flexible statistical inference methods in the context of GLMs, survival models, and joint models with covariates censored due to DLs.

 Though we focus on lower DLs throughout this dissertation, the ideas we discuss can be generalized to covariates censored due to upper DLs. While lower DLs are most common in biological examples, Austin and Brunner (2003) gave an example of covariates subject to upper DLs in the context of surveys meant to measure overall health status. In these surveys, indicators such as alcohol intake, smoking frequency, and drug use are only recorded accurately up to a certain level in order to encourage more honest reporting. As another example of upper DLs, in economics, exact salaries and expenditures are generally only reported up to a certain amount (Rigobon and Stoker, 2007; Melenberg and van Soest, 1996).

1.2 Dissertation Outline

The remainder of the paper is organized as follows. In Chapter 2, we develop several modeling techniques for GLMs with censored covariates and focus specifically on our proposed “improper” multiple imputation method. We briefly introduce previously developed approaches in Section 2.1. In Section 2.2, we explain GLMs without censored covariates and summarize the necessary notation. In Section 2.3, we outline several conventional methods that can be used to derive
parameter estimates in GLMs with censored covariates. In Section 2.4, we explain our proposed improper multiple imputation method. In Section 2.5, we provide simulations comparing all of the suggested methods as well as standard error estimation methods for the proposed improper multiple imputation method. Finally, in Section 2.6, we analyze data from the GenIMS study using several of the discussed methods. We review our proposals and discuss avenues for further research and improvement in Section 2.7.

In Chapter 3, we propose methods for analyzing survival models with censored covariates, again focusing on an improper multiple imputation technique. In Section 3.1, we review methods which have already been developed for this modeling situation. In Section 3.2, we explain the accelerated failure time (AFT) model for survival data as well as the seminonparametric distribution which we propose for modeling the error in the AFT model. In Sections 3.4 and 3.5, we develop the improper multiple imputation and iterated improper multiple imputation algorithms for analyzing survival models with covariates subject to DLs. In Section 3.6, we provide numerical studies to demonstrate the performance of the proposed methods, and in Section 3.7, we again apply these methods to the GenIMS dataset. Lastly, we summarize the benefits of our approach and discuss some limitations in Section 3.8.

In Chapter 4, we develop a computationally-efficient, approximate EM algorithm for fitting joint models for a binary response and multiple longitudinal covariates subject to DLs. In Section 4.1, we review the development of joint models over the past 20 years, focusing on research that relates to our problem. In Section 4.2, we introduce the joint model with a logistic model for the binary response and a linear mixed model for each of the longitudinal covariates. We also propose an approximate version of an EM algorithm to find maximum likelihood estimates for the joint model. In Section 4.3, we extend the proposed approach to situations where the longitudinal covariates are subject to censoring at lower DLs. We study the proposed methods with numerical studies in Section 4.4. Finally, in Section 4.5 we analyze the GenIMS dataset using the proposed joint modeling technique. We discuss possible extensions for our joint modeling proposals in Section 4.6.
In the final chapter, Chapter 5, we discuss methods for flexibly modeling the distribution of the covariates subject to DLs. In Section 5.1, we explain the difficulties of modeling the distributions of the covariates subject to DLs and emphasize the need for flexibility. We suggest several methods for flexible modeling in Section 5.2, and in Section 5.3 we discuss a few simple ways of checking distributional assumptions that are made. Finally, in Section 5.4, we review the difficulty of modeling with censored covariates and stress that while we must be careful when assuming parametric distributions, the benefits generally outweigh the costs when dealing with covariates censored due to DLs.

We provide additional details in five appendices. Appendix A and Appendix C give proofs for the theorems appearing in Chapters 2 and 3. Appendix B provides additional simulations comparing several of the methods discussed in Chapter 2. In Appendices D and E, we lay out the algebraic and technical details for applying our proposed joint modeling method discussed in Chapter 4.
Chapter 2

Generalized Linear Models with Covariates Subject to Detection Limits

2.1 Introduction

In the Genetic and Inflammatory Markers of Sepsis (GenIMS) study, which motivated this research, the primary goal was to study the relationship between three biomarkers subject to DLs and the development of severe sepsis and eventual death. In Section 1.1, we outlined several traditionally used methods for handling covariates subject to DLs and noted that these methods are not ideal since they lead to either incorrect inference or inefficient estimators. Over the past decade, several authors have developed more sophisticated methods for handling variables subject to censoring in linear and generalized linear models (GLMs).

A few researchers have proposed maximum likelihood methods for dealing with covariates censored due to DLs. Lynn (2001), Lyles et al. (2001), and Piepho et al. (2002) used maximum likelihood methods to estimate the parameters in a bivariate normal distribution with fixed censoring on one or both variables. Lynn (2001) also considered maximizing the likelihood for
logistic regression models with a single covariate that is subject to censoring while Austin and Brunner (2003) and Austin and Hoch (2004) explored maximum likelihood methods in normal linear models with two covariates, one of which is censored. Giovanini (2008) developed a Monte Carlo expectation-maximization (EM) algorithm for maximizing the likelihood of a longitudinal logistic mixed model with a single censored baseline covariate. More recently May et al. (2011) developed a general Monte Carlo EM algorithm for maximizing the likelihood of a GLM with multiple covariates subject to DLs. Though maximum likelihood methods lead to consistent and efficient parameter estimates, computation can be very intensive since representations of the censored-data likelihood involve potentially multi-dimensional integrations and generally do not have closed forms.

Some authors have considered alternative methods for finding parameter estimates in GLMs with censored covariates. Tsimikas et al. (2012) proposed an estimating equation approach based on likelihood methods which is computationally simple and does not require a strict parametric form for the distribution of the response. However, the method is only described for a single censored covariate. Lyles et al. (2001) used multiple imputation to estimate the parameters of a bivariate normal distribution when one of the variables is censored, and Lynn (2001) considered several ad-hoc multiple imputation methods for a logistic regression model with a single censored covariate. Taking a different approach, Lubin et al. (2004) developed a multiple imputation algorithm for models with a single covariate subject to a DL which uses bootstrapping to account for uncertainty in initial parameter estimates. Recently, Lee et al. (2012) proposed a multiple imputation algorithm which allows for multiple covariates censored due to DLs, though they assumed a multivariate normal model for generating imputations regardless of the response model.

The methods mentioned above are computationally intensive or are designed for a limited number of covariates, and traditional methods such as complete case analyses and substitution suffer from inefficiency or estimation bias. Our work advances the field in three major ways. First, we study the theoretical validity of complete case and mean imputation methods in GLMs
with covariates subject to DLs. To our knowledge, this is the first time such properties have been studied theoretically in this context. Second, we extend several methods from missing and censored data literature to the context of GLMs with multiple covariates subject to DLs. We describe the basic details of the methods and compare their attributes and faults. Third, and most substantially, we develop a new and computationally-efficient “improper” multiple imputation algorithm. We establish the asymptotic properties of the improper multiple imputation estimator and propose a convenient variance estimation method. Even though improper multiple imputation has been studied for missing data, our development for censored covariates is nontrivial. The special feature of censoring due to DLs makes it technically challenging to incorporate the censoring mechanism in a practical way to achieve consistency and efficiency improvement. Through numerical studies, we demonstrate that the proposed improper multiple imputation estimator performs comparably to maximum likelihood, fully Bayesian, and “proper” multiple imputation methods while requiring much less computational effort.

The remainder of this chapter is organized as follows. In Section 2.2, we describe the GLM framework and introduce the notation for the scenario of censored covariates. In Section 2.3, we investigate several conventional approaches for dealing with censored covariates within the context of GLMs. In Section 2.4, we present our proposed improper multiple imputation method, as well as a consistent variance estimator for the parameter estimates. In Section 2.5, we carry out an extensive simulation study to compare the performance of the methods discussed in this chapter, focusing on the proposed improper multiple imputation method. In Section 2.6, we apply the proposed methods to the dataset from the GenIMS study. Finally, in Section 2.7, we discuss the limitations of the improper multiple imputation method and some avenues for further research. The technical details for the proposition and theorems appearing in this chapter are provided in Appendix A.
2.2 Model Set-up

Suppose we observe independent samples \((Y_i, W_i), i = 1, \ldots, n\), where \(Y_i\) is a univariate response and \(W_i\) is a \(p\)-dimensional vector of covariates. We assume that the distribution of \(Y_i\) given \(W_i\) belongs to the exponential dispersion family,

\[
f(y_i|w_i, \xi_i, \phi) = \exp\left\{\frac{y_i \xi_i - b(\xi_i)}{a_i(\phi)} + c(y_i, \phi)\right\},
\]

(2.1)

where \(\xi_i\) is the natural parameter, \(\phi\) is a possible dispersion parameter, and \(a_i(\cdot), b(\cdot), \) and \(c(\cdot, \cdot)\) are known functions determining the specific distribution. For simplicity, we assume that \(\phi\) is known, though most of the results in this chapter could easily be extended to situations where \(\phi\) is unknown.

The exponential dispersion family has the property that \(\mu_i = E(Y_i|W_i) = b'(\xi_i)\) and \(h_i(\mu_i) = \text{Var}(Y_i|W_i) = b''(\xi_i)a_i(\phi)\). A GLM then relates \(\mu_i\) to the covariates by assuming

\[
g(\mu_i) = \sum_{j=1}^{p} W_{ij} \beta_j,
\]

(2.2)

where \(g(\cdot)\) is referred to as the link function and \(W_{ij}\) and \(\beta_j\) are the \(j^{th}\) elements in the covariate vector \(W_i\) and the parameter vector \(\beta = (\beta_1, \ldots, \beta_p)^T\), respectively.

When there is no censoring, \(\beta\) can be estimated by solving the estimating equations

\[
n^{-1} \sum_{i=1}^{n} w_{ij}(y_i - \mu_i) \frac{h_i(\mu_i) g'(\mu_i)}{h_i(\mu_i) g'(\mu_i)} = 0, \quad j = 1, \ldots, p,
\]

(2.3)

where \(g'(\mu_i) = \partial g(\mu_i)/\partial \mu_i\). For a GLM with \(a_i(\phi) = a(\phi), i = 1, \ldots, n\), and the canonical link function \(g(\mu_i) = \xi_i\), (2.3) simplifies to

\[
n^{-1} \sum_{i=1}^{n} w_{ij}(y_i - \mu_i) = 0, \quad j = 1, \ldots, p.
\]
For the remainder of this chapter, we assume that some covariates in \( W_i \) are subject to censoring due to lower DLs. Thus, for the \( i^{th} \) individual, we let \( W_i = (Z_i^T, X_i^T)^T \), where \( Z_i = (Z_{i1}, \ldots, Z_{i(p-q)})^T \) is the \((p-q)\)-dimensional vector of covariates fully observed for each individual, and \( X_i = (X_{i1}, \ldots, X_{iq})^T \) is the \(q\)-dimensional vector of covariates subject to censoring below \( d = (d_1, \ldots, d_q)^T \), the vector of DLs. For \( X_{ij}, j = 1, \ldots, q \), we only observe \( X_{ij}^* = \max(X_{ij}, d_j) \) and \( \rho_{ij} = I(X_{ij} \geq d_j) \), so that the complete set of observed data for the \( i^{th} \) individual is \((Y_i, Z_i, X_i^*, \rho_i)\), where \( X_i^* = (X_{i1}^*, \ldots, X_{iq}^*)^T \) and \( \rho_i = (\rho_{i1}, \ldots, \rho_{iq})^T \). Lastly, for the \( i^{th} \) subject, we define \( X_i^c \) as the subset of \( X_i \) for which \( \rho_{ij} = 0 \) and \( X_i^0 \) as the subset of \( X_i \) for which \( \rho_{ij} = 1 \). Similarly, we let \( d_i^c \) be the subset of \( d \) corresponding to \( X_i^c \). For notational simplicity, we assume that the DLs do not change with subjects. However, all the methods discussed in this dissertation can be applied to data with subject-specific DLs, which occur for instance in experiments where measurements are taken using different instruments.

For several of the theorems and technical details in this dissertation, we assume a multivariate distribution for \( f(x_i, \mathbf{x}_i) \). However, we note that in practice it is only really necessary to model the distribution of covariates subject to censoring, \( X_i \), conditional on fully observed covariates, \( Z_i \). Thus, for the remainder of this dissertation, we assume that the distribution \( f(x_i|z_i; \gamma) \) is a known \( q \)-variate distribution indexed by the parameter vector \( \gamma \), from which we can obtain \( f(x_i^c|z_i, x_i^0; \gamma) \) for all possible subsets of censored covariates, \( X_i^c \). For example, if \( f(x_i|z_i; \gamma) \) is a \( q \)-variate normal distribution, then \( f(x_i^c|z_i, x_i^0; \gamma) \) is also a normal distribution of dimension \( \leq q \). Though this distributional assumption may seem restrictive, it provides us with a useful way to extrapolate information into the censored data region. For positive random variables, we suggest using a Box-Cox transformation so that the multivariate normal distribution can be employed. If such an assumption is unreasonable, we recommend modeling \( f(x_i|z_i; \gamma) \) using a series of conditional univariate distributions \( f(x_{i1}|x_{i2}, x_{i3}, \ldots, x_{iq}, z_i) f(x_{i2}|x_{i3}, \ldots, x_{iq}, z_i) \cdots f(x_{iq}|z_i) \). Each of these univariate distributions can then be modeled by a regression model with a flexible error term. In Chapter 5, we describe flexible modeling of the distribution of \( f(x_i|z_i; \gamma) \) in more detail. We henceforth denote \( \mathbf{\theta} \) as
the vector of all parameters in a particular model, though we emphasize that the regression parameter vector $\beta$, the subset of $\theta$ defined through (2.1) and (2.2), is of primary interest for statistical inference.

In the next section, we describe several conventional estimation methods and discuss their limitations. Though we label these methods as conventional, with the exception of maximum likelihood and substitution methods, we do not believe they have previously been employed for censored covariate data.

### 2.3 Conventional Analysis Methods

#### 2.3.1 Complete Case Estimation

In the presence of censored predictors, one simple method for estimating the parameters in a GLM is the complete case approach, where the statistical analysis is restricted to those individuals for whom all covariates are completely observed. That is, the complete case estimator, ${\hat{\beta}}_{CC}$, is the solution to

$$n^* - \frac{1}{n^*} \sum_{i=1}^{n^*} \rho_i^* \frac{w_{ij}(y_i - \mu_i)}{h_i(\mu_i)g'(\mu_i)} = 0, \quad j = 1, \ldots, p,$$

where $\rho_i^* = \prod_{j=1}^{q} \rho_{ij}$ and $n^* = \sum_{i=1}^{n} \rho_i^*$ is the effective sample size.

**Proposition 1.** Under the GLM defined by (2.1) and (2.2) and the regularity conditions A.1-A.3 outlined in Appendix A, ${\hat{\beta}}_{CC} \xrightarrow{p} \beta$ as $n^* \to \infty$, where $\beta$ is the true parameter vector.

Proposition 1 is a direct extension from missing-data literature, as the censoring in our context does not depend on the response. In contrast to cases with censored responses, the complete case estimator is still consistent when the covariates are subject to fixed censoring. However, as it does not take into account any data obtained for individuals with censored covariates, the complete case estimator is inefficient. If a large percentage of individuals have censored covariates, the loss in efficiency can be substantial.
2.3.2 Simple Imputation Methods

An alternative approach for dealing with censored covariates is to use imputation methods, where censored data are filled-in with reasonable values. A variety of imputation methods have been proposed for missing-data problems (refer to Little, 1992 for a comprehensive review), but only a few have been explored in the context of censored predictors. A simple method for dealing with censored predictors is to impute, or substitute, the censored values by DL, DL/2, or DL/√2. However, Helsel (2012) and others have concluded that using functions of the DL is inappropriate, resulting in biased parameter and standard error estimators.

An alternative imputation method is conditional mean imputation, where censored covariates are imputed with (i) $E(X_i^c | X_i^c < d_i; \theta)$, (ii) $E(X_i^c | Z_i, X_i^o, X_i^c < d_i; \theta)$, or (iii) $E(X_i^c | Y_i, Z_i, X_i^o, X_i^c < d_i; \theta)$, and the unknown $\theta$ can be replaced by some consistent initial estimator. The imputed data can then be used to obtain estimates of $\beta$ via solving (2.3). We refer to the resulting estimators based on the three types of imputations as $\bar{\beta}_{M1}$, $\bar{\beta}_{M2}$, and $\bar{\beta}_{M3}$, respectively.

Calculating the conditional expectations may require numerical integration. However, unlike maximum likelihood methods which we describe in Section 2.3.3, conditional mean imputation only requires one integration for each individual with censored covariates as opposed to one integration at each step in a maximization algorithm. Thus, this method is relatively less computationally intensive. On the other hand, unlike maximum likelihood, where the observed Fisher information matrix can be used to obtain standard errors, variance estimation for $\bar{\beta}_{M1}$, $\bar{\beta}_{M2}$, and $\bar{\beta}_{M3}$ is difficult. Even in the case of linear regression with missing at random covariates, Little (1992) explained that the standard error estimates based on information matrices will be biased without complicated variance corrections.

To our knowledge, the use of conditional mean imputation has not been explored thoroughly in the context of GLMs with censored covariates. We demonstrate in Theorem 1 that conditional mean imputation generally gives biased estimates. We focus on the case where only one covariate, $X$, is subject to censoring below a DL $d$. We consider this special case for notational
ease and because it illustrates that conditional mean imputation usually gives biased estimates even in the simplest censoring scenario.

**Theorem 1.** Consider a GLM defined by (2.1) and (2.2) with covariate vector \( W = (Z^T, X)^T \).

Under the regularity conditions A.1-A.3 outlined in Appendix A,

1. \( \hat{\beta}_{M1} \xrightarrow{p} \beta \) as \( n \to \infty \) if and only if
   \[
   \int_Z \int_{-\infty}^d w f(x) dx \frac{f_X(\mu | x < d) - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} f(z) dz = 0,
   \]
   where \( \bar{\mu} = \mu(Z, X) \), \( \bar{X} = E(X|X < d; \theta) \), and \( Z \) is the support of \( Z \);

2. \( \hat{\beta}_{M2} \xrightarrow{p} \beta \) as \( n \to \infty \) if and only if
   \[
   \int_Z \int_{-\infty}^d w f(x|z) dx \frac{f_X(\mu | z, x < d) - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} f(z) dz = 0,
   \]
   where \( \bar{X} = E(X|Z, X < d; \theta) \);

3. \( \hat{\beta}_{M3} \xrightarrow{p} \beta \) as \( n \to \infty \) if and only if
   \[
   \int_Z \int_{-\infty}^d w \left( \int_{-\infty}^y \frac{(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} f(y|z, x) dy \right) f(z, x) dx dz = 0,
   \]
   where \( \bar{X} = E(X|Y, Z, X < d; \theta) \).

**Remark 1.** The consistency of \( \hat{\beta}_{M2} \) will be fulfilled if either \( \int_{-\infty}^d w f(x|z) dx = 0 \) or \( E_X(\mu | Z, X < d) = \bar{\mu} \). Now \( \int_{-\infty}^d Z f(x|z) dx = Z \int_{-\infty}^d f(x|z) dx = Z P(X < d|Z) \neq 0 \) unless \( Z = 0 \) with probability 1 since we implicitly assume in this paper that \( P(X < d|Z) > 0 \) for at least some values of \( Z \). Thus, to show consistency regardless of the distribution for \( Z \), we must have that \( E_X(\mu | Z, X < d) = \bar{\mu} \). For a GLM defined by the identity link, \( \mu = W^T \beta \), the estimator \( \hat{\beta}_{M2} \) is consistent for \( \beta \) since \( E_X(\mu | Z, X < d) = E_X((Z^T, X)\beta | Z, X < d) = (Z^T, \bar{X})\beta = \bar{\mu} \). When the identity link is not used, \( \hat{\beta}_{M2} \) will generally not be consistent unless the \( \beta \) parameter associated
with $X$ is zero. For example, for Poisson regression with a single covariate that is subject to censoring and the canonical link, $\mu = \exp(X\beta)$, we have $E(\mu | X < d) = E(\exp(X\beta) | X < d) \geq \exp\{E(X | X < d)\beta\} = \bar{\mu}$ by the Conditional Jensen’s Inequality, with equality only when $\beta = 0$. While it may be possible to find a distribution for $Z$ such that $\widehat{\beta}_{M2} \overset{p}{\to} \beta$ even when $EX (\mu | Z, X < d) \neq \bar{\mu}$, this will generally not be the case.

**Remark 2.** Using the same arguments as given in Remark 1, the consistency of $\widehat{\beta}_{M1}$ will be fulfilled, in general, only if $EX (\mu | X < d) = \bar{\mu}$, which again is only true with the identity link.

**Remark 3.** For the canonical link, we can use similar arguments as those in Remark 1 to argue that $\widehat{\beta}_{M3}$ is consistent for $\beta$ generally only if $EX (\mu | Z, X < d) = \bar{\mu}$, where $\bar{\mu}$ is now based on $\bar{X} = E(X | Y, Z, X < d)$. Now, even with the identity link, $\mu = W^{T}\beta$, $EX (\mu | Z, X < d) = EX \{(Z^{T}, X)\beta | Z, X < d\} \neq (Z^{T}, \bar{X})\beta = \bar{\mu}$ unless $X$ is conditionally independent of $Y$ so that $\bar{X} = E(X | Y, Z, X < d) = E(X | Z, X < d)$. Thus, when considering GLMs based on a canonical link, $\widehat{\beta}_{M3}$ will generally only be consistent if the response $Y$ is conditionally independent of the covariate $X$ given $Z$, so that the $\beta$ parameter associated with $X$ will be zero. Again, we note that while it may be possible to find a distribution for $Z$ such that $\widehat{\beta}_{M3} \overset{p}{\to} \beta$ even when $EX (\mu | Z, X < d) \neq \bar{\mu}$, this will generally not be the case.

### 2.3.3 Maximum Likelihood Estimation

Another method to estimate $\beta$ is through maximizing the likelihood contributed by all the observed data. We represent this likelihood as

$$\prod_{i=1}^{n} \int_{-\infty}^{d_i} f(y_i, x_i^c, x_i^c | z_i; \theta)dx_i^c,$$

(2.4)

where $\int_{-\infty}^{d_i}$ is an integral whose dimension corresponds to the length of $X_i^c$ and $\theta = (\beta^{T}, \gamma^{T})^{T}$ is the vector of parameters indexing $f(y_i, x_i^c, x_i^c | z_i; \theta) = f(y_i | z_i, x_i; \beta) \times f(x_i^c, x_i^c | z_i; \gamma)$, with $f(y_i | z_i, x_i; \beta)$ distributed according to (2.1) and (2.2) and $f(x_i^c, x_i^c | z_i; \gamma)$ as the conditional density of $(X_i^c, X_i^c)$ given $Z_i$. Note that in (2.4), $X_i^c$ is a subject-specific set of covariates
whose length corresponds to the number of censored covariates for the $i^{th}$ individual.

For illustrative purposes, suppose that $Z = (Z_1, Z_2)^T$ and $X = (X_1, X_2, X_3)^T$. The contribution to (2.4) for the $i^{th}$ individual for whom we observe the DLs $d_1$ and $d_3$ rather than $X_i^1$ and $X_i^3$ is

$$
\int_{-\infty}^{d_3} \int_{-\infty}^{d_1} f(y_i, x_{i1}, x_{i2}, x_3|z_{i1}, z_{i2}; \theta) dx_1 dx_3.
$$

The likelihood in (2.4) generally does not have a closed form, and for this reason we often must resort to approximations or numerical integration when maximizing this likelihood. In general, approximations are difficult to derive while numerical integration is notoriously difficult in higher dimensions. Recently, May et al. (2011) proposed a Monte Carlo EM algorithm for maximizing (2.4) which was shown to work well but is also computationally intensive because it uses Monte Carlo methods to approximate numerical integration.

### 2.3.4 Bayesian Methods

Bayesian methods have been considered as a practical method for handling missing covariate data. Until a recent work by Wu et al. (2012), Bayesian methods had not been explored in the specific context of GLMs with covariates subject to DLs. In the following, we summarize a basic Bayesian approach, which is an extension of the missing data methods described in detail by Ibrahim et al. (2002, 2005).

In a fully Bayesian analysis, inference for $\beta$ is carried out using the observed data posterior, given by

$$
\pi(\theta|y, w^o) \propto \left\{ \prod_{i=1}^{n} \int_{-\infty}^{d^c_i} f(y_i, x_{i1}^c, x_{i2}^c | z_i, \theta) dx_{i1}^c \right\} \pi(\theta),
$$

(2.5)

where $\pi(\theta)$ is a prior for $\theta = (\beta^T, \gamma^T)^T$, $w^o = \{z_i, x_{i1}^0, \rho_i\}_{i=1}^{n}$, and all other notation is identical to that in Section 2.3.3.

After specifying a prior, $\pi(\theta)$, we can obtain samples from (2.5) using Gibbs sampling. That is, we iteratively sample from the full conditional distributions $f(x_{i1}^c|\beta, \gamma, y_i, z_i, x_{i1}^0, x_{i1}^c < d_i^c), i = 1, \cdots, n$, $\pi(\beta|\gamma, y, z, x^o, x^c)$, and $\pi(\gamma|\beta, y, z, x^o, x^c)$, where $z = \{z_i\}_{i=1}^{n}$, $x^o = \{x_{i1}^o\}_{i=1}^{n}$.
\(x^c = \{x^c_i\}_{i=1}^n\) and initial values for \(\beta\) at the first iteration can be chosen as complete case estimates while initial values for \(\gamma\) can be chosen as the maximum likelihood estimates or any other reasonable values. Note that since we do not actually observe \(x^c\), we are treating each element in this set as a parameter vector. After obtaining \(N\) samples for \(\beta, \beta^{(1)}, \ldots, \beta^{(N)}\), we can estimate \(\beta\) and its standard error by the posterior mode and standard deviation. Standard Markov chain Monte Carlo (MCMC) sampling methods can be used to obtain draws from the full conditional distributions above when closed forms are not available. Though we do not discuss fully Bayesian methods further, we refer to Ibrahim et al. (2005) for more details and additional references in the context of missing data.

As an alternative to fully Bayesian methods, multiple imputation has become increasingly popular over the past several decades. The multiple imputation estimator was originally proposed by Rubin (1987) for parameter estimation in the presence of missing data. The idea is based on obtaining imputations for missing data by making \(m\) random draws for each missing value from the conditional distribution of the missing variable(s) given the other variables in the model. These \(m\) sets of draws can then be used to construct \(m\) complete datasets, based on which conventional methods designed for complete data can be used to calculate a final parameter estimate,

\[
\tilde{\theta} = m^{-1} \sum_{k=1}^m \tilde{\theta}^{(k)},
\]

where \(\tilde{\theta}^{(k)}\) is the estimate based on the \(k^{th}\) imputed dataset.

To obtain \(m\) imputations in a Bayesian framework, we must assume a prior for \(\theta, \pi(\theta)\), and then make random draws from the posterior predictive distribution of the censored data,

\[
f(x^c_i \mid y_i, z_i, x^o_i, x^c_i < d_i^o) \propto \int f(x^c_i \mid y_i, z_i, x^o_i, x^c_i < d_i^o; \theta) \pi(\theta \mid y, w^o) d\theta,
\]

where \(\pi(\theta \mid y, w^o)\) is the posterior distribution given by (2.5).

Bayesian multiple imputation is also referred to as “proper” imputation in the literature (Rubin, 1987; Schaubel and Cai, 2006) because it leads to valid large sample inferences using
the following simple variance estimator (Rubin, 1987),

$$\text{Var}(\hat{\theta}) = V_W + (1 + m^{-1})V_B,$$  \hspace{1cm} (2.6)

where $V_W = m^{-1} \sum_{k=1}^{m} \bar{V}^{(k)}$, with $\bar{V}^{(k)}$ representing the covariance matrix for the $k^{th}$ imputed dataset, is the “within” imputation variance, and $V_B = (m-1)^{-1} \sum_{k=1}^{m} (\bar{\theta}^{(k)} - \bar{\theta})(\bar{\theta}^{(k)} - \bar{\theta})^T$ is the “between” imputation variance. We again refer to Ibrahim et al. (2005) for more details regarding proper (Bayesian) multiple imputation in the context of missing covariates.

All of the methods described thus far have clear disadvantages. Complete case analyses lead to inefficient estimates while conditional mean imputation and substitution methods generally lead to inconsistent estimates. Though Bayesian methods are flexible, like maximum likelihood methods they often require a great deal of computational effort, and they also necessitate the user to define priors and monitor the convergence of sampling chains. In the following section, we present a frequentist-based, “improper” multiple imputation estimator that does not suffer from many of these limitations.

### 2.4 Improper Multiple Imputation Method

Alternative to the proper (Bayesian) multiple imputation, we can conduct multiple imputation in a frequentist framework by obtaining $m$ draws from the distribution $f(x_i^c|y_i, z_i, x_i^p, x_i^c < d_i^c; \hat{\theta}_I)$, where $\hat{\theta}_I$ is a consistent initial estimator for $\theta$. This type of multiple imputation is generally referred to as “improper” multiple imputation (Rubin, 1987), for which (2.6) is no longer consistent because the variation in $\hat{\theta}_I$ is not taken into account. However, we believe that the improper multiple imputation approach is advantageous over proper multiple imputation for two primary reasons: (i) obtaining draws from the posterior predictive distribution can be difficult and computationally intensive and (ii) improper multiple imputation generally produces more efficient estimates for a finite number of imputations (Wang and Robins, 1998). A similar discussion of the computational advantage of improper multiple imputation was provided in
Schaubel and Cai (2006) for a different set-up.

In the remainder of this section, we develop an improper multiple imputation method for GLMs with covariates subject to DLs and propose a consistent variance estimator using the theory of Wang and Robins (1998) and Robins and Wang (2000). We show that the method is practical and straightforward and that obtaining standard errors is only minimally more difficult than for proper multiple imputation.

2.4.1 Estimation of $\theta$

Our proposed multiple imputation algorithm is as follows:

**Step 1.** Using all the data $(z_i; x_i^*, \rho_i)$, $i = 1, \ldots, n$, and the DLs $d$, obtain the maximum likelihood estimate, $\hat{\gamma}^I$, as an initial estimate for $\gamma$, the parameters indexing $f(x_i|z_i)$.

**Step 2.** Using the complete cases, obtain the maximum likelihood estimate, $\hat{\beta}^I$, as an initial estimate for $\beta$, the parameters indexing $f(y_i|z_i, x_i)$.

**Step 3.** Form the 1rst imputed dataset by generating $\tilde{x}_i^c$ for each individual $i$ with censoring, $i = 1, \ldots, n$, from $f(x_i^c|y_i, z_i, x_i^o; \hat{\theta}^I)$ to replace the censored covariates $x_i^c$, where $\hat{\theta}^I = (\hat{\beta}^IT, \hat{\gamma}^IT)^T$. Denote $\tilde{x}_i$ as the resulting covariate vector for the $i^{th}$ subject that includes both $x_i^o$ and $\tilde{x}_i^c$.

**Step 4.** Using the complete dataset, $(y_i; z_i, \tilde{x}_i)$, $i = 1, \ldots, n$, obtain an estimate $\hat{\beta}^{(1)}$ for $\beta$ by solving the estimating equations given by (2.3), and an estimate $\hat{\gamma}^{(1)}$ for $\gamma$ by maximum likelihood estimation. Let $\hat{\theta}^{(1)} = (\hat{\beta}^{(1)T}, \hat{\gamma}^{(1)T})$.

**Step 5.** Repeat Steps 3 and 4 $m$ times, resulting in $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(m)}$.

**Step 6.** Finally, the improper multiple imputation estimator is defined as

$$\hat{\theta}_{IMI} = m^{-1} \Sigma_{k=1}^m \hat{\theta}^{(k)}.$$

In Step 3 of the above algorithm, the method for obtaining draws from the truncated distribution of $f(x_i^c|y_i, z_i, x_i^o; \hat{\theta}^I)$ may differ for various GLMs and distributions $f(x_i|z_i; \gamma)$. In
normal linear models where \( f(x_i | z_i) \) is also normal, we can sample directly from a truncated normal distribution. Most generally, by noting that Bayes theorem gives

\[
f(x_i^c | y_i, z_i, x_i^0) = \frac{f(y_i | z_i, x_i^0, x_i^c) f(x_i^c | z_i, x_i^0)}{f(y_i | z_i, x_i^0)},
\]

the Metropolis-Hastings algorithm or other MCMC methods may be used. For the simulation in Section 2.5 and application in Section 2.6, we consider a logistic model for a binary \( Y_i \), \( Y_i | X_i, Z_i \sim \text{Bernoulli}(\text{exp}(- (Z_i^T X_i^T \beta)])^{-1} \), and a multivariate normal distribution for \( X | Z, X_i | Z_i \sim N(\Lambda Z_i, \Omega) \), where \( \Lambda \) is a \( q \times (p-q) \) matrix of mean parameters. Then, after obtaining the initial estimates of \( \beta \) and \( \gamma = (\Lambda, \Omega) \), \( \beta^I \) and \( \gamma^I = (\Lambda^I, \Omega^I) \), we can use the following acceptance-rejection algorithm to obtain an imputation draw for the \( i \)th subject:

1. generate a candidate imputation, \( \tilde{x}_i^c \), for the vector \( X_i \) from the truncated conditional normal distribution \( f(x_i^c | z_i, x_i^0, x_i^c < d_i^c; \gamma^I) \);

2. generate \( U \) from a Uniform(0,1) distribution;

3. let \( \tilde{x}_i^c \) be the imputed vector for \( X_i \) if \( U < f(y_i | z_i, x_i^0, \tilde{x}_i^c) \), and repeat step (1) otherwise.

This acceptance-rejection algorithm works for logistic regression because

\[
\frac{f(y_i | z_i, x_i^0, x_i^c) f(x_i^c | z_i, x_i^0)}{f(y_i | z_i, x_i^0)} \leq \frac{f(x_i^c | z_i, x_i^0)}{f(y_i | z_i, x_i^0)},
\]

since \( f(y_i | z_i, x_i^0, x_i^c) \leq 1 \) and \( f(y_i | z_i, x_i^0) \) does not depend on \( X_i^c \).

As an additional note, we admit that it might be tempting to impute censored values using the conditional distribution \( f(x_i^c | z_i, x_i^0, x_i^c < d_i^c) \), ignoring the response \( Y_i \), because it is usually much easier to sample from \( f(x_i^c | z_i, x_i^0, x_i^c < d_i^c) \) than from \( f(x_i^c | y_i, z_i, x_i^0, x_i^c < d_i^c) \). For example, if \( f(x_i | z_i) \) is a normal distribution, then we can easily sample from \( f(x_i^c | z_i, x_i^0, x_i^c < d_i^c) \) because this distribution is simply a truncated normal. However, even if \( f(x_i | z_i) \) is normal, for most GLMs, \( f(x_i^c | y_i, z_i, x_i^0) \) is not, and obtaining random draws for \( X_i^c \) becomes more difficult, as explained above.

19
We define the regression parameter estimator based on using \( f(x_i^e | z_i, x_i^c, x_i^e < d_i^e) \) to obtain \( \tilde{x}_i^e \) in step 3 of the proposed improper multiple imputation algorithm as \( \tilde{\beta}_{IMI_w} \), and we provide the following theorem cautioning against the use of this estimator. For simplicity, we restrict this theorem to the case when only one covariate, \( X \), is subject to censoring below the DL \( d \).

**Theorem 2.** Consider a GLM defined by (2.1) and (2.2) with covariate vector \( W = (Z^T, X)^T \). Under the regularity conditions A.1-A.3 outlined in Appendix A, \( \tilde{\beta}_{IMI_w} \overset{P}{\rightarrow} \beta \) if and only if

1. \( E\{S_Z(\beta)\} = -E_Z \left[ Z \Pr(X < d|Z) \text{Cov} \left\{ \frac{1}{\mu(\mu)^\tau(\mu)}, \mu \right\} \right] = 0; \)

2. \( E\{S_X(\beta)\} = -E_Z \left[ \Pr(X < d|Z) \text{Cov} \left\{ \frac{X}{\mu(\mu)^\tau(\mu)}, \mu \right\} \right] = 0, \)

where the covariance is taken with respect to the distribution \( f(x|z, x < d; \gamma) \) and \( S_Z(\beta) \) and \( S_X(\beta) \) are the estimating equations in (2.3) related to \( Z \) and \( X \), respectively.

**Remark 4.** With the canonical link, \( a_i(\phi) = \phi/\tau \) (\( \phi, \tau \) known and positive), and only one covariate \( X \) subject to censoring, \( E\{S_Z(\beta)\} = 0 \) and \( E\{S_X(\beta)\} = -E_Z \{\tau^{-1} \Pr(X < d|Z) \text{Cov}(X, \mu)\} \).

Since we assume \( \Pr(X < d|Z) > 0 \) for a subset of the support of \( Z \) with positive measure, \( E\{S_X(\beta)\} = 0 \) only if \( \beta_x = 0 \) so that \( \tilde{\beta}_{IMI_w} \overset{P}{\rightarrow} \beta = (\beta_Z^T, \beta_X^T)^T \) only if \( \beta_X = 0 \), where \( \beta_X \) is the coefficient parameter for the covariate \( X \). This follows since for the canonical link, \( \mu \) is an increasing function of \( X_\beta X \) for any given \( Z \) so that \( \text{Cov}(X, \mu) > 0 \) for all \( Z \) if \( \beta_X > 0 \), \( \text{Cov}(X, \mu) < 0 \) for all \( Z \) if \( \beta_X < 0 \), and \( \text{Cov}(X, \mu) = 0 \) for all \( Z \) if \( \beta_X = 0 \).

We can also determine the direction of asymptotic bias for \( \tilde{\beta}_{IMI_w,X} \), the estimated coefficient parameter associated with covariate \( X \), when \( \beta_X \neq 0 \). By expanding the estimating functions evaluated at \( \tilde{\beta}_{IMI_w} \) about the true parameter \( \beta \), we find that

\[
\begin{bmatrix}
\tilde{\beta}_{IMI_w,z} - \beta_Z \\
\tilde{\beta}_{IMI_w,X} - \beta_X
\end{bmatrix} = -A(\tilde{\beta}^\dagger)^{-1} \begin{bmatrix} S_Z(\beta) \\ S_X(\beta) \end{bmatrix} = -A(\tilde{\beta}^\dagger)^{-1} \begin{bmatrix} E\{S_Z(\beta)\} + o_p(1) \\ E\{S_X(\beta)\} + o_p(1) \end{bmatrix},
\]

where \( A(\tilde{\beta}^\dagger) \) is the derivative matrix of \( \{S_Z(\beta)^T, S_X(\beta)^T\} \) evaluated at \( \tilde{\beta}^\dagger \), a point on the line segment joining \( \tilde{\beta}_{IMI_w} \) and \( \beta \). If we define the lower-right element in \( A(\tilde{\beta}^\dagger)^{-1} \) as \( A^{22} \), then \( E(\tilde{\beta}_{IMI_w,X} - \beta_X) \approx E[-A^{22}E\{S_X(\beta)\}] = E[-A^{22}E\{S_X(\beta)\}] \) for large \( n \) since \( E\{S_Z(\beta)\} = 0 \).
It is straightforward to show that $-A^{22}$ is a positive random variable so that asymptotically
$$\text{sign}\{E(\bar{\beta}_{IMI} - \beta_X]\} = \text{sign}\{E(S_X(\beta)] \} = -\text{sign}\{\text{Cov}(X, \mu)\} = -\text{sign}(\beta_X)$.

### 2.4.2 Estimation of $\text{Var}(\hat{\theta})$

For improper multiple imputation, Rubin’s variance formula (2.6) is not consistent for the variance of $\hat{\theta}_{IMI}$ because it does not take into account variation in the initial estimates $\hat{\beta}_I$ and $\hat{\gamma}_I$. Extending the theory in Wang and Robins (1998), Robins and Wang (2000), and Tsiatis (2006), we obtain the following asymptotic properties of $\hat{\theta}_{IMI}$.

**Theorem 3.** Consider a GLM defined by (2.1) and (2.2). Under the regularity conditions A.1-A.3 outlined in Appendix A,

$$n^{1/2}(\hat{\theta}_{IMI} - \theta) \xrightarrow{d} N(0, \Sigma),$$

where

$$\Sigma = \{I^F(\theta)\}^{-1} + (1 + m^{-1})\{I^F(\theta)\}^{-1}D\{I^F(\theta)\}^{-1} + \{I^F(\theta)\}^{-1}D\text{Var}(\bar{\theta}^I)D\{I^F(\theta)\}^{-1}, \quad (2.7)$$

$$D = \{I^F(\theta) - I(\theta)\}, \quad I^F(\theta) = E\{S^F(Y, W; \theta)S^{FT}(Y, W; \theta)\} \text{ is the full-data information matrix,}$$

$$\text{and } I(\theta) = E\{S(Y, Z, X^*, \rho; \theta)S^{T}(Y, Z, X^*, \rho; \theta)\} \text{ is the observed-data information matrix,}$$

**evaluated at the true parameter vector, $\theta$.**

To clarify, the full-data information matrix and score vector for $\theta$, $I^F$ and $S^F$, are based on the complete data with no censoring while the observed-data information matrix and score vector for $\theta$, $I$ and $S$, are based on the censored data. These matrices and vectors are derived based on the models for the response and covariates. While the theory here is driven based on a joint model for $(Y_i, Z_i, X_i)$, we note that in this dissertation we have only suggested modeling $f(y_i|z_i, x_i)$ and $f(x_i|z_i)$ in practice.

To estimate (2.7), Wang and Robins (1998) proposed estimating $I^F(\theta_0)$ and $I(\theta_0)$ using
the consistent estimators

$$
\hat{I}(\theta_0) = n^{-1} \sum_{i=1}^{n} \left\{ m(m-1) \right\}^{-1} \sum_{k,k'=1,\ldots,m, k \neq k'} S^{F}\{ (y_i, z_i, \bar{x}_{ik}) ; \hat{\theta}^{(k')} \} S^{F}\{ (y_i, z_i, \bar{x}_{ik}) ; \hat{\theta}^{(k)} \},
$$

where $S^{F} = \frac{\partial}{\partial \theta} \{ \ell(y_i, z_i, \bar{x}_{ik} ; \theta) \}$ is the score vector from the full data with log-likelihood $\ell(y_i, z_i, \bar{x}_{ik} ; \theta)$, $\bar{x}_{ik}$ denotes the $k$th vector of imputed values for the $i$th individual, and $\hat{\theta}^{(k)}$ denotes the estimated parameter vector based on the $k$th dataset. The $\text{Var}(\hat{\theta}^{I})$ can be estimated by the observed information matrix for the initial parameter estimates $\hat{\theta}^{I}$.

It is often laborious to estimate the asymptotic variance (2.7) using the consistent estimators given by (2.8). For this reason, we propose an alternative easy-to-calculate approximate variance formula,

$$
\hat{\Sigma} = V_W + (1 + m^{-1})V_B + V_B V_W^{-1} \text{Var}(\hat{\theta}^{I}) V_W^{-1} V_B,
$$

where $V_W$ and $V_B$ are the within- and between-imputation variances, respectively. This estimator can be derived by noting that Rubin’s variance formula (2.6) is consistent for the first two terms of (2.7) as $n \to \infty$ and $m \to \infty$ (Wang and Robins, 1998). Though (2.9) is not a consistent estimator for finite $m$, we have found that it works well even for a high percentage of censoring (> 50%) and a moderate number of imputed data sets, say $m \in [15, 30]$.

### 2.5 Simulation Study

We conducted a simulation study to assess the performance of the methods discussed in Sections 2.3.1-2.3.4 as well as the proposed improper multiple imputation estimator discussed in Section 2.4. We considered a situation with two covariates, $X_1$ and $X_2$, both subject to censoring, and a single binary response variable, $Y$. We let $X_1$ and $X_2$ be bivariate normal with zero mean, unit variance, and correlation $\rho = 0.5$. We assumed that $Y|X_1, X_2$ follows a logistic regression model,
$Y|X_1, X_2 \sim \text{Bernoulli}([1 + \exp\{-\beta_0 - X_1\beta_1 - X_2\beta_2\}]^{-1})$. We let \( \beta = (3, 1.5, -3)^T \) and considered three levels of censoring percentage, 20\%, 40\%, and 60\%, where the censoring percentage is defined as the proportion of cases with at least one covariate below the DL $d = d_1 = d_2$. For each scenario, the simulation was repeated 2000 times with 200 subjects. We provide results for other simulation set-ups in Appendix B.

2.5.1 Estimation of $\beta$

We obtained estimates for $\beta$ using the following methods:

1. omniscient method (Omni) in which all observations are known as if no censoring occurred;
2. complete case analysis (CC);
3. substitution method (SUB) using $DL/\sqrt{2}$ to replace the censored values;
4. conditional mean imputation (Mean1) conditioning only on whether the covariate is less than the DL;
5. conditional mean imputation (Mean2) conditioning only on the uncensored covariates;
6. conditional mean imputation (Mean3) conditioning on both the response and the uncensored covariates;
7. maximum likelihood method (ML) using a Monte Carlo EM algorithm;
8. fully Bayesian method (FB);
9. proper (Bayesian) multiple imputation (PMI);
10. improper (frequentist) multiple imputation (IMI).

The omniscient method serves as a gold standard while the IMI method is the proposed improper multiple imputation method described in Section 2.4. For the ML method, we estimated $\beta$ using the Monte Carlo EM algorithm described in May et al. (2011) with 250 samples.
for each censored observation in the E-step. For the FB and PMI methods we used Jeffrey’s prior for $\beta$ and a multivariate normal-Wishart conjugate prior for the $\gamma$ parameters indexing the distribution of the covariates. We used the Metropolis-Hastings algorithm to obtain 1000 draws from the full conditional for $\beta$. For the ML, FB, PMI, and IMI methods, we implemented the acceptance-rejection algorithm described in Section 2.4 to draw imputations for censored observations. For both the PMI and IMI methods, we obtained $m = 15$ imputed data sets. For each of the methods except the FB method we used the bias correction methods described by Firth (1993) to eliminate the first-order bias in the $\beta$ estimates for logistic regression. For the FB method, we estimated $\beta$ by the posterior mode.

To estimate the standard errors for the SUB, Mean1, Mean2, and Mean3 estimates, we used the observed Fisher’s information matrix based on the filled-in dataset. For the maximum likelihood method, we employed bootstrapping as described in May et al. (2011), while for the fully Bayesian method we used the posterior standard deviation. For the PMI and IMI method, we applied the formulas (2.6) and (2.9), respectively.

Table 2.1 summarizes the average bias, simulation (Monte Carlo) standard deviation, average estimated standard error, and empirical 95% coverage probability of the parameter estimates based on the above estimation methods for 20%, 40%, and 60% censoring. The last column of Table 2.1 shows the average elapsed time (in seconds) needed to compute the estimates for a single simulated dataset. For fair comparison, we used R (R Core Team, 2012) to program each of the methods.

The CC, ML, PMI, and IMI methods yielded estimates for $\beta$ with minimal or no bias. The CC and IMI estimates are asymptotically unbiased as proven by Proposition 1 and Theorem 2, while the unbiasedness of the ML estimates is expected since they are derived by directly maximizing the likelihood of the available data. The FB method has a slight bias while the SUB, Mean1, Mean2, and Mean3 methods are more clearly biased, as was indicated by previous literature (Helsel, 2012) and Theorem 1.
Table 2.1: Simulation results for a logistic regression on 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Time (s): average elapsed computational time in seconds for a single simulated dataset; Omni: omniscient method; CC: complete case analysis; SUB: substitution by DL/$\sqrt{2}$; Mean1: conditional mean imputation method conditioning only on the DL; Mean2: conditional mean imputation method conditioning on the covariates only; Mean3: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper (Bayesian) multiple imputation method; IMI: improper (frequentist) multiple imputation method.

<table>
<thead>
<tr>
<th>Cens. %</th>
<th>Method</th>
<th>$\beta_0 = 3$</th>
<th>$\beta_1 = 1.5$</th>
<th>$\beta_2 = -3$</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>SE</td>
<td>CP</td>
</tr>
<tr>
<td>20</td>
<td>Omni</td>
<td>0.00</td>
<td>0.43</td>
<td>0.44</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.01</td>
<td>0.48</td>
<td>0.48</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>-0.04</td>
<td>0.43</td>
<td>0.43</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>0.03</td>
<td>0.43</td>
<td>0.44</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td>0.01</td>
<td>0.45</td>
<td>0.45</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.02</td>
<td>0.45</td>
<td>0.45</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.01</td>
<td>0.43</td>
<td>0.43</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.03</td>
<td>0.45</td>
<td>0.46</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.01</td>
<td>0.44</td>
<td>0.45</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>-0.01</td>
<td>0.44</td>
<td>0.45</td>
<td>0.95</td>
</tr>
<tr>
<td>40</td>
<td>Omni</td>
<td>0.00</td>
<td>0.43</td>
<td>0.44</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.02</td>
<td>0.58</td>
<td>0.58</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>-0.14</td>
<td>0.41</td>
<td>0.41</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>-0.11</td>
<td>0.43</td>
<td>0.42</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td>-0.07</td>
<td>0.46</td>
<td>0.44</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.02</td>
<td>0.50</td>
<td>0.45</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.01</td>
<td>0.46</td>
<td>0.47</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.04</td>
<td>0.47</td>
<td>0.47</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.01</td>
<td>0.47</td>
<td>0.46</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.00</td>
<td>0.47</td>
<td>0.46</td>
<td>0.95</td>
</tr>
<tr>
<td>60</td>
<td>Omni</td>
<td>0.00</td>
<td>0.43</td>
<td>0.44</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.03</td>
<td>0.77</td>
<td>0.78</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>-0.14</td>
<td>0.37</td>
<td>0.38</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>-0.31</td>
<td>0.42</td>
<td>0.39</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td>-0.31</td>
<td>0.44</td>
<td>0.39</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.13</td>
<td>0.54</td>
<td>0.43</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.01</td>
<td>0.49</td>
<td>0.49</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.03</td>
<td>0.52</td>
<td>0.51</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.02</td>
<td>0.53</td>
<td>0.51</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>-0.02</td>
<td>0.50</td>
<td>0.50</td>
<td>0.94</td>
</tr>
</tbody>
</table>
We note that among the approximately unbiased methods, the ML, FB, PMI, and IMI methods are clearly more efficient than the CC analysis. For example, with 60% censoring, the relative efficiency gains for these estimators compared to the CC estimator range from 54% to 59% for $\hat{\beta}_0$, 55% to 63% for $\hat{\beta}_1$, and 36% to 43% for $\hat{\beta}_2$, where the relative efficiency gain is calculated by \( \frac{\text{Var}(\hat{\beta}_{CC}) - \text{Var}(\hat{\beta})}{\text{Var}(\hat{\beta}_{CC})} \), and $\hat{\beta}_{CC}$ and $\hat{\beta}$ are the estimators for the CC and unbiased methods, respectively. While the ML method is the most efficient of these methods, the FB, PMI, and IMI methods perform very comparably.

For the majority of the methods, the average of the standard error estimates is a relatively accurate estimate of the simulation standard deviation, with the exception of the Mean3 estimator. However, while coverage probabilities are good for the CC, ML, FB, PMI, and IMI methods, they are poor for the SUB, Mean1, Mean2, and Mean3 methods, especially with higher censoring.

Each of the methods was implemented in R (R Core Team, 2012) on a computer with an Intel Core i7 870 @ 2.93 GHz processor and 8 GB of RAM. We found that the IMI is much faster and easier to program than both the PMI and ML methods. Additionally, since the PMI method relies on Bayesian techniques, convergence of all the parameters must be monitored. The PMI method is shown to be faster than the FB method since we only need $m$ independent draws from the posterior predictive distribution for the data rather than the entire posterior distribution of the parameters. We also note that we included standard error estimation in the calculation for the elapsed computational time for all the methods except the ML method. Thus the ML method actually takes longer than indicated in Table 2.1.

In summary, the CC, ML, FB, PMI, and IMI methods all produce approximately unbiased parameter estimates while the SUB, Mean1, Mean2, and Mean3 methods lead to biased estimates. Among the unbiased methods, the CC analysis is inferior as it leads to estimates with higher variance. The ML, FB, PMI, and IMI estimators behave similarly, though the ML estimator slightly outperforms the others in terms of efficiency while the proposed IMI method is computationally superior.
Table 2.2: A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.001 and 0.003. SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}_{IMI}$; SE$_C$: average standard error estimated using (2.8); SE$_A$: average standard error estimated using (2.9); SE$_R$: average standard error estimated using (2.6); SE$_B$: average standard error estimated via a bootstrap procedure.

<table>
<thead>
<tr>
<th></th>
<th>20% Censoring</th>
<th></th>
<th>40% Censoring</th>
<th></th>
<th>60% Censoring</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_0$</td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
<td>$\hat{\beta}_0$</td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
</tr>
<tr>
<td>SD</td>
<td>0.440</td>
<td>0.377</td>
<td>0.520</td>
<td>0.469</td>
<td>0.407</td>
<td>0.554</td>
</tr>
<tr>
<td>SE$_C$</td>
<td>0.449</td>
<td>0.382</td>
<td>0.526</td>
<td>0.464</td>
<td>0.402</td>
<td>0.550</td>
</tr>
<tr>
<td>SE$_A$</td>
<td>0.449</td>
<td>0.382</td>
<td>0.525</td>
<td>0.464</td>
<td>0.403</td>
<td>0.550</td>
</tr>
<tr>
<td>SE$_R$</td>
<td>0.447</td>
<td>0.381</td>
<td>0.522</td>
<td>0.458</td>
<td>0.392</td>
<td>0.542</td>
</tr>
<tr>
<td>SE$_B$</td>
<td>0.469</td>
<td>0.401</td>
<td>0.538</td>
<td>0.512</td>
<td>0.433</td>
<td>0.605</td>
</tr>
</tbody>
</table>

2.5.2 Variance Estimation of $\hat{\beta}_{IMI}$

We considered four methods for estimating the variance of the proposed improper multiple imputation estimator, $\hat{\beta}_{IMI}$: (i) a consistent estimator using (2.8), (ii) the approximate estimator given by (2.9), (iii) Rubin’s variance formula (2.6), and (iv) a bootstrap procedure. We label the corresponding standard error estimates as SE$_C$, SE$_A$, SE$_R$, and SE$_B$, respectively. To implement the bootstrap procedure, we randomly resampled the individuals in the simulated dataset 50 times, calculated $\hat{\beta}_{IMI}$ for each bootstrapped dataset, and found the standard deviation of these 50 estimates. Table 2.2 gives the simulation standard deviations (SDs) of $\hat{\beta}_{IMI}$ and the standard error estimates SE$_C$, SE$_A$, SE$_R$, and SE$_B$ averaged across the 2000 simulated data sets.

As we would expect by the theory outlined in Section 2.4.2, the SE$_R$ estimates are smaller than the simulation SDs, confirming that Rubin’s formula (2.6) underestimates the variation in the improper multiple imputation estimator. The SE$_C$ and SE$_A$ estimates are very close to the simulation SDs for all three parameters, validating the corrected asymptotic variance given by (2.7) and both estimation methods (2.8) and (2.9). The SE$_B$ estimates tend to be overly high, possibly due to heavy censoring in some bootstrap samples.
2.6 Application to the GenIMS Study

We illustrate several of the analysis methods discussed in Sections 2.3 and 2.4 by applying them to the Genetic and Inflammatory Markers of Sepsis (GenIMS) dataset. One of the main purposes of the GenIMS study was to identify relationships between cytokine levels in the body and the development of severe sepsis, defined in this study as community acquired pneumonia (CAP) plus organ dysfunction. A second purpose of this study, which we focus on in this chapter, was to study the relationship between cytokine levels and 90-day survival (event of surviving at least 90 days) for patients with CAP (Kellum et al., 2007).

Cytokines are cell-signaling protein molecules that are sent out by cells in the immune system. Three cytokines were measured in this study: tumor necrosis factor (TNF), interleukin-6 (IL-6), and interleukin-10 (IL-10). The TNF and IL-6 cytokines serve as biomarkers of pro-inflammatory responses to CAP while IL-10 serves as a biomarker of anti-inflammatory responses to CAP. It has been thought that pro- and anti-inflammatory responses in the body help explain the development of severe sepsis and resulting deaths, and that understanding these relationships could be important for developing medical treatments.

The data for the GenIMS study were obtained by first enrolling individuals with CAP immediately after admission to a hospital and then collecting biological measurements and demographic information for each individual. In our analysis, we considered the 1418 patients who actually acquired CAP, necessitated a hospital stay, and had TNF, IL-6, and IL-10 measurements taken on the first day of hospitalization.

We used a logistic regression model to explain the relationship between 90-day survival and six covariates: the levels of the three cytokines, TNF, IL-6, and IL-10, on the first day of hospitalization, sex (1 representing males, 0 representing females), race (1 representing Caucasians, 0 representing all other races), and age. The three biomarkers, TNF, IL-6, and IL-10, were all subject to censoring below the detection thresholds 4, 2 or 5, and 5 pg/ml, respectively, with censoring proportions of 35.54%, 13.40%, and 46.83%, respectively. A total of 64% of the individuals had at least one of these measurements censored.
We assumed that given the sex, race, and age covariates, the natural logarithms of TNF, IL-6, and IL-10 are multivariate normal. Though our analysis was based on the logarithms of TNF, IL-6, and IL-10, we henceforth continue to refer to the log-cytokines as TNF, IL-6, and IL-10. Specifically, we assumed \((\text{TNF}_i,\text{IL-6}_i,\text{IL-10}_i)^T \overset{ind.}{\sim} N_3(\Lambda z_i, \Omega)\), where \(\Lambda\) is a 3 \times 4 matrix of mean parameters, \(z_i = (1, \text{sex}_i, \text{race}_i, \text{age}_i)^T\), and \(\Omega\) is a 3 \times 3 variance-covariance matrix. To informally verify this normality assumption, we plotted normal Q-Q plots for the log of the three cytokines, shown in Figure 2.1. These Q-Q plots are only partially complete because they show only the sample quantiles above the \(p_0^{th}\) quantile, where \(p_0\) is the sample censoring proportion, as the lower quantiles cannot be estimated nonparametrically. The Q-Q plots indicate that the marginal distributions of the log-transformed TNF, IL-6, and IL-10 covariates are approximately normal.

![Graphs showing Q-Q plots for log(TNF), log(IL-6), and log(IL-10).](image)

**Figure 2.1:** Normal Q-Q plots for censored cytokines: log(TNF), log(IL-6), and log(IL-10). The Y-axis corresponds to the sample quantiles above the \(p_0^{th}\) quantile of the cytokine biomarkers, where \(p_0\) is the sample censoring proportion. The X-axis corresponds to the quantiles of a \(N(0, 1)\).

With the multivariate normal distributional assumption for the censored covariates, we conducted a logistic regression analysis for 90-day survival using the proposed improper multiple imputation method (IMI), maximum likelihood method (ML), fully Bayesian method (FB), proper multiple imputation method (PMI), and complete case analysis (CC; based on 511
Table 2.3: Parameter and standard error estimates for coefficients of log-cytokine and demographic covariates in the logistic regression model for 90-day survival, based on the GenIMS dataset. Par.: parameter estimate; SE: standard error estimate; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th></th>
<th>ML</th>
<th></th>
<th>FB</th>
<th></th>
<th>PMI</th>
<th></th>
<th>IMI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par.</td>
<td>SE</td>
<td>Par.</td>
<td>SE</td>
<td>Par.</td>
<td>SE</td>
<td>Par.</td>
<td>SE</td>
<td>Par.</td>
<td>SE</td>
</tr>
<tr>
<td>Int.</td>
<td>8.48</td>
<td>1.11</td>
<td>6.71</td>
<td>0.54</td>
<td>6.89</td>
<td>0.56</td>
<td>6.72</td>
<td>0.56</td>
<td>6.71</td>
<td>0.56</td>
</tr>
<tr>
<td>TNF</td>
<td>-0.25</td>
<td>0.16</td>
<td>-0.08</td>
<td>0.07</td>
<td>-0.10</td>
<td>0.08</td>
<td>-0.09</td>
<td>0.09</td>
<td>-0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>IL-6</td>
<td>-0.08</td>
<td>0.07</td>
<td>-0.05</td>
<td>0.04</td>
<td>-0.04</td>
<td>0.04</td>
<td>-0.05</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>IL-10</td>
<td>-0.05</td>
<td>0.11</td>
<td>-0.08</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.06</td>
<td>-0.07</td>
<td>0.06</td>
<td>-0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.18</td>
<td>0.25</td>
<td>-0.42</td>
<td>0.15</td>
<td>-0.41</td>
<td>0.16</td>
<td>-0.42</td>
<td>0.16</td>
<td>-0.42</td>
<td>0.16</td>
</tr>
<tr>
<td>Race</td>
<td>-0.29</td>
<td>0.50</td>
<td>-0.31</td>
<td>0.27</td>
<td>-0.34</td>
<td>0.28</td>
<td>-0.32</td>
<td>0.28</td>
<td>-0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>Age</td>
<td>-0.07</td>
<td>0.01</td>
<td>-0.05</td>
<td>0.01</td>
<td>-0.06</td>
<td>0.01</td>
<td>-0.05</td>
<td>0.01</td>
<td>-0.05</td>
<td>0.01</td>
</tr>
</tbody>
</table>

All of the analysis methods produced similar parameter estimates, but the standard errors for the CC method are significantly higher. As an example, the parameter estimates based on the IMI method indicate that TNF, IL-6, and IL-10 are only minimally important for predicting the 90-day survival of patients, with p-values of 0.347, 0.263, and 0.251, respectively. However, the sex and age covariates are statistically significant in the model at the 0.05 level, with older males having lower probabilities of 90-day survival. For the CC analysis, age is the only statistically significant covariate. The computation times for the PMI, FB, and ML methods were approximately 12, 15, and 40 times longer than the proposed IMI method.

Though our analysis suggested that none of the three cytokines, TNF, IL-6, and IL-10, are
statistically significant in the logistic regression, we note that there are moderate correlations between these variables, ranging from 0.22 to 0.40 in absolute value. We repeated the analysis including only one of the cytokines in the model at a time. The associated \( p \)-values for TNF, IL-6, and IL-10 using the proposed IMI method are 0.049, 0.014, and 0.030, respectively. With a CC analysis, the \( p \)-values are 0.007, 0.028, and 0.364, respectively. Thus, based on the IMI analysis, we found that the levels of the three cytokines are important for predicting 90-day survival, but including all three in the model leads to some prediction redundancy. For the CC analysis, we were still unable to conclude that IL-10 is significant.

### 2.7 Discussion

We proposed an improper multiple imputation method for handling covariates censored due to DLs in GLMs. We have proven that the proposed estimator is consistent and asymptotically normal. We demonstrated that Rubin’s variance formula (2.6) is inadequate for improper imputation and a corrected variance estimator is needed to account for variation in the initial parameter estimates. We suggested two variance estimation methods for the improper multiple imputation estimator and showed empirically that these estimators perform well.

We also extended several missing data methods to the context of GLMs with covariates censored due to DLs and compared these and other competing methods to the proposed improper multiple imputation estimator. We showed both theoretically and empirically that conditional mean imputation methods may give biased parameter estimates. We also showed that substitution methods are inadequate while complete case analyses are inefficient. Maximum likelihood, fully Bayesian, proper multiple imputation, and improper multiple imputation methods all perform similarly with regard to estimation, but the improper multiple imputation estimator is computationally more efficient.

In this chapter, we generally assumed a multivariate normal model for \( f(x_i|z_i) \), which eases the computational burden. The implementation of the proposed methods is more challenging for other distributional models, especially when multiple covariates are subject to censoring.
However, this comment applies to each of the methods. In Chapter 5, we discuss a few approaches for dealing with non-normally distributed covariates, though future research could focus on ways to use more flexible multivariate distributions or semiparametric techniques.

In cases where censoring is on the response $Y_i$, the complete case estimator is generally both invalid and inefficient. In contrast, when censoring is on the covariates $X_i$, as in this dissertation, we showed that the complete case estimator is still valid. Additionally, different from the imputation and maximum likelihood methods, the validity of the complete case estimator in our set-up does not rely on any parametric assumption for $f(x_i|z_i)$. Then, taking into account that in practice it is difficult to assess the parametric distributional assumptions for observations below the DL, if the censoring level on $X_i$ is not high, the complete case analysis would be a good option since it is robust, valid, and easy to apply. However, in cases with heavy censoring on $X_i$, more complicated imputation methods should be considered to improve the efficiency.

Lastly, we note that if in addition to the censored covariates there is also data that is missing at random, the fully Bayesian, multiple imputation, and maximum likelihood methods can all easily be extended using standard methods. However, the complete case estimator and the improper multiple imputation method as explained in Section 2.4 may no longer be appropriate if missingness depends on the response.
Chapter 3

Survival Models with Covariates Subject to Detection Limits

3.1 Introduction

In Chapter 2, we discussed that for the motivating GenIMS study, it was of primary interest to model the binary event of survival for patients with community acquired pneumonia using several biomarkers and demographic covariates. However, it may also be of interest to study how these biomarkers are related to survival times. The survival times are subject to right censoring while the three important biomarkers are left-censored below known DLs, as discussed in Section 2.6. Our main objective in this chapter is to develop a computationally-efficient procedure for conducting inference in the context of survival models with covariates subject to DLs.

Very little research exists for survival models with covariates subject to censoring. Pugh et al. (1993), Lipsitz and Ibrahim (1996), Chen and Little (1999), and Marshall et al. (2010), among others, have considered survival models with missing covariate data. Also, Nielsen (2002) considered maximum likelihood, imputation, and weighted estimating equation approaches for survival models with coarsened at random covariates. However, only a few authors have suggested methods for modeling survival data with covariates censored at DLs.
Langohr et al. (2004) and Sattar et al. (2012) proposed methods based on using fully-parametric survival models for the case of a single interval censored predictor. Lee et al. (2003) considered survival models with a single covariate subject to DLs, though they proposed using semiparametric Cox proportional hazards models in which the relative risk function for the censored covariates is replaced by a nonparametric estimate of its expected value. More recently, D’Angelo and Weissfeld (2008) developed an indexing approach where censored covariate values are directly replaced by their conditional expectation given a linear combination of the fully observed covariates. While their method performs reasonably well, it is somewhat ad-hoc and limited to cases when no more than two covariates are subject to DLs. The traditionally used methods discussed in Chapters 1 and 2 – complete case, substitution, and conditional mean imputation analyses – can also be applied in the context of survival data, though they suffer from the same limitations.

In this chapter, we develop a straightforward, computationally-efficient improper multiple imputation method for handling multiple covariates subject to DLs in the context of accelerated failure time (AFT) models for censored survival data. So that the AFT survival model is suitably flexible, we recommend using the seminonparametric (SNP) distribution to model the error term. We establish the asymptotic consistency and normality of the improper multiple imputation estimator and propose a convenient variance estimation method. We additionally suggest an iterative version of this estimator which improves efficiency with only a few updates. Through numerical studies, we demonstrate that our proposed estimators lead to unbiased estimates that are potentially more efficient than several competing methods. We additionally show that using the flexible SNP distribution is more robust than typical parametric methods.

The remainder of this chapter is organized as follows. In Section 3.2, we review AFT models and the SNP distribution. We also explain how to fit AFT models with an SNP error term. In Section 3.3, we briefly review several conventional methods, which were explained in greater detail in Section 2.3 for GLMs, as they apply to survival AFT models. In Sections 3.4 and 3.5, we develop the proposed improper multiple imputation and iterated improper multiple imputation
methods and establish their asymptotic properties. In Section 3.6, we carry out simulations to compare the performance of the proposed methods with several simpler approaches. In Section 3.7, we apply the proposed methods to the dataset from the GenIMS study. Finally, in Section 3.8, we discuss the limitations of the proposed multiple imputation method and some avenues for further research. The technical details for the proposition and theorems appearing in this chapter are provided in Appendix C.

3.2 Seminonparametric Accelerated Failure Time Model

In this section, we present the seminonparametric accelerated failure time (SNP-AFT) model for modeling survival data, and we discuss an algorithm for fitting the model when covariates are fully observed. While we refer to the time-to-event variable as a “survival” variable, we recognize that the event of interest may in fact be different than or even unrelated to survival. Additionally, in the following, we only consider the case where the survival time is a univariate response subject to right censoring. The univariate assumption is taken since univariate survival models are most common and also because the censored covariate problem is already potentially computationally laborious. The assumption of right censoring on the survival time is for convenience only.

3.2.1 Accelerated Failure Time Model with Seminonparametric Error

The accelerated failure time (AFT) model provides a useful way of relating a $q$-dimensional vector of covariates, $W_i$, to a survival time, $T_i$, $i = 1, \ldots, n$. Specifically, the AFT model assumes that

$$\log T_i = W_i^T \beta + \sigma \epsilon_i,$$

where $\beta$ is a vector of coefficient parameters relating $W_i$ to $T_i$, $\sigma$ is a scale parameter, and $\epsilon_i$, $i = 1, \ldots, n$, are independent and identically distributed errors. Due to right censoring of the survival times, we only observe $\tilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, where $C_i$ is a censoring
Table 3.1: Common distributions for AFT models

| Distribution of $\epsilon$                  | Distribution of $T|W$       |
|---------------------------------------------|-----------------------------|
| extreme value (2 parameters)                | Weibull                     |
| extreme value (1 parameter)                 | exponential                 |
| logistic                                    | log-logistic                |
| normal                                      | log-normal                  |

random variable.

Traditionally, fitting an AFT model has required a fully-parametric approach, and thus semiparametric Cox proportional hazards models are often preferred in practice. However, AFT models have several advantages over Cox proportional hazards models. Perhaps most significantly, AFT models allow us to directly interpret the parameters $\beta$ as representing the conditional effect of the covariates on survival time. Additionally, with AFT models we can easily derive the distribution of $T_i$, which is useful for accommodating covariates subject to DLs in the improper multiple imputation algorithm described in Section 3.4.

When using AFT models, researchers generally fit a variety of distributions for the AFT model, such as those in Table 3.1, and chose a final distribution based on which fit gives the lowest value for an information criterion, such as the Akaike information criterion (AIC, Akaike, 1974) or the Bayesian information criterion (BIC, Schwarz, 1978). However, we may still wish to be more flexible. We propose using the seminonparametric (SNP) distribution to flexibly model the error term $\epsilon_i$ in (3.1). A univariate random variable $V$ is distributed SNP if its density function can be expressed as

$$f_K(v) = P_K^2(v)g(v),$$

(3.2)

where $P_K(v) = a_0 + a_1v + \cdots + a_Kv^K$ is a polynomial of order $K$, $g(v)$ is a kernel density with a moment generating function, and $\int f_K(v)dv = 1$. In this dissertation, we assume that $g(v)$ is either a standard normal or standard exponential distribution for reasons we discuss further in Section 3.2.2. When $K = \infty$, Gallant and Nychka (1987) showed that any “smooth” density
may be represented by an SNP distribution, where smooth is defined by certain smoothness and differentiability conditions. Even when considering a finite value of $K$, the SNP can be used to effectively model a wide variety of smooth distributions including skewed, heavy-or light-tailed, or multi-modal distributions.

The degree of the polynomial, $K$, in (3.2) can be thought of as a tuning parameter. For larger values of $K$, the SNP distribution is more flexible but also more complicated. In practice, we often wish to balance the goodness-of-fit of the distribution with the number of parameters involved during estimation. When $K = 0$, the SNP distribution is identical to that of the kernel density $g(v)$. In this paper, we recommend choosing $K$ using the BIC because it favors simpler densities compared to most other well-known information criteria, and the computational requirements for fitting an SNP distribution are greatly reduced with smaller values of $K$. It has been our practice to only consider $K \leq 2$ since it was shown by Tang (2008) that $K > 2$ is usually not chosen by selection criterion for most common distributional shapes.

While the SNP density is flexible, it can be challenging to maximize the likelihood given the observed data when $K > 0$. Zhang and Davidian (2001) suggested an alternative parameterization to make maximization more stable. Since we only consider a univariate survival variable $T_i$, we limit the discussion here to the one-dimensional case. They first noted that requiring $\int f_K(v)dv = 1$ is equivalent to requiring $E\{P^2_K(U)\} = 1$, where $U \sim g(u)$. Then, noting that

$$P_K(v) = \sum_{i=0}^{K} a_i v^i = a^T \cdot (1, v, v^2, ..., v^K)^T = a^T s,$$

where $a = (a_0, a_1, \cdots, a_K)^T$ and $s = (1, v, v^2, ..., v^K)^T$, it is straightforward to show that

$$E\{P^2_K(U)\} = E(a^T ss^T a) = a^T E(ss^T) a = a^T A a,$$

where $A$ can be written as $BB^T$, with $B$ an upper triangular matrix, since it is positive definite. But then we have that

$$E\{P^2_K(U)\} = a^T BB^T a = c^T c,$$
where \( \mathbf{c} = B^{T} \mathbf{a} \). So, rather than estimate the \( K+1 \) parameters in \( \mathbf{a} \), we can use a polar coordinate transformation to estimate the \( K \) parameters in \( \mathbf{c}(\phi) \), where

\[
\begin{align*}
    c_1 &= \sin(\phi_1) \\
    c_2 &= \cos(\phi_1) \sin(\phi_2) \\
    &\vdots \\
    c_K &= \cos(\phi_1) \cos(\phi_2) \cdots \sin(\phi_K),
\end{align*}
\]

and \( \phi_i \in (-\pi/2, \pi/2] \), \( i = 1, \ldots, K \). With this transformation, we still have a unique mapping from \( \mathbf{c}(\phi) \) to \( \mathbf{a} \), and we are able to focus the maximization search to a range of angles rather than the entire real line.

To demonstrate how to obtain the required matrices for finding \( \mathbf{a} \), consider the following examples. First, let \( K = 2 \) so that \( f_K(v) = (a_0 + a_1v + a_2v^2)g(v) \), subject to \( (a_0, a_1, a_2) \) giving \( \int f_K(v)dv = \int (a_0 + a_1v + a_2v^2)g(v)dv = 1 \). Then, if we let \( U \sim g(v) \), where \( g(v) \) is normal kernel,

\[
A = \mathbb{E}(ss^T) = \begin{pmatrix}
1 & U & U^2 \\
U & U^2 & U^3 \\
U^2 & U^3 & U^4
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 3
\end{pmatrix},
\]

so that by Cholesky’s decomposition we find

\[
B = \begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & \sqrt{2}
\end{pmatrix},
\]

\[
B^{-1} = \begin{pmatrix}
1 & 0 & -1/\sqrt{2} \\
0 & 1 & 0 \\
0 & 0 & 1/\sqrt{2}
\end{pmatrix}.
\]
and
\[
c = \begin{pmatrix}
\sin(\phi_1) \\
\cos(\phi_1) \sin(\phi_2) \\
\cos(\phi_1) \cos(\phi_2)
\end{pmatrix},
\]
so that
\[
a = B^{-1} c(\phi) = \begin{pmatrix}
\sin(\phi_1) - \cos(\phi_1) \cos(\phi_2) / \sqrt{2} \\
\cos(\phi_1) \sin(\phi_2) \\
\cos(\phi_1) \cos(\phi_2) / \sqrt{2}
\end{pmatrix}.
\]

Similarly, if we let \(U \sim g(v)\), where \(g(v)\) is an exponential kernel,
\[
A = E(ss^T) = E \begin{pmatrix}
1 & U & U^2 \\
U & U^2 & U^3 \\
U^2 & U^3 & U^4
\end{pmatrix} = \begin{pmatrix}
1 & 1 & 2 \\
1 & 2 & 6 \\
2 & 6 & 24
\end{pmatrix},
\]
so that
\[
B = \begin{pmatrix}
1 & 1 & 2 \\
0 & 1 & 4 \\
0 & 0 & 2
\end{pmatrix},
\]
\[
B^{-1} = \begin{pmatrix}
1 & -1 & 1 / \sqrt{2} \\
0 & 1 & -2 \\
0 & 0 & 1 / 2
\end{pmatrix},
\]
and
\[
a = B^{-1} c(\phi) = \begin{pmatrix}
\sin(\phi_1) + \cos(\phi_1) \{ \cos(\phi_2) / \sqrt{2} - \sin(\phi_2) \} \\
\cos(\phi_1) \{ \sin(\phi_2) - 2 \cos(\phi_2) \} \\
\cos(\phi_1) \cos(\phi_2) / 2
\end{pmatrix}.
\]

While in this dissertation, for reasons explained above, we only consider \(K \leq 2\), the examples above can be followed to derive \(A, B,\) and \(B^{-1}\), as well as how \(a\) relates to \(c(\phi)\), for \(K \geq 2\). For \(K < 2\), \(A\) is simply the upper left \(K \times K\) submatrix of those given in the examples above.
3.2.2 Fitting a Seminonparametric Accelerated Failure Time Model

To fit an SNP distribution for the error in (3.1), it is necessary to choose a kernel for the SNP, \(g(v)\). Since, when \(K = 0\), the SNP distribution is identical to that of the kernel density \(g(v)\), it is logical to let \(g(v)\) be a standard survival distribution. In this paper, we follow the strategy of Zhang and Davidian (2007) and Doehler and Davidian (2008) by considering two possibilities for \(g(v)\), the standard normal and standard exponential densities. These two distributional choices require slightly different approaches for fitting an SNP-AFT model. Specifically, the error term \(\epsilon_i\) in (3.1) can be modeled directly using a normal-based SNP, but for the exponential-based SNP we must model \(\exp(\epsilon_i)\), since the exponential distribution only takes positive values. Then, for either modeling choice, making the usual assumption that \(T_i/\text{upmodels} C_i \text{given} W_i\), we can estimate the parameter vector \(\theta = (\beta^T, \sigma, \phi^T)^T\) by maximizing the likelihood

\[
\mathcal{L}(\theta; \bar{T}_i, \delta_i, w_i) = \prod_{i=1}^{n} \left\{ f_T(t_i|w_i) \right\}^{\delta_i} \left\{ S_T(c_i|w_i) \right\}^{1-\delta_i},
\]

where \(f_T\) and \(S_T\) are the conditional density and survival functions of \(T_i\), respectively, and \(\bar{T}_i = \text{min}(T_i, C_i)\). Using standard transformation techniques, it is easy to show that for the normal-based SNP,

\[
f_T(t|w) = (\sigma t)^{-1} P^2_K(Q) g(Q),
\]

and

\[
S_T(c|w) = \int_{L}^{\infty} P^2_K(z) g(z) dz,
\]

where \(Q = \{\log(t) - w\beta\}/\sigma\), \(g\) is a standard normal density, and \(L = \{\log(c) - w\beta\}/\sigma\). Similarly, for the exponential-based SNP, we find that

\[
f_T(t|w) = (\sigma t)^{-1} Q P^2_K(Q) g(Q),
\]
and
\[ S_T(c|w) = \int_L^\infty P_K^2(z)g(z)dz, \]
where \( Q = \exp\left[\{\log(t) - w\beta\}/\sigma\right] \), \( g \) is the standard exponential density, and \( L = \exp\left[\{\log(c) - w\beta\}/\sigma\right] \).

Maximizing the likelihood (3.3) when \( K \geq 1 \) is not easy, largely because the SNP distribution is so flexible and the vector of parameters, \( \theta \), is potentially high-dimensional. For this reason, we suggest obtaining several sets of good starting values for \( \theta = (\beta^T, \sigma, \phi^T)^T \) and then conducting maximization algorithms from each of these sets of values. We recommend the following steps to obtain these starting values:

1. Fit the AFT model (3.1) with normal errors to get the initial estimates \( \beta^* \) and \( \sigma^* \);
2. Using \( \beta^* \) and \( \sigma^* \), evaluate the log of the likelihood given in (3.3) at a grid of points in \((-\pi/2, \pi/2)^K\), a \( K \)-dimensional hypercube covering the range of values for \( \phi \);
3. Using \( \phi^* \) from the \( r \) highest log-likelihood values in step 2, update the estimates for \( \sigma \) and the intercept \( \beta_0 \) by equating the moments of the normal AFT model fit in Step 1 with the SNP-AFT model:
   
   (a) if the SNP is normal-based, re-estimate \( \sigma \) as \( \sigma^{**} = \sigma^*/\text{SD}(V) \) and \( \beta_0 \) as \( \beta_0^{**} = \beta_0^* - \sigma^{**}\text{E}(V) \), where \( \text{E} \) and \( \text{SD} \) are the mean and standard deviation of a variable \( V \) distributed as a normal-based SNP;
   
   (b) if the SNP is exponential-based, re-estimate \( \sigma \) as \( \sigma^{**} = \sigma^*/\text{SD}\{\log(V)\} \) and \( \beta_0 \) as \( \beta_0^{**} = \beta_0^* - \sigma^{**}\text{E}\{\log(V)\} \), where \( V \) is distributed as an exponential-based SNP.

Using these \( r \) sets of starting values, \((\beta_0^{**}, \beta_1^*, \ldots, \beta_p^*, \sigma^{**}, \phi^*)\), standard maximization procedures can be used to find \( r \) sets of potential maximum likelihood estimates. The estimates that give the highest log-likelihood are then considered to be the actual maximum likelihood estimates. We recommend letting the grid search in Step 2 be fairly dense, with at least \( 10^K \) points, and selecting \( r \) values in Step 3 that are not “close” to each other, where close is defined by the
distance between points on the grid. This prevents the procedure described above from using $r$ sets of starting values that are relatively similar. Based on our personal experience, $3 \leq r \leq 5$ provides a good choice. This procedure should then be repeated for each value of $K$ and for both the exponential and normal kernels. As we suggested in Section 3.2.1, BIC can be used to choose an appropriate kernel and value for $K$.

While this maximization procedure appears complicated, it can be automated quite easily and is reasonably fast. We note that it is not necessary to use numerical integration techniques to obtain $S_T(c|z)$ since there exist closed form representations for both the normal- and exponential-based SNP. Specifically, for the normal-based SNP with $K = 2$,

$$
S_T(c|\mathbf{w}) = a_0^2 \Psi(L) + 2a_0a_1g(L) + (a_1^2 + 2a_0a_2)\{Lg(L) + \Psi(L)\} + 2a_1a_2(L^2 + 2)g(L)
$$

$$
+ a_2^2\{L^3g(L) + 3Lg(L) + 3\Psi(L)\},
$$

(3.6)

where $\Psi$ is the survival function for the standard normal distribution, $g$ is the standard normal density, and $L$ is defined as before. Similarly, for the exponential-based SNP with $K = 2$,

$$
S_T(c|\mathbf{w}) = a_0^2 \Psi(L; 1) + 2a_0a_1\Psi(L; 2) + 2(a_1^2 + 2a_0a_2)\Psi(L; 3) + 12a_1a_2\Psi(L; 4) + 24a_2^2\Psi(L; 5),
$$

(3.7)

where $\Psi(\cdot; j)$ is the survival function of a gamma density with scale parameter 1 and shape parameter $j$ and $L$ is defined as before. To obtain $S_T(c|\mathbf{w})$ for $K = 1$, we can simply let $a_2 = 0$ in the expressions above.

To summarize, we have described how to fit an AFT model with an SNP error term when covariates are fully observed. While we already provided above some motivation for exploring flexible methods for modeling survival data, we note that the SNP distribution is particularly useful in the problem at hand of censored covariates since the likelihood-based methods described for GLMs in Sections 2.3 and 2.4 rely on a parametric distribution for the response variable. That is, by using SNP-AFT models, we can provide straightforward extensions of many of the methods discussed in Sections 2.3 and 2.4, particularly “improper” multiple im-
putation. Several other flexible modeling methods, such as a mixture of normals, are much more difficult to sample from, making multiple imputation methods more computationally burdensome. Also, while the often used generalized gamma distribution may be useful for a wide variety of observed survival datasets, the SNP is much more flexible. Lastly, though we do not consider higher dimensional survival data, the SNP distribution has the advantage of being very easily extended to multiple dimensions.

In the next section, we briefly review several of the methods we discussed in Section 2.3, extending them to the context of AFT survival modeling. While we label the section “Conventional Analysis Methods,” we note that as with Section 2.3, the majority of these methods have not previously appeared in literature. We resume the assumption that a subset of the covariates in $W_i$, $Z_i$, are fully observed while the remaining covariates, $X_i$, are subject to censoring at a lower DL. We use the same notations as described in Section 2.2.

3.3 Conventional Analysis Methods

3.3.1 Complete Case Estimation

As with GLMs, the simplest way to handle data subject to DLs is to consider only the complete cases. In the context of survival data, we define the complete cases with respect to the covariates, not the survival variable. Thus, the complete case estimator, $\hat{\beta}_{CC}$, is found by maximizing the likelihood

$$\prod_{i=1}^{n} \rho_i^* \{f_T(t_i|\mathbf{w}_i)\}^{\delta_i} \{S_T(c_i|\mathbf{w}_i)\}^{1-\delta_i},$$

where $\rho_i^* = \prod_{j=1}^{q} \rho_{ij}$ and $\rho_{ij} = I(X_{ij} > d_j)$.

**Proposition 2.** Suppose assumptions C.1-C.3 in Appendix C hold and that $T_i$ and $C_i$ are conditionally independent with distributions that do not share any common parameters. Then $\hat{\beta}_{CC} \xrightarrow{p} \beta$ as $n^* \to \infty$, where $\beta$ is the true parameter vector.

Proposition 2 is true for the same reason as in GLMs, because the censoring on $X_i$ does not
depend on the response $T_i$. However, if a large percentage of individuals have covariates which are censored, the efficiency loss by using $\hat{\beta}_{CC}$ can be substantial. A complete case analysis has the advantage that the use of (3.8) does not require AFT modeling, and, instead, Cox proportional hazards models or other robust strategies could also be used.

3.3.2 Simple Imputation Methods

As described in Section 2.3.2, a simple method for dealing with censored predictors is to substitute the censored values by DL, DL/2, or DL/$\sqrt{2}$. As with complete case methods, any survival methods of choice can be employed after making these substitutions. However, this method has no statistical foundation and generally leads to biased estimates in survival modeling (D’Angelo and Weissfeld, 2008).

Alternatively, conditional mean imputation methods can be used to handle censored covariates. If imputation is done by using $E(X_i^c | X_i^o < d_i^c; \theta)$ or $E(X_i^c | Z_i, X_i^o, X_i^c < d_i^c; \theta)$, where the unknown $\theta$ is replaced by a consistent estimator, any standard survival methods can be employed since the imputation does not depend on the response variable $T_i$. However, if imputations are made by using $E(X_i^c | T_i, \Delta_i, Z_i, X_i^o, X_i^c < d_i^c; \theta)$, a distribution for the response also needs to be assumed, making AFT models the more obvious choice for analysis. After imputing the data, estimates for $\beta$ can be obtained via standard likelihood methods. However, as with GLMs, the resulting estimates and associated standard errors are generally inconsistent.

3.3.3 Maximum Likelihood Method

Another method to estimate $\beta$ is through maximizing the likelihood contributed by all the observed data. We represent this likelihood as

$$
\prod_{i=1}^{n} \int_{-\infty}^{d_i^c} \{f_T(t_i | z_i, x_i^o, x_i^c; \beta, \sigma, \phi) \}^{\delta_i} \{S_T(c_i | z_i, x_i^o, x_i^c; \beta, \sigma, \phi) \}^{1-\delta_i} f(x_i^o, x_i^c | z_i; \gamma) dx_i^c, \quad (3.9)
$$
where \( f_T \) and \( S_T \) are the conditional density and survival functions of \( T_i \) and \( \int_{-\infty}^{d_i^c} \) is an integral whose dimension corresponds to the length of \( X_i^c \). We note that in (3.9), \( X_i^c \) is a subject-specific set of covariates whose length corresponds to the number of censored covariates for the \( i^{th} \) individual. To maximize (3.9), the density and survival functions (3.4) and (3.6), or (3.5) and (3.7), may be used, with initial estimates for \( \beta^T, \sigma, \) and \( \phi \) found using the complete cases.

The likelihood (3.9) potentially involves high-dimensional integration and generally does not have a closed form. For this reason we often must resort to numerical integration when maximizing (3.9), which can be extremely slow. Additionally, since the survival function \( S_T \) in (3.9) may not have a closed form for some distributional choices for \( T_i \), the dimension of integration may be one higher than as described in Section 2.3.3 for GLMs. The Monte Carlo EM algorithm proposed by May et al. (2011) in the context of GLMs with covariates subject to DLs can be extended to survival data, though maximization is even slower since one additional variable, \( T_i \), is subject to censoring.

### 3.3.4 Bayesian Methods

In a fully Bayesian analysis of survival data, inference for \( \beta \) is carried out using the observed data posterior, given by

\[
\pi(\theta | \bar{t}, \delta, w^o) \propto \prod_{i=1}^{n} \int_{-\infty}^{d_i^c} f(\bar{t}_i | z_i, x_i^o, x_i^c; \beta, \sigma, \phi) \delta_i S(\bar{t}_i | z_i, x_i^o, x_i^c; \beta, \sigma, \phi)^{1 - \delta_i} f(x_i^o, x_i^c | z_i; \gamma) dx_i^c \pi(\gamma),
\]  

(3.10)

where \( \pi(\theta) \) is a prior for \( \theta = (\beta^T, \sigma, \gamma^T, \phi^T)^T, \ w^o = \{z_i, x_i^o, \rho_i\}_{i=1}^{n}, \) and all other notation is identical to that in Section 3.3.3.

After specifying a prior, \( \pi(\theta) \), samples from (3.10) may be obtained using Gibbs sampling. That is, we can iteratively sample from the full conditional distributions \( f(x_i^c | \beta, \sigma, \phi, \bar{t}_i, \delta_i, z_i, \ x_i^o, x_i^c < d_i^c), \ i = 1, \ldots, n, \ \pi(\beta, \sigma, \phi | \gamma, \bar{t}, \delta, z, x^o, x^c), \) and \( \pi(\gamma | \beta, \sigma, \phi, \bar{t}, \delta, z, x^o, x^c), \) where \( z = \{z_i\}_{i=1}^{n}, \ x^o = \{x_i^o\}_{i=1}^{n}, \ x^c = \{x_i^c\}_{i=1}^{n}, \) and initial values for \( \beta, \sigma, \) and \( \phi \) can be chosen as complete case estimates while initial values for \( \gamma \) can be chosen as the maximum likelihood estimates or any other reasonable values. Note that since we do not actually observe \( x^c \), we are treating
each element in this set as a parameter vector. While we suggest assuming that \( \pi(\beta, \sigma, \phi) \) is independent of \( \pi(\gamma) \), these priors could be chosen in many ways. Parameter estimates and standard errors can be obtained exactly as described in Section 2.3.4.

Rather than deal with the survival function in the likelihood part of (3.10), we may also sample from the full conditional distribution \( f(t_i|\beta, \sigma, \phi, \gamma, z_i, x_i^c, d_i < c_i) \) for those individuals where \( T_i \) is not observed, essentially treating \( T_i \) as another variable subject to known, but varying, DLs. Though this approach may sound easier, in reality it may be more computationally intensive since it is often harder to sample from this potentially complicated distribution than to deal with the survival function in the likelihood. For example, when using the exponential- or normal-based SNP distribution to model the error of the AFT model (3.1), the truncated distribution of \( T_i \) conditional on the covariates is fairly cumbersome to sample from. However, the survival function for \( T_i \) has a closed form, making it fairly straightforward to sample from the full conditionals for \( X_i^c, i = 1, \ldots, n, (\beta^T, \sigma, \phi^T) \), and \( \gamma \) conditional on \( \bar{T} \) and \( \Delta \).

As an alternative to fully Bayesian methods, we could use proper multiple imputation methods, which were described in detail in Section 2.3.4. To obtain \( m \) imputations, we must make random draws from the posterior predictor distribution of the censored data

\[
f(x_i^c|\bar{t}_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c) \propto \int f(x_i^c|\bar{t}_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \theta) \pi(\theta|\bar{t}, \delta, w^o) d\theta,
\]

where \( \pi(\theta|\bar{t}, \delta, w^o) \) is the posterior distribution given by (2.5). For each set of imputations, we can estimate \( \theta \) by maximizing the likelihood (3.3) and its variance by computing (2.6).

If we impute values for censored \( T_i \)’s in addition to censored \( X_i \)’s, then for an individual for which \( T_i > C_i \), we make draws from

\[
f(x_i^c, t_i|z_i, x_i^o, t_i > c_i, x_i^c < d_i^c) \propto \int f(x_i^c, t_i|z_i, x_i^o, t_i > c_i, x_i^c < d_i^c; \theta) \pi(\theta|\bar{t}, \delta, w^o) d\theta.
\]

Final estimates for \( \theta \) are then obtained by maximizing the likelihood (3.3) with \( \delta_i = 1, i = 1, \ldots, n \), and the variance of \( \bar{\theta} \) is estimated by (2.6). As with fully Bayesian methods, we generally do
not recommend imputing censored values of $T_i$. In the context of multiple imputation, imputing censored values of $T_i$ would not only likely increase the computational burden, but it would also decrease the efficiency of the final estimator unless the number of imputations was very large.

While the methods described in Sections 3.3.1-3.3.4 each have their individual appeal, they have various disadvantages. The complete case estimation method is inefficient, the substitution and conditional mean imputation methods are generally biased, and the maximum likelihood and Bayesian methods are computationally intensive. The computational requirements in survival modeling are even greater than for GLMs since integration may be necessary when computing the survival function and also since we suggest using the SNP distribution to model the distribution of the survival variable $T_i$. In the next section, we present our proposed improper multiple imputation algorithm for handling covariates subject to DLs in SNP-AFT survival models. This method reduces the computational burden compared to the maximum likelihood and Bayesian methods and increases the efficiency compared to the complete case method.

### 3.4 Improper Multiple Imputation Method

Our proposed improper multiple imputation algorithm is rooted in the fact that the complete case estimates in the AFT model are consistent. This allows us to obtain a consistent set of initial estimates for $\theta = (\beta^T, \sigma^T, \gamma^T)^T, \hat{\theta}^I$, which can then be used to generate $m$ sets of imputed data for the censored covariates from the distribution $f(x_i^c|\tilde{t}_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \hat{\theta}^I)$. Using the $m$ sets of imputed data for the censored covariates, $X_i^c, i = 1, \ldots, n$, final estimates for $\theta$ can be obtained using standard techniques.

Our proposed improper multiple imputation algorithm is as follows:

**Step 1.** Using all the data $(z_i, x_i^o, \rho_i), i = 1, \ldots, n$, and the DLs $d$, obtain the maximum likelihood estimate, $\hat{\gamma}^I$, as an initial estimate for $\gamma$, the parameters indexing $f(x_i|z_i)$.

**Step 2.** Using the complete cases, maximize the likelihood given by (3.3) using the SNP distribution for $f(\tilde{t}_i, \delta_i|z_i, x_i)$ with $K = 0, 1, 2$ for both the normal and exponential kernels.
Obtain the estimates, $\hat{\beta}^\ast$, $\hat{\sigma}^\ast$, and $\hat{\phi}^\ast$, for each kernel and $K$. Set $\hat{\beta}'$, $\hat{\sigma}'$, and $\hat{\phi}'$ equal to the estimates that minimize the BIC across $K$ and the two different kernel models.

**Step 3.** Form the 1st imputed dataset by generating $\bar{x}_i^c$ for each individual $i$ with censoring, $i = 1, \ldots, n$, from $f(x_i^c|\bar{t}_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \hat{\theta}'\!T\!)$ to replace the censored covariates $x_i^c$, where $\hat{\theta}' = (\hat{\beta}'T, \hat{\sigma}'T, \hat{\phi}'T, \hat{\gamma}'T)T$. Denote $\bar{x}_i$ as the resulting covariate vector for the $i$th subject that includes both $x_i^o$ and $\bar{x}_i^c$.

**Step 4.** Using the complete dataset, $(\bar{t}_i, \delta_i, z_i, \bar{x}_i)$, $i = 1, \ldots, n$, obtain estimates for $\theta = (\beta^T, \sigma, \phi^T, \gamma^T)^T$, $\hat{\theta}(1)$, by maximizing the likelihood (3.3).

**Step 5.** Repeat Steps 3 and 4 $m$ times, resulting in $\hat{\theta}(1), \ldots, \hat{\theta}(m)$.

**Step 6.** Finally, the improper multiple imputation estimator is defined as

$$\bar{\theta}_{IMI} = m^{-1} \sum_{k=1}^{m} \hat{\theta}^{(k)}.$$

Steps 2-4 of this algorithm warrant additional explanation. In Step 2, the maximum likelihood estimates for the complete cases can be obtained exactly as explained in Section 3.2.2. The algorithm for finding starting values needs to be employed for both the normal and exponential kernel as well as $K = 0, 1, 2$. In Step 4, while $K$ and the kernel for the SNP have already been fixed, several sets of starting values still need to be used to ensure that the maximum likelihood estimates are found. These starting values can be found similarly as in the algorithm described in Section 3.2.2, except that the SNP-AFT model estimated by the complete cases should be used for obtaining initial estimates rather than the normal AFT model.

In Step 3, we can obtain samples from $f(x_i^c|\bar{t}_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \hat{\theta}'\!T\!)$ using standard Monte Carlo sampling methods such as the Metropolis-Hastings algorithm. However, we must be careful in how we do this. Specifically, in order to avoid modeling the distribution of the censoring variable $C_i$, we need to make the following assumptions:

(A) $T_i \perp C_i$ given $W_i$;

(B) $C_i \perp X_i$ given $Z_i$;
the conditional distributions for $T_i$ and $C_i$ given $W_i$ have no common parameters,

where $\perp$ indicates statistical independence.

Then, if $\Delta_i = 1$ ($T_i \leq C_i$),

$$f(x_i^c|t_i, \delta_i = 1, z_i, x_i^o) = \frac{f_T(t_i|z_i, x_i)S_C(t_i|z_i, x_i^o) f(x_i^c|z_i, x_i^o)}{f(t_i, \delta_i = 1|z_i, x_i^o)} \propto f_T(t_i|z_i, x_i) f(x_i^c|z_i, x_i^o),$$

and if $\Delta_i = 0$ ($T_i > C_i$),

$$f(x_i^c|c_i, \delta_i = 0, z_i, x_i^o) = \frac{S_T(c_i|z_i, x_i) f_C(c_i|z_i, x_i^o) f(x_i^c|z_i, x_i^o)}{f(c_i, \delta_i = 0|z_i, x_i^o)} \propto S_T(c_i|z_i, x_i) f(x_i^c|z_i, x_i^o).$$

While (A) and (C) are common assumptions in survival analyses and allow us to avoid modeling the censoring variable $C_i$ when the covariates are fully observed, we now also need (B) so that $f_C$ and $S_C$ no longer involve $X_i$. In many cases, it may be reasonable to believe that $X_i$ is not related to $C_i$. However, even if they are somehow dependent, we can hope that the fully observed covariates $Z_i$ completely explain the extent of this relationship. For example, suppose we are interested in the survival time of patients with lung cancer on a particular drug. We may believe that $X_i$, a covariate of interest which is subject to DLs, is related to $C_i$ since individuals with particular levels of $X_i$ are more likely to have negative side effects and stop taking the drug. However, if we were to have a fully observed covariate, $Z_i$, which is a time-independent proxy for the overall level of side effects then it is certainly seems plausible that given $Z_i$, $X_i$ and $C_i$ are independent. If (B) is questionable, we suggest fitting a flexible parametric model for $C_i$ so that the assumption is no longer necessary.

We note that these same assumptions (A)-(C) are also necessary for the maximum likelihood and Bayesian methods described in Section 3.3, or any parametric method. Lastly, as explained in Section 3.3.4, we do not recommend imputing values for censored $T_i$’s since this is generally more computational work than fitting (3.3) which includes individuals for which $\Delta_i = 0$.

As explained in Section 2.4.2, Rubin’s variance formula (2.6) is not consistent for the variance of $\tilde{\theta}_{IMI}$ because it does not take into account variation in the initial estimates $\tilde{\beta}^I, \tilde{\sigma}^I, \tilde{\phi}^I,$ and
Let \( \hat{\gamma} \). The following theorem provides the asymptotic properties of \( \hat{\Theta}_{IMI} \) and allows us to construct a consistent variance estimator for our multiple imputation algorithm.

**Theorem 4.** Suppose assumptions (A)-(C) and C.1-C.3 given in Appendix C hold. Then, for an AFT model (3.1) with a correctly specified error distribution,

\[
n^{1/2}(\hat{\Theta}_{IMI} - \theta) \xrightarrow{d} N(0, \Sigma),
\]

where

\[
\Sigma = \{I^F(\theta)\}^{-1} + (1 + m^{-1})\{I^F(\theta)\}^{-1} D\{I^F(\theta)\}^{-1} + \{I^F(\theta)\}^{-1} D\text{Var}(\hat{\Theta}^I) D\{I^F(\theta)\}^{-1}, \tag{3.11}
\]

\[
D = \{I^F(\theta) - I(\theta)\}, \quad \hat{I}^F(\theta) = E\{S^F(\bar{T}, \Delta, W; \theta) S^T(\bar{T}, \Delta, W; \theta)\}
\]

is the full-data information matrix derived from the likelihood without censoring, and

\[
I(\theta) = E\{S(\bar{T}, \Delta, Z, X^*, \rho; \theta) S^T(\bar{T}, \Delta, Z, X^*, \rho; \theta)\}
\]

is the observed-data information matrix derived from the censored-data likelihood, evaluated at the true parameter vector, \( \theta \). Here \( S^F \) and \( S \) denote the score functions.

While we note that we rarely can *a priori* know the true error distribution for the AFT model (3.1), using the SNP distribution as described in Section 3.2.1 allows us to well approximate the error distribution and derive approximately consistent estimates for the parameters. To estimate (3.11), we can use consistent estimators for \( \hat{I}^F(\theta_0) \) and \( I(\theta_0) \) similar to those proposed in Section 2.4,

\[
\hat{I}^F(\theta) = \left( -m^{-1} \sum_{k=1}^{m} \left[ n^{-1} \sum_{i=1}^{n} \frac{\partial S^F\{(\tilde{t}_i, \delta_i, z_i, \tilde{x}_{ik}); \hat{\theta}^{(k)}\}}{\partial \hat{\theta}^T} \right] \right)^{-1}, \tag{3.12}
\]

\[
\hat{I}(\theta) = n^{-1} \sum_{i=1}^{n} (m(m - 1))^{-1} \sum_{k,k'} \sum_{i=1}^{n} S^F\{(\tilde{t}_i, \delta_i, z_i, \tilde{x}_{ik}); \hat{\theta}^{(k)}\} S^T\{(\tilde{t}_i, \delta_i, z_i, \tilde{x}_{ik'}); \hat{\theta}^{(k')}\},
\]

where \( S^F = \frac{\partial}{\partial \theta} \{\ell(\tilde{t}_i, \delta_i, z_i, \tilde{x}_{ik}; \theta)\} \) is the score vector from the full data with log-likelihood...
\( \ell(\bar{t}_i, \delta_i, z_i, \bar{x}_{ik}; \theta) \), \( \bar{x}_{ik} \) denotes the \( k^{th} \) vector of imputed values for the \( i^{th} \) individual, and \( \bar{\theta}^{(k)} \) denotes the estimated parameter vector based on the \( k^{th} \) dataset. Also, \( \text{Var}(\bar{\theta}^{(k)}) \) can be estimated by the observed information matrix for the initial parameter estimates \( \bar{\theta}^{I} \). It may seem counterintuitive to estimate \( I^F(\theta_0) \) by the inverse of the average of the inverse information matrices for the \( m \) imputed datasets rather than the average of the information matrices. However, the suggested estimator is more stable when using the SNP distribution because the estimated \( \phi \) parameters may possibly vary widely across imputed datasets. By taking the inverse before averaging, we are able to maintain very good final estimates for the sub-information matrix for the \( \beta \) parameters, which are of primary interest.

Since it is often laborious to estimate the asymptotic variance (3.12) using the consistent estimators given above, we note that the approximate variance formula given by (2.9) works in the context of survival modeling as well. We remind the reader that using \( m \in [15, 30] \) should be sufficient in most scenarios to give very good variance estimates.

### 3.5 Iterated Improper Multiple Imputation Method

As \( m \to \infty \), the multiple imputation estimator described in the previous section is equivalent to a one-step update of the EM algorithm (Robins and Wang, 2000; Tsiatis, 2006). Thus, we can iterate the multiple imputation procedure with a reasonably large \( m \) to approximate maximum likelihood estimation of the model parameters, \( \theta \). In practice, we may wish to only perform a few iterations in order to balance computational time with the efficiency gained by iterating the multiple imputation procedure.

In the following iterated improper multiple imputation procedure, define an \( s \)-dimensional vector \( m \), where \( s \) corresponds to the number of updates for the procedure, \( m_j \) is the number of imputed datasets used for each update, \( j = 1, \ldots, s \). To guarantee increased efficiency with each update, we let \( m_1 \leq m_2 \leq \cdots \leq m_s \).

**Step 1.** To find the 1st update for \( \theta, \bar{\theta}_1 \), conduct the complete improper multiple imputa-
tion procedure described in Section 3.4 with \( m_1 \) imputed datasets. Obtain the associated variance-covariance matrix, \( \text{Var}(\tilde{\theta}_1) \).

**Step 2.** To find the 2nd update for \( \theta, \tilde{\theta}_2 \), repeat the improper multiple imputation procedure with \( m_2 \) imputed datasets, but use \( \tilde{\theta}_1 \) and \( \text{Var}(\tilde{\theta}_1) \) as the initial parameter and variance estimates.

**Step 3.** To find the \( i^{th} \) update for \( \theta, \tilde{\theta}_i \), repeat the improper multiple imputation procedure with \( m_i \) imputed datasets, using \( \tilde{\theta}_{i-1} \) and \( \text{Var}(\tilde{\theta}_{i-1}) \) as the initial parameter and variance estimates.

**Step 4.** Let \( \tilde{\theta}_{IIMI} = \tilde{\theta}_s \) and \( \text{Var}(\tilde{\theta}_{IIMI}) = \text{Var}(\tilde{\theta}_s) \) be the final iterated improper multiple imputation parameter and variance estimates.

In the procedure we proposed above, the final variance estimator is easy to calculate since the estimated variance for the parameter estimates at each step is simply an update of the estimate from the previous step. Either (3.12) or (2.9) can be used to estimate the variance of the updated improper multiple imputation estimator at each step. The following result gives theoretical justification for using this estimator.

**Theorem 5.** Suppose assumptions (A)-(C) and C.1-C.3 given in Appendix C hold. Then, for an AFT model (3.1) with a correctly specified error distribution,

\[
n^{1/2}(\tilde{\theta}_{IIMI} - \theta) \xrightarrow{d} N(0, \Sigma_s),
\]

where

\[
\Sigma_s = \{I^F(\theta)\}^{-1} + (1 + m^{-1})\{I^F(\theta)\}^{-1}D\{I^F(\theta)\}^{-1} + \{I^F(\theta)\}^{-1}D\text{Var}(\tilde{\theta}_{s-1})D\{I^F(\theta)\}^{-1}, \tag{3.13}
\]

\( \text{Var}(\tilde{\theta}_{s-1}) \) is the variance of \( \tilde{\theta}_{s-1} \), the estimator for \( \theta \) at the \((s - 1)^{th}\) iteration, and all other quantities are defined as in Theorem 4.
While the iterated multiple imputation procedure and associated theory described above are based on a known number of iterations \( s \), it is not required for \( s \) to be predetermined. That is, the same procedure and theory applies if instead choosing \( s \) before conducting the analysis, we continue iterating the multiple imputation procedure until the difference between consecutive estimates is negligible. In this case, as described previously, the iterated multiple imputation estimator will simply produce the maximum likelihood estimates since it is equivalent to an EM algorithm. However, the computational requirements for obtaining the maximum likelihood estimates via this iterated multiple imputation procedure may be burdensome if \( s \) is large, especially since the number of iterations \( m \) at each step also needs to be large in order to guarantee convergence. For this reason, we recommend pre-choosing \( s \) to be relatively small unless computational time is not an issue.

3.6 Simulation Study

We conducted numerical studies to assess the performance of the proposed improper multiple imputation and iterated improper multiple imputation estimators. We set up the simulations to represent a situation similar to that in the GenIMS dataset described in Sections 2.6 and 3.7. Specifically, we generated the covariates \( Z \sim \text{beta}(3, 2) \cdot 84 + 18 \), to represent an “age” variable, and

\[
\begin{pmatrix}
X_1 \\
X_2
\end{pmatrix} \sim N\left( \begin{pmatrix}
1.4 + 0.0087Z \\
2.4 - 0.0091Z
\end{pmatrix}, \begin{pmatrix}
2.5 & 1 \\
1 & 5
\end{pmatrix} \right),
\]

to represent the log-cytokines TNF and IL-10. Observations of \( X_1 \) and \( X_2 \) were censored at the DLs \( d_1 = \log(4) \) and \( d_2 = \log(5) \), with censoring rates of 35% and 47%. Finally, we generated the survival time based on the model \( \log(T) = (1, X_1, X_2, Z)\beta + \sigma \epsilon \), where \( \beta = (5, -0.2, 0.2, -0.1)^T \).

For the first simulation design, we generated \( \exp(\sigma \epsilon) \sim \text{Weibull}(1.5) \), where \( \sigma \) is the scale parameter in the AFT model (3.1). For the second simulation design, we generated \( \exp(\sigma \epsilon) \sim p \text{Weibull}(2) + (1 - p) \text{lognormal}(0, 1) \), a mixture distribution with \( p \sim \text{Bernoulli}(0.6) \). For both simulation designs, we generated a censoring variable \( C \) from a uniform distribution so that
approximately 60% of the survival times were censored. We simulated 1000 datasets with 500
individuals per dataset.

We looked at several methods for analyzing the simulated data:

1. complete case method (CC);

2. substitution method (SUB) using $DL/\sqrt{2}$ to replace the censored covariate values;

3. improper imputation method (IMI);

4. iterated improper multiple imputation method (IIMI).

For the IMI estimator, we obtained $m = 15$ imputed datasets while for the IIMI estimator,
we performed $s = 4$ iterations with $m = (15, 15, 30, 50)^T$. We chose $s = 4$ to illustrate the
potential efficiency gains by iterating the multiple imputation procedure while still remaining
computationally reasonable. We did not compare the conditional mean imputation, maximum
likelihood, and Bayesian methods here since conditional mean imputation was shown to be
biased in GLMs and the maximum likelihood and Bayesian methods are very computationally
intensive.

For each of the methods, we considered three strategies for obtaining final estimates. First,
we analyzed the survival data assuming a Weibull distribution for the error in the AFT model.
Second, we considered fitting the AFT model with Exponential, Weibull, lognormal, and loglo-
gistic distributed error terms and choosing between these distributions using the BIC. We label
this strategy the “parametric” method. Finally, we analyzed the survival data using the SNP
distribution for the error term. We estimated the model parameters using the CC, SUB, IMI,
and IIMI methods for each of these three strategies. For the CC method, we maximized the
likelihood given by (3.8) while for the SUB, IMI, and IIMI methods we maximized the likelihood
(3.3) after obtaining imputations.

Table 3.2 gives the average bias, simulation (Monte Carlo) standard deviation (SD), and
average standard error (SE) estimates for the 1000 simulated datasets for both the Weibull and
mixture errors. These estimates are given for the CC, SUB, IMI, and IIMI analysis methods based on the Weibull, parametric, and SNP distributional assumptions described above.

For the simulation design with Weibull errors, Table 3.2 shows that the CC analysis leads to inefficient estimates while the SUB method leads to biased estimates. The parameter estimates are unbiased for the CC, IMI, and IIMI methods for all three modeling strategies. While correctly specifying the error distribution as Weibull gives lower standard errors, the parametric and SNP fitting strategies perform competitively. For the non-intercept parameters of the AFT model, using the Weibull only offers 1-6% relative efficiency gains over the SNP, with even less improvement over the parametric method, where the relative efficiency gain is calculated by \( \frac{\text{var}(\hat{\beta}_{\text{SNP}}) - \text{var}(\hat{\beta}_{\text{Weib}})}{\text{var}(\hat{\beta}_{\text{SNP}})} \). Regardless of the strategy that is used for fitting the error term, the IMI method offers substantial efficiency gains over both the MI and CC analysis methods. For example, with an SNP analysis the relative efficiency gains are 73-79% and 9-30% over the CC and MI methods, respectively. Additionally, note that the efficiency of the IMI method for the SNP strategy is nearly identical to assuming the correct distributional model.

For the simulation design with mixture errors, the CC method again gives unbiased estimates for the parametric and SNP fitting methods, but the intercept term is biased when incorrectly assuming the Weibull model. The CC estimator is inefficient for all three fitting strategies and the standard error estimation is poor when taking the Weibull or parametric approach. As before, the SUB method gives poor parameter and standard error estimates in all scenarios. As before, the SUB method gives poor parameter and standard error estimates in all scenarios. Since the Weibull distribution inaccurately models the error in the AFT model, it does poorly in terms of bias and standard error estimation even with the MI and IMI methods. Interestingly, both the parametric and SNP analysis strategies do fairly well in terms of bias for the MI and IMI methods, though the parametric approach leads to higher standard errors. Finally, we note that the IMI procedure outperforms both the CC and MI analysis in terms of efficiency. For example, with an SNP analysis the relative efficiency gains are 72-87% and 14-32% over the CC and MI methods, respectively.
Table 3.2: Simulation results for an AFT regression model with Weibull errors, left, and a mixture of Weibull and lognormal errors, right, for 1000 datasets. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CC: complete case analysis; SUB: Substitution by DL/√2; IMI: improper multiple imputation method; IIMI: iterated improper multiple imputation method

<table>
<thead>
<tr>
<th>Method</th>
<th>Weibull Errors</th>
<th>Mixture Errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β₀</td>
<td>β₁</td>
<td>β₂</td>
</tr>
<tr>
<td>Weibull CC Bias</td>
<td>-0.001</td>
<td>-0.002</td>
<td>-0.009</td>
</tr>
<tr>
<td>SD</td>
<td>0.847</td>
<td>0.140</td>
<td>0.110</td>
</tr>
<tr>
<td>SE</td>
<td>0.850</td>
<td>0.144</td>
<td>0.114</td>
</tr>
<tr>
<td>Parametric CC Bias</td>
<td>-0.038</td>
<td>-0.002</td>
<td>-0.009</td>
</tr>
<tr>
<td>SD</td>
<td>0.865</td>
<td>0.142</td>
<td>0.112</td>
</tr>
<tr>
<td>SE</td>
<td>0.855</td>
<td>0.139</td>
<td>0.110</td>
</tr>
<tr>
<td>SNP CC   Bias</td>
<td>-0.082</td>
<td>-0.002</td>
<td>-0.008</td>
</tr>
<tr>
<td>SD</td>
<td>0.939</td>
<td>0.144</td>
<td>0.113</td>
</tr>
<tr>
<td>SE</td>
<td>0.931</td>
<td>0.141</td>
<td>0.115</td>
</tr>
<tr>
<td>Weibull SUB Bias</td>
<td>5.572</td>
<td>-0.285</td>
<td>-1.551</td>
</tr>
<tr>
<td>SD</td>
<td>1.043</td>
<td>0.236</td>
<td>0.178</td>
</tr>
<tr>
<td>SE</td>
<td>0.984</td>
<td>0.154</td>
<td>0.104</td>
</tr>
<tr>
<td>Parametric SUB Bias</td>
<td>4.384</td>
<td>-0.314</td>
<td>-1.332</td>
</tr>
<tr>
<td>SD</td>
<td>0.930</td>
<td>0.190</td>
<td>0.257</td>
</tr>
<tr>
<td>SE</td>
<td>0.931</td>
<td>0.151</td>
<td>0.131</td>
</tr>
<tr>
<td>SNP SUB  Bias</td>
<td>5.888</td>
<td>-0.088</td>
<td>-0.210</td>
</tr>
<tr>
<td>SD</td>
<td>0.753</td>
<td>0.084</td>
<td>0.179</td>
</tr>
<tr>
<td>SE</td>
<td>0.654</td>
<td>0.089</td>
<td>0.079</td>
</tr>
<tr>
<td>Weibull IMI Bias</td>
<td>-0.050</td>
<td>-0.000</td>
<td>0.003</td>
</tr>
<tr>
<td>SD</td>
<td>0.430</td>
<td>0.066</td>
<td>0.059</td>
</tr>
<tr>
<td>SE</td>
<td>0.438</td>
<td>0.065</td>
<td>0.059</td>
</tr>
<tr>
<td>Parametric IMI Bias</td>
<td>-0.077</td>
<td>-0.000</td>
<td>0.003</td>
</tr>
<tr>
<td>SD</td>
<td>0.461</td>
<td>0.067</td>
<td>0.060</td>
</tr>
<tr>
<td>SE</td>
<td>0.440</td>
<td>0.066</td>
<td>0.059</td>
</tr>
<tr>
<td>SNP IMI  Bias</td>
<td>-0.035</td>
<td>-0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>SD</td>
<td>0.470</td>
<td>0.068</td>
<td>0.062</td>
</tr>
<tr>
<td>SE</td>
<td>0.469</td>
<td>0.066</td>
<td>0.061</td>
</tr>
<tr>
<td>Weibull IIMI Bias</td>
<td>0.002</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>SD</td>
<td>0.417</td>
<td>0.065</td>
<td>0.052</td>
</tr>
<tr>
<td>SE</td>
<td>0.431</td>
<td>0.063</td>
<td>0.050</td>
</tr>
<tr>
<td>Parametric IIMI Bias</td>
<td>-0.022</td>
<td>0.000</td>
<td>0.002</td>
</tr>
<tr>
<td>SD</td>
<td>0.440</td>
<td>0.065</td>
<td>0.052</td>
</tr>
<tr>
<td>SE</td>
<td>0.432</td>
<td>0.063</td>
<td>0.050</td>
</tr>
<tr>
<td>SNP IIMI Bias</td>
<td>-0.031</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>SD</td>
<td>0.430</td>
<td>0.065</td>
<td>0.052</td>
</tr>
<tr>
<td>SE</td>
<td>0.430</td>
<td>0.064</td>
<td>0.050</td>
</tr>
</tbody>
</table>
Overall, it is clear that the improper multiple imputation and iterative improper multiple imputation methods provide significant efficiency gains over the complete case method. Also, fitting the SNP-AFT model provides bias and efficiency improvements over more restrictive parametric approaches when the survival times do not follow a usual parametric form.

The simulation above was implemented in R on a computer with an Intel Core i7 870 @ 2.93 GHz processor and 8 GB of RAM. For the scenario when the true survival distribution is Weibull, the average computational times per dataset were 0.19 and 0.45 minutes for the parametric and SNP distributional assumptions, respectively, for the MI analysis method, and 0.73 and 1.25 minutes, respectively, for the IMI analysis method. For the scenario when the true survival distribution is a mixture distribution, the average computational times per dataset were 0.20 and 2.10 minutes for the parametric and SNP distributional assumptions, respectively, for the MI analysis method, and 0.74 and 14.6 minutes, respectively, for the IMI analysis method. The computational times when assuming a Weibull distribution were very similar to the parametric case, and all of the average computational times for the CC and SUB analysis methods were less than 0.2 minutes.

3.7 Application to the GenIMS Study

We illustrate the proposed analysis methods discussed in Sections 3.4 and 3.5 by applying them to the Genetic and Inflammatory Markers of Sepsis (GenIMS) dataset. As explained in Section 2.6, one of the main purposes of the GenIMS study was to identify relationships between cytokine levels in the body and the development of severe sepsis, defined in this study as community acquired pneumonia (CAP) plus organ dysfunction. A second purpose of this study, which we focus on in this chapter, was to study the relationship between cytokine levels and survival time for patients with CAP (Kellum et al., 2007).

Three cytokines were measured in this study: tumor necrosis factor (TNF), interleukin-6 (IL-6), and interleukin-10 (IL-10). The TNF and IL-6 cytokines serve as biomarkers of pro-inflammatory responses to CAP while IL-10 serves as a biomarker of anti-inflammatory re-
responses to CAP. The three biomarkers, TNF, IL-6, and IL-10, were all subject to censoring below the detection thresholds 4, 2 or 5, and 5 pg/ml, respectively, with censoring proportions of 35.54%, 13.40%, and 46.83%, respectively. A total of 64% of the individuals had at least one of these measurements censored. We used the AFT model described in Section 3.2 to explain the relationship between survival time and six covariates: the levels of the three cytokines, TNF, IL-6, and IL-10, on the first day of hospitalization, sex (1 representing males, 0 representing females), race (1 representing Caucasians, 0 representing all other races), and age.

As in Section 2.6, we assumed that given the sex, race, and age covariates, the natural logarithms of TNF, IL-6, and IL-10 are multivariate normal (refer to Figure 2.1 for an informal justification). Though our analysis was based on the logarithms of TNF, IL-6, and IL-10, we henceforth continue to refer to the log-cytokines as TNF, IL-6, and IL-10. Specifically, we assumed \((\text{TNF}_i, \text{IL-6}_i, \text{IL-10}_i)^T \sim \text{ind. } N_3(\Lambda z_i, \Omega)\), where \(\Lambda\) is a \(3 \times 4\) matrix of mean parameters, \(z_i = (1, \text{sex}_i, \text{race}_i, \text{age}_i)^T\), and \(\Omega\) is a \(3 \times 3\) variance-covariance matrix.

With the multivariate normal distributional assumption for the censored covariates, we conducted an AFT analysis using the proposed improper multiple imputation method (IMI) and iterated improper multiple imputation method (IIMI), as well as using a complete case analysis (CC; 511 individuals). We conducted these analysis using both the parametric and SNP distributional assumptions. The original intent of the GenIMS study was to study survival at 90 days since pneumonia is generally a recoverable disease and deaths after 90 days are likely unrelated to the original sickness. For this reason, our main analysis below is based on treating all survival times as censored if they are alive at 90 days. However, we also conducted an analysis where censoring is not observed until as many as 1162 days after initial hospitalization.

Table 3.3 summarizes the parameter and standard errors estimates for each of the six analysis methods, with the first set of estimates based on censoring all times after 90 days and the second set based on censoring only at the end of patient enrollment in the study. The content of our discussion below is based completely on the 90-day censoring results because these results are of primary interest.
Table 3.3: Parameter and standard errors estimates for coefficients of log-cytokine and demographic covariates in the AFT model for survival time, based on the GenIMS dataset; first set of estimates is based on using fixed 90-day censoring times, second set of estimates is based on using censoring times from last check-up time. Par.: parameter estimate; SE: standard error estimate; CC: complete case analysis; IMI: improper multiple imputation method; IIMI: iterated improper multiple imputation method.

<table>
<thead>
<tr>
<th>Parametric</th>
<th>SNP</th>
<th>Parametric</th>
<th>SNP</th>
<th>Parametric</th>
<th>SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>CC</td>
<td>IMI</td>
<td>IMI</td>
<td>IIMI</td>
<td>IIMI</td>
</tr>
<tr>
<td>Par.</td>
<td>SE</td>
<td>Par.</td>
<td>SE</td>
<td>Par.</td>
<td>SE</td>
</tr>
<tr>
<td>Int.</td>
<td>15.82</td>
<td>1.64</td>
<td>13.49</td>
<td>1.80</td>
<td>14.91</td>
</tr>
<tr>
<td>TNF</td>
<td>-0.44</td>
<td>0.23</td>
<td>-0.46</td>
<td>0.24</td>
<td>-0.19</td>
</tr>
<tr>
<td>IL-6</td>
<td>-0.15</td>
<td>0.10</td>
<td>-0.14</td>
<td>0.10</td>
<td>-0.13</td>
</tr>
<tr>
<td>IL-10</td>
<td>-0.01</td>
<td>0.17</td>
<td>0.01</td>
<td>0.17</td>
<td>-0.11</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.46</td>
<td>0.36</td>
<td>-0.49</td>
<td>0.36</td>
<td>-0.55</td>
</tr>
<tr>
<td>Race</td>
<td>-0.13</td>
<td>0.66</td>
<td>-0.02</td>
<td>0.65</td>
<td>-0.26</td>
</tr>
<tr>
<td>Age</td>
<td>-0.09</td>
<td>0.02</td>
<td>-0.08</td>
<td>0.02</td>
<td>-0.08</td>
</tr>
<tr>
<td>Int.</td>
<td>18.11</td>
<td>1.36</td>
<td>15.67</td>
<td>1.42</td>
<td>15.24</td>
</tr>
<tr>
<td>TNF</td>
<td>-0.51</td>
<td>0.22</td>
<td>-0.54</td>
<td>0.22</td>
<td>-0.22</td>
</tr>
<tr>
<td>IL-6</td>
<td>-0.05</td>
<td>0.09</td>
<td>-0.05</td>
<td>0.09</td>
<td>-0.04</td>
</tr>
<tr>
<td>IL-10</td>
<td>-0.02</td>
<td>0.15</td>
<td>-0.02</td>
<td>0.15</td>
<td>-0.04</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.71</td>
<td>0.34</td>
<td>-0.72</td>
<td>0.34</td>
<td>-0.77</td>
</tr>
<tr>
<td>Race</td>
<td>0.09</td>
<td>0.56</td>
<td>0.14</td>
<td>0.55</td>
<td>0.17</td>
</tr>
<tr>
<td>Age</td>
<td>-0.12</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.01</td>
<td>-0.09</td>
</tr>
</tbody>
</table>
We observe in Table 3.3 that all of the analysis methods produced similar parameter estimates, but the standard errors for the CC method are clearly higher. The parametric and SNP approaches for modeling the error term of the AFT model give nearly identical estimates in all cases, perhaps due to the fact that the fitted distributions for the AFT model based on the parametric and SNP approaches were very similar. Specifically, a normal distribution was chosen with the parametric approach, and a normal kernel with $K = 1$ was chosen with the SNP approach. All the methods except the CC analysis indicate that sex and age are important covariates, with males and older patients having lower survival. The SNP IIMI analysis method, for example, gives p-values of 0.028 and less than 0.001, respectively, for these two covariates.

All three of the cytokine biomarkers, TNF, IL-6, or IL-10, are moderately significant, with p-values of 0.159, 0.091, and 0.097, respectively, given by the SNP IIMI analysis. Additionally, the p-value for the test that all three biomarkers have no effect on survival, $H_0 : \beta_{TNF} = \beta_{IL-6} = \beta_{IL-10} = 0$, is less than 0.001. Thus, there is significant evidence that the three biomarkers are associated with survival. Compared to the results in Section 2.6, these p-values are much lower than in the logistic regression model for the event of survival after 90 days.

Since there are moderate correlations between the cytokines, ranging from 0.22 to 0.40 in absolute value, we repeated the analysis including only one of the biomarkers in the model at a time and found that the associated p-values for TNF, IL-6, and IL-10 for the SNP IIMI method are all less than 0.001.

### 3.8 Discussion

In this chapter, we proposed an improper multiple imputation and iterated improper multiple imputation method for handling covariates censored due to DLs in SNP-AFT survival models. We have proven that the proposed estimators are consistent and asymptotically normal, with standard errors that are reasonable to estimate. We additionally proposed complete case, conditional mean imputation, fully Bayesian, and proper multiple imputation approaches for dealing
with censored covariates in survival models. However, we explained that similar to Chapter 2, these methods lead to inefficient estimates, biased estimates, or estimates that are extremely computationally intensive to obtain.

As in Chapter 2, we generally assumed a multivariate normal model for \( f(x_i | z_i) \), which eases the computational burden. The implementation of the proposed methods is more challenging for other distributional models, especially when multiple covariates are subject to censoring. However, this comment applies to each of the methods. In Chapter 5, we discuss a few approaches for flexibly modeling covariates.

The validity of the complete case estimator in our set-up does not rely on any parametric assumption on the distribution \( f(x_i | z_i) \). We note that in practice it is difficult to assess the parametric distributional assumptions for observations below the DL. Therefore, if the censoring level on \( X_i \) is not high, the complete case analysis would be a good option since it is robust, valid, and easy to apply. However, in cases with heavy censoring on \( X_i \), the proposed multiple imputation methods are preferred to improve the efficiency.

Lastly, we note that if in addition to the censored covariates there is also data that is missing at random, the complete case estimator and the improper multiple imputation methods may no longer be appropriate if missingness depends on the response. Fortunately, the iterated improper multiple imputation estimator does not require a consistent initial estimate since, for a large number of iterations and a large number of imputations at each iteration, it is equivalent to an EM algorithm. However, when the initial estimator is not consistent, a different formula for variance estimation than that suggested in Section 3.4 must be used, such as that suggested by Louis (1982).
Chapter 4

Joint Modeling of a Binary Response and Multiple Longitudinal Covariates Subject to Detection Limits

4.1 Introduction

In biomedical studies, biomarkers are commonly measured repeatedly over time in order to capture changes within patients over the progression of a certain disease. When it is of interest to use these longitudinal biomarkers as covariates for the purpose of describing a response, joint models have become a popular modeling strategy. Seminal papers proposing joint modeling were written by Self and Pawitan (1992) and DeGruttola and Tu (1994) in the context of using CD4 counts and viral load to predict time to the development of AIDS.

A joint model is a type of model comprised of two or more submodels that are related through a common set of latent variables. Generally, linear or non-linear mixed models are proposed for describing the covariates that are measured over time and an outcome model is proposed for the
response. The random effects in the longitudinal mixed models are then incorporated into the outcome model in order to capture the effect of the covariates on the response. By including the random effects in the outcome model rather than the actual longitudinally-observed covariate values, joint models also naturally account for measurement error on the longitudinal covariates. Most commonly, the response of interest for joint modeling is a time-to-event variable, which may be described by a survival model such as the Cox proportional hazards model.

Fitting joint models is very computationally intensive and a great deal of research for joint models has revolved around computational feasibility. Original methods for obtaining parameter estimates in joint models involved two-stage approaches for which the random effects for the longitudinal covariates were first estimated and then plugged into the survival model (Self and Pawitan, 1992; Tsiatis et al., 1995). However, two-stage approaches generally lead to biased results (Dafni and Tsiatis, 1998; Tsiatis and Davidian, 2001) and therefore are not ideal. Wulfsohn and Tsiatis (1997) developed a maximum likelihood method that overcomes the issues of bias by jointly maximizing the two submodels. In order to make the maximization computationally feasible, they suggested using an expectation-maximization (EM) algorithm in which Gaussian-Hermite quadrature methods are used to compute expectations with respect to the conditional distribution of the random effects given the data. This strategy has become the standard approach for fitting joint models, but can be very computationally expensive for joint models with a large number of random effects.

In the motivating GenIMS study, biological measurements were collected daily on patients admitted to the hospital for community acquired pneumonia, generally over a one to eight day period. One of the main purposes of the study was to investigate the relationship between the event of survival after 90 days and three cytokine biomarkers, which were longitudinally measured and subject to lower detection limits (DLs). Since the response of interest, survival after 90 days, is binary, we consider fitting a logistic submodel rather than a survival submodel. This challenge, together with the fact that there are multiple longitudinal covariates of interest and these covariates are subject to censoring at DLs, makes the GenIMS dataset somewhat
Several papers have considered extensions of the standard joint model which individually address one of these three data challenges. With respect to handling multiple longitudinal covariates, Lin et al. (2002) suggested a maximum likelihood method based on using an EM algorithm where expectations in the E-step are computed using Monte Carlo integration. However, since the number of random effects in the model – corresponding to the dimension of integration in the joint model likelihood – increases with more longitudinal covariates, obtaining parameter estimates may be computationally cumbersome. For this reason, Ye et al. (2008) and Rizopoulos et al. (2009) proposed using Laplace approximations to calculate the expectations in the E-step of the EM algorithm while Proust-Lima et al. (2009) suggested a different modeling strategy altogether, where the longitudinal covariates are described by one latent process but with a single additional random effect specific to each covariate.

A few researchers have considered fitting a binary outcome submodel rather than a survival submodel. Wang et al. (2000) suggested regression calibration and estimating equation approaches to fitting a logistic or probit model with random effects for covariates. They showed that replacing the random effects by least squares estimators in the logistic model leads to biased parameter estimates in that model, the same observation as was made previously with a survival submodel. Li et al. (2004) used similar estimating equation approaches based on sufficient and conditional scores that can be applied to any GLM with subject-specific random effects and yield consistent estimators regardless of the distribution for the random effects. Li et al. (2007b) proposed a semiparametric version of the joint model in which normality of the random effects was not assumed, and then suggested using an EM algorithm or two-stage pseudo-likelihood approach to estimate the parameters in the model. Li et al. (2007a) extended the sufficient and conditional score methods suggested in Li et al. (2004) for multiple longitudinal covariates. Recently, De la Cruz et al. (2011) considered a joint model with logistic regression and a non-linear mixed effects model, and Hwang et al. (2011) extended the EM algorithm approach developed by Wulfsohn and Tsiatis (1997) to the context of a logistic regression submodel for
There exists very limited work for joint modeling with longitudinal covariates subject to DLs. Wu (2002) considered a joint model for a censored longitudinal outcome described by a non-linear mixed effects model with covariates measured with error. In a recent work, May (2011) modeled survival time for AIDS patients using CD4 counts that were repeatedly measured. Though the CD4 counts were not censored, the longitudinal model for these CD4 counts included viral load, which was subject to lower DLs. May (2011) proposed a Bayesian joint model that fits a truncated prior to the viral loads that were observed below the DL. Dong et al. (2013) also recently considered a joint model with a single longitudinal covariate subject to a lower DL. They jointly modeled this covariate and patient survival time using an accelerated failure time model and obtained estimates using Gaussian-Hermite quadrature and a quasi-Newton method for maximization.

While several authors have considered joint modeling of a binary outcome, possibly with multiple longitudinal variables or covariates subject to DLs, no author has considered all three of these data aspects simultaneously. The censoring on the covariates makes it technically challenging to extend existing methods for fitting joint models with binary outcomes, such as those based on estimating equations, while also making an EM algorithm approach more computationally intensive since quadrature integration methods are more difficult to apply with censored covariate data. Incorporating multiple longitudinal covariates subject to DLs exacerbates the computational challenges.

In this paper, we propose a likelihood-based approach for fitting the joint model. We suggest using a censored linear mixed model with normal errors for describing the longitudinal covariates. We then include the associated censored-data distributions for the longitudinal covariates in the overall joint likelihood also containing the binary response and random effects distributions. In order to overcome the computational difficulties for fitting this joint likelihood, we propose a new EM algorithm approach that uses a normal density to approximate the distribution of the random effects given the data in the E-step of the algorithm. This approximation
reduces the dimension of necessary integration to one, regardless of the number of random effects, and thus improves computational efficiency. While we focus on the special case when covariates are subject to DLs, we also describe the proposed approximate EM algorithm for joint models without censoring on the covariates. We show through simulations that our approach for fitting the joint model of a binary response and multiple longitudinal covariates subject to DLs leads to approximately consistent parameter and variance estimates while greatly reducing computational times compared to an EM algorithm based on standard Monte Carlo integration methods. Though the methods discussed in this chapter are focused on a binary response in the context of logistic regression, they could easily be extended to the class of GLMs discussed in Section 2.2.

The remainder of this chapter is organized as follows. In Section 4.2, we layout the joint modeling framework when there is no censoring on the covariates and present the approximate EM algorithm to obtain maximum likelihood estimates. We also describe two straightforward methods for variance estimation. In Section 4.3, we extend this framework to include censoring on the covariates due to DLs. In Section 4.4, we evaluate our proposed normal approximation through simulations, both when censoring is and is not present in the covariates. In Section 4.5, we apply our proposed method to the GenIMS dataset. Section 4.6 concludes the paper with discussions of limitations and avenues for further research.

### 4.2 Joint Modeling of a Binary Response and Multiple Longitudinal Covariates

In the following we present the longitudinal and logistic submodels for the covariates and binary response, as well as the joint likelihood that we wish to maximize to obtain parameter estimates. We then describe an approximate EM algorithm for maximizing the joint likelihood. In this section, we assume that the longitudinal covariates are not subject to DLs.
4.2.1 Models and Notation

Suppose we observe independent samples \( \{Y_i, X_{i1}(t_{i1}), \ldots, X_{iq}(t_{iq}), Z_i\}_{i=1}^n \), where \( Y_i \) is a binary response, \( X_{ik}(t_{ik}) = (X_{i1k}(t_{i1k}), \ldots, X_{in_{ik}k}(t_{in_{ik}k}))^T, k = 1, \ldots, q, \) are longitudinal covariate vectors with \( n_{ik} \) repeated observations at times \( t_{ik} = (t_{i1k}, \ldots, t_{in_{ik}k})^T, \) and \( Z_i \) is a \((p-q)\)-dimensional vector of baseline covariates. For convenience, we generally refer to \( X_{ik}(t_{ik}) \) simply as \( X_{ik} \), though we note that the model for \( X_{ik} \) is actually conditional on \( t_{ik} \). We assume the following linear mixed model for the \( k \)th longitudinal covariate:

\[
X_{ijk} = r_{ijk}^T \gamma_k + s_{ijk}^T b_i + \epsilon_{ijk}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_{ik}, \quad (4.1)
\]

where \( \epsilon_{ijk} \sim \text{iid} N(0, \tau_k^2) \), \( r_{ijk} \) is a possibly time-dependent, fixed-effects design vector, potentially including the baseline covariates \( Z_i \), and \( s_{ijk} \) is a possibly time-dependent random-effects design vector. We assume that the longitudinal covariates \( X_{ik}, k = 1, \ldots, q, \) are conditionally independent given the vector of baseline covariates \( Z_i \) and the vector of random effects \( b_i \), which we assume has a multivariate normal distribution,

\[
b_i = (b_{i1}^T, b_{i2}^T, \ldots, b_{iq}^T)^T \sim N(B, D),
\]

where \( B \) is the mean vector for the random effects and \( D \) is an unstructured covariance matrix. The dimensions of \( B \) and \( D \) correspond directly to the length of \( b_i \), and since we do not force \( B = 0, r_{ijk} \) should not include any of the associated terms present in \( s_{ijk} \).

We assume that the binary response \( Y_i \) is related to the longitudinal covariates \( X_{ik}, k = 1, \ldots, q, \) only through the random effects \( b_i \) and possibly \( Z_i \). Specifically, we assume the following logistic model for \( \pi(Z_i, b_i) = \Pr(Y_i = 1|Z_i, b_i) \) :

\[
\logit \{\pi(Z_i, b_i)\} = \beta_0 + \beta_1^T Z_i + \beta_2^T b_i, \quad (4.2)
\]

where \( \beta = (\beta_0, \beta_1^T, \beta_2^T)^T \) is the \((p+1)\)-dimensional vector of regression coefficients.
The joint likelihood for all the data is given by

\[
\prod_{i=1}^{n} \int_{-\infty}^{\infty} f(y_i | z_i, b_i; \beta) \left\{ \prod_{k=1}^{q} f(x_{ik} | z_i, b_{ik}; \gamma_k, \tau_k^2) \right\} f(b_i; B, D) db_i,
\]

where \( f(y_i | z_i, b_i; \beta) \) is a Bernoulli \( \{ 1 + \exp(-\beta_0 - \beta_1^T z_i - \beta_2^T b_i) \}^{-1} \) and

\[
f(x_{ik} | z_i, b_{ik}; \gamma_k, \tau_k^2) = \prod_{j=1}^{n_{ik}} f(x_{ijk} | z_i, b_{ik}; \gamma_k, \tau_k^2) = \prod_{j=1}^{n_{ik}} \frac{1}{\sqrt{2\pi\tau_k}} \exp \left\{ -\frac{(x_{ijk} - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik})^2}{2\tau_k^2} \right\}.
\]

We henceforth denote all the parameters by \( \theta = (\beta, \gamma, \tau^2, B, D) \), where \( \gamma = (\gamma_1^T, \ldots, \gamma_q^T)^T \) and \( \tau^2 = (\tau_1^2, \ldots, \tau_q^2)^T \). Also, we denote both the probability density and probability mass functions of discrete and continuous distributions by \( f \). Note that because we assume mutual conditional independence of \( Y_i \) and \( X_{ik}, k = 1, \ldots, q \), given \( b_i \) and \( Z_i \), each parameter is only involved in a single piece of the complete-data log-likelihood. In order to maximize the likelihood (4.3), we suggest using the EM algorithm, originally developed by Dempster et al. (1977).

### 4.2.2 Maximization via the EM Algorithm

The use of the EM algorithm for obtaining parameter estimates in joint models was originally proposed by Wulfsohn and Tsiatis (1997) in the context of a Cox proportional hazards submodel for a survival outcome. The details for the EM algorithm with a logistic submodel for the response variable were given by Hwang et al. (2011). While we propose a similar set-up as that given by Hwang et al. (2011), our longitudinal submodel (4.1) is more general by allowing for \( r_{ijk} \) to be non-empty and also permitting multiple longitudinal covariates subject to DLs. In this section, we describe a more traditional EM algorithm approach for obtaining parameter estimates, and, in Section 4.2.3, we propose a new approximate version of the EM algorithm that significantly increases computational efficiency by reducing the dimension of integration to one in the E-step of the algorithm.
The EM algorithm is most useful for maximizing the observed-data log-likelihood in the presence of missing data, and it does this by iterating between two steps, the E-step and M-step. In the E-step, the expected value of the log-likelihood of the complete data is calculated with respect to the missing data conditional on all of the observed data as well as a set of current parameter estimates. In the M-step, this expected log-likelihood is maximized to obtain new parameter estimates. In the context of the joint model, the observed data for each individual is \((Y_i, X_{i1}(t_{i1}), \cdots, X_{iq}(t_{iq}), Z_i)\), while the random effects in \(b_i\) are treated as missing data since these values are not directly observed. The expected value of the complete data log-likelihood in the E-step is

\[
Q(\theta|\theta^{(v)}) = \sum_{i=1}^{n} E \{ \log f(y_i|z_i, b_i; \beta) + \log f(x_{i1}|z_i, b_{i1}; \gamma_1, \tau_1^2) + \cdots + \log f(x_{iq}|z_i, b_{iq}; \gamma_q, \tau_q^2) \\
+ \log f(b_i; B, D) \},
\]

where \(\theta^{(v)} = (\beta^{(v)}, \gamma^{(v)}, \tau^{2(v)}, B^{(v)}, D^{(v)})\) is the current set of parameter estimates and the expectation in \((4.4)\), henceforth denoted as \(E_i\) since it is different for each individual \(i\), is with respect to the distribution \(f(b_i|y_i, x_{i1}, \cdots, x_{iq}, z_i; \theta^{(v)})\). To maximize \((4.4)\), most of the parameters have closed form updates at each step in the EM algorithm. Specifically, we have that

\[
\begin{align*}
\hat{B} &= \frac{1}{n} \sum_{i=1}^{n} E_i(b_i)/n, \\
\hat{D} &= \frac{1}{n} \sum_{i=1}^{n} E_i(b_i - \hat{B}^{(v)})(b_i - \hat{B}^{(v)})^{T}/n, \\
\hat{\gamma}_k &= \sum_{i=1}^{n} (R_{ik}^{T}R_{ik})^{-1} \sum_{i=1}^{n} R_{ik}^{T}E_i(x_{ik} - S_{ik}b_{ik}), \quad k = 1, \cdots, q, \\
\hat{\tau}_k^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n_{ik}} E_i(x_{ijk} - r_{ijk}^{T}\hat{\gamma}_k^{(v)} - s_{ijk}^{T}b_{ik})^2/\sum_{i=1}^{n} n_{ik}, \quad k = 1, \cdots, q.
\end{align*}
\]
where

\[
R_{ik} = \begin{bmatrix}
    r_{i1k}^T \\
r_{i2k}^T \\
    \vdots \\
r_{i\text{in}_ik}^T
\end{bmatrix}, \quad \text{and} \quad S_{ik} = \begin{bmatrix}
    s_{i1k}^T \\
s_{i2k}^T \\
    \vdots \\
s_{i\text{in}_ik}^T
\end{bmatrix}.
\]

There is no closed form update for \( \beta \), the parameter vector of primary interest. Instead, at each iteration in the maximization procedure an update for \( \beta \) can be obtained by a one-step Newton-Raphson algorithm,

\[
\bar{\beta} = \bar{\beta}^{(v)} - \left\{ Q''_{\beta}\left(\bar{\beta}^{(v)}\right)\right\}^{-1} Q'_{\beta}\left(\bar{\beta}^{(v)}\right),
\]

(4.6)

where \( Q'_{\beta}\left(\bar{\beta}^{(v)}\right) \) is the vector of first partial derivatives of (4.4) with respect to \( \beta \) and \( Q''_{\beta}\left(\bar{\beta}^{(v)}\right) \) is the matrix of second partial derivatives of (4.4) with respect to \( \beta \), both evaluated at \( \bar{\beta}^{(v)} \).

Refer to Appendix D for the derivation of this update formula and additional details.

Each estimate in (4.5) and (4.6) involves \( n \) expectations, which are taken with respect to the distributions \( f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \), \( i = 1, \ldots, n \). This conditional density for the \( i \)th individual can be written as

\[
f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) = \frac{f(b_i, y_i|x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)})}{f(y_i|x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)})} = \frac{f(y_i|x_{i1}, \ldots, x_{iq}, z_i, b_i; \theta^{(v)}) f(b_i|x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)})}{\int_{-\infty}^{\infty} f(y_i|x_{i1}, \ldots, x_{iq}, z_i, b_i; \theta^{(v)}) f(b_i|x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \, db_i},
\]

so that each of the expectations in (4.5) and (4.6) can be written as

\[
\frac{\int_{-\infty}^{\infty} h(b_i) f(y_i|x_{i1}, \ldots, x_{iq}, z_i, b_i; \theta^{(v)}) f(b_i|x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \, db_i}{\int_{-\infty}^{\infty} f(y_i|x_{i1}, \ldots, x_{iq}, z_i, b_i; \theta^{(v)}) f(b_i|x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \, db_i},
\]

(4.7)

where \( h(b_i) \) is a function of \( b_i \). The conditional distribution of \( f(b_i|x_{i1}, \ldots, x_{iq}, z_i) \) can be shown to be multivariate normal so that the expectations of form (4.7) may be calculated using adaptive Gaussian-Hermite quadrature methods (refer to Appendix E for an overview.
of Gaussian-Hermite quadrature as applied to the joint modeling problem). While adaptive Gaussian-Hermite quadrature is very computationally efficient for calculating the integrals of the form (4.7) when \( b_i \) has relatively few dimensions, it becomes very computationally expensive as the dimension of \( b_i \) increases. When the dimension of \( b_i \geq 6 \), it is generally recommended to compute the expectation using Monte Carlo methods (for example, James, 1980).

For Monte Carlo integration, we must make repeated draws from the distribution

\[
f (b_i | y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \propto f (b_i, y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \\
= f (y_i | z_i, b_i; \beta^{(v)}) \left\{ \prod_{k=1}^q f (x_{ik} | z_i, b_i; \gamma_k^{(v)}, \tau_k^{2(v)}) \right\} \\
\times f (b_i; B^{(v)}, D^{(v)}).
\]

These draws can be made using standard sampling techniques (see Gilks et al., 1996, for example). While a large number of draws must be obtained in order to well approximate the necessary expectations, Monte Carlo integration is comparatively more efficient for computing high-dimensional integrals than Gaussian quadrature methods.

### 4.2.3 Approximate EM Algorithm

When the number of longitudinal covariates or the number of random effects in (4.3) is high, both adaptive Gaussian-Hermite quadrature methods and Monte Carlo methods can be extremely slow. For this reason, we propose approximating \( f (b_i | y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \) by a multivariate normal distribution. Then, by taking advantage of the fact that linear combinations of \( b_i \) would also be normal, we show that the dimension of integration in our set-up can be reduced to one regardless of the number of random effects. Specifically, we assume that

\[
f (b_i | y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta^{(v)}) \approx N (\hat{b}_i, \hat{\Sigma}_i),
\]
where \( \hat{b}_i = \arg\max_{b_i} \{ \log f(y_i, x_i, z_i, b_i; \theta^{(v)}) \} \) and

\[
\widehat{\Sigma}_i = \left\{ -\frac{\partial^2 \log f(y_i, x_i, z_i, b_i; \theta^{(v)})}{\partial b_i \partial b_i^T} \right|_{b_i = \hat{b}_i} \right\}^{-1}.
\]

That is, we propose approximating the conditional distribution for \( b_i \) using a multivariate normal centered at the posterior mode of \( f(b_i|y_i, x_i, z_i; \theta^{(v)}) \) and scaled by the estimated variance of this mode estimate.

Heuristically, for a single longitudinal covariate, the normal approximation (4.8) follows because as \( n_i \to \infty \), the density for \( f(b_i|y_i, x_i, z_i; \theta) \), given by

\[
f(b_i|y_i, x_i, z_i; \theta) \propto f(y_i|z_i, b_i; \beta) \left\{ \prod_{j=1}^{n_i} f(x_{ij}|z_i, b_i; \gamma, \tau^2) \right\} f(b_i; B, D),
\]

is dominated by the term \( \prod_{j=1}^{n_i} f(x_{ij}|z_i, b_i; \gamma, \tau^2) \). When there are multiple longitudinal covariates, multivariate normality follows as \( n_{ik} \to \infty \ \forall \ k = 1, \ldots, q \), though in simulations we have found that the approximation does well even when \( n_{ik} = 1 \ \forall \ k = 1, \ldots, q \).

Normal approximations for the conditional posterior of \( b_i \) have been proposed previously in a variety of contexts, including linear and generalized linear mixed models (Baghishani and Mohammadzadeh, 2012) and joint models with a survival outcome (Rizopoulos, 2012a). In the latter paper, Rizopoulos (2012a) stated that using the Bayesian central limit theorem, with a single longitudinal covariate, as \( n_i \to \infty \), \( f(b_i|\bar{t}_i, \delta_i, x_i; \theta) \overset{p}{\to} N(\bar{b}_i, \Sigma_i) \), where \( \bar{t}_i \) is the minimum of a survival and censoring time, \( \delta_i \) is a censoring indicator, \( \bar{b}_i = \arg\max_{b_i} \{ \log f(x_i|b_i) \} \), and \( \Sigma_i = \left\{ -\partial^2 \log f(x_i|b_i)/\partial b_i \partial b_i^T \right|_{b_i = \hat{b}_i} \right\}^{-1} \). The purpose of this approximation was to eliminate the need to update quadrature points at each step in an EM algorithm when using adaptive Gaussian-Hermite quadrature methods and instead calculate them only once by fitting a linear mixed model. We propose using a similar approximation, (4.8), in the context of a joint model with a binary outcome described by a logistic model. However, rather than using the normal approximation to calculate quadrature points, we use it to describe the entire distribution \( f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta) \) at each step in the EM algorithm. For this reason, instead of using the
asymptotic posterior mode suggested by Rizopoulos (2012a), \( \hat{\theta}_i = \arg\max_{\theta} \{ \log f(x_i|b_i) \} \), we suggest using \( \hat{b}_i = \arg\max_{b_i} \{ \log f(y_i, x_i, z_i, b_i; \theta) \} \) since it is more accurate for finite \( n_i \).

With this normal approximation (4.8), the expectations in (4.5) and (4.6) are calculated with respect to one-dimensional normal distributions. In the update for \( \gamma_k \), the expectation \( E_i(S_{ik}b_{ik}) = E_i \{ (s^T_{ik}b_{ik}, \ldots, s^T_{im_{ik}}b_{ik})^T \} \) is now computed component-wise with respect to the one-dimensional normal distributions \( N(s^T_{ijk}\hat{b}_{ik}, s^T_{ijk}\hat{\Sigma}_{ik}s_{ijk}) \), \( j = 1, \ldots, n_{ik} \), where \( \hat{\Sigma}_{ik} \) is the sub-matrix of \( \hat{\Sigma} \) associated with the random effects vector \( b_{ik} \). Similarly, in the update for \( \tau^2_k \), the expectations \( E_i(r^T_{ijk}\hat{\gamma}_k + s^T_{ijk}b_{ik}) \), \( j = 1, \ldots, n_{ik} \), are calculated with respect to the one-dimensional normal distributions \( N(r^T_{ijk}\hat{\gamma}_k + s^T_{ijk}\hat{b}_{ik}, s^T_{ijk}\hat{\Sigma}_{ik}s_{ijk}) \), \( j = 1, \ldots, n_{ik} \). We can also show that in the update for \( \beta \) (refer to Appendix D for details), we need to obtain the expectation \( E_i \left[ \log \{ 1 + \exp (\beta_0 + \beta_1^T Z_i + \beta_2^T b_i) \} \right] \), which using (4.8) is with respect to the one-dimensional normal distribution \( N(\beta_0 + \beta_1^T Z_i + \beta_2^T b_i, \beta_1^T \hat{\Sigma}_i \beta_1, \beta_2^T \hat{\Sigma}_i \beta_2) \). Most simply, with the normal approximation, updates for \( B \) and \( D \) do not require any integration since \( E_i(b_i) = \hat{b}_i \) and \( E_i(b_i - B)(b_i - B)^T = \hat{\Sigma}_i + (\hat{b}_i - B)(\hat{b}_i - B)^T \). Gaussian-Hermite quadrature or Monte Carlo methods can be applied to quickly calculate each of the one-dimensional expectations.

To summarize, we propose the following algorithm for obtaining approximate maximum likelihood estimates, with calculation tips based on R software (R Core Team, 2012):

1. Using all the data, \( \{ y_i, x_{i1}(t_{i1}), \ldots, x_{iq}(t_{iq}), z_i \}_{i=1}^n \), obtain an initial estimate for \( \theta, \hat{\theta}^{(1)} \).

   This can be done by fitting separate linear mixed models for each longitudinal covariate to get estimates for \( \gamma_k \) and \( \tau^2_k \), \( k = 1, \ldots, q \). The matrix \( D \) can then be estimated as the sample covariance matrix of the posterior estimates for the random effects. Initial estimates for \( \beta \) can be obtained by fitting a logistic model for \( Y_i \) based on \( Z_i \) and the posterior estimates for the random effects from the fitted linear mixed models. With R, the “lmer” function in the “lme4” package (Bates et al., 2012) can be used to obtain the linear mixed model parameter estimates and the “glm” function can be used to obtain the logistic model parameter estimates.

2. Maximize \( f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \hat{\theta}^{(1)}) \) with respect to \( b_i \), \( i = 1, \ldots, n \), in order to obtain \( \hat{b}_i \).
and \( \hat{\Sigma}_i \). With R, the “nlm” or “optim” function can be used to obtain these estimates.

3. Update \( \hat{\theta}^{(1)} \) as \( \hat{\theta}^{(2)} \) by using (4.5) and (4.6), obtaining the expected values by assuming that \( f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \hat{\theta}^{(1)}) \approx N(\hat{b}_i, \hat{\Sigma}_i), \ i = 1, \ldots, n. \)

4. Repeat Steps 2-3, each time obtaining \( \hat{\theta}^{(v+1)} \) based on \( \hat{\theta}^{(v)} \).

5. When \( \hat{\theta}^{(v+1)} \) is sufficiently close to \( \hat{\theta}^{(v)} \), define \( \hat{\theta} = \hat{\theta}^{(v+1)} \) as the vector of approximate maximum likelihood estimates.

In Step 5, we suggest defining consecutive estimates for \( \theta \) as close when

\[
\max \left( \frac{\left| \theta^{(v+1)}_1 - \theta^{(v)}_1 \right|}{\left| \theta^{(v)}_1 \right|}, \ldots, \frac{\left| \theta^{(v+1)}_s - \theta^{(v)}_s \right|}{\left| \theta^{(v)}_s \right|} \right) < \epsilon,
\]

where \( s \) is the number of elements in \( \theta \) and \( \epsilon \) is a predefined distance. In practice, we have found that this approximate EM algorithm works well. However, this algorithm could also be used to quickly find an initial estimate for \( \theta \), which could then be used as the starting value in an EM algorithm where the true distribution \( f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \hat{\theta}) \) is used in each E-step.

While several methods are available for estimating the variances of the maximum likelihood estimator \( \hat{\theta} \), we suggest using the approach forwarded by Rizopoulos (2012b) based on directly calculating derivatives of the observed-data likelihood functions. To obtain the variance estimate, first note that the observed-data score can be represented by

\[
S(\theta) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} A(\theta, b_i) f(b_i|y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta) db_i, \quad (4.9)
\]

where

\[
A(\theta, b_i) = \frac{\partial}{\partial \theta} \left\{ \log f(y_i|z_i, b_i; \beta) + \log f(x_{i1}|z_i, b_{i1}; \gamma_1, \tau_1^2) + \cdots + \log f(x_{iq}|z_i, b_{iq}; \gamma_q, \tau_q^2) + \log f(b_i; B, D) \right\}.
\]
Then, the contribution to the Hessian matrix by the $i^{th}$ individual can be written as

$$
\frac{S_i(\theta)}{\partial \theta^T} = \int_{-\infty}^{\infty} \frac{\partial A(\theta, b_i)}{\partial \theta^T} f(b_i | y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta) \, db_i \\
+ \int_{-\infty}^{\infty} A(\theta, b_i) \{ A(\theta, b_i) - S_i(\theta) \}^T f(b_i | y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta) \, db_i. \tag{4.10}
$$

Finally, we can calculate the variance of the parameter estimates as

$$
\widehat{\text{Var}}(\hat{\theta}) = \left\{ -\sum_{i=1}^{n} \frac{S_i(\theta)}{\partial \theta^T} \right\}^{-1} . \tag{4.11}
$$

We suggest using the normal approximation (4.8) to approximate the integrals in (4.9) and (4.10), though since these calculations only need to be made once, Monte Carlo integration based on the true distribution $f(b_i | y_i, x_{i1}, \ldots, x_{iq}, z_i; \theta)$ is also computationally reasonable. Since the estimates obtained via the proposed approximate EM algorithm are not technically maximum likelihood estimates, a sandwich (robust) variance estimator may be obtained as

$$
\widehat{\text{Var}}(\hat{\theta}) = \left\{ -\sum_{i=1}^{n} \frac{S_i(\theta)}{\partial \theta^T} \right\}^{-1} \left\{ -\sum_{i=1}^{n} \frac{S_i(\hat{\theta}) S_i(\hat{\theta})^T}{\partial \theta^T} \right\}^{-1} . \tag{4.12}
$$

Alternatively, we could use the variance estimation method proposed by Louis (1982) which is very useful in the context of Monte Carlo EM algorithms. To obtain this variance estimate, we can first generate $r$ values, $b_i^{(1)}, \ldots, b_i^{(r)}$, from the final approximated posterior distribution of $f(b_i | y_i, x_{i1}, x_{i2}, \ldots, x_{iq}, z_i; \hat{\theta})$, which we may assume is approximately $N(\bar{b}_i, \bar{\Sigma}_i)$, with $\bar{b}_i$ and $\bar{\Sigma}_i$ obtained after the final iteration. After obtaining a set of samples for each individual, $i = 1, \ldots, n$, we may estimate the covariance matrix of the parameters by

$$
\widehat{\text{Var}}(\theta) = Q''(\hat{\theta}) + \frac{\sum_{i=1}^{n} \{ Q'_i(\hat{\theta}) - \bar{Q}_i(\hat{\theta}) \}^2}{n - 1}
$$

where $Q''$ is the matrix of second partial derivatives of (4.4) and $Q'_i$ is the vector of first partial derivatives of (4.4) for the $i^{th}$ individual, both of which are estimated using the $r$ Monte Carlo
samples. Since we are sampling from a normal distribution and the standard error only needs to be estimated once, these computations are relatively fast. We could also construct a sandwich variance estimator similar to (4.12) by noting that the $i^{th}$ component of the observed-data score vector is the same as $Q'_i(\hat{\theta})$.

### 4.3 Joint Modeling of a Binary Response and Multiple Covariates Subject to Detection Limits

When the longitudinal covariates are subject to DLs, the joint model described in Section 4.2 no longer applies. Additionally, the methods for dealing with censoring due to DLs described in Chapters 2 and 3 do not extend directly, though we could still naively impute censored values using functions of the DLs. Specifically, multiple imputation and conditional mean imputation methods would no longer be straightforward since imputations would need to be based on both current and past observed data. Also, while fully Bayesian methods can be applied (see May et al., 2011 in the context of joint models for a single longitudinal covariate and a survival outcome), they are computationally intensive. A complete case analysis could be performed by excluding all individuals with any censored observations over time or by excluding only those observations which are censored, but it is not clear when such approaches would lead to consistent parameter estimates, especially with the presence of additional missing data that is very common in longitudinal studies.

In the following, we recommend an extension of the joint likelihood approach described in Section 4.2 to incorporate the censoring on the longitudinal covariates in the linear mixed submodels. We also propose an adaptation of the approximate EM algorithm described in Section 4.2.3 that is still much faster than a traditional EM approach.
4.3.1 Longitudinal Model with Censoring

Suppose we observe independent samples \( \{ Y_i, X^*_i(t_{i1}), \ldots, X^*_i(t_{iq}), \rho_{i1}, \ldots, \rho_{iq}, Z_i \}_{i=1}^n \) where \( Y_i \) is a binary response, \( X^*_i(t_{ik}) = (\max\{ X_{i1k}(t_{i1k}), d_k \}, \ldots, \max\{ X_{in_{ik}k}(t_{in_{ik}k}), d_k \})^T, \) \( k = 1, \ldots, q, \) are longitudinal covariate vectors with \( n_{ik} \) repeated observations at times \( t_{ik} = (t_{i1k}, \ldots, t_{in_{ik}k})^T, \) \( \rho_{ik} = (\rho_{i1k}, \ldots, \rho_{in_{ik}k})^T, \) \( k = 1, \ldots, q, \) are vectors of censoring indicators, \( Z_i \) is a \((p-q)\)-dimensional vector of baseline covariates, and \( d = (d_1, \ldots, d_q)^T \) is a vector of lower DLs corresponding to the \( q \) longitudinal covariates. For the same reason as stated in Section 4.2.1, we generally refer to \( X^*_i(t_{ik}) \) simply as \( X^*_{ik}. \)

We continue to assume the longitudinal and logistic submodels (4.1) and (4.2) so that the joint likelihood for the longitudinal covariates and binary response can still be represented by (4.3). However, we propose that the contribution to the likelihood for the \( k^{th} \) longitudinal covariate should be represented by

\[
f(x^*_{ik} | \rho_{ik}, z_i, b_{ik}; \gamma_k, \tau_k^2) = \prod_{j=1}^{n_{ik}} \left[ \frac{1}{\tau_k} \phi \left( \frac{x_{ijk} - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik}}{\tau_k} \right) \right]^\rho_{ijk} \times \left[ \Phi \left( \frac{d_k - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik}}{\tau_k} \right) \right]^{1-\rho_{ijk}}, \tag{4.13}
\]

where \( \phi \) is the standard normal density and \( \Phi \) is the cdf of the standard normal distribution. That is, we propose using the cdf of the standard normal to represent the contribution to the overall likelihood for all observations that are observed to be below the DLs. This approach has been taken by numerous authors in the context of linear mixed models with censored responses (Pettitt, 1986; Lyles et al., 2000; Vock et al., 2012, among others).

4.3.2 Approximate EM Algorithm with Multiple Covariates Subject to Detection Limits

The likelihood given by (4.3) together with (4.13) can be maximized using an EM algorithm very similar to that described in Section 4.2.2. The main difference is that there are no closed
form representations for $\hat{\gamma}$ and $\hat{\tau}^2$ at each step in the algorithm since the part of the likelihood involving these parameters includes the standard normal cdf. However, an update for $(\gamma, \tau^2)$ may still be obtained by using the one-step Newton-Raphson algorithm,

$$
\begin{pmatrix}
\hat{\gamma} \\
\hat{\tau}^2
\end{pmatrix}
= \begin{pmatrix}
\hat{\gamma}^{(v)} \\
\hat{\tau}^{2(v)}
\end{pmatrix}
- \left\{ Q''_{\hat{\gamma}, \hat{\tau}^2} (\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)}) \right\}^{-1}
\begin{pmatrix}
Q'_{\hat{\gamma}} (\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)}) \\
Q'_{\hat{\tau}^2} (\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)})
\end{pmatrix},
$$

(4.14)

where $Q'_{\hat{\gamma}} (\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)})$ and $Q'_{\hat{\tau}^2} (\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)})$ are the vectors of first partial derivatives of (4.4) with respect to $\gamma$ and $\tau^2$, respectively, and $Q''_{\hat{\gamma}, \hat{\tau}^2} (\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)})$ is the matrix of second partial derivatives of (4.4) with respect to $\gamma$ and $\tau^2$, all of which are evaluated at $(\hat{\gamma}^{(v)}, \hat{\tau}^{2(v)})$. Refer to Appendix D for the derivation of this updating formula and additional details.

As before, in order to compute the expectations in the E-step of the EM algorithm, Monte Carlo integration may be used. While Gaussian quadrature methods may also still be applied, the conditional distribution $f(b_i| x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i)$ is no longer normal, and thus is less convenient and may require more evaluation points for similar accuracy.

With either Monte Carlo or Gaussian quadrature integration methods, calculating the expectations is very slow for high-dimensional random effects. For this reason, we propose using the normal approximation suggested in Section 4.2.3. With a single longitudinal covariate subject to censoring, the normal approximation follows when the number of uncensored values grows large compared to the number of censored values, since then the density for $f(b_i| x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \theta)$ is dominated by the term $\prod_{j=1}^{n_i} f(x_{ij}| z_i, b_i)$. With multiple longitudinal covariates, the normal approximation follows as the number of uncensored values grows large for each of the covariates. Unfortunately, for those individuals with very few or no observations above the DLs, $f(b_i| y_i, x_i^*, \rho, z_i; \theta)$ will not necessarily be approximately normal. Thus, we must modify the approximate EM algorithm described in Section 4.2.3. Specifically, we propose using the normal approximation when reasonable and Monte Carlo integration for those individuals with very few covariates observed above the DLs.

To empirically get a better idea at how the normal approximation performs at various levels
Figure 4.1: Distribution of the random effects conditional on the data in a joint model with a single longitudinal covariate described by a linear mixed model with a random intercept, $b_0$, and a random slope, $b_1$, using simulated data and various degrees of censoring.

of censoring, we plotted the marginal posterior distributions of the random effects in a few scenarios. Figure 4.1 displays the marginal distributions of $f(b_i|y_i, x_i^*, \rho, z_i; \theta)$ for a case with a single covariate and a random intercept and a random slope, $b_0$ and $b_1$, for varying degrees of censoring. To generate the distributions in Figure 4.1, we considered data from individuals in the simulated dataset described in Section 4.4 with the true parameter values plugged into $f(b_i|y_i, x_i^*, \rho, z_i; \theta)$. We see that even for high degrees of censoring the normal approximation appears quite reasonable. However, when all longitudinal observations are censored at the DLs, the marginal distribution of the random slope conditional on the data becomes very skewed.
Figure 4.2: Distribution of the random effects conditional on the data in a joint model with three longitudinal covariates described by linear mixed models with a random intercept and slope term, using simulated data and various degrees of censoring. The intercepts and slope random effects for the three covariates are represented by \((b_0, b_2, b_4)\) and \((b_1, b_3, b_5)\), respectively.
Similarly, Figure 4.2 displays the marginal distributions of \( f(b_i \mid y_i, x_{i1}^*, x_{i2}^*, x_{i3}^*, \rho, z_i; \theta) \) for a case with three covariates, each having a random intercept and a random slope for a total of six random effects, \((b_0, b_1, b_2, b_3, b_4, b_5)\), for varying degrees of censoring. We again considered data from individuals in the simulated dataset described in Section 4.4 to generate these distributions. We see that the normal approximation seems reasonable when there are at least a few covariate values observed above the DLs, even if no values for one or two of the covariates are observed to be above the DLs.

Based on our experience and empirical examples such as those displayed in Figures 4.1 and 4.2, we recommend using the normal approximation if at least 20% of the total number of longitudinal observations for an individual are observed above the DLs, regardless of the number of longitudinal covariates \(q\). Then, to summarize, we propose the following algorithm for obtaining approximate maximum likelihood estimates in the presence of censoring on the longitudinal covariates, with calculation tips based on R software (R Core Team, 2012):

1. Using all the data, \( \{y_i, x_{i1}^*(t_{i1}), \ldots, x_{iq}^*(t_{iq}), \rho_{i1}, \ldots, \rho_{iq}, z_i\}_{i=1}^n \), obtain an initial estimate for \( \hat{\theta}^{(1)} \). This can be done by fitting separate censored linear mixed models for each longitudinal covariate to get estimates for \( \gamma_k \) and \( \tau_k^2, k = 1, \ldots, q \). The matrix \( D \) can then be estimated as the sample covariance matrix of the posterior estimates for the random effects. Initial estimates for \( \beta \) can be obtained by fitting a logistic model for \( Y_i \) based on \( Z_i \) and the posterior estimates for the random effects from the fitted censored linear mixed models. With R, the “lme” package (Vaida and Liu, 2012) can be used to obtain the censored linear mixed model parameter estimates and the “glm” function can be used to obtain the logistic model parameter estimates.

2. For those individuals with at least 20% of the total number of longitudinal observations observed above the DLs, maximize \( f(b_i \mid y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \hat{\theta}^{(1)}) \) with respect to \( b_i, i = 1, \ldots, n \), in order to obtain \( \hat{b}_i \) and \( \hat{\Sigma}_i \). With R, the “nlm” or “optim” function can be used to obtain these estimates. For those individuals with less than 20%
of the total number of longitudinal observations above the DLs, obtain a sample from $f(b_i|y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \hat{\theta}^{(1)})$ using Monte Carlo sampling techniques. With R, the function “Metrop1R” in the package “MCMCpack” (Martin et al., 2011) can be used to obtain these samples.

3. Update $\tilde{\theta}^{(1)}$ as $\tilde{\theta}^{(2)}$ by using (4.5) for $B$ and $D$, (4.6) for $\beta$, and (4.14) for $\gamma$ and $\tau^2$, obtaining the expected values by assuming $f(b_i|y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \hat{\theta}^{(1)}) \approx N(\hat{b}_i, \hat{\Sigma}_i)$ for each individual $i$ with at least 20% of the total number of longitudinal observations observed above the DLs and using Monte Carlo methods based on the true posterior distribution otherwise.

4. Repeat Steps 2-3, each time obtaining $\tilde{\theta}^{(v+1)}$ based on $\tilde{\theta}^{(v)}$.

5. When $\tilde{\theta}^{(v+1)}$ is sufficiently close to $\tilde{\theta}^{(v)}$, define $\tilde{\theta} = \tilde{\theta}^{(v+1)}$ as the vector of approximate maximum likelihood estimates.

While in our experience this algorithm has worked well in practice, if desired, the normal approximation can be suspended after a certain number of iterations and the EM algorithm can be completed using the true conditional distribution of the random effects. This should generally be faster than using the EM algorithm without the approximation, and the final estimate for $\theta$ will be the maximum likelihood estimate. Variance estimation can be conducted as described in Section 4.2.3, but for those individuals with greater than 80% censoring, we suggest using the true distribution $f(b_i|y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \tilde{\theta}^{(v)})$ to calculate the integrals in (4.9) and (4.10). Since the variance only needs to be calculated once, rather than at each step in the EM algorithm, the additional computation requirements are minimal.

4.4 Simulation Study

We conducted a series of simulations to assess the performance of the proposed approximate EM algorithm with and without censoring on the covariates. In Section 4.4.1, we use numerical
studies to assess the effectiveness of the proposed approximate EM algorithm when there is a single longitudinal covariate. In Section 4.4.2, we use numerical studies to investigate the approximate EM algorithm in the case when there are three longitudinal covariates.

4.4.1 Joint Model with One Longitudinal Covariate

We generated 500 datasets of 500 individuals with a binary response, baseline covariate, and up to eight repeated observations on a single longitudinal covariate. We generated data to mimic the GenIMS dataset that we analyze in Section 4.5. For each individual, we first generated a baseline “age” covariate \( Z_i \) as \( \beta(3, 2) \cdot 84 + 18 \). We then generated a random number from one to eight, each with equal probability, corresponding to the number of observations for an individual. We also generated intercept and slope random effects as

\[
\begin{pmatrix}
    b_{0i} \\
    b_{1i}
\end{pmatrix}
\sim N\left(\begin{pmatrix}
    1 \\
    0.2
\end{pmatrix}, \begin{pmatrix}
    1 & 0.2 \\
    0.2 & 0.3
\end{pmatrix}\right),
\]

We then let \( X_{ij} \sim N(Z_i \gamma + b_{0i} + b_{1i}t_{ij}, \tau^2) \), where \( \gamma = 1/70, \tau^2 = 1 \), and \( t_{ij} \) is the \( j^{th} \) observed time for the \( i^{th} \) individual, where \( 0 \leq t_{ij} \leq 7 \). To mimic the GenIMS dataset, which has unbalanced repeated measurements, we randomly deleted 10% of the longitudinal observations. Finally, we generated a binary response by

\[
Y_i | Z_i, b_{0i}, b_{1i} \sim \text{Bernoulli}\left\{1 + \exp\left(- (1, Z_i) \beta_1 - (b_{0i}, b_{1i}) \beta_2\right)\right\},
\]

where \( \beta_1 = (-2, 0.035)^T \) and \( \beta_2 = (0.3, -0.45)^T \).

Table 4.1 gives the bias, simulation (Monte Carlo) standard deviation, average standard error estimate, and empirical 95% coverage probability for each of the parameters in \( \beta \) and the bias and simulation standard deviation for \( \gamma, \tau^2, B, \) and \( D \). We only report standard errors and coverage probabilities for the parameters in \( \beta \) since they are of primary interest. The standard errors are estimated using (4.11) rather than (4.12) because we found that it performed slightly better. We emphasize that there is no censoring on the covariates for the results in Table 4.1.

None of the parameter estimates have statistically significant biases at the \( \alpha = 0.01 \) level, where the standard error of \( \hat{\beta} \) is calculated as \( \text{SD}/\sqrt{500} \). Standard error estimation for the
Table 4.1: Simulation results for the joint model of a binary response and a single longitudinal covariate not subject to censoring. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bias</th>
<th>SD</th>
<th>SE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{11}$</td>
<td>-0.024</td>
<td>0.475</td>
<td>0.469</td>
<td>0.966</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.000</td>
<td>0.006</td>
<td>0.006</td>
<td>0.966</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>0.017</td>
<td>0.215</td>
<td>0.209</td>
<td>0.942</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-0.029</td>
<td>0.326</td>
<td>0.331</td>
<td>0.962</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.000</td>
<td>0.004</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>-0.001</td>
<td>0.038</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_1$</td>
<td>-0.014</td>
<td>0.276</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_2$</td>
<td>0.001</td>
<td>0.029</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>-0.007</td>
<td>0.134</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>-0.001</td>
<td>0.043</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>-0.002</td>
<td>0.027</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

parameters in $\beta$ seems good and empirical coverage probabilities are maintained near the 95% level.

We repeated the simulation as described above, but we introduced a DL $d$ so that 20%, 40%, and 60% of the longitudinal observations were censored below $d$. Table 4.2 gives the bias, simulation standard deviation, average standard error estimate, and empirical 95% coverage probability for each of the parameters in $\beta$ as well as the bias and simulation standard deviation for the remaining parameters for simulations with 20%, 40%, and 60% censoring due to DLs.

The estimates for all the parameters in $\beta$ are unbiased at the $\alpha = 0.01$ level, though, with the exception of $\tau^2$ and $D_{12}$ for 20% censoring and $D_{12}$ for 60% censoring, each of the remaining parameter estimates is slightly biased. Standard error estimates and empirical coverage probabilities for the parameters in $\beta$ are good for all the levels of censoring. Thus, while several of the parameter estimates are slightly biased, estimation and inference for the parameters of main interest seems satisfactory.
Table 4.2: Simulation results for the joint model of a binary response and a single longitudinal covariate subject to 20%, 40%, and 60% censoring. Par.: Parameter; SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability.

<table>
<thead>
<tr>
<th>Par.</th>
<th>20% Censoring</th>
<th>40% Censoring</th>
<th>60% Censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>SE</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>-0.019</td>
<td>0.476</td>
<td>0.474</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.000</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>-0.002</td>
<td>0.213</td>
<td>0.210</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-0.039</td>
<td>0.357</td>
<td>0.364</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-0.001</td>
<td>0.004</td>
<td>-</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>-0.001</td>
<td>0.041</td>
<td>-</td>
</tr>
<tr>
<td>$B_1$</td>
<td>0.089</td>
<td>0.287</td>
<td>-</td>
</tr>
<tr>
<td>$B_2$</td>
<td>0.019</td>
<td>0.029</td>
<td>-</td>
</tr>
<tr>
<td>$D_{11}$</td>
<td>0.041</td>
<td>0.140</td>
<td>-</td>
</tr>
<tr>
<td>$D_{12}$</td>
<td>0.002</td>
<td>0.042</td>
<td>-</td>
</tr>
<tr>
<td>$D_{22}$</td>
<td>-0.026</td>
<td>0.027</td>
<td>-</td>
</tr>
</tbody>
</table>

4.4.2 Joint Model with Three Longitudinal Covariates

We conducted a second set of simulations to assess the performance of the approximate EM algorithm in the presence of multiple longitudinal covariates, since, in practice, the proposed normal approximation described in Section 4.2.3 is most useful when the dimension of the random effects is high. We generated 300 datasets of 500 individuals with a binary response, baseline covariate, and up to eight repeated observations on three longitudinal covariates. We again modeled this simulation after the GenIMS dataset, first generating a baseline “age” covariate as described in Section 4.4.1. We then generated a random number from one to eight, each with equal probability, corresponding to the number of observations for each covariate for an individual. We generated an intercept and slope random effect for each of the three covariates as $b_i \sim N(B, D)$, with $B = (1, 0.2, 3, -0.1, -2, 0.3)^T$ and $D_{11} = D_{33} = D_{55} = 1$, $D_{22} = D_{44} = D_{66} = 0.3$, $D_{12} = D_{21} = D_{34} = D_{43} = D_{56} = D_{65} = 0.05$, and all other elements of $D$ set equal to 0.1, where $D_{ij}$ is the element in the $i^{th}$ row and $j^{th}$ column of $D$. The first and second elements of $B$ correspond to the first longitudinal covariate, the third and fourth to the second
Table 4.3: Simulation results for the joint model of a binary response and three longitudinal covariates not subject to censoring. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bias</th>
<th>SD</th>
<th>SE</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{11}$</td>
<td>-0.015</td>
<td>0.887</td>
<td>0.939</td>
<td>0.970</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.000</td>
<td>0.006</td>
<td>0.006</td>
<td>0.940</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>0.009</td>
<td>0.195</td>
<td>0.184</td>
<td>0.947</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-0.009</td>
<td>0.282</td>
<td>0.286</td>
<td>0.967</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>0.035</td>
<td>0.231</td>
<td>0.234</td>
<td>0.963</td>
</tr>
<tr>
<td>$\beta_{24}$</td>
<td>-0.021</td>
<td>0.325</td>
<td>0.332</td>
<td>0.953</td>
</tr>
<tr>
<td>$\beta_{25}$</td>
<td>0.029</td>
<td>0.203</td>
<td>0.196</td>
<td>0.957</td>
</tr>
<tr>
<td>$\beta_{26}$</td>
<td>-0.033</td>
<td>0.310</td>
<td>0.310</td>
<td>0.963</td>
</tr>
<tr>
<td>$\tau_{1}^{2}$</td>
<td>-0.006</td>
<td>0.041</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_{2}^{2}$</td>
<td>0.006</td>
<td>0.059</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_{3}^{2}$</td>
<td>-0.002</td>
<td>0.052</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_{1}$</td>
<td>-0.007</td>
<td>0.064</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_{2}$</td>
<td>-0.001</td>
<td>0.029</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_{3}$</td>
<td>-0.007</td>
<td>0.073</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_{4}$</td>
<td>0.002</td>
<td>0.032</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_{5}$</td>
<td>-0.011</td>
<td>0.062</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$B_{6}$</td>
<td>-0.004</td>
<td>0.029</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

longitudinal covariate, and the fifth and sixth to the third longitudinal covariate. We let $X_{ijk} \sim N((1, t_{ijk})b_{ik}, \tau_{k}^{2})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_{ik}$, $k = 1, 2, 3$, where $\tau^{2} = (1, 1.5, 1.25)^{T}$, where $\tau^{2} = (1, 1.5, 1.25)^{T}$ and $t_{ijk}$ is the $j^{th}$ observed time for the $i^{th}$ individual and $k^{th}$ covariate, where $0 \leq t_{ijk} \leq 7$. To mimic the GenIMS data, we randomly deleted 10% of all the observations at a given time for an individual and an additional 5% of the remaining observed longitudinal observations. Finally, we generated a binary response by $Y_{i} | Z_{i}, b_{i} \sim Bernoulli \left\{ 1 + \exp\left( - (1, Z_{i})\beta_{1} - b_{i}^{T}\beta_{2} \right) \right\}$, where $\beta_{1} = (-1, 0.01)^{T}$ and $\beta_{2} = (0.2, -0.1, 0.4, -0.3, 0.25, -0.3)^{T}$.

Table 4.3 gives the bias, simulation standard deviation, average standard error estimate, and empirical 95% coverage probability for each of the parameters in $\beta$ and the bias and simulation standard deviation for each of the remaining parameters except $D$, which we do not report due to the large number of parameters in $D$, though we note the estimator generally performed very well. As before, we only report standard error estimates and coverage probabilities for the
Table 4.4: Simulation results for the joint model of a binary response and three longitudinal covariates subject to 25% and 50% total censoring. SD: simulation (Monte Carlo) standard deviation; SE: average estimated standard error; CP: empirical 95% coverage probability.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>25% Censoring</th>
<th>50% Censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0.173</td>
<td>0.808</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>0.000</td>
<td>0.197</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>-0.027</td>
<td>0.319</td>
</tr>
<tr>
<td>$\beta_{23}$</td>
<td>-0.024</td>
<td>0.195</td>
</tr>
<tr>
<td>$\beta_{24}$</td>
<td>-0.001</td>
<td>0.331</td>
</tr>
<tr>
<td>$\beta_{25}$</td>
<td>0.013</td>
<td>0.205</td>
</tr>
<tr>
<td>$\beta_{26}$</td>
<td>-0.042</td>
<td>0.352</td>
</tr>
<tr>
<td>$\tau_1^2$</td>
<td>-0.006</td>
<td>0.046</td>
</tr>
<tr>
<td>$\tau_2^2$</td>
<td>0.006</td>
<td>0.061</td>
</tr>
<tr>
<td>$\tau_3^2$</td>
<td>-0.002</td>
<td>0.062</td>
</tr>
<tr>
<td>$B_1$</td>
<td>-0.007</td>
<td>0.066</td>
</tr>
<tr>
<td>$B_2$</td>
<td>-0.001</td>
<td>0.030</td>
</tr>
<tr>
<td>$B_3$</td>
<td>-0.007</td>
<td>0.074</td>
</tr>
<tr>
<td>$B_4$</td>
<td>0.002</td>
<td>0.032</td>
</tr>
<tr>
<td>$B_5$</td>
<td>-0.011</td>
<td>0.069</td>
</tr>
<tr>
<td>$B_6$</td>
<td>-0.004</td>
<td>0.030</td>
</tr>
</tbody>
</table>

parameters in $\beta$ since they are of primary interest. We note that for the results in Table 4.3 none of the covariates are subject to censoring.

All of the parameter estimates in Table 4.3 are unbiased at the $\alpha = 0.01$ level except $\beta_{23}$, which has a slightly significant bias. The standard error estimates and coverage probabilities are good for all of the parameters in $\beta$, including $\beta_{23}$.

We repeated the simulation as described above, but we introduced censoring due to DLs on each of the longitudinal covariates. We defined the DLs $d_1$, $d_2$, and $d_3$ so that censoring on the three covariates corresponded to approximately 25%, 12.5%, and 37.5%, respectively, for a total censoring rate of 25%, as well as approximately 50%, 25%, and 75%, respectively, for a total censoring rate of 50%.
Table 4.4 gives the bias, simulation standard deviation, average standard error estimate, and empirical 95% coverage probability for each of the parameters in $\beta$ and the bias and simulation standard deviation for the remaining parameters except $D$ for simulations with 25% and 50% total censoring due to DLs. All of the parameters in $\beta$ are unbiased at the 0.01 $\alpha$-level, with the exception of the intercept in the model with 25% censoring. While most of the parameters in $\tau^2$ and $B$ have statistically significant biases, they are generally small relative to the size of the true parameter values. Standard error estimates and coverage probabilities are fairly good for both levels of censoring.

### 4.4.3 Computational Advantages

All of simulations above were conducted using the statistical software program R (R Core Team, 2012). Based on this software, we compared the computational time required for fitting the joint models in the simulations using our proposed approximation versus Monte Carlo integration in the E-step of the EM algorithm. For the case when there is a single longitudinal covariate with no censoring, the proposed approximate EM algorithm is about 231.4 times faster than an EM algorithm based on Monte Carlo integration using 10,000 random draws. When 20%, 40%, or 60% of the longitudinal covariate values are censored, the proposed EM algorithm is about 10.5, 3.6, and 2.1 times faster, respectively. For the case when there are three longitudinal covariates and 0%, 25%, or 50% total censoring, the proposed EM algorithm is 226.4, 38.5, and 6.6 times faster, respectively, than an EM algorithm based on Monte Carlo integration using 10,000 random draws. With joint models often taking several hours or days to fit, this computational improvement could potentially be very significant.

### 4.5 Application to the GenIMS Study

We illustrate the proposed approximate EM algorithm for maximizing the joint model of a binary variable and several longitudinal covariates subject to DLs by applying it to the Genetic and Inflammatory Markers of Sepsis (GenIMS) dataset. As explained in Section 2.6, one of the
main purposes of the GenIMS study was to identify relationships between cytokine levels in the body and the event of survival after 90-days for patients with CAP.

We remind the reader that three cytokines were measured in this study: tumor necrosis factor (TNF), interleukin-6 (IL-6), and interleukin-10 (IL-10). The TNF and IL-6 cytokines serve as biomarkers of pro-inflammatory responses to CAP while IL-10 serves as a biomarker of anti-inflammatory responses to CAP. The three biomarkers, TNF, IL-6, and IL-10, were all subject to censoring below the detection thresholds 4, 2 or 5, and 5 pg/ml, respectively, and measured repeatedly for up to eight days for the majority of patients but for as many as 30 days for a few. We only considered the 1875 patients that truly had CAP, required a hospital stay, and had at least one, possibly censored, measurement for each cytokine. The censoring proportions across all time-repeated measurements for the three biomarkers are 39.74%, 28.20%, and 70.59%, respectively, and 46.87% overall. A total of 98.53% of the individuals had at least one censored measurement.

We used the joint model described in Sections 4.2 and 4.3 to explain the relationship between 90-day survival and six covariates: the levels of the three cytokines, TNF, IL-6, and IL-10, sex (1 representing males, 0 representing females), race (1 representing Caucasians, 0 representing all other races), and age. As with our analyses in Sections 2.6 and 3.7, we took a log-transformation before analysis and henceforth refer to the log-cytokines as TNF, IL-6, and IL-10.

We assumed the longitudinal model (4.1) with no elements in $r_{ijk}$ and $s_{ijk} = (1, t_{ijk})^T$, $i = 1, \ldots, 1875$, $j = 1, \ldots, n_{ik}$, $k = 1, \ldots, 3$. While we left $r_{ijk}$ empty in our analyses, we assumed that the associated random effects had non-zero means. This is equivalent to defining $\mathbf{r}_{ijk} = (1, t_{ijk})^T$ and forcing the random effects to have zero means. Using the longitudinal plots shown in Figure 4.3 as a guide, we decided to include an intercept and slope random effect term for each of the cytokine biomarkers. We assumed that the joint distribution of the six random effects is multivariate normal. We allowed for unique variances $\tau^2_{TNF}$, $\tau^2_{IL-6}$, and $\tau^2_{IL-10}$, but we assumed that these variances are constant over time for each covariate. Finally, we used the logistic model (4.2) to relate 90-day survival to the random effects and baseline covariates sex, race,
Figure 4.3: Longitudinal plots of log(TNF), log(IL-6), and log(IL-10) for a random sample of patients in the GenIMS study.
and age.

In the GenIMS dataset, the longitudinal covariates were subject to missingness in addition to the censoring at DLs. Specifically, 48.5%, 15.3%, and 15.4% of the TNF, IL-6, and IL-10 observations, respectively, were missing completely. The high level of missingness for the TNF biomarker was by design, as TNF measurements were only collected after day one on a random subset of the individuals. Most of the remaining missingness resulted from the fact that no measurements were immediately taken on individuals arriving at the hospital during certain times on weekends and also due to reduced measurement taking during holidays (Kellum et al., 2007). Based on these explanations for missingness, we felt that it was reasonable to conclude that the missing values are missing at random. That is, we have no reason to believe that the missingness is in any way related to unobserved factors.

Table 4.5 summarizes the parameter and standard error estimates for the intercept, age, race, and sex covariates, as well as the intercept and random effects associated with the TNF, IL-6, and IL-10 covariates. We only include the parameters in the logistic submodel of the joint model since they are of primary interest for the original GenIMS study.

The sex and race covariates are both significantly related to 90-day survival at the $\alpha = 0.05$ level, with p-values of 0.012 and 0.027, respectively, and the age covariate is very significant, with a p-value of less than 0.001. Two of the three cytokine biomarkers, IL-6 and IL-10, are significantly related to 90-day survival. Specifically, the baseline levels of IL-6 and IL-10, as well as the rates of change over time for IL-6 and IL-10 levels, are very strongly negatively related to 90-day survival with p-values all less than 0.001. The p-values for the baseline TNF levels and rate of change over time are 0.077 and 0.418.

The results in Table 4.5 more strongly indicate that the cytokine biomarkers are related to survival than the results given in Section 2.6 when the longitudinal information was ignored. There are several possible reasons for this. First, by using the longitudinal information via a linear mixed model, we take into account the measurement error in biomarker observations. Ignoring measurement error generally causes an attenuation of estimated covariate effects. Sec-
Table 4.5: Parameter and standard error estimates for the coefficients of log-cytokine and demographic covariates in the joint model for 90-day survival, based on the GenIMS dataset. Each pair of estimates for the log-cytokines represents the baseline effect of the covariate as well as the effect corresponding to the longitudinal trajectory.

<table>
<thead>
<tr>
<th>Par.</th>
<th>Est.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.446</td>
<td>0.794</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.330</td>
<td>0.131</td>
</tr>
<tr>
<td>Race</td>
<td>-0.266</td>
<td>0.120</td>
</tr>
<tr>
<td>Age</td>
<td>-0.053</td>
<td>0.005</td>
</tr>
<tr>
<td>TNF, Intercept</td>
<td>0.267</td>
<td>0.151</td>
</tr>
<tr>
<td>TNF, Slope</td>
<td>-0.590</td>
<td>1.577</td>
</tr>
<tr>
<td>IL-6, Intercept</td>
<td>-0.382</td>
<td>0.105</td>
</tr>
<tr>
<td>IL-6, Slope</td>
<td>-1.943</td>
<td>0.448</td>
</tr>
<tr>
<td>IL-10, Intercept</td>
<td>-0.363</td>
<td>0.110</td>
</tr>
<tr>
<td>IL-10, Slope</td>
<td>-1.334</td>
<td>0.293</td>
</tr>
</tbody>
</table>

Second, by incorporating the longitudinal information, we are no longer assuming that the first day of data collection is a good representative of all of the days. Finally, and perhaps most substantially, by not requiring that a patient has an observation on the first day, the number of patients in the analysis increased from 1418 to 1875.

The direction of the effects mostly remains the same as in the analyses in Sections 2.6 and 3.7. That is, those who are older, male, white, and have higher levels of the cytokines are associated with a lower probability of survival than those who are younger, female, non-white, and have lower levels of the cytokines. The exception to this statement is that having higher baseline levels of TNF, a biomarker of pro-inflammatory responses in the body, is estimated to increase the probability of survival, though this effect is not statistically significant. This may seem to run contrary to the findings in the previous analyses, but we note that this baseline effect of TNF is conditional on the other cytokine effects. Almost 80% of the available longitudinal information is for the IL-6 and IL-10 biomarkers, so it may be that with this increased information on the these biomarkers, we now have indication that the conditional effect of TNF is positive at baseline.
4.6 Discussion

We have proposed an approximate EM algorithm for maximizing the joint model with a binary outcome and multiple longitudinal covariates. This algorithm performs very well, giving unbiased or only slightly biased estimates with good coverage probabilities. We extended this approximate algorithm to handle censoring on the covariates due to DLs and we again showed that using this algorithm produces good estimates and reasonable coverage probabilities, especially for the parameters of main interest in the logistic submodel. The main advantage of this approximate algorithm is that it is much less computationally intensive than traditional maximization methods. Most importantly, the necessary computational time does not increase exponentially as the number of random effects increases since only one-dimensional integrations are required.

The joint model that we have proposed is more flexible than those presented by several authors, specifically by allowing for multiple longitudinal covariates, all of which may be subject to DLs, and also by allowing the inclusion of baseline covariates in the linear mixed models for the longitudinal covariates. However, some of our assumptions may be restrictive in some scenarios. We presumed that the trajectories for the longitudinal covariates can be explained by a linear mixed model, though in some contexts, a non-linear mixed model may be more appropriate. Additionally, we have assumed independent and identically distributed errors in the longitudinal model where sometimes allowing for heterogeneous errors is more accurate. Lastly, we presupposed that the random effects are normally distributed. Several authors have shown that assuming normality may lead to serious biases when the random effects distribution is not normal, and thus it may be useful to consider more flexible distributions for the random effects. Fortunately, as long as we assume normality on the error term in the longitudinal model for $X_{ik}$, $i = 1, \ldots, n$, $k = 1, \ldots, q$, the main idea of the approximate EM algorithm that we proposed should still be valid since the conditional distribution $f(b_i \mid y_i, x_{i1}, x_{i2}, \ldots, x_{iq}, z_i; \theta)$ still converges to a normal as the number of longitudinal observations gets large.

Finally, we note that our approximate EM algorithm for fitting joint models is similar to the
Laplace approximation approach suggested by Rizopoulos et al. (2009). In both cases, the goal is to reduce the computational requirements for finding parameter estimates by simplifying the calculation of multidimensional expectations in the E-step of the EM algorithm. With Laplace approximations, the expectations themselves are approximated rather than the distribution for which the expectations are calculated with respect to, as we suggest. Our approach is simpler to apply in practice and more straightforward to extend to handling longitudinal covariates censored due to DLs, though future research could focus on comparing the two methods directly.
Chapter 5

Flexible Models for the Conditional Distribution of the Covariates Subject to Detection Limits

5.1 Introduction

Throughout this dissertation, we have assumed that the distributional model for the covariates subject to censoring conditional on the fully observed covariates, \( f(x_i | z_i) \), is known. In practice, this assumption may be unrealistic. Gomez et al. (2003), Nie et al. (2010), Arunajadai and Rauh (2012), and a few others have tackled this issue by proposing nonparametric or flexible modeling strategies. Gomez et al. (2003) proposed methods for a single, randomly interval-censored covariate in the context of linear regression. In this scenario, more information is available than when censoring is at fixed DLs. The statistical validity of flexible strategies for DL problems is less clear and has generally been examined only through approximations and simulations. For example, Arunajadai and Rauh (2012) proposed using conditional mean imputation to replace values below the DL by assuming a generalized gamma distribution for a single censored covariate. While we showed in Section 2.3.2 that conditional mean imputation may be valid in linear
models, this will not be the case for most other modeling scenarios. Additionally, this method is still rooted in a distributional assumption and statistical inference is not clear once values are imputed. Thus, there has been relatively little success at semiparametric or nonparametric attempts of handling covariates censored due to DLs.

In this chapter, we discuss some ways to parametrically or semiparametrically model the covariates subject to DLs while remaining realistically flexible. We feel that assuming a parametric or semiparametric model for $f(x_i|z_i)$ is advantageous for two reasons: (i) nonparametric methods still often require unverifiable assumptions and do not take advantage of the distributional evidence above the DL; (ii) the parametric and semiparametric methods we suggest provide straightforward methods for variance estimation and statistical inference. With regard to the first reason, we note that most nonparametric approaches that have been proposed for dealing with covariates subject to DLs in regression models have still made assumptions about the values of the covariate below the DL without taking into account the only available information - the data above the DL. In most biological scenarios, it is often very reasonable to assume that the distributional patterns above a DL continue below that limit, though this is inevitably a judgment call. In either case, it seems reasonable that the best way to handle data below the DL is to use the available data above the DL. The second reason should be clear, since the maximum likelihood methods and the multiple imputation methods we proposed in the preceding chapters allow for straightforward variance estimation.

In the following section, we suggest several ways for flexibly modeling the distribution $f(x_i|z_i)$ that allow for the straightforward application of the Bayesian, maximum likelihood, and multiple imputation procedures we have discussed in Chapters 2 and 3, focusing mainly on the proposed improper multiple imputation procedures. We do not discuss the modeling of $f(x_i|z_i, b_i)$ in the joint modeling situation described in Chapter 4, which is a slightly different problem. In this context, we model the variation of a longitudinal covariate around a particular trajectory and need to assume that a normal distribution is valid whether or not the trajectory is below the DL.
5.2 Flexible Modeling Methods

In this section, we briefly review several ways of handling the modeling of \( f(x_i | z_i) \). We begin by exploring the consequences of assuming the most convenient distribution, a multivariate normal, when it is not appropriate.

5.2.1 Multivariate Normal

In practice, researchers often struggle with flexibly modeling real-world data, especially when the data is multivariate in nature. For this reason, a multivariate normal distribution is regularly assumed when it is necessary to fit a parametric model to multi-dimensional data.

To assess the sensitivity of various methods against the misspecification of the parametric distribution, specifically assuming a multivariate normal, we conducted a simulation where the assumed distribution of the censored covariates is incorrect. We let the covariates \( X_1 \) and \( X_2 \) be bivariate gamma, as defined by a Clayton copula (Clayton, 1978) with each variable having marginal shape and scale parameters 2 and 0.5. We generated samples from this distribution using the Copula package in R (Hofert et al., 2012). We naively modeled the joint distribution of \( \log(X_1) \) and \( \log(X_2) \) by a bivariate normal even though the Q-Q plots shown in Figure 5.1 do not seem to support this assumption. The remainder of the simulation set-up was not changed from Section 2.5. That is, we assumed that \( Y | X_1, X_2 \) follows a logistic regression model, \( Y | X_1, X_2 \sim \text{Bernoulli}(1 + \exp(-\beta_0 - X_1 \beta_1 - X_2 \beta_2))^{-1} \), with \( \beta = (3, 1.5, -3)^T \). We again considered three levels of censoring percentage, 20%, 40%, and 60%. The simulation was repeated 2000 times with 200 subjects for Table 5.1 and 500 subjects for Table 5.2.

Tables 5.1 and 5.2 summarize the average bias, simulation (Monte Carlo) standard deviation, average estimated standard error, and empirical 95% coverage probability of the parameter estimates for the Omni, CC, ML, FB, PMI, and IMI methods for 20%, 40%, and 60% censoring (refer to Section 2.5 for notation). Because the normal assumption is incorrect, all of the methods except the Omni and CC methods have clear biases on the parameter estimates. Interestingly, for both \( n = 200 \) and 500, standard error estimation is reasonable and coverage probabilities
Table 5.1: Simulation results for a logistic regression on 2000 datasets with 200 individuals and an incorrect distributional assumption on the covariates. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

<table>
<thead>
<tr>
<th>Cens. %</th>
<th>Method</th>
<th>$\beta_0 = 3$</th>
<th>$\beta_1 = 1.5$</th>
<th>$\beta_2 = -3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>SE</td>
<td>CP</td>
</tr>
<tr>
<td>20</td>
<td>Omni</td>
<td>0.00</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>CC</td>
<td>-0.01</td>
<td>0.48</td>
<td>0.46</td>
<td>0.94</td>
</tr>
<tr>
<td>ML</td>
<td>0.01</td>
<td>0.46</td>
<td>0.45</td>
<td>0.94</td>
</tr>
<tr>
<td>FB</td>
<td>0.04</td>
<td>0.47</td>
<td>0.46</td>
<td>0.94</td>
</tr>
<tr>
<td>PMI</td>
<td>0.01</td>
<td>0.46</td>
<td>0.45</td>
<td>0.94</td>
</tr>
<tr>
<td>IMI</td>
<td>0.01</td>
<td>0.46</td>
<td>0.45</td>
<td>0.94</td>
</tr>
<tr>
<td>40</td>
<td>Omni</td>
<td>0.00</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>CC</td>
<td>0.00</td>
<td>0.59</td>
<td>0.59</td>
<td>0.95</td>
</tr>
<tr>
<td>ML</td>
<td>0.05</td>
<td>0.47</td>
<td>0.44</td>
<td>0.94</td>
</tr>
<tr>
<td>FB</td>
<td>0.08</td>
<td>0.50</td>
<td>0.47</td>
<td>0.94</td>
</tr>
<tr>
<td>PMI</td>
<td>0.04</td>
<td>0.47</td>
<td>0.45</td>
<td>0.95</td>
</tr>
<tr>
<td>IMI</td>
<td>0.04</td>
<td>0.47</td>
<td>0.45</td>
<td>0.95</td>
</tr>
<tr>
<td>60</td>
<td>Omni</td>
<td>0.00</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>CC</td>
<td>-0.01</td>
<td>0.82</td>
<td>0.78</td>
<td>0.94</td>
</tr>
<tr>
<td>ML</td>
<td>0.07</td>
<td>0.47</td>
<td>0.45</td>
<td>0.94</td>
</tr>
<tr>
<td>FB</td>
<td>0.12</td>
<td>0.52</td>
<td>0.50</td>
<td>0.94</td>
</tr>
<tr>
<td>PMI</td>
<td>0.08</td>
<td>0.50</td>
<td>0.46</td>
<td>0.94</td>
</tr>
<tr>
<td>IMI</td>
<td>0.07</td>
<td>0.48</td>
<td>0.47</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Figure 5.1: Normal Q-Q plots for $\log(X_1)$ and $\log(X_2)$ using data from a single simulated dataset.

are close to 95%, even at high censoring percentages.

While this simulation demonstrates that making the incorrect distributional assumption may not be overly costly, this is largely due to the nature of the set-up, for which the censoring on the untransformed covariates only occurs in a region between 0 and the DL. With a somewhat reasonable model, it is hard to be overly wrong in such a relatively small region, especially when the DL is close to 0 in relation to the range of the data. Of course, most situations in practice follow this set-up, such as in biomedical studies where measurements are taken on biological variables, which are often strictly positive. With a very large degree of censoring, an unbounded lower range, or upper DLs, using a multivariate normal when it is unreasonable may lead to estimators which perform more poorly.

5.2.2 Box-Cox Transformations

When a random variable subject to a DL is strictly positive, as is often the case with biological variables, Box-Cox transformations are usually appropriate (Box and Cox, 1964). A Box-Cox transformation is a type of power transformation where the order of the data are maintained
Table 5.2: Simulation results for a logistic regression on 2000 datasets with 500 individuals and an incorrect distributional assumption on the covariates. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

<table>
<thead>
<tr>
<th>Cens. %</th>
<th>Method</th>
<th>$\beta_0 = 3$</th>
<th></th>
<th></th>
<th>$\beta_1 = 1.5$</th>
<th></th>
<th></th>
<th>$\beta_2 = -3$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Omni</td>
<td>0.00 0.26 0.26 0.95</td>
<td>-0.01 0.38 0.39 0.95</td>
<td>0.00 0.43 0.44 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>-0.01 0.27 0.28 0.95</td>
<td>0.01 0.43 0.44 0.95</td>
<td>0.01 0.46 0.47 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td>0.01 0.26 0.26 0.95</td>
<td>-0.03 0.38 0.38 0.94</td>
<td>-0.01 0.43 0.43 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>0.04 0.27 0.27 0.95</td>
<td>-0.01 0.40 0.40 0.94</td>
<td>-0.04 0.46 0.45 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI</td>
<td>0.01 0.26 0.26 0.95</td>
<td>-0.02 0.39 0.40 0.96</td>
<td>-0.01 0.43 0.44 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMI</td>
<td>0.01 0.26 0.26 0.95</td>
<td>-0.03 0.39 0.40 0.95</td>
<td>-0.01 0.43 0.44 0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Omni</td>
<td>0.00 0.26 0.26 0.95</td>
<td>-0.01 0.38 0.39 0.95</td>
<td>0.00 0.43 0.44 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>-0.01 0.32 0.33 0.96</td>
<td>0.00 0.54 0.54 0.95</td>
<td>0.03 0.53 0.54 0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td>0.03 0.26 0.26 0.94</td>
<td>-0.06 0.38 0.38 0.94</td>
<td>-0.05 0.43 0.43 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>0.06 0.28 0.28 0.95</td>
<td>-0.03 0.40 0.41 0.94</td>
<td>-0.07 0.47 0.46 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI</td>
<td>0.03 0.27 0.27 0.95</td>
<td>-0.04 0.39 0.41 0.95</td>
<td>-0.04 0.45 0.45 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMI</td>
<td>0.03 0.26 0.27 0.96</td>
<td>-0.06 0.39 0.41 0.95</td>
<td>-0.05 0.44 0.46 0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Omni</td>
<td>0.00 0.26 0.26 0.95</td>
<td>-0.01 0.38 0.39 0.95</td>
<td>0.00 0.43 0.44 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.00 0.50 0.48 0.95</td>
<td>-0.01 0.76 0.74 0.95</td>
<td>0.01 0.69 0.69 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML</td>
<td>0.06 0.27 0.26 0.93</td>
<td>-0.10 0.40 0.41 0.93</td>
<td>-0.08 0.45 0.44 0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>0.09 0.28 0.29 0.94</td>
<td>-0.06 0.43 0.44 0.94</td>
<td>-0.09 0.48 0.48 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI</td>
<td>0.06 0.28 0.28 0.94</td>
<td>-0.06 0.43 0.43 0.94</td>
<td>-0.07 0.47 0.48 0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMI</td>
<td>0.07 0.27 0.28 0.94</td>
<td>-0.06 0.44 0.43 0.94</td>
<td>-0.09 0.46 0.47 0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
but re-distributed to more closely fit a normal distribution. Specifically, for a univariate random variable $X$, the Box-Cox transformation transforms $X$ to $X^{(\lambda)}$ using the following formula:

$$X^{(\lambda)} = \begin{cases} \frac{X^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(X) & \text{if } \lambda = 0 \end{cases}$$

Several extensions and adaptations have been proposed, including multivariate Box-Cox methods and techniques for handling random variables which can take negative values, and are summarized in a review article by Sakia (1992). Poirier (1978), Han and Kromal (2004), and Cai et al. (2005), among others, extended Box-Cox transformations to handle censored data, which we can apply for covariates subject to DLs. We repeated the simulations described in Section 5.2.1, but instead of taking log-transformations, we applied univariate Box-Cox transformations on each of the two covariates separately. We only conducted simulations for the $n = 200$ case.

Tables 5.3 summarizes the average bias, simulation standard deviation, average estimated standard error, and empirical 95% coverage probability of the parameter estimates for the Omni, CC, ML, FB, PMI, and IMI methods for 20%, 40%, and 60% censoring after having applied univariate Box-Cox transformations to each variable separately in order to use a multivariate normal assumption for $f(x_i|z_i)$. The biases are reduced from those observed in Tables 5.1 and 5.2 and coverage probabilities and standard error estimation are also improved.

### 5.2.3 Seminonparametric Distribution

In Chapter 3, we proposed using the seminonparametric (SNP) distribution to model the distribution of a survival variable in the context of the proposed improper multiple imputation method. The SNP distribution could also be used to model the covariate distribution $f(x_i|z_i)$. While we only described the univariate SNP in Section 3.2.1, it has straightforward multivariate extensions which are useful for modeling a variety of multivariate distributional shapes. The
Table 5.3: Simulation results for a logistic regression on 2000 datasets with 200 individuals and a Box-Cox transformation applied to the covariates. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimate standard error; CP: empirical 95% coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

<table>
<thead>
<tr>
<th>Cens. %</th>
<th>Method</th>
<th>$\beta_0 = 3$</th>
<th>$\beta_1 = 1.5$</th>
<th>$\beta_2 = -3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>SE</td>
</tr>
<tr>
<td>20</td>
<td>Omni</td>
<td>0.00</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>-0.01</td>
<td>0.48</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.00</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.04</td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.01</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.00</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td>40</td>
<td>Omni</td>
<td>0.00</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.00</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.01</td>
<td>0.46</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.07</td>
<td>0.49</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.03</td>
<td>0.47</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.03</td>
<td>0.47</td>
<td>0.45</td>
</tr>
<tr>
<td>60</td>
<td>Omni</td>
<td>0.00</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>-0.01</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.05</td>
<td>0.50</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.10</td>
<td>0.55</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.06</td>
<td>0.54</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.06</td>
<td>0.54</td>
<td>0.51</td>
</tr>
</tbody>
</table>
q-variate SNP density for \( \mathbf{v} \) can represented by

\[
f_K(\mathbf{v}) = P_K^2(\mathbf{v})g(\mathbf{v}),
\]

where

\[
P_K(\mathbf{v}) = \sum_{\lambda_1,\ldots,\lambda_q = 0,1,\ldots,q: 0 \leq \lambda_1 + \cdots + \lambda_q \leq K} \sum_{\lambda_1\cdots\lambda_q} a_{\lambda_1\cdots\lambda_q} v_1^{\lambda_1} \cdots v_q^{\lambda_q}
\]

is a truncated polynomial of degree \( K \), \( g(\mathbf{v}) \) is a multivariate kernel density with a moment generating function, and \( \int f_K(\mathbf{v})d\mathbf{v} = 1 \). For example, if we let \( K = 2 \) and \( q = 2 \),

\[
P_K(\mathbf{v}) = a_{00} + a_{10}v_1 + a_{01}v_2 + a_{11}v_1v_2 + a_{20}v_1^2 + a_{02}v_2^2.
\]

A natural choice for \( g(\mathbf{v}) \) is a \( q \)-dimensional normal distribution, \( N_q(\mathbf{0}, I_q) \).

When \( \mathbf{X}_i \) is one-dimensional, the SNP is fairly easy to fit. Unfortunately, as the dimension \( q \) increases, it becomes much harder to find the appropriate maximum likelihood estimates, especially if there are other conditioning variables \( \mathbf{Z}_i \). However, as described in Section 3.2.1, the SNP distribution is extremely flexible and is useful for modeling symmetric, multi-modal, skewed, and heavy-tailed variables. We refer to Tang (2008) for a more thorough discussion of the multivariate SNP and finding maximum likelihood estimates for this distribution.

### 5.2.4 Series of Univariate Regressions

In Section 2.2, we briefly introduced the idea of modeling \( f(x_i|z_i) \) using a series of conditional univariate distributions \( f(x_{i1}|x_{i2}, x_{i3}, \ldots, x_{iq}, z_i) f(x_{i2}|x_{i3}, \ldots, x_{iq}, z_i) \cdots f(x_{iq}|z_i) \). Each univariate conditional distribution can be modeled in a variety of ways, such as with the flexible SNP distribution discussed in Section 5.2.3, though typical regression methods will generally suffice. If it was deemed appropriate, each of the univariate models could simply be fit by a multivariate linear regression. Of course, if the distribution of the errors for one or all of these distributions was not reasonably normal, generalized linear models, discussed in Section 2.2, could be used.
This might apply specifically if one of the variables subject to censoring was discrete. It would be tough to model a joint distribution of continuous and discrete variables, but by using a series of conditional distributions, it becomes feasible.

While this method for modeling $f(x_i|z_i)$ is very flexible, we note that it has a few drawbacks. For one, there are many modeling assumptions which must be checked simultaneously. Residual plots could be used on each univariate regression, though there is still no way of checking the distributional assumptions below the DL. Additionally, since each of the variables $X_{i1}, \ldots, X_{iq}$ is subject to censoring, we would need to estimate censored regression models. Finally, the number of parameters involved in fitting a series of distributions may be very large, especially if the dimension of $Z_i$ is large. Thus, we would have a large number of nuisance parameters to deal with in the overall model.

### 5.2.5 Other Methods

There are many other possible methods for modeling $f(x_i|z_i)$ which may be suitably flexible. However, dealing with censoring in these methods as well as incorporating them into a multiple imputation algorithm may be challenging. For example, a mixture of normals is often used for flexible modeling in both univariate or multivariate cases. However, dealing with the censoring when fitting $f(x_i|z_i)$ would be challenging. Additionally, when considering GLMs, for example, it would be more work obtaining samples from the distribution $f(x_i^c|y_i, z_i, x_i^p, x_i^c < d)$.

A popular distributional choice for flexibly modeling positive random variables is the generalized gamma distribution (Stacy, 1962), represented by

$$f(x; a, b, c) = \left( \frac{c}{a} \right) x^{b-1} e^{-\left(\frac{x}{a}\right)^c / \Gamma\left(\frac{b}{c}\right)}.$$

Unfortunately, multivariate extensions of the generalized gamma require special techniques to induce correlation among the variables, and it is therefore challenging to deal with censoring as well as sampling from this distribution.
Undoubtedly, more ways already exist or could be developed for flexibly modeling \( f(x_i|z_i) \), some of which may be suitable in the context of the methods described in Chapters 2 and 3. However, we feel that the methods described in Sections 5.2.1-5.2.3 will generally be suitable, especially in biological examples when censoring due to DLs most often occurs.

5.3 Model Checking

After having made distributional assumptions for \( f(x_i|z_i) \), it would seem prudent to have some methods for checking the validity of these assumptions. We provide a few easy ways in the following five subsections.

5.3.1 Distributional Tests

When \( f(x_i|z_i) \) is assumed to follow a particular distribution, we may employ goodness-of-fit tests to check the fit to that distribution. While distributional tests cannot prove that a distribution is correct for describing data, they can give evidence that it is incorrect. Two of the most famous distributional tests are the Kolmogorov-Smirnov test for any distribution \( f \) and the Shapiro-Wilk test for normality (Shapiro and Wilk, 1965). Barr and Davidson (1973) and Royston (1993) gave extensions for the Kolmogorov-Smirnov test and Shapiro-Wilk test, respectively, to handle censored data.

Unfortunately, distributional tests suffer from several weaknesses. For one, they are known to have low power for small sample sizes, while for large samples sizes, distributional tests detect even slight deviations from the distribution assumed under the null hypothesis. Additionally, distributional tests for censored samples have not been thoroughly considered for the multivariate case. Perhaps most significantly, distributional tests are designed for checking the marginal distribution of a variable and thus are less useful when modeling \( X_i \) conditional on additional covariates \( Z_i \).
5.3.2 Information Criterion Methods

Whenever we fit a model to data, we have the option of refitting other models in an attempt to improve the fit. We can fit any set of data perfectly, so we often like to penalize our model based on the number of parameters it involves. The most popular methods for comparing several fitted models are the Akaike information criterion (AIC, Akaike, 1974) and the Bayesian information criterion (BIC, Schwarz, 1978), two criteria which we briefly discussed in Section 3.2.1. The idea behind these criteria is that we want to minimize the likelihood and number of parameters simultaneously. As applied to the problem of censored covariates, it may be beneficial to fit several censored parametric models and then compare these models using a method such as BIC in order to choose the best option.

5.3.3 Truncated vs. Censored Percentages

When a model is fit to censored data, an easy way to check whether the fit is reasonable is to check if the percentage of data that is censored in the dataset corresponds to the percentage of the fit distribution that lies below the DL (a truncated portion of the fit distribution). The proportion of the fit distribution that is below the DL can be found by simply evaluating the cumulative density function of the fit distribution at the DL. If the percentage of observed censoring is not close to this proportion, then the chances are that the assumed distribution is unreasonable.

Unfortunately, deciding the definition of close is not easy. As an example, for the data in Figure 5.2 appearing in the next subsection, the proportion of censored data is 50%, and if we assume a normal distribution, the cumulative density function at the DL for a normal fitted to the data is 0.473. While 0.5 and 0.473 may seem close, if the data were truly normally distributed with mean -1.388 and variance 3.375, as was estimated by the censored data maximum likelihood, we would observe this large of a difference with probability $3.16 \times 10^{-19}$.

We suggest the following parametric bootstrap algorithm for estimating the probability that the difference between these two quantities is due to random chance. First, based on the
observed censored data, calculate the maximum likelihood estimates for the fitted distribution. Next, simulate datasets from this distribution, which for the example above would be a normal with mean -1.388 and variance 3.375. For each dataset, find the proportion of values below the DL as well as the estimated value of the cdf at the DL. Calculate the difference between these two quantities for each dataset and construct the empirical distribution of the differences. Finally, if the distributional assumption is correct, then the observed difference should fall in the center region of this empirical distribution of the differences. The p-value is simply the proportion of observed differences which are greater in absolute value than the observed difference.

5.3.4 Graphical Checks

One simple way to visually verify whether an assumed distribution for $f(x_i | z_i)$ is reasonable, when $X_i$ is univariate, is to look at histograms or Q-Q plots of the marginal distribution $f(x_i)$. We admit that only considering graphs for the marginal distribution of $X_i$ is not ideal since we are interested in modeling $f(x_i | z_i)$. However, except in the case where $Z_i$ only takes a few values or can be sensibly stratified, it is challenging to consider the histogram of one variable conditional on another variable or set of variables. In relation to Q-Q plots, we do describe a technique for graphically checking the distribution $f(x_i | z_i)$ rather than $f(x_i)$, but most of the discussion below focuses on checking marginal distributions for $X_i$.

A histogram is useful for visually inspecting whether the observed data above the DL agree with the assumed distribution. However, this may sometimes be hard to check and we recommend overlaying the curve for the assumed distribution fitted to the censored data. This fitted curve can be obtained by maximizing the likelihood of the censored data to obtain parameter estimates for the distribution. For example, Figure 5.2 displays a histogram of the data in a simulated dataset observed above the DL along with the curve of a censored normal fitted to the data. Without the curve, we may assume that a normal distribution seems reasonable. However, 50% of the data are censored in this example, and the overlayed fitted normal curve makes it clear that a normal distribution is not reasonable. The data for this histogram were
Figure 5.2: Histogram of a sample with 50% censoring below the DL = $-1.5$. The curve of a normal distribution fitted to the censored data by maximum likelihood is overlayed on the histogram, with probability rather than frequency plotted on the Y-axis.

generated from a very left-skewed seminonparametric distribution.

A Q-Q plot, which plots the ordered data versus the quantiles of a distribution, is generally more straightforward for assessing whether an assumed distribution is reasonable, both with censored and uncensored data. When the data closely follow the assumed distributed, the Q-Q plot will strongly resemble a 45-degree line through the origin. While a Q-Q plot for data subject to censoring at a lower DL only includes the upper $1 - p_0$ percent of the values, where $p_0$ is the fraction of censored data, since we do not observe any values below the DL, the Q-Q plot does take into account the censored data. This is because the lowest $p_0$ percent of the data contribute to the Q-Q plot by contributing the $0^{th}$ to $p_0^{th}$ quantiles of the observed data. Figure 5.3 gives two Q-Q plots based on a normal distributional assumption, with the left plot representing the ordered data above the DL plotted versus the quantiles of a truncated standard normal and the right plot representing the ordered data above the DL plotted versus the upper $1 - p_0$ quantiles of a standard normal. The plot on the left does not take into account the censored data, and in this example implies a good fit. However, just because the uncensored data fit a truncated
normal does not mean that the entire set of data would fit an untruncated normal. The plot on
the right takes into account the censored data by only plotting versus the upper quantiles of
a standard normal. The right Q-Q plot clearly shows that the normal distribution is not good
for the entire dataset.

While a standard Q-Q plot based on censored data generally gives a good sense of whether
a distribution is suitable for describing a dataset, using the standardized distribution along the
X-axis, such as in Figure 5.3, is sometimes misleading. Datasets often have non-zero means
and various levels of variation. So, if data is normally distributed with mean 12 and variance
1.5, then a normal Q-Q plot will follow a straight line, but one that is not 45 degrees. This
generally is not an issue since when we speak of assuming a distribution, we generally are only
assuming a family of distributions (i.e. a standard normal versus a normal with any mean and
scale). However, with censored data, a reference line is more important since the Q-Q plot is
not available for the entire dataset, making it harder to know whether the Q-Q plot actually
follows a reference line. For this reason, it makes more sense to use an adapted Q-Q plot, where
we first obtain maximum likelihood estimates for the assumed censored distribution, and then

Figure 5.3: Q-Q plots of the data observed above the DL = −1.5 versus quantiles of an assumed
distribution. The left Q-Q plot incorrectly plots the quantiles of a truncated distribution on
the X-axis while the right Q-Q plot correctly plots the upper 1 − p0 quantiles of an untruncated
distribution.
Figure 5.4: Q-Q plot of the data observed above the DL = −1.5 versus the upper 1−p0 quantiles of a fitted normal distribution. The data should follow the 45 degree reference line through the origin if the normal distribution accurately describes the data.

plot the ordered data above the DL versus the upper 1−p0 quantiles of the distribution based on the maximum likelihood estimates. Using this strategy, the points on the Q-Q plot for the censored data should follow a 45 degree line through the origin. Figure 5.4 is based on the same data used in Figure 5.3, now using the strategy of plotting against a fitted normal. We see that the plotted points look the same but that with the reference line added the departure from normality is more obvious compared to the right Q-Q plot in Figure 5.3.

For location-scale families of distributions, we can use the idea of an adapted Q-Q plot to check whether the distribution \( f(x_i|z_i) \) is reasonable rather than just the marginal distribution \( f(x_i) \). For example, if we assume \( X_i \sim N(\gamma_0 + \gamma_1 Z_{i1} + \cdots + \gamma_{p-q} Z_{i(p-q)}, \sigma^2) \), then to construct an adapted Q-Q plot we would first find maximum likelihood estimates for \( \gamma = (\gamma_1, \cdots, \gamma_{p-q})^T \) and \( \sigma^2, \widehat{\gamma} \) and \( \widehat{\sigma}^2 \). However, the fitted normal distribution is the not the same for each individual and thus we cannot follow the exact strategy described above for adapted Q-Q plots. Instead, since \( Z_i \) is only present in the location of this distribution, we can first center each of the data observations \( X_i, i = 1, \cdots, n \), as \( X_i - \widehat{\gamma}_0 - \widehat{\gamma}_1 Z_{i1} - \cdots - \widehat{\gamma}_{p-q} Z_{i(p-q)} \). Then, the modified adapted Q-Q plot can be constructed by plotting these centered data observations versus the quantiles.
of a $N(0, \sigma^2)$. If $Z_i$ was also involved in the scale term, we could order the centered and scaled observations and plot them versus the quantiles of a $N(0, 1)$.

While graphical checks are valuable tools for assessing whether an assumed distribution is reasonable, they get more complicated as the number of variables subject to DLs increases. To get around this, we can make marginal plots for each of the censored variables, possibly conditional on additional covariates $Z_i$. However, even if marginal plots look good, there is still no guarantee that the multivariate distribution $f(x_i|z_i)$ is appropriate.

### 5.3.5 Expected Value of the Response

Perhaps the best way to check whether the assumed distribution $f(x_i|z_i)$ is reasonable is after the analysis has been completed. Specifically, we can compare whether the predicted values of the response in the censored-data region match up with the values which are observed in the original data set which have censored covariates. For example, in GLMs if we have only one covariate subject to censoring $X_i$, we can compare $E(Y_i|Z_i, X_i < d)$ based on the final fitted GLM with the average of the observed $Y_i$’s for all individuals for which we observe $X_i < d$. If these two are not close, then we have an indication that the model for $f(x_i|z_i)$ is poor in the censored-data region.

### 5.4 Discussion

In this chapter, we have reviewed several strategies for modeling $f(x_i|z_i)$ with $X_i$ subject to DLs in the context of GLMs and survival models outlined in Chapters 2 and 3. We have also discussed several methods for checking these assumptions before and after fitting the final covariate distribution or the final GLM or survival model.

In practice, we recognize that, even with flexible modeling strategies and adequate checking, we still have to make an unverifiable assumption by extrapolating information from the observed-data region into the censored-data region. As is always the case with extrapolation, it must be done carefully and still may be dangerous. We stress that in many biological ex-
amples, where data censored due to DLs most often occur, we cannot go too wrong since the
censored-data region is bounded.

As we discussed in the introduction to this chapter, avoiding distributional assumptions on
the covariates generally means that we have to make a different assumption about the data
in the censoring region, such as that replacing censored values by the $DL/\sqrt{2}$ is reasonable.
This is because we ultimately want to assess the relationship between a response and a set of
covariates for which some values are not known. Of course, a complete case analysis, as applied
in Chapter 2 and Chapter 3, avoids distributional assumptions on the censored covariates but
sacrifices efficiency.

One of the committee members reviewing this dissertation work suggested two alternative
approaches which truly avoid making assumptions about the distribution of $X_i$. First, in the
case where prediction of a response is of concern, and not the relationship between the response
and the covariates, we can treat the problem as a measurement error problem. Even when
measurement error is an issue, it is generally best to use the observed data for the purposes of
predicting a response. That is, we may simply use the values at the DL in whatever model we fit.
Second, even when we are concerned with the relationship between the covariates and a response,
if we deal directly with the estimating equations for the model, it may be still be possible to
avoid distributional assumptions for the covariates. Specifically, for those individuals whose data
are fully observed, we can use the typical estimating equations and for those individuals whose
data are censored, we can use an expected estimating equation conditional on the censoring.
REFERENCES


Bates, D., Maechler, M., and Bolker, B. (2012), *lme4: Linear mixed-effects models using S4 classes*, R package version 0.999999-0.


115


APPENDICES
Appendix A

Chapter 2 Proofs and Technical Details

In this appendix, we give proofs for Proposition 1, and Theorems 1-3 in Chapter 2. For notational simplicity, we do not differentiate between random variables and their realized values by using upper and lower case letters, as is typical. Additionally, for Theorem 1, we assume for simplicity that $h_i(\mu_i) = h(\mu_i), i = 1, \ldots, n$.

A.1 Regularity Conditions

Define $\frac{w_i(y_i-\mu_i)}{h(\mu_i)g(\mu_i)} = \psi(y_i, w_i, \beta)$, where $\psi$ and $w_i$ are $p$-dimensional. We consider the following regularity conditions:

A.1 The $p \times 1$ parameter vector $\beta$ lies in a compact subset $B$ of $\mathbb{R}^p$, and $E\{\psi(y, w, \beta)\}$ exists for all $\beta \in B$.

A.2 If $E\{\psi(y, w, \beta)\} = 0$, then $E\{\psi(y, w, \beta')\} \neq 0$ unless $\beta = \beta'$.

A.3 $\psi_j(y, w, \beta)$ is continuous in $\beta$ and bounded by an integrable function of $(y, w)$ that does not depend on $\beta$ for $j = 1, \ldots, p$. 
A.2 Proof of Proposition 1

Let \((y_i, w_i), i = 1, \ldots, n,\) be independent samples with \(f(y_i|w_i)\) belonging to the exponential dispersion family, given by (2.1), \(g(\mu_i) = \sum_{j=1}^p w_{ij} \beta_j,\) and \(w_i = (z_i^T, x_i^T)^T.\) Defining \(\rho^*_i = \prod_{j=1}^q \rho_{ij} = I(x_{i1} \geq d_1, \ldots, x_{iq} \geq d_q),\) the expected value of the \(j^{th}\) complete case estimating function, suppressing the index \(i,\) is

\[
E \left\{ \rho^*_i \frac{w_j(y - \mu)}{h(\mu)g'(\mu)} \right\} = E \left[ E \left\{ I(x_1 \geq d_1, \ldots, x_q \geq d_q) \frac{w_j(y - \mu)}{h(\mu)g'(\mu)} \right| w \right] \\
= E \left[ I(x_1 \geq d_1, \ldots, x_q \geq d_q) \frac{w_j}{h(\mu)g'(\mu)} E \{(y - \mu) | w\} \right] \\
= E \left[ I(x_1 \geq d_1, \ldots, x_q \geq d_q) \frac{w_j}{h(\mu)g'(\mu)} (\mu - \mu) \right] \\
= 0.
\]

Since the expectation is 0 for all \(p\) estimating functions, under regularity conditions A.1-A.3 it can be shown with standard M-estimation theory (see Serfling, 1980, for example) that

\[
\hat{\beta}_{CC} \xrightarrow{p} \beta \quad \text{as} \quad n^* \to \infty,
\]

where \(n^* = \sum_{i=1}^n \rho^*_i\) is the number of individuals with non-zero contribution to the complete case estimating functions.

A.3 Proof of Theorem 1

Let \((y_i, z_i, x_i, \rho_i), i = 1, \ldots, n,\) be independent samples with \(f(y_i|z_i, x_i)\) belonging to the exponential dispersion family, given by (2.1), \(g(\mu_i) = \sum_{j=1}^p w_{ij} \beta_j,\) where \(w_i = (z_i^T, x_i^T)^T.\) Let \(x_i\) be univariate and \(f(x_i|z_i)\) be a known distribution. For the first part of this proof, let \(\bar{x}_i = E(x_i|x_i < d; \theta)\) be the imputed values of \(x_i\) when \(x_i < d.\) Additionally, define \(S_2(\beta) = 0\) as
the first $p - 1$ estimating equations in (2.3) and $S_x(\beta)$ as the final estimating equation in (2.3), 
with both based on imputed values.

Then, suppressing the index $i$, the expected value of the estimating function $S_x(\beta)$ is

$$E_{y,z,x} \left\{ I(x \geq d) \frac{x(y - \mu)}{h(\mu)g'(\mu)} \right\} + E_{y,z,x} \left\{ I(x < d) \frac{\overline{x}(y - \mu)}{h(\overline{\mu})g'(\overline{\mu})} \right\}
= E_{z,x} \left\{ I(x \geq d) \frac{x}{h(\mu)g'(\mu)} E \{ (y - \mu) \mid z, x \} \right\} + E_{z,x} \left\{ E \left\{ I(x < d) \frac{\overline{x}(y - \mu)}{h(\overline{\mu})g'(\overline{\mu})} \mid z, x \right\} \right\}
= 0 + E_{z,x} \left\{ I(x < d) \frac{\overline{x}(\mu - \overline{\mu})}{h(\overline{\mu})g'(\overline{\mu})} \right\}
= E_z \left\{ \frac{E(x \mid x < d)}{h(\overline{\mu})g'(\overline{\mu})} E \{ I(x < d) (\mu - \overline{\mu}) \} \right\}
= E_z \left\{ \frac{1}{h(\overline{\mu})g'(\overline{\mu})} \int_{-\infty}^{d} x f(x) dx \left\{ \int_{-\infty}^{d} \mu f(x) dx - \overline{\mu} \int_{-\infty}^{d} f(x) dx \right\} \right\}
= E_z \left\{ \frac{\int_{-\infty}^{d} x f(x) dx}{h(\overline{\mu})g'(\overline{\mu})} \left\{ \int_{-\infty}^{d} \mu f(x) dx - \overline{\mu} \int_{-\infty}^{d} f(x) dx \right\} \right\}
= E_z \left\{ \frac{\int_{-\infty}^{d} x f(x) dx}{h(\overline{\mu})g'(\overline{\mu})} \left\{ E_x(\mu \mid x < d) - \overline{\mu} \right\} \right\}.

Similarly, the expected value of the estimating function $S_x(\beta)$ is

$$E_z \left\{ \frac{z \int_{-\infty}^{d} f(x) dx}{h(\overline{\mu})g'(\overline{\mu})} \left\{ E_x(\mu \mid x < d) - \overline{\mu} \right\} \right\}.
$$

Thus, under regularity conditions A.1-A.3, it can be shown with standard M-estimation theory
that $\beta_{M1} \xrightarrow{p} \beta$ if and only if

$$\int_{Z} \int_{-\infty}^{d} w f(x) dx h(\overline{\mu})g'(\overline{\mu}) \left\{ E_x(\mu \mid x < d) - \overline{\mu} \right\} f(z) dz = 0,
$$

where $w = (z^T, x^T)$.

We now consider conditional mean imputation using $\overline{x} = E(x \mid z, x < d)$. Now $\overline{x}$, $\overline{\mu}$, $g'(\overline{\mu})$, 
124
and \( h(\bar{\mu}) \) all depend on \( z \). It follows that the expectation of the estimating function \( S_x(\beta) \) is

\[
E_{y,z,x} \left\{ I(x \geq d) \frac{x(y - \mu)}{h(\mu)g'(\mu)} \right\} + E_{y,z,x} \left\{ I(x < d) \frac{\bar{x}(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \right\} = E_{z,x} \left[ I(x \geq d) \frac{x}{h(\mu)g'(\mu)} E \{ (y - \mu) \mid z, x \} \right] + E_{z,x} \left[ E \left\{ I(x < d) \frac{\bar{x}(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \mid z, x \right\} \right] = 0 + E_{z,x} \left\{ I(x < d) \frac{\bar{x}(\mu - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \right\} = E_{z} \left[ E \left\{ I(x < d) \frac{\bar{x}(\mu - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \mid z \right\} \right] = E_{z} \left[ \frac{1}{h(\bar{\mu})g'(\bar{\mu})} \int_{-\infty}^{d} x f(x \mid z) dx \int_{-\infty}^{d} f(x \mid z) dx \right] = E_{z} \left[ \frac{1}{h(\bar{\mu})g'(\bar{\mu})} \left\{ \frac{1}{h(\bar{\mu})g'(\bar{\mu})} \int_{-\infty}^{d} f(x \mid z) dx \right\} \right] = E_{z} \left[ \frac{1}{h(\bar{\mu})g'(\bar{\mu})} \{ E_x (\mu \mid z, x < d) - \bar{\mu} \} \right].
\]

Similarly, the expected value of the estimating function \( S_z(\beta) \) is

\[
E_{z} \left[ \frac{z \int_{-\infty}^{d} f(x \mid z) dx}{h(\bar{\mu})g'(\bar{\mu})} \{ E_x (\mu \mid z, x < d) - \bar{\mu} \} \right].
\]

Thus, under regularity conditions A.1-A.3, it can be shown with standard M-estimation theory that \( \tilde{\theta}_{M2} \to \beta \) if and only if

\[
\int_{Z} \int_{-\infty}^{d} \frac{w f(x \mid z) dz}{h(\bar{\mu})g'(\bar{\mu})} \{ E_x (\mu \mid z, x < d) - \bar{\mu} \} f(z) dz = 0,
\]

where \( w = (z^T, x)^T \).

Finally, we consider conditional mean imputation using \( \bar{x} = E(x \mid y, z, x < d) \). Now \( \bar{x}, \bar{\mu}, g'(\bar{\mu}) \),
and $h(\bar{\mu})$ all depend on $y$. It follows that the expectation of the estimating function $S_z(\beta)$ is

$$
E_{y,z,x} \left\{ I(x \geq d) \frac{x(y - \mu)}{h(\mu)g'(\mu)} \right\} + E_{y,z,x} \left\{ I(x < d) \frac{\bar{x}(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \right\} 
= E_{z,x} \left[ I(x \geq d) \frac{x}{h(\mu)g'(\mu)} E \{ (y - \mu) | z, x \} \right] + E_{y,z} \left[ E \left\{ I(x < d) \frac{\bar{x}(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \right| y, z \right]\]
= 0 + E_{y,z} \left[ \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} E \{ x | y, z, x < d \} E \{ I(x < d) \mid y, z \} \right]\]
= E_{y,z} \left\{ \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} \int_{-\infty}^{d} x f(x | y, z) dx \int_{-\infty}^{d} f(x | y, z) dx \right\}
= E_{y,z} \left\{ \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} \int_{-\infty}^{d} x f(x | y, z) dx \right\}
= \int_{y \times z} \left\{ \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} \int_{-\infty}^{d} x f(x | y, z) dx \right\} f(y, z) d(y, z).$$

By the regularity condition A.3 and Fubini’s theorem, we can write this expression as

$$
\int_{z} \int_{-\infty}^{d} x \left\{ \int_{y} \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} f(y | z, x) dy \right\} f(z, x) dxdz.
$$

Similarly, the expected value of the estimating function $S_z(\beta)$ is

$$
\int_{z} \int_{-\infty}^{d} \left\{ \int_{y} \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} f(y | z, x) dy \right\} f(z, x) dxdz.
$$

Thus, under regularity conditions A.1-A.3, it can be shown with standard M-estimation theory that $\hat{\beta}_{M3} \sim^{p} \beta$ if and only if

$$
\int_{z} \int_{-\infty}^{d} w \left\{ \int_{y} \frac{y - \bar{\mu}}{h(\bar{\mu})g'(\bar{\mu})} f(y | z, x) dy \right\} f(z, x) dxdz = 0,
$$

where $w = (z^T, x^T)$. □
A.4 Proof of Theorem 2

Let \((y_i, z_i, x_i, \rho_i), i = 1, \ldots, n\), be independent samples with \(f(y_i|z_i, x_i)\) belonging to the exponential dispersion family, given by (2.1), and \(g(\mu_i) = \sum_{j=1}^{p} w_{ij} \beta_j\), where \(w_i = (z_i^T, x_i)^T\). Let \(x_i\) be univariate and \(f(x_i|z_i)\) be a known distribution. When \(x_i < d\), define imputed values of \(x_i\) by \(\bar{x}_i\), a draw from the distribution \(f(x_i|z_i, x_i < d; \gamma)\).

Then, suppressing the index \(i\), the expected value of the estimating function \(S_x(\beta)\) is

\[
E_{y,z,x} \left\{ I(x \geq d) \frac{x(y - \mu)}{h(\mu)g'(\mu)} \right\} + E_{y,z,x} \left\{ I(x < d) \frac{\bar{x}(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \right\} = E_{z,x} \left\{ I(x \geq d) \frac{x}{h(\mu)g'(\mu)} E\{ (y - \mu) | z, x \} \right\} + E_{z,x} \left\{ I(x < d) E\left\{ \frac{\bar{x}(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} | z, x \right\} \right\} = 0 \quad E_{z,x} \left\{ I(x < d) \frac{\bar{x}}{h(\bar{\mu})g'(\bar{\mu})} \{ \mu - \bar{\mu} \} \right\} = E_z \left\{ E \left\{ I(x < d) \frac{\bar{x} \mu}{h(\bar{\mu})g'(\bar{\mu})} | z \right\} - E \left\{ I(x < d) \frac{\bar{x} \mu}{h(\bar{\mu})g'(\bar{\mu})} | z \right\} \right\} = E_z \left\{ E \{ I(x < d) \mu \} E \left\{ \frac{\bar{x}}{h(\bar{\mu})g'(\bar{\mu})} | z \right\} - E \{ I(x < d) \} E \left\{ \frac{\bar{x} \mu}{h(\bar{\mu})g'(\bar{\mu})} | z \right\} \right\} = E_z \left[ \int_{-\infty}^{d} f(x|z) dx E \left\{ \frac{x}{h(\mu)g'(\mu)} | z, x < d \right\} \right] - \int_{-\infty}^{d} f(x|z) dx E \left\{ \frac{x \mu}{h(\mu)g'(\mu)} | z, x < d \right\} \right\} = E_z \left[ \frac{Pr(x < d|z)}{I(x < d) \mu \} E \left\{ \frac{x}{h(\mu)g'(\mu)} | z, x < d \right\} \right] - E \left\{ \frac{x \mu}{h(\mu)g'(\mu)} | z, x < d \right\} \right\} \right\} = -E_z \left[ \frac{Pr(x < d|z)}{I(x < d) \mu \} Cov \left\{ \frac{x}{h(\mu)g'(\mu)} \right\} \right],
\]

where the covariance is taken with respect to the distribution \(f(x|z, x < d; \gamma)\). If we consider the estimating function for the \(\beta\) parameters associated with \(z, S_x(\beta)\), following a similar argument
it is easy to show
\[
E\{S_{z}(\beta)\} = -E_{z}\left[ zPr(x < d|z)\text{Cov}\left(\frac{1}{h(\mu)g'(\mu)}, \mu\right) \right],
\]
where the covariance is again with respect to the distribution \(f(x|z, x < d; \gamma)\). Under the regularity conditions A.1-A.3, it can be shown with standard M-estimation theory that \(\hat{\beta}_{IML} \xrightarrow{p} \beta\) if and only if
\[
E_{z}\left[ Pr(x < d|z)\text{Cov}\left(\frac{x}{h(\mu)g'(\mu)}, \mu\right) \right] = 0 \quad \text{and} \quad E_{z}\left[ zPr(x < d|z)\text{Cov}\left(\frac{1}{h(\mu)g'(\mu)}, \mu\right) \right] = 0. \]

A.5 Proof Sketch of Theorem 3

Let \((y_{i}, z_{i}, x_{i}, \rho_{i})\), \(i = 1, \ldots, n\), be independent samples with \(f(y_{i}|z_{i}, x_{i})\) belonging to the exponential dispersion family, given by (2.1), and \(g(\mu_{i}) = \sum_{j=1}^{p} w_{ij} \beta_{j}\), where \(w_{i} = (z_{i}^{T}, x_{i}^{T})^{T}\).

For a random individual, suppressing the index \(i\), there are as many as \(2^{q}\) potential censoring patterns for \(q\)-dimensional \(x\). Define the possible subsets of observed and censored covariates as \(x_{i}^{o}\) and \(x_{i}^{c}\), respectively, \(l = 1, \ldots, 2^{q}\). Also, define \(d_{i}^{c}\) as the vector of DLs associated with the covariates \(x_{i}^{c}\). Let \(x_{i}^{c}|z, x_{i}^{o}, x_{i}^{c} < d_{i}^{c} \sim f(x_{i}^{c}|z, x_{i}^{o}, x_{i}^{c} < d_{i}^{c}; \gamma)\) be known for all possible \(x_{i}^{c}\), \(l = 1, \ldots, 2^{q}\), and let \(\bar{x}_{i}^{c}\), the imputed vector for \(x_{i}^{c}\), be drawn from this distribution. Define \(\bar{x}_{i}\) as the vector consisting of \(x_{i}^{o}\) and \(\bar{x}_{i}^{c}\) together.

Then, the expected value of the estimating functions given by (2.3), using imputed values, is
\[
\sum_{l=1}^{2^{q}} E_{\bar{x}_{i}}^{c} \left( \frac{y - \bar{\mu}_{l}}{h(\bar{\mu}_{l})g'(\bar{\mu}_{l})} I(x_{i}^{c} < d_{i}^{c}) \right).
\]
We now consider a single censoring combination \(l\), and thus drop the subscript \(l\) henceforth. Then, the expected value of the \(l^{th}\) part of the estimating function for \(\beta_{j}\) associated with
covariate \( x_j \in x^c \) is

\[
E \left\{ \frac{x_j(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} I(x^c < d^c) \right\}
= E \left[ E \left\{ \frac{x_j(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} I(x^c < d^c) \bigg| y, z, x^o \right\} \right]
= E \left[ I(x^c < d^c) \left\{ \frac{x_j(y - \bar{\mu})}{h(\bar{\mu})g'(\bar{\mu})} \bigg| y, z, x^o \right\} \right]
= E \left[ \int_{-\infty}^{d^c} f(x^c|y, z, x^o) dx^c E \left\{ \frac{x_j(y - \mu)}{h(\mu)g'(\mu)} \bigg| y, z, x^o, x^c < d^c \right\} \right]
= E \left[ \int_{-\infty}^{d^c} \frac{x_j(y - \mu)}{h(\mu)g'(\mu)} f(x^c|y, z, x^o) dx^c \right]
= \int_{y \times z \times x^o} \left\{ \int_{-\infty}^{d^c} \frac{x_j(y - \mu)}{h(\mu)g'(\mu)} f(x^c|y, z, x^o) dx^c \right\} f(y, z, x^o) d(y, z, x^o).

By regularity condition A.3 and Fubini’s Theorem, we can write this expression as

\[
\int_{z} \int_{x^o} \int_{-\infty}^{d^c} \frac{x_j(y - \mu)}{h(\mu)g'(\mu)} \left\{ \int_{y} (y - \mu)f(y|z, x^o, x^c)dy \right\} f(z, x^o, x^c) dx^c dx^o dz
= \int_{z} \int_{x^o} \int_{-\infty}^{d^c} \frac{x_j(\mu - \mu)}{h(\mu)g'(\mu)} f(z, x^o, x^c) dx^c dx^o dz
= 0.
\]

This is true for any censoring combination \( l \) and is also straightforward to show for the estimating functions associated with the elements of \( z \) and \( x^o \). Thus, the expectation of the estimating functions is 0. Note that for the parameters \( \gamma \), which are essentially nuisance parameters in this problem, the score functions with complete data are independent of those based on \( \beta \). Using a similar approach as above, it is easy to show the derivative of the likelihood equations for \( \gamma \) also have expected value 0. We now outline the remainder of this proof following similar arguments as in Tsiatis (2006, Ch. 14).

Until now we have assumed that the multiple imputations were drawn based on the true
parameter vector, $\theta$. However, in practice, the imputations are based on $\widehat{\theta}'$, a consistent initial estimate of $\theta$, so that standard M-estimation theory does not directly give the consistency of $\widehat{\theta}_{IMI}$. For the remainder of the proof, assume that $\widehat{\beta}'$ and $\widehat{\gamma}'$ in $\widehat{\theta}' = (\widehat{\beta}'^T, \widehat{\gamma}'^T)^T$ are obtained via complete case and maximum likelihood estimation, respectively, as described in Section 3.1 of the dissertation. Then $\widehat{\theta}'$ is a regular and asymptotically linear estimator (RAL) for the true parameter $\theta$ as $n^* \to \infty$ (see Ch. 3 in Tsiatis, 2006 for an explanation of RAL estimators). For any RAL estimator, we can derive its unique influence function, as is necessary for Lemma 2 below.

For the remainder of the proof, let $\theta$ be the true parameter vector, $\widehat{\theta}_{IMI}$ be the improper multiple imputation estimator, and $\widehat{\theta}^{(k)}$ be the maximum likelihood estimator for the $k^{th}$ imputed dataset. We now derive an approximation to the influence function of $\widehat{\theta}_{IMI}$ in order to derive its asymptotic properties. We henceforth represent the score equations based on the imputed data as $S^F\{y_i, z_i, \bar{x}_{ik}(\widehat{\theta}'); \theta\}$.

**Lemma 1.** The improper multiple imputation estimator, $\widehat{\theta}_{IMI}$, can be approximated as

$$n^{1/2}(\widehat{\theta}_{IMI} - \theta) = n^{1/2} \sum_{i=1}^{n} \{I^F(\theta)\}\sum_{k=1}^{m} S^F\{y_i, z_i, \bar{x}_{ik}(\widehat{\theta}'); \theta\} + o_p(1),$$

where $I^F(\theta) = E\{S^F(y_i, w_i; \theta)S^{F^T}(y_i, w_i; \theta)\}$ is the full-data information matrix and $S^F\{y_i, z_i, \bar{x}_{ik}(\widehat{\theta}'); \theta\} = \frac{\partial}{\partial \theta} \{\ell\{y_i, z_i, \bar{x}_{ik}(\widehat{\theta}'); \theta\}\}$ is the full-data score vector for the $i^{th}$ individual ($\ell$ is the log-likelihood of the full data), based on the $k^{th}$ imputed vector $\bar{x}_{ik}$, which in turn is based on the initial parameter estimate $\widehat{\theta}'$.

**Proof.** Refer to the proof of Lemma 14.1 in Tsiatis (2006) for details. \qed

Since the score vector $S^F\{y_i, z_i, \bar{x}_{ik}(\widehat{\theta}'); \theta\}$ is dependent on the random quantity $\widehat{\theta}'$, $n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{m} S^F\{y_i, z_i, \bar{x}_{ik}(\widehat{\theta}'); \theta\}$ is not the sum of i.i.d. terms, and thus the approximation given in Lemma 1 does not give the influence function for $\widehat{\theta}_{IMI}$. However, we can write
\[ n^{-1/2} \sum_{i=1}^{n} \left[ m^{-1} \sum_{k=1}^{m} S^F \left\{ y_i, z_i, \bar{x}_{ik}(\theta^I); \theta \right\} \right] \]

\[ + n^{-1/2} \sum_{i=1}^{n} \left[ m^{-1} \sum_{k=1}^{m} S^F \left\{ y_i, z_i, \bar{x}_{ik}(\theta^I); \theta \right\} - m^{-1} \sum_{k=1}^{m} S^F \left\{ y_i, z_i, \bar{x}_{ik}(\theta); \theta \right\} \right]. \]  

(A.1)

The first term of expression (A.1) is the sum of mean-zero (shown in the first part of proof in Appendix D) i.i.d. random vectors that converge asymptotically to a normal distribution by the central limit theorem under general regularity conditions.

**Lemma 2.** The second term in expression (A.1) above is equal to

\[ n^{1/2} \sum_{i=1}^{n} \left\{ I^F(\theta) - I(\theta) \right\} q(y_i, z_i, x_i^*, \rho_i) + o_p(1), \]

where \( I^F(\theta) = E\{S^F(y_i, w_i; \theta)S^{F^T}(y_i, w_i; \theta)\} \) is the full-data information matrix,

\( I(\theta) = E\{S(y_i, z_i, x_i^*, \rho_i; \theta)S^T(y_i, z_i, x_i^*, \rho_i; \theta)\} \) is the observed-data information matrix, and

\( q(y_i, z_i, x_i^*, \rho_i) \) is the \( i \)th influence function of \( \hat{\theta}^I \).

**Proof.** Refer to the proofs of Theorem 14.3 and Lemmas 14.2-5 in Tsiatis (2006) for details. \( \square \)

Together, Lemma 1 and Lemma 2 give the influence function for \( \hat{\theta}_{IMI} \). Then, by the central limit theorem, under general regularity conditions, we have that

\[ n^{-1/2}(\hat{\theta}_{IMI} - \theta) \rightarrow N(0, \Sigma), \]

where

\[ \Sigma = \text{Var} \left\{ I^F(\theta) \right\}^{-1} \left[ m^{-1} \sum_{k=1}^{m} S^F \left\{ y_i, z_i, \bar{x}_{ik}(\theta); \theta \right\} + \left\{ I^F(\theta) - I(\theta) \right\} q(y_i, z_i, x_i^*, \rho_i) \right\}. \]
Some algebra gives

$$\text{Var} \left( m^{-1} \sum_{k=1}^{m} S^F \{ y_i, z_i, \bar{x}_{ik}(\theta); \theta \} \right) = m^{-1} \{ I^F(\theta) - I(\theta) \} + I(\theta),$$

$$\text{Cov} \left( m^{-1} \sum_{k=1}^{m} S^F \{ y_i, z_i, \bar{x}_{ik}(\theta); \theta \}, \{ I^F(\theta) - I(\theta) \} q(y_i, z_i, x_i^*, \rho_i) \right) = I^F(\theta) - I(\theta),$$

and

$$\text{Var} \left( \{ I^F(\theta) - I(\theta) \} q(y_i, z_i, x_i^*, \rho_i) \right) = \{ I^F(\theta) - I(\theta) \} \text{Var} \{ q(y_i, z_i, x_i^*, \rho_i) \} \{ I^F(\theta) - I(\theta) \}. $$

Together this gives

$$\Sigma = \{ I^F(\theta) \}^{-1} \left[ I(\theta) + m^{-1} \{ I^F(\theta) - I(\theta) \} + 2 \{ I^F(\theta) - I(\theta) \} \right]$$

$$+ \{ I^F(\theta) - I(\theta) \} \text{Var} \{ q(y_i, z_i, x_i^*, \rho_i) \} \{ I^F(\theta) - I(\theta) \} \{ I^F(\theta) \}^{-1}. $$

With a little algebraic manipulation, we get

$$\Sigma = \{ I^F(\theta) \}^{-1} + (1 + m^{-1}) \{ I^F(\theta) \}^{-1} \{ I^F(\theta) - I(\theta) \} \{ I^F(\theta) \}^{-1}$$

$$+ \{ I^F(\theta) \}^{-1} \{ I^F(\theta) - I(\theta) \} \text{Var} \{ q(y_i, z_i, x_i^*, \rho_i) \} \{ I^F(\theta) - I(\theta) \} \{ I^F(\theta) \}^{-1}. $$

For the initial estimators used in the paper, $\text{Var} \{ q(y_i, z_i, x_i^*, \rho_i) \} = \text{Var}(\hat{\theta}^I).$
Appendix B

Additional Simulations for Generalized Linear Models with Covariates Subject to Detection Limits

We provide several additional tables of simulation results to demonstrate the effect of changing parameters in the simulations appearing in Section 2.5. Tables B.1 and B.2 give the results for a simulation with the same parameters as in the Section 2.5 but with 500 observations per dataset. The results in Table B.1 largely echo the results from the Section 2.5, though the results in Table B.2 less clearly demonstrate the disadvantage of using Rubin’s variance formula and bootstrap techniques for estimating the variance of $\hat{\beta}_{IMI}$.

Tables B.3 and B.4 give the results for a simulation with 200 observations per dataset and a lower set of parameter values, $\beta = (0.5, -0.45, 0.3)$. There are very similar patterns of bias, efficiency, and computational cost as those in Table 2.1, though we note that the coverage probabilities for the naive estimators are slightly better.
Table B.1: Simulation results for a logistic regression on 2000 datasets with 500 observations and two covariates of correlation $\rho = 0.5$. Biases which are statistically significant at the $\alpha = 0.99$ level are shown in bold. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95% coverage probability; Time (s): average elapsed computational time in seconds for a single simulated dataset; Omni: omniscient method; CC: complete case analysis; SUB: substitution by DL/$\sqrt{2}$; Mean1: conditional mean imputation method conditioning only on the DL; Mean2: conditional mean imputation method conditioning on the covariates only; Mean3: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

<table>
<thead>
<tr>
<th>Cens. %</th>
<th>Method</th>
<th>$\beta_0 = 3$</th>
<th>$\beta_1 = 1.5$</th>
<th>$\beta_2 = -3$</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias SD SE CP</td>
<td>Bias SD SE CP</td>
<td>Bias SD SE CP</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Omni</td>
<td>0.01 0.27 0.28 0.95</td>
<td>0.01 0.23 0.23 0.94</td>
<td>-0.02 0.33 0.33 0.94</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.01 0.31 0.30 0.95</td>
<td>0.01 0.28 0.27 0.94</td>
<td>-0.01 0.36 0.35 0.95</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td><strong>-0.04</strong> 0.28 0.27 0.94</td>
<td><strong>0.02</strong> 0.25 0.23 0.94</td>
<td><strong>0.02</strong> 0.34 0.32 0.93</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>-0.02 0.28 0.28 0.94</td>
<td><strong>0.03</strong> 0.24 0.23 0.94</td>
<td>-0.04 0.34 0.33 0.94</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td><strong>0.02</strong> 0.29 0.28 0.94</td>
<td>-0.08 0.23 0.22 0.91</td>
<td>0.01 0.35 0.33 0.94</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.02 0.29 0.28 0.94</td>
<td><strong>0.07</strong> 0.26 0.24 0.93</td>
<td>-0.05 0.36 0.33 0.94</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>-0.01 0.28 0.29 0.95</td>
<td>-0.01 0.25 0.25 0.95</td>
<td>0.01 0.34 0.34 0.94</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td><strong>0.03</strong> 0.29 0.29 0.95</td>
<td><strong>0.02</strong> 0.25 0.24 0.94</td>
<td>-0.04 0.35 0.34 0.94</td>
<td>249.2</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.01 0.29 0.28 0.95</td>
<td>0.01 0.25 0.24 0.94</td>
<td>-0.02 0.35 0.33 0.94</td>
<td>125.8</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>-0.01 0.29 0.28 0.95</td>
<td>-0.01 0.25 0.24 0.94</td>
<td>0.00 0.34 0.33 0.94</td>
<td>1.9</td>
</tr>
</tbody>
</table>

| 40     | Omni   | 0.01 0.28 0.28 0.95 | 0.01 0.24 0.23 0.94 | -0.01 0.33 0.33 0.94 | < 0.1     |
|        | CC     | 0.02 0.37 0.36 0.95 | 0.00 0.32 0.32 0.95 | -0.01 0.40 0.40 0.95 | < 0.1     |
|        | SUB    | **-0.13** 0.26 0.26 0.89 | **0.06** 0.26 0.25 0.94 | **0.08** 0.32 0.31 0.92 | < 0.1     |
|        | Mean1  | **-0.11** 0.28 0.27 0.90 | **-0.10** 0.23 0.22 0.88 | **0.16** 0.33 0.31 0.83 | 0.3       |
|        | Mean2  | **-0.07** 0.29 0.28 0.91 | **-0.14** 0.23 0.22 0.86 | **0.13** 0.35 0.32 0.88 | 1.0       |
|        | Mean3  | 0.00 0.31 0.28 0.92 | **0.15** 0.30 0.25 0.88 | -0.04 0.38 0.33 0.92 | 2.9       |
|        | ML     | 0.00 0.28 0.29 0.96 | -0.02 0.26 0.25 0.94 | 0.00 0.34 0.34 0.95 | 171.3     |
|        | FB     | **0.02** 0.30 0.29 0.95 | **-0.03** 0.26 0.26 0.94 | -0.02 0.35 0.35 0.94 | 259.6     |
|        | PMI    | 0.00 0.29 0.30 0.96 | **-0.02** 0.26 0.26 0.94 | -0.01 0.35 0.35 0.96 | 162.1     |
|        | IMI    | 0.00 0.28 0.29 0.95 | -0.02 0.26 0.25 0.94 | 0.00 0.35 0.35 0.95 | 2.1       |

| 60     | Omni   | 0.01 0.28 0.28 0.95 | 0.00 0.24 0.23 0.94 | -0.01 0.33 0.33 0.95 | < 0.1     |
|        | CC     | 0.02 0.48 0.48 0.95 | -0.02 0.41 0.41 0.95 | -0.01 0.48 0.47 0.95 | < 0.1     |
|        | SUB    | **-0.25** 0.23 0.24 0.79 | **0.14** 0.30 0.28 0.93 | **0.09** 0.30 0.30 0.93 | < 0.1     |
|        | Mean1  | **-0.31** 0.26 0.24 0.67 | **-0.22** 0.22 0.21 0.78 | **0.44** 0.30 0.27 0.58 | 0.3       |
|        | Mean2  | **-0.31** 0.27 0.25 0.68 | **-0.22** 0.22 0.21 0.77 | **0.41** 0.31 0.28 0.62 | 1.4       |
|        | Mean3  | **-0.17** 0.30 0.26 0.82 | **0.20** 0.33 0.26 0.84 | **0.10** 0.37 0.31 0.88 | 3.9       |
|        | ML     | -0.02 0.28 0.28 0.95 | **-0.02** 0.25 0.25 0.94 | 0.02 0.34 0.35 0.95 | 309.5     |
|        | FB     | 0.01 0.30 0.30 0.95 | **-0.04** 0.27 0.26 0.94 | 0.00 0.36 0.36 0.94 | 291.5     |
|        | PMI    | -0.01 0.29 0.31 0.95 | -0.02 0.27 0.27 0.94 | 0.02 0.36 0.37 0.96 | 230.9     |
|        | IMI    | -0.01 0.29 0.30 0.96 | -0.01 0.28 0.26 0.94 | 0.02 0.35 0.36 0.96 | 2.7       |
Table B.2: A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 datasets with 500 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.0005 and 0.0015. SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}_{IMI}$; SE$_C$: average standard error estimated using (2.8); SE$_A$: average standard error estimated using (2.9); SE$_R$: average standard error estimated using (2.6); SE$_B$: average standard error estimated via a bootstrap procedure.

<table>
<thead>
<tr>
<th></th>
<th>20 % Censoring</th>
<th>40 % Censoring</th>
<th>60 % Censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>SD</td>
<td>0.279</td>
<td>0.249</td>
<td>0.344</td>
</tr>
<tr>
<td>SE$_C$</td>
<td>0.281</td>
<td>0.248</td>
<td>0.342</td>
</tr>
<tr>
<td>SE$_A$</td>
<td>0.281</td>
<td>0.248</td>
<td>0.342</td>
</tr>
<tr>
<td>SE$_R$</td>
<td>0.280</td>
<td>0.246</td>
<td>0.339</td>
</tr>
<tr>
<td>SE$_B$</td>
<td>0.289</td>
<td>0.250</td>
<td>0.349</td>
</tr>
</tbody>
</table>
Table B.3: Simulation results for a logistic regression on 2000 datasets with 200 observations and two covariates of correlation \( \rho = 0.5 \). Biases which are statistically significant at the \( \alpha = 0.99 \) level are shown in bold. Cens. %: percent of individuals with at least one censored covariate; SD: simulation (Monte Carlo) standard deviation of \( \hat{\beta} \); SE: average estimated standard error; CP: empirical 95% coverage probability; Time (s): average elapsed computational time in seconds for a single simulated dataset; Omni: omniscient method; CC: complete case analysis; SUB: substitution by \( \text{DL}/\sqrt{2} \); Mean1: conditional mean imputation method conditioning only on the DL; Mean2: conditional mean imputation method conditioning on the covariates only; Mean3: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method.

<table>
<thead>
<tr>
<th>Cens. %</th>
<th>Method</th>
<th>( \beta_0 = 0.5 )</th>
<th>( \beta_1 = -0.45 )</th>
<th>( \beta_2 = 0.3 )</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias  SD  SE  CP</td>
<td>Bias  SD  SE  CP</td>
<td>Bias  SD  SE  CP</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Omni</td>
<td>0.00  0.14 0.15 0.96</td>
<td>0.01  0.30 0.30 0.95</td>
<td>0.01  0.30 0.30 0.95</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.00  0.17 0.18 0.96</td>
<td>0.00  0.41 0.40 0.95</td>
<td>0.00  0.40 0.41 0.96</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.08  0.26 0.26 0.94</td>
<td>-0.05  0.25 0.26 0.95</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.01  0.31 0.31 0.96</td>
<td>0.00  0.31 0.31 0.96</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td>0.00  0.14 0.15 0.96</td>
<td>0.01  0.30 0.31 0.96</td>
<td>0.00  0.30 0.31 0.95</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.02  0.31 0.31 0.95</td>
<td>0.02  0.31 0.31 0.95</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.00  0.14 0.14 0.95</td>
<td>0.01  0.30 0.30 0.96</td>
<td>0.00  0.30 0.30 0.95</td>
<td>58.7</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.01  0.15 0.15 0.95</td>
<td>0.00  0.31 0.31 0.96</td>
<td>0.01  0.31 0.31 0.96</td>
<td>65.5</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.01  0.30 0.31 0.96</td>
<td>0.00  0.31 0.31 0.95</td>
<td>32.2</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.01  0.30 0.31 0.95</td>
<td>0.00  0.31 0.31 0.95</td>
<td>1.1</td>
</tr>
<tr>
<td>40</td>
<td>Omni</td>
<td>0.00  0.14 0.15 0.96</td>
<td>0.01  0.30 0.30 0.95</td>
<td>0.01  0.30 0.30 0.95</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.00  0.24 0.24 0.96</td>
<td>0.00  0.52 0.52 0.96</td>
<td>0.02  0.51 0.53 0.97</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>0.01  0.15 0.15 0.96</td>
<td>0.11  0.24 0.24 0.93</td>
<td>-0.07  0.24 0.24 0.94</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>0.00  0.14 0.15 0.96</td>
<td>0.01  0.32 0.32 0.95</td>
<td>0.00  0.33 0.33 0.95</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.02  0.30 0.31 0.96</td>
<td>0.01  0.31 0.31 0.95</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.00  0.15 0.15 0.96</td>
<td>-0.03  0.32 0.31 0.95</td>
<td>0.03  0.33 0.32 0.94</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.01  0.31 0.32 0.95</td>
<td>0.00  0.31 0.32 0.95</td>
<td>76.1</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.01  0.15 0.15 0.95</td>
<td>0.00  0.32 0.32 0.95</td>
<td>0.01  0.32 0.32 0.95</td>
<td>71.2</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.01  0.31 0.32 0.96</td>
<td>0.00  0.31 0.32 0.95</td>
<td>50.1</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.00  0.31 0.31 0.95</td>
<td>0.01  0.31 0.32 0.95</td>
<td>1.1</td>
</tr>
<tr>
<td>60</td>
<td>Omni</td>
<td>0.00  0.14 0.15 0.96</td>
<td>0.01  0.30 0.30 0.95</td>
<td>0.01  0.30 0.30 0.95</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.01  0.39 0.40 0.96</td>
<td>0.00  0.71 0.73 0.97</td>
<td>0.01  0.72 0.74 0.96</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>-0.01  0.15 0.15 0.96</td>
<td>0.12  0.24 0.24 0.92</td>
<td>-0.08  0.24 0.24 0.94</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td></td>
<td>Mean1</td>
<td>0.00  0.16 0.15 0.96</td>
<td>0.03  0.34 0.35 0.95</td>
<td>0.00  0.36 0.36 0.95</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Mean2</td>
<td>0.00  0.14 0.15 0.96</td>
<td>0.02  0.31 0.32 0.95</td>
<td>0.02  0.32 0.33 0.95</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Mean3</td>
<td>0.00  0.15 0.15 0.96</td>
<td>-0.04  0.35 0.32 0.93</td>
<td>0.04  0.36 0.33 0.93</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>0.00  0.15 0.15 0.95</td>
<td>0.01  0.32 0.32 0.95</td>
<td>0.01  0.32 0.32 0.95</td>
<td>102.2</td>
</tr>
<tr>
<td></td>
<td>FB</td>
<td>0.01  0.15 0.15 0.96</td>
<td>0.00  0.33 0.33 0.95</td>
<td>0.01  0.33 0.33 0.96</td>
<td>78.6</td>
</tr>
<tr>
<td></td>
<td>PMI</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.00  0.33 0.33 0.95</td>
<td>0.01  0.33 0.33 0.96</td>
<td>63.4</td>
</tr>
<tr>
<td></td>
<td>IMI</td>
<td>0.00  0.15 0.15 0.96</td>
<td>0.00  0.33 0.33 0.95</td>
<td>0.01  0.33 0.33 0.95</td>
<td>1.3</td>
</tr>
</tbody>
</table>

136
Table B.4: A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 datasets with 200 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.00009 and 0.0006. SD: simulation (Monte Carlo) standard deviation of $\hat{\beta}_{IMI}$; SE$_C$: average standard error estimated using (2.8); SE$_A$: average standard error estimated using (2.9); SE$_R$: average standard error estimated using (2.6); SE$_B$: average standard error estimated via a bootstrap procedure.

<table>
<thead>
<tr>
<th></th>
<th>20 % Censoring</th>
<th>40 % Censoring</th>
<th>60 % Censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_0$</td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
</tr>
<tr>
<td>SD</td>
<td>0.440</td>
<td>0.377</td>
<td>0.520</td>
</tr>
<tr>
<td>SE$_C$</td>
<td>0.449</td>
<td>0.382</td>
<td>0.526</td>
</tr>
<tr>
<td>SE$_A$</td>
<td>0.449</td>
<td>0.382</td>
<td>0.525</td>
</tr>
<tr>
<td>SE$_R$</td>
<td>0.447</td>
<td>0.381</td>
<td>0.522</td>
</tr>
<tr>
<td>SE$_B$</td>
<td>0.469</td>
<td>0.401</td>
<td>0.538</td>
</tr>
</tbody>
</table>
Appendix C

Chapter 3 Proofs and Technical Details

In this appendix, we give proofs for Proposition 2, and Theorems 4-5 in Chapter 3. For notational simplicity, we do not differentiate between random variables and their realized values by using upper and lower case letters, as is typical.

C.1 Regularity Conditions

Define \( \frac{\partial}{\partial \theta} \{ \ell(\theta | \bar{t}_i, \delta_i, w_i) \} = \psi(\bar{t}_i, \delta_i, w_i, \theta) \), where \( \ell \) is the log-likelihood and \( \psi \) and \( w_i \) are \( s \)-dimensional. We consider the following regularity conditions:

C.1 The \( p \times 1 \) parameter vector \( \theta \) lies in a compact subset \( \Theta \) of \( \mathbb{R}^p \), and \( E\{\psi(\bar{t}, \delta, w, \theta)\} \) exists for all \( \theta \in \Theta \).

C.2 If \( E\{\psi(\bar{t}, \delta, w, \theta)\} = 0 \), then \( E\{\psi(\bar{t}, \delta, w, \theta')\} \neq 0 \) unless \( \theta = \theta' \).

C.3 The \( j^{th} \) component of \( \psi(\bar{t}, \delta, w, \theta) \), \( \psi_j(\bar{t}, \delta, w, \theta) \), is continuous in \( \theta \) and bounded by an integrable function of \( (\bar{t}, \delta, w) \) that does not depend on \( \theta \) for \( j = 1, \ldots, s \).
C.2 Proof of Proposition 2

Let \((\tilde{t}_i, \delta_i, w_i), i = 1, \ldots, n\), be independent samples with \(f(t_i|w_i)\) known and \(w_i = (z_i^T, x_i^T)^T\). Defining \(\rho_i^* = \prod_{j=1}^q \rho_{ij} = I(x_{i1} \geq d_1, \ldots, x_{iq} \geq d_q)\) and \(\ell(\beta|\tilde{t}_i, \delta_i, w_i)\) as the log of the likelihood given by (3.3), the expected value of the \(j\)th score function for the parameter \(\beta\) is

\[
E \left\{ \rho^*_i \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\tilde{t}_i, \delta, w) \right\} \right\} = E \left[ E \left\{ I(x_1 \geq d_1, \ldots, x_q \geq d_q) \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\tilde{t}_i, \delta, w) \right\} \right| w \right] \\
= E \left[ I(x_1 \geq d_1, \ldots, x_q \geq d_q) E \left\{ \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\tilde{t}_i, \delta, w) \right\} \right| w \right] \\
= E \left\{ I(x_1 \geq d_1, \ldots, x_q \geq d_q) \cdot 0 \right\} \\
= 0.
\]

Since the expectation is 0 for all \(p\) score functions for \(\beta\), under the regularity conditions C.1-C.3 it can be shown with standard M-estimation theory (see Serfling, 1980, for example) that

\[
\bar{\beta}_{CC} \xrightarrow{p} \beta \quad \text{as} \quad n^* \to \infty,
\]

where \(n^* = \sum_{i=1}^n \rho_i^*\) is the number of individuals with non-zero contribution to the complete case score functions. ■

C.3 Proof Sketch of Theorem 4

Let \((\tilde{t}_i, \delta_i, z_i, x_i, \rho_i), i = 1, \ldots, n\), be independent samples with \(f_T(t_i|z_i, x_i)\) known. In order to prove the consistency of the multiple imputation estimator \(\bar{\theta}_{IMI}\), we prove that the score functions are unbiased for \(0\) at the true parameter value. Under the regularity conditions C.1-C.3, consistency would then generally be established. However, because some of the data used for solving the score equations is imputed using a consistent initial estimator, we must then show that using this imputed data does not effect the consistency of the multiple imputation estimator.
We first prove that the score functions are unbiased for \( \mathbf{0} \). To do this, note that for a random individual, suppressing the index \( i \), there are as many as \( 2^q \) potential censoring patterns for \( q \)-dimensional \( \mathbf{x} \). Define the possible subsets of observed and censored covariates as \( \mathbf{x}_i^o \) and \( \mathbf{x}_i^c \), respectively, \( l = 1, \ldots, 2^q \). Also, define \( \mathbf{d}_i^c \) as the vector of DLs associated with the covariates \( \mathbf{x}_i^c \). Let \( x_i^c|z, \mathbf{x}_i^o, \mathbf{x}_i^c < \mathbf{d}_i^c \sim f(x_i^c|z, \mathbf{x}_i^o, \mathbf{x}_i^c < \mathbf{d}_i^c; \gamma) \) be known for all possible \( \mathbf{x}_i^c \), \( l = 1, \ldots, 2^q \), and let \( \overline{x}_i^c \), the imputed vector for \( \mathbf{x}_i^c \), be drawn from this distribution. Define \( \overline{x}_i \) as the vector consisting of \( \mathbf{x}_i^o \) and \( \overline{x}_i^c \) together.

Then, the expected value of the score function using imputed values, is

\[
\sum_{l=1}^{2^q} E \left[ \frac{\partial}{\partial \beta} \left\{ \ell(\beta|\overline{t}, \delta, z, \overline{x}_l) \right\} I(x_i^c < \mathbf{d}_i^c) \right],
\]

where \( \ell(\beta|\overline{t}, \delta, z, \overline{x}_l) \) is the log of the likelihood given by (3.3). We now consider a single censoring combination \( l \), and thus drop the subscript \( l \) henceforth. Then, the expected value of the \( l^{th} \) part of the score function for \( \beta_j \) associated with covariate \( x_j \in \mathbf{x}^c \) is

\[
E \left[ \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\overline{t}, \delta, z, \overline{x}) \right\} I(\mathbf{x}^c < \mathbf{d}^c) \right]
\]

\[
= E \left[ E \left\{ \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\overline{t}, \delta, z, \overline{x}) \right\} I(\mathbf{x}^c < \mathbf{d}^c) \right\} \right]
\]

\[
= E \left[ E \left\{ I(\mathbf{x}^c < \mathbf{d}^c) \left\{ \ell(\beta|\overline{t}, \delta, z, \overline{x}) \right\} \right\} E \left\{ \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\overline{t}, \delta, z, \overline{x}) \right\} \right\} \right]
\]

\[
= E \left[ \int_{-\infty}^{\mathbf{d}^c} f(\mathbf{x}^c|\overline{t}, \delta, z, \mathbf{x}^o) d\mathbf{x}^c \left\{ \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\overline{t}, \delta, z, \mathbf{x}) \right\} \right\} \int_{-\infty}^{\mathbf{d}^c} f(|\mathbf{x}^c|\overline{t}, \delta, z, \mathbf{x}^o) d\mathbf{x}^c \right]
\]

\[
= E \left[ \int_{-\infty}^{\mathbf{d}^c} \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\overline{t}, \delta, z, \mathbf{x}) \right\} f(\mathbf{x}^c|\overline{t}, \delta, z, \mathbf{x}^o) d\mathbf{x}^c \right]
\]

\[
= \int_{\overline{t}, \delta \times \mathbf{x}^c} \left\{ \int_{-\infty}^{\mathbf{d}^c} \frac{\partial}{\partial \beta_j} \left\{ \ell(\beta|\overline{t}, \delta, z, \mathbf{x}) \right\} f(\mathbf{x}^c|\overline{t}, \delta, z, \mathbf{x}^o) d\mathbf{x}^c \right\} f(\overline{t}, \delta, z, \mathbf{x}^o) d(\overline{t}, \delta, z, \mathbf{x}^o).
\]
By regularity condition C.3 and Fubini’s Theorem, we can write this expression as

\[
\int_{\mathcal{Z}} \int_{\chi^0} \int_{-\infty}^{d^c} \left\{ \int_{\mathcal{F}, \delta} \frac{\partial}{\partial \beta_j} \{ \ell(\beta|\mathcal{I}, \delta, z, x) \} f(\mathcal{I}, \delta|z, x^0, x^c) d(\mathcal{I}, \delta) \right\} f(z, x^0, x^c) dx^c dx^0 dz = 0.
\]

This is true for any censoring combination \( l \) and is also straightforward to show for the score functions associated with the elements of \( z \) and \( x^0 \). Thus, the expectation of the score functions is \( 0 \).

Since the multiple imputations were not drawn based on the true parameter vector, \( \theta \), but rather \( \hat{\theta}^I \), a consistent initial estimate of \( \theta \), standard M-estimation theory does not directly give the consistency of \( \hat{\theta}_{IMI} \). The remainder of this proof establishes consistency under the assumption that the imputations were drawn using \( \hat{\theta}^I \) and follows similarly to the results presented in Wang and Robins (1998), Robins and Wang (2000), and Tsiatis (2006, Ch. 14).

We now assume that \( \hat{\beta}^I \) and \( \gamma^I \) in \( \hat{\theta}^I = (\hat{\beta}^{IT}, \sigma, \phi^{IT}, \hat{\gamma}^{IT})^T \) are obtained via complete case and maximum likelihood estimation, respectively, as described in Section 3.2 of the paper. Then \( \hat{\theta}^I \) is a regular and asymptotically linear estimator (RAL) for the true parameter \( \theta \) as \( n^* \to \infty \) (see Ch. 3 in Tsiatis, 2006 for an explanation of RAL estimators). For any RAL estimator, we can derive its unique influence function, as is necessary for the following Lemmas 1 and 2 below.

For the remainder of the proof, let \( \theta \) be the true parameter vector, \( \hat{\theta}_{IMI} \) be the proposed multiple imputation estimator, and \( \hat{\theta}^{(k)} \) be the maximum likelihood estimator for the \( k \)th imputed dataset. In order establish the asymptotic properties of \( \hat{\theta}_{IMI} \), we derive an approximation to its influence function of \( \hat{\theta}_{IMI} \).

**Lemma 3.** The multiple imputation estimator, \( \hat{\theta}_{IMI} \), can be approximated as

\[
n^{1/2}(\hat{\theta}_{IMI} - \theta) = n^{1/2} \sum_{i=1}^{n} (I^F(\theta))^{-1} \left[ m^{-1} \sum_{k=1}^{m} S^F(\bar{t}_i, \delta_i, z_i, \bar{x}_{ik}(\hat{\theta}^I); \theta) \right] + o_p(1),
\]

where \( I^F(\theta) = E\{S^F(\bar{t}_i, \delta_i, w_i; \theta) S^{FT}(\bar{t}_i, \delta_i, w_i; \theta) \} \) is the full-data information matrix and \( S^F(\bar{t}_i, \delta_i, z_i, \bar{x}_{ik}(\hat{\theta}^I); \theta) = \frac{\partial}{\partial \theta}[\ell(\bar{t}_i, \delta_i, z_i, \bar{x}_{ik}(\hat{\theta}^I); \theta)] \) is the full-data score vector for the \( i \)th individual (\( \ell \) is the log-likelihood of the full data), based on the \( k \)th imputed vector \( \bar{x}_{ik} \), which in

Since the score vector $S^F\{\tilde{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta'); \theta\}$ is dependent on the random quantity $\theta'$, $n^{-1/2}\sum_{i=1}^n \left[ m^{-1}\sum_{k=1}^m S^F \{\tilde{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta'); \theta\} \right]$ is not the sum of i.i.d. terms, and thus the approximation given in Lemma 3 does not give the influence function for $\hat{\theta}_{IMI}$. However, we can write $n^{-1/2}\sum_{i=1}^n \left[ m^{-1}\sum_{k=1}^m S^F \{\tilde{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta'); \theta\} \right]$ as

$$n^{-1/2}\sum_{i=1}^n \left[ m^{-1}\sum_{k=1}^m S^F \{\tilde{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta); \theta\} \right]$$

$$+ n^{-1/2}\sum_{i=1}^n \left[ m^{-1}\sum_{k=1}^m S^F \{\tilde{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta'); \theta\} - m^{-1}\sum_{k=1}^m S^F \{\tilde{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta); \theta\} \right]. \quad (C.1)$$

The first term of expression (A.1) is the sum of mean-zero i.i.d. random vectors (as shown in the first part of this proof) that converge asymptotically to a normal distribution by the central limit theorem under general regularity conditions.

**Lemma 4.** The second term in expression (A.1) above is equal to

$$n^{1/2}\sum_{i=1}^n \left\{ I^F(\theta) - I(\theta) \right\} q(\tilde{t}_i, \delta_i, z_i, x_{i}^*, \rho_i) + o_p(1),$$

where $I^F(\theta) = E\{S^F(\tilde{t}_i, \delta_i, w_i; \theta)S^{FT}(\tilde{t}_i, \delta_i, w_i; \theta)\}$ is the full-data information matrix, $I(\theta) = E\{S(\tilde{t}_i, \delta_i, z_i, x_{i}^*; \rho_i, \theta)S^T(\tilde{t}_i, \delta_i, z_i, x_{i}^*; \rho_i; \theta)\}$ is the observed-data information matrix, and $q(\tilde{t}_i, \delta_i, z_i, x_{i}^*, \rho_i)$ is the $i^{th}$ influence function of $\theta'$.


Together, Lemma 3 and Lemma 4 give the influence function for $\hat{\theta}_{IMI}$. Then, by the central limit theorem, under general regularity conditions, we have that

$$n^{-1/2}(\hat{\theta}_{IMI} - \theta) \rightarrow N(0, \Sigma),$$

142
where

$$\Sigma = \text{Var}\left\{ I^F(\theta) \right\}^{-1} \left[ m^{-1} \sum_{k=1}^{m} S^F \{ \bar{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta); \theta \} + \{ I^F(\theta) - I(\theta) \} q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \right]\right\}.$$ 

Some algebra gives

$$\text{Var}\left[ m^{-1} \sum_{k=1}^{m} S^F \{ \bar{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta); \theta \} \right] = m^{-1} \{ I^F(\theta) - I(\theta) \} + I(\theta),$$

$$\text{cov}\left[ m^{-1} \sum_{k=1}^{m} S^F \{ \bar{t}_i, \delta_i, z_i, \bar{x}_{ik}(\theta); \theta \} , \{ I^F(\theta) - I(\theta) \} q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \right] = I^F(\theta) - I(\theta),$$

and

$$\text{Var}\left[ \{ I^F(\theta) - I(\theta) \} q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \right] = \{ I^F(\theta) - I(\theta) \} \text{Var}\{ q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \} \{ I^F(\theta) - I(\theta) \}.$$

Together this gives

$$\Sigma = \{ I^F(\theta) \}^{-1} \left[ I(\theta) + m^{-1} \{ I^F(\theta) - I(\theta) \} + 2 \{ I^F(\theta) - I(\theta) \} \right. + \left. \{ I^F(\theta) - I(\theta) \} \text{Var}\{ q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \} \{ I^F(\theta) - I(\theta) \} \right] \{ I^F(\theta) \}^{-1}.$$

With a little algebraic manipulation, we get

$$\Sigma = \{ I^F(\theta) \}^{-1} + (1 + m^{-1}) \{ I^F(\theta) \}^{-1} \{ I^F(\theta) - I(\theta) \} \{ I^F(\theta) \}^{-1} \left. + \right. \{ I^F(\theta) \}^{-1} \{ I^F(\theta) - I(\theta) \} \text{Var}\{ q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \} \{ I^F(\theta) - I(\theta) \} \{ I^F(\theta) \}^{-1}.$$

For the initial estimators used in the paper, \( \text{Var}\{ q(\bar{t}_i, \delta_i, z_i, x_i^*, \rho_i) \} = \text{Var}(\hat{\theta}) \).
C.4 Proof of Theorem 5

By Theorem 4, using \( \hat{\theta}' \) and \( \text{Var}(\hat{\theta}') \) as initial estimates,

\[
\bar{\theta}_1 \to N(\theta, \Sigma_1) \quad \text{as} \quad n \to \infty,
\]

where \( \bar{\theta}_1 \) is the parameter estimate at the first iteration, \( \theta \) is the true parameter vector, and \( \Sigma_1 \) is the true variance at the first iteration given by (3.11). Then, using (3.12) to estimate the variance \( \bar{\Sigma}_1 \), we have that \( \bar{\theta}_1 \) and \( \bar{\Sigma}_1 \) are consistent estimators for \( \theta \) and \( \Sigma_1 \).

At the second iteration, using \( \bar{\theta}_1 \) and \( \bar{\Sigma}_1 \) as initial estimates, by Theorem 4 we have that

\[
\bar{\theta}_2 \to N(\theta, \Sigma_2) \quad \text{as} \quad n \to \infty,
\]

where \( \bar{\theta}_2 \) is the parameter estimate at the second iteration, \( \theta \) is the true parameter vector, and \( \Sigma_2 \) is the true variance at the second iteration given by (3.11). Then, using (3.12) to estimate the variance \( \bar{\Sigma}_2 \), we have that \( \bar{\theta}_2 \) and \( \bar{\Sigma}_2 \) are consistent estimators for \( \theta \) and \( \Sigma_2 \).

At the \( i^{th} \) iteration, using \( \bar{\theta}_{i-1} \) and \( \bar{\Sigma}_{i-1} \) as initial estimates, by Theorem 4 we have that

\[
\bar{\theta}_i \to N(\theta, \Sigma_i) \quad \text{as} \quad n \to \infty,
\]

where \( \bar{\theta}_i \) is the parameter estimate at the \( i^{th} \) iteration, \( \theta \) is the true parameter vector, and \( \Sigma_i \) is the true variance at the \( i^{th} \) iteration given by (3.11). Then, using (3.12) to estimate the variance \( \bar{\Sigma}_i \), we have that \( \bar{\theta}_i \) and \( \bar{\Sigma}_i \) are consistent estimators for \( \theta \) and \( \Sigma_i \).

For the final iteration \( s \), using \( \bar{\theta}_{s-1} \) and \( \bar{\Sigma}_{s-1} \) as initial estimates, by Theorem 4 we have that

\[
\bar{\theta}_s \to N(\theta, \Sigma_s) \quad \text{as} \quad n \to \infty,
\]

where \( \bar{\theta}_s \) is the parameter estimate at the \( s^{th} \) iteration, \( \theta \) is the true parameter vector, and \( \Sigma_s \) is the true variance at the \( s^{th} \) iteration given by (3.11). Then, using (3.12) to estimate the
variance $\hat{\Sigma}_a$, we have that $\hat{\theta}_s$ and $\hat{\Sigma}_a$ are consistent estimators for $\theta$ and $\Sigma_a$. □
Appendix D

Technical Details of the Approximate EM Algorithm Procedure

We present here the necessary derivations for finding the updates of the parameters when using the approximate EM algorithm proposed in Sections 4.2.3 and 4.3.2. Specifically, we derive the updates for $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\tau}^2$ for the scenario with censoring on some of the longitudinal covariates. When there is no censoring, the updates formulas for $\hat{\gamma}$ and $\hat{\tau}$ derived below may still be used with $\rho_{ijk} = 1$, $i = 1, \ldots, n$, $j = 1, \ldots, n_{ik}$, $k = 1, \ldots, q$.

To derive the updates, we must take derivatives of the $Q$-function (4.4) with respect to the corresponding parameters. We rewrite the $Q$-function here in the context of covariates subject to DLs:

\[
Q(\theta|\theta^{(v)}) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left( \log f(y_i|z_i, b_i; \beta) + \log f(x_{i1}^*|\rho_{i1}, z_i, b_{i1}; \gamma_1, \tau_1^2) + \cdots + \log f(x_{iq}^*|\rho_{iq}, z_i, b_{iq}; \gamma_q, \tau_q^2) \\
+ \log f(b_i; B, D) \right) f(b_i|y_i, x_{i1}, \cdots, x_{iq}^*, \rho_{i1}, \cdots, \rho_{iq}, z_i; \theta^{(v)}) \, db_i.
\]  

(D.1)
For the subsequent derivations, we make the following normal approximation:

\[
f(b_i|y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \theta^{(v)}) \approx N(\tilde{b}_i, \tilde{\Sigma}_i),
\]

(D.2)

where \(\tilde{b}_i\) is the mode of \(f(b_i|y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \theta^{(v)})\) found by maximizing over \(b_i\) given the current set of parameter estimates \(\theta^{(v)}\) and the observed data for the \(i^{th}\) individual, and \(\tilde{\Sigma}_i\) is the inverse of the negative second derivative of \(\log\{f(b_i|y_i, x_{i1}^*, \ldots, x_{iq}^*, \rho_{i1}, \ldots, \rho_{iq}, z_i; \theta^{(v)})\}\), where the derivative is taken with respect to \(b_i\) and evaluated at \(\tilde{b}_i\).

We first derive the updating formula for \(\beta\), which only appears in (D.1) through \(f(y_i|z_i, b_i; \beta)\). To simplify the notation, note that we include the intercept \(\beta_0\) in the parameter vector \(\beta_1\). Then, we have that

\[
f(y_i|z_i, b_i; \beta) = \left(\frac{1}{1 + e^{-\beta_1^T z_i - \beta_2^T b_i}}\right)^{y_i} \left(1 - \frac{1}{1 + e^{-\beta_1^T z_i - \beta_2^T b_i}}\right)^{1-y_i}
\]

\[
= \left(e^{\beta_1^T z_i + \beta_2^T b_i}\right)^{y_i} \left(1 + e^{\beta_1^T z_i + \beta_2^T b_i}\right)^{-y_i} \left(1 + e^{\beta_1^T z_i + \beta_2^T b_i}\right)^{-1} \left(1 + e^{\beta_1^T z_i + \beta_2^T b_i}\right)^{y_i}
\]

\[
= \left(e^{\beta_1^T z_i + \beta_2^T b_i}\right)^{y_i} \left(1 + e^{\beta_1^T z_i + \beta_2^T b_i}\right)^{-1},
\]

so that

\[
E\{\log f(y_i|z_i, b_i; \beta)\} = E\left\{y_i \log \left(e^{\beta_1^T z_i + \beta_2^T b_i}\right) - \log \left(1 + e^{\beta_1^T z_i + \beta_2^T b_i}\right)\right\}
\]

\[
= y_i \{\beta_1^T z_i + \beta_2^T E(b_i)\} - E\left\{\log \left(1 + e^{\beta_1^T z_i + \beta_2^T b_i}\right)\right\}.\quad (D.3)
\]

Using (D.2), we have that

\[
\beta_1^T z_i + \beta_2^T b_i \overset{\text{a}}{\sim} N(\beta_1^T z_i + \beta_2^T \tilde{b}_i, \beta_2^T \tilde{\Sigma}_i \beta_2) = N(\mu_i, \sigma_i^2),
\]

where \(\overset{\text{a}}{\sim}\) indicates “is approximately distributed as.” Then, we can apply the Gaussian-Hermite
quadrature methods described in Appendix E to approximate (D.3) as

\[
\widetilde{E}(\ell_\beta) = y_i \{ \beta_1^T z_i + \beta_2^T \tilde{b}_1 \} + \sum_{l=1}^r \frac{u_l}{\sqrt{2\pi}} e^{-\frac{u_l^2}{2}} \log \left( 1 + e^{\mu_l + \sigma_l v_l} \right),
\]

where \( \widetilde{E}(\ell_\beta) \) is a shorthand notation for the approximate expected value of the part of the log-likelihood involving \( \beta \) and the quadrature points, \( v_l \), and weights, \( u_l, l = 1, \ldots, r \), can be found using the program provided in Appendix E. For convenience, we henceforth denote

\[
u_l^* = \frac{u_l}{\sqrt{2\pi}} e^{-\frac{u_l^2}{2}}.
\]

Taking the partial derivatives of \( \widetilde{E}(\ell_\beta) \) with respect to \( \beta_1 \) and \( \beta_2 \), we find that

\[
Q'_{\beta_1} \approx \frac{\partial \widetilde{E}(\ell_\beta)}{\partial \beta_1} = \sum_{i=1}^n \left( y_i z_i - \sum_{l=1}^r u_l^* \frac{e^{\mu_l + \sigma_l v_l}}{1 + e^{\mu_l + \sigma_l v_l}} z_i \right), \quad \text{(D.4)}
\]

and

\[
Q'_{\beta_2} \approx \frac{\partial \widetilde{E}(\ell_\beta)}{\partial \beta_2} = \sum_{i=1}^n \left( y_i \tilde{b}_i - \sum_{l=1}^r u_l^* \frac{e^{\mu_l + \sigma_l v_l}}{1 + e^{\mu_l + \sigma_l v_l}} \left( \tilde{b}_i + \sigma_i^{-1/2} \tilde{\Sigma}_i \beta_2 \right) \right), \quad \text{(D.5)}
\]

where \( Q'_{\beta_1} \) and \( Q'_{\beta_2} \) are the partial derivatives of (D.1) with respect to \( \beta_1 \) and \( \beta_2 \). The negative second partial derivatives of (D.1) with respect to \( \beta_1 \) and \( \beta_2 \) are given by

\[
-\frac{\partial Q'_{\beta_1}}{\partial \beta_1} \approx -\frac{\partial^2 \widetilde{E}(\ell_\beta)}{\partial \beta_1 \partial \beta_1^T} = \sum_{i=1}^n \sum_{l=1}^r \frac{u_l^*}{1 + e^{\mu_l + \sigma_l v_l}} \left( 1 - \frac{e^{\mu_l + \sigma_l v_l}}{1 + e^{\mu_l + \sigma_l v_l}} \right) z_i z_i^T, \quad \text{(D.6)}
\]

\[
-\frac{\partial Q'_{\beta_2}}{\partial \beta_2} \approx -\frac{\partial^2 \widetilde{E}(\ell_\beta)}{\partial \beta_2 \partial \beta_2^T} = \sum_{i=1}^n \sum_{l=1}^r \frac{u_l^*}{1 + e^{\mu_l + \sigma_l v_l}} \left\{ \sigma_i^{-1/2} \tilde{\Sigma}_i - \sigma_i^{-3/2} \tilde{\Sigma}_i \beta_2 \tilde{\Sigma}_i \beta_2^T \right\} v_l + \left( \tilde{b}_i + \sigma_i^{-1/2} \tilde{\Sigma}_i \beta_2 \right) \left( \tilde{b}_i + \sigma_i^{-1/2} \tilde{\Sigma}_i \beta_2 \right)^T
\]

\[-\frac{\partial Q'_{\beta_2}}{\partial \beta_2} \approx \frac{e^{\mu_l + \sigma_l v_l}}{1 + e^{\mu_l + \sigma_l v_l}} \left( \tilde{b}_i + \sigma_i^{-1/2} \tilde{\Sigma}_i \beta_2 \right) \left( \tilde{b}_i + \sigma_i^{-1/2} \tilde{\Sigma}_i \beta_2 \right)^T, \quad \text{(D.7)}
\]

and

\[
-\frac{\partial Q'_{\beta_1}}{\partial \beta_2} \approx -\frac{\partial^2 \widetilde{E}(\ell_\beta)}{\partial \beta_1 \partial \beta_2^T} = \sum_{i=1}^n \sum_{l=1}^r \frac{u_l^*}{1 + e^{\mu_l + \sigma_l v_l}} \left( 1 - \frac{e^{\mu_l + \sigma_l v_l}}{1 + e^{\mu_l + \sigma_l v_l}} \right) z_i \left( \tilde{b}_i + \sigma_i^{-1/2} \tilde{\Sigma}_i \beta_2 \right)^T \quad \text{(D.8)}
\]
Then, noting that
\[ -Q''_{\beta} = - \left( \begin{array}{l}
\frac{\partial Q'_{\beta_1}}{\partial \beta_1} \\
\frac{\partial Q'_{\beta_1}}{\partial \beta_2} \\
T \\
\frac{\partial Q'_{\beta_2}}{\partial \beta_1} \\
\frac{\partial Q'_{\beta_2}}{\partial \beta_2}
\end{array} \right), \]
equation (4.6), together with equations (D.4)-(D.8), can be used to update \( \beta \) at each iteration of the approximate EM algorithm.

We now derive the updates for \( \gamma_k \) and \( \tau_k^2 \), which only appear in (D.1) through
\[ f(x_{ik}^* | \rho_{ik}, z_i, b_{ik}; \gamma_k, \tau_k^2) \]. Then, we have that
\[
\log f(x_{ik}^* | \rho_{ik}, z_i, b_{ik}; \gamma_k, \tau_k^2) = \sum_{i=1}^{n} \sum_{j=1}^{n^{ij}} \rho_{ijk} \left\{ -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \tau_k^2 - \frac{(x_{ijk} - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik})^2}{2\tau_k} \right\} \\
+ (1 - \rho_{ijk}) \left\{ \log \Phi \left( \frac{d_k - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik}}{\tau_k} \right) \right\},
\]
so that
\[
E \{ \log f(x_{ik}^* | \rho_{ik}, z_i, b_{ik}; \gamma_k, \tau_k^2) \} = \sum_{i=1}^{n} \sum_{j=1}^{n^{ij}} \rho_{ijk} \left\{ -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \tau_k^2 - \frac{E(x_{ijk} - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik})^2}{2\tau_k} \right\} \\
+ (1 - \rho_{ijk}) E \left\{ \log \Phi \left( \frac{d_k - r_{ijk}^T \gamma_k - s_{ijk}^T b_{ik}}{\tau_k} \right) \right\}.
\]
Then using the normality assumption (D.2), we have that
\[
r_{ijk}^T \gamma_k + s_{ijk}^T b_{ik} \overset{\text{a}}{\sim} N(r_{ijk}^T \gamma + s_{ijk}^T \bar{b}_{ik}, s_{ijk}^T \Sigma_{ik} s_{ijk}^T) = N(\mu_{ijk}, \sigma^2_{ijk}),
\]
where \( \Sigma_{ik} \) is the submatrix of \( \Sigma_i \) associated with the random effects vector \( b_{ik} \), and thus
\[
\bar{E}\{ \ell_{\gamma_k, \tau_k^2} \} = \sum_{i=1}^{n} \sum_{j=1}^{n^{ij}} \rho_{ijk} \left\{ -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \tau_k^2 - \frac{\sum_{l=1}^{r} u_{il}^T (x_{ijk} - \mu_{ijk} - \sigma_{ijk} v_l)^2}{2\tau_k} \right\} \\
+ (1 - \rho_{ijk}) \sum_{l=1}^{r} u_{il}^T \log \Phi \left( \frac{d_k - \mu_{ijk} - \sigma_{ijk} v_l}{\tau_k} \right),
\]
where \( \bar{E}\{ \ell_{\gamma_k, \tau_k^2} \} \) is the approximate expected value of the log-likelihood involving \( \gamma_k \) and \( \tau_k^2 \).
Taking the partial derivatives of $\bar{E}(\ell_{\gamma_k, \tau_k^2})$ with respect to $\gamma_k$ and $\tau_k^2$, we find

$$Q'_{\gamma_k} \approx \frac{\partial \bar{E}(\ell_{\gamma_k, \tau_k^2})}{\partial \gamma_k} = \sum_{i=1}^{n} \sum_{j=1}^{n_k} \left\{ \frac{\rho_{ijk}}{2\tau_k^2} (x_{ijk} - \mu_{ijk}) r_{ijk} \right\} + (1 - \rho_{ijk}) \sum_{l=1}^{r} u_l^* \phi \left( \frac{d_k - \mu_{ijk} - \sigma_{ijk} v_l}{\tau_k} \right) \Phi \left( \frac{d_k - \mu_{ijk} - \sigma_{ijk} v_l}{\tau_k} \right)^{-1} r_{ijk}$$

and

$$Q'_{\tau_k^2} \approx \frac{\partial \bar{E}(\ell_{\gamma_k, \tau_k^2})}{\partial \tau_k^2} = \sum_{i=1}^{n} \sum_{j=1}^{n_k} \left\{ \frac{\rho_{ijk}}{2\tau_k^2} \left[ \frac{\rho_{ijk}}{2\tau_k^2} \left( x_{ijk} - \mu_{ijk} - \sigma_{ijk} v_l \right)^2 + (1 - \rho_{ijk}) \sum_{l=1}^{r} u_l^* \phi \left( \frac{d_k - \mu_{ijk} - \sigma_{ijk} v_l}{\tau_k} \right) \right] \right\} \times \Phi \left( \frac{d_k - \mu_{ijk} - \sigma_{ijk} v_l}{\tau_k} \right)^{-1} \frac{d_k - \mu_{ijk} - \sigma_{ijk} v_l}{2\tau_k^3}.$$
with (D.9)-(D.13), can be used to update
simply the block diagonal of the matrices $Q^T \tau^2$, $k = 1, \ldots, q$.

Then, we have that

$$-Q''_{\gamma_k, \tau^2} = - \begin{pmatrix} \frac{\partial Q'_{\gamma_k}}{\partial \gamma_k} & \frac{\partial Q'_{\tau_k}}{\partial \tau_k} \\ \frac{\partial Q'_{\tau_k}}{\partial \gamma_k} & \frac{\partial Q'_{\tau_k}}{\partial \tau_k} \end{pmatrix}^T, \quad k = 1, \ldots, q.$$  

Since we assumed conditional independence of $X_{ik}, k = 1, \ldots, q, Q''_{\gamma_k, \tau^2}$ in equation (4.14) is simply the block diagonal of the matrices $Q''_{\gamma_k, \tau^2}, k = 1, \ldots, q$. Then, equation (4.14), together with (D.9)-(D.13), can be used to update $(\gamma^T, \tau^2)^T$ at each iteration of the approximate EM algorithm.

151
Appendix E

An Overview of Adaptive Gaussian-Hermite Quadrature Methods

In many statistical applications, we must compute integrals which have no closed forms. The most basic way to approximate any integral of the form

\[ \int_{a}^{b} f(x) dx \]

is to apply a simple algorithm such as the trapezoidal or Simpson’s rule. However, these rules generally require many evaluations of \( f(x) \) across the range of \( x \) and thus can be very computationally challenging when a high degree of accuracy is desired. Gaussian quadrature greatly reduces the number of evaluation points by carefully weighting \( f(x) \) at particular points over the range of \( x \). Specifically, with Gaussian quadrature, we make the following approximation:

\[ \int_{a}^{b} f(x) dx \approx \frac{b-a}{2} \sum_{i=1}^{n} w_{i} f \left( \frac{b-a}{2} x_{i} + \frac{a+b}{2} \right), \]
where the weights $w_i$ and evaluation points $x_i$, $i = 1, \cdots, n$, are chosen optimally (see, for example, Golub and Welsch, 1969).

Statisticians often deal specifically with integrals of the form

$$\int_{-\infty}^{\infty} f(x) e^{-x^2} dx,$$

(E.1)

for which Gaussian-Hermite quadrature is superior. The main idea of Gaussian-Hermite quadrature remains the same. Specifically we wish to use the approximation

$$\int_{-\infty}^{\infty} f(x) e^{-x^2} dx \approx \sum_{i=1}^{n} w_i f(x_i),$$

where we only need to evaluate the $f(x)$ part of the integrand at specific points $x_i$ weighted by $w_i$. Because this approximation is most accurate when the location of the quadrature points is centered at the main mass of the integrand and an appropriate scaling is used, adaptive Gaussian-Hermite quadrature methods are more useful than simple Gaussian-Hermite quadrature. Adaptive Gaussian-Hermite quadrature methods allow for very accurate computation of integrals of the form (E.1) with very few quadrature points.

Most simply, adaptive Gaussian-Hermite quadrature applies as follows. If we want to find an integral of the form (E.1) where $f(x) = p(x)h(x)$, with $p(x)$ behaving like a polynomial and $h(x)$ behaving like $\exp \left\{ -\frac{(x - \mu)}{\sigma^2} \right\}$, we may make the following approximation:

$$\int_{-\infty}^{\infty} f(x) dx \approx \sigma \sum_{i=1}^{n} w_i f(\mu + \sigma x_i),$$

(E.2)

where $w_i$ and $x_i$, $i = 1, \cdots, n$, are the weight and points of evaluation, $\mu$ is the mode of $h(x)$, and

$$\sigma^2 = \left[ -\frac{d^2 \log \{h(x)\}}{dx^2} \right]_{x=\mu}^{-1}.$$
Similarly, if we need to evaluate a $p$-dimensional integral of the form

$$\int_{\mathbb{R}^p} f(x) dx,$$

with $f(x) = p(x)h(x)$, we can make the following approximation:

$$\int_{\mathbb{R}^p} f(x) dx \approx |V|^{1/2} \sum_{i=1}^n \sum_{j=1}^n w_i w_j f\{\mu + V^{1/2} (x_i, \ldots, x_j)^T\},$$  \hspace{1cm} (E.3)

where there are $p$ sums in (E.3), $\mu$ is the $p$-dimensional mode of $h(x)$, and

$$V = -\left. \frac{d^2 \log \{h(x)\}}{dx dx^T} \right|_{x=\mu}^{-1},$$

where $V^{1/2}$ is the Cholesky factor of $V$. We refer to Hildebrand (1956) for more details on Gaussian quadrature methods and to Rizopoulos (2012b) for details on adaptive Gaussian-Hermite quadrature as it applies to joint models with a survival submodel.

To calculate the quadrature points $w_i$ and $x_i$, $i = 1, \ldots, n$, needed for the evaluation of (E.2) and (E.3), and thus the expectations derived in Appendix D, we provide the following function in R (translated from a SAS program written by Daowen Zhang):

```r
GaussianHermite <- function(n){
  esp <- 3.0e-14
  pim4 <- 1/sqrt(sqrt(pi))
  maxit <- 1000

  x <- w <- rep(0,n)
  m <- (n+1 - ((n+1)%%2))/2
  p <- 2*n
  q <- 2*n+1

  for(i in 1:m){
    if(i==1) z <- sqrt(q) - 1.85575*q^(-.1667)
    if(i==2) z <- z - 1.14^((.425)/z)
    if(i==3) z <- 1.86*z - 0.86*x[1]
  }
}
```

154
if(i==4) z <- 1.91*z - 0.91*x[2]
if(i!=1 & i!=2 & i!=3 & i!=4) z <- 2*z - x[i-2]

its <- 1
dif <- 1
while(its<maxit & dif>esp){
  p1 <- pim4
  p2 <- 0
  for(j in 1:n){
    p3 <- p2
    p2 <- p1
    p1 <- z*sqrt(2/j)*p2 - sqrt((j-1)/j)*p3
  }
  pp <- sqrt(p)*p2
  z <- z - p1/pp
  dif <- abs(p1/pp)
  its <- its +1
}

x[i] <- z
x[n+1-i] <- -z
w[i] <- 2/(pp^2)
w[n+1-i] <- w[i]
}
w <- sqrt(2)*w*exp(x^2)
x <- sqrt(2)*x
return(list(round(w,10),round(x,10)))

#Set the number of quadrature points & call function
n <- 9
GaussianHermite(n)