9-24-2019

Joint Analysis of Self-Reported and Biological Measurements

Di Zheng
University of Connecticut - Storrs, di.zheng@uconn.edu

Joint Analysis of Self-Reported and Biological Measurements

Di Zheng, Ph.D.
University of Connecticut, 2019

ABSTRACT

Joint analysis of self-report and biomarker measurements provides new opportunities to understand and characterize human behaviors. Self-report measures are the most common way to assess human behavior, because they are quick, straightforward, and inexpensive. But they are easily limited by factors such as recall bias toward under-report. Thus a wide variety of biological measurements have been developed to objectively assess human behaviors. However, the accuracy of biological measurement can also vary between studies, not just through chance, but also with changes in the study setting, the spectrum of disease, and definition of the target condition. Henceforth, self-report measures and biological marker together are likely to provide the basis for a more accurate estimate of participants' behavior than either does alone. This is the reason why simultaneous analysis of self-report measures and biomarker is appealing. There are two major research issues with such joint analysis. First, when researchers intend to combine biological marker and self-report measures as explanatory variables in the longitudinal analysis, the problem of multicollinearity arises. Second, in longitudinal

studies, variables which are recorded over the course of study are easily subject to missing observations.

The data motivating our research arise from a longitudinal cohort study of an HIV clinic in southwestern Uganda who were not yet eligible for antiretroviral therapy (ART). Beginning in 2011, 447 patients were recruited with follow-up visits every 6 months for up to 3 years. The objective of the study is to examine the relationship between alcohol use and HIV disease progression measured by CD4 cell count among ART naive HIV-infected Ugandans. Self-report measures on the Alcohol Use Disorders Identification Test-Consumption (AUDIT-C), and biological markers-phosphatidylethanol (PEth), are both used to measure alcohol use.

To address the correlation between AUDIT-C score and PEth, we propose Bayesian shrinkage prior in the setting of linear mixed model. In light of missing observations in response and time-dependent covariates, we propose a two-stage multiple imputation for the missing response and missing time-varying covariates in longitudinal data. Last, we extend the two-stage multiple imputation approach by introducing Bayesian shrinkage prior into the imputation process to account for partly-observed response and partly-observed correlated time-dependent covariates simultaneously.

We carry out a detailed analysis of the data using the proposed approaches. Simulation studies are conducted to compare the proposed approaches to existing approaches.

Joint Analysis of Self-Reported and Biological Measurements

Di Zheng

B.A., Wuhan University, Wuhan, China, 2012

M.A., the State University of New Jersey, NJ, USA, 2013

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2019

APPROVAL PAGE

Doctor of Philosophy Dissertation

Joint Analysis of Self-Reported and Biological Measurements

Presented by

Di Zheng, B.A., M.A.

Major Advisor _____

Ofer Harel

Associate Advisor _____

Haiying Wang

Associate Advisor _____

Victor Hugo Lachos Davila

University of Connecticut

2019

Acknowledgements

Over the past four year, I would first like to express my deep gratitude to Professor Ofer Harel, my supervisor, for his consistent support, patient guidance and constructive critiques of my research work. He infuses my research with curiosity and enthusiasm and always instructs me to dig into the problem to explore the underlying science. His guidance has made this a thoughtful and rewarding journey.

I am very appreciative of the useful comments from my advisory dissertation committee, Professor Victor Hugo Lachos and Professor Haiying Wang. Their knowledge and insights are helpful and inspiring to my research.

I would like to thank all other professors in the department for delivering the valuable science of statistics courses from various aspects. My special thanks are extended to the administrative staff in the our department, Tracy Burke, Anthony Luis and Megan Pesta for providing assistance and answering questions.

I would also like to thank Dr. Judy Han and Dr. Winnie Muyindike for their plentiful support and extraordinary collaboration. This project would have been impossible without the support of the NIH NIAAA Grant #U01AA020776 funding.

My grateful thanks are also extended to all my colleagues in the department for their wonderful support and collaboration.

Last but not least, I would like to thank my parents for their support and encouragement throughout my study. My heartfelt appreciation goes to my husband for his undeviating support. They give me the strength of my heart.

Contents

Acknowledgements **iii**

1 Introduction **1**

1.1 Literature Review 1

1.2 The Motivating Data 6

1.3 Dissertation Narratives 13

2 Application of Bayesian Shrinkage Priors in Linear Mixed Effect Model 17

2.1 Introduction 17

2.2 Methodology 19

2.2.1 Multicollinearity in linear regression 19

2.2.2 Bayesian Gaussian Shrinkage Priors 20

2.2.3 Linear mixed model 22

2.3 Bayesian Shrinkage Mixed Estimator 22

2.3.1 Normal-gamma prior for fixed effect 23

2.3.2 Shrinkage for multicollinearity 25

2.4 Simulations 26

2.4.1 Simulation Results 28

2.5 Data Application 33

2.6 Discussion 37

3 Application of Two-stage Multiple Imputation for Missing Response and Missing Time-varying Covariates 40

3.1 Introduction 40

3.2 Methodology 42

3.2.1 Missing Data Mechanism 42

3.2.2 Two-stage Multiple Imputation 44

3.3 Multiple Imputation for Missing Covariates in Linear Mixed Model . . . 46

3.3.1 Multivariate Multiple Imputation 48

3.3.2 Two-stage Multiple Imputation 49

3.4 Simulation Studies 56

3.4.1 Simulation Results 58

3.5 Data Application 64

3.6 Discussions 67

4 Application of Two-stage Multiple Imputation for Missing Correlated Time-varying Covariates 69

4.1 Introduction 69

4.2 Two-Stage Multiple Imputation for Incomplete Correlated Time-varying Covariates 71

4.2.1 Specification of Imputation Models 71

4.3 Simulation studies 77

4.3.1 Simulation Results 78

4.4 Data Example 82

4.5 Discussion 83

5 Conclusion and Future Work **85**

5.1 Overview 85

5.2 Extension to Non-continuous Data 86

5.3 Extension to Non-ignorable Missing Data 86

5.3.1 Extension of Two-Stage Multiple Imputation to Partial Ignorability 88

Bibliography **96**

List of Tables

1 HIV-infected persons in southwestern Uganda: participants characteristics at baseline (N=447) 14

2 Results from simulation study. Percentage bias (Bias %), MSE, 95% credible interval coverage (Cov), and 95% credible interval width (CI width) from different values of expected value of λ, a_λ 31

3 Results from simulation study with expected value of λ, a_λ being 0.1. Percentage bias (Bias %), MSE, 95% credible interval coverage (Cov), and 95% credible interval width (CI Width) from the proposed method BSME, along with traditional approaches MLE and RR. Results are based on 1000 replications each 32

4 HIV-infected persons in southwestern Uganda: participants characteristics at baseline (N=145) 34

5 Application to the HIV data: Estimated coefficients obtained by fitting linear mixed model, using maximum likelihood estimation and Bayesian shrinkage mixed estimation to incorporate related AUDIT-C and PEth . 36

6 Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval width (CI Width) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with ($\sigma = 1, \psi = 1, M = 5, N = 2$) 62

7 Simulation Results - comparisons of standardized bias (%) and RMSE of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with ($\sigma = 1, \psi = 1, M = 50, N = 2$) 63

8 Application to the HIV study: estimated coefficients by fitting linear mixed model, using various approach to handle missing response and missing time-varying covariates 66

9 Simulation Results - comparison standardized bias (%), root mean squared error (RMSE), and 95% confidence interval coverage rate (cov) of estimated coefficients between multivariate multiple imputation (MULT) and two-stage multiple imputation (TSMI) 81

10 Application to the HIV data: Estimated coefficients obtained by fitting linear mixed model, using multivariate multiple imputation (MULT) and two-stage multiple imputation (TSMI) for missing values 83

11 Results from simulation study with different expected value of λ, a_λ for moderate correlation level 0.6. Percentage bias (Bias %), MSE, 95% confidence interval coverage (Cov), and 95% confidence interval length (CI Length) from the proposed method BSME along with traditional method MS. Results are based on 1000 replications each 90

12 Results from simulation study with different expected value of λ, a_λ for moderate correlation level 0.8. Percentage bias (Bias %), MSE, 95% confidence interval coverage (Cov), and 95% confidence interval length (CI Length) from the proposed method BSME along with traditional method MS. Results are based on 1000 replications each 91

13 Results from simulation study with different expected value of λ, a_λ for moderate correlation level 0.99. Percentage bias (Bias %), MSE, 95% confidence interval coverage (Cov), and 95% confidence interval length (CI Length) from the proposed method BSME along with traditional method MS. Results are based on 1000 replications each 92

14 Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval length (CI Length) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with sample size 100, $\sigma = 1, \psi = 1, M = 5, N = 2$ 93

15 Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval length (CI Length) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with $\sigma = 2, \psi = 1, M = 5, N = 2$ 94

16 Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval length (CI Length) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with $\sigma = 1, \psi = 4, M = 5, N = 2$ 95

List of Figures

1 Aggregation plot of CD4 cell count in log scale, AUDIT-C and PEth over the course the study 15

2 Average CD4 cell count in log scale for available cases 35

Chapter 1

Introduction

1.1 Literature Review

Longitudinal research [Diggle et al., 2002] has been experiencing increasingly popularity in medical science, psychiatry, biology and social sciences. The analysis of longitudinal data requires a particular statistical technique such as mixed effect models [Diggle et al., 2002, Verbeke and Molenberghs, 2009, , Fitzmaurice et al., 2012] or generalized estimating equations [Zeger et al., 1988, Hardin and Hilbe, 2012].

When some covariates are highly-correlated, the problem of multicollinearity arises. In the setting of multiple linear regression, the multicollinearity would lead to unreliable inference of the parameter. The multicollinearity in linear mixed model is a more complex issue. Multicollinearity can occur in the covariates within groups, the covariates between groups, and cross-level interactions. Multicollinearity at different levels have different impact on estimation [Yu et al., 2015]. Kreft and De Leeuw [1998] showed that multicollinearity can make the interpretation of model coefficients difficult, especially when dealing with cross-level interactions. Slight changes in the model led to very different results, with coefficients for correlated variables changing over models and

standard errors changing even more. Therefore, multicollinearity in linear mixed model must be proceed with caution.

Remedies to multicollinearity have been discussed extensively in the context of linear regression, such as ridge regression and principal component regression. Ridge regression [Hoerl and Kennard, 1970a,b] is a classical shrinkage approach to overcome the multicollinearity, by modifying the ordinary least squares approach allowing bias on the regression estimates. Although the estimators are biased, the biases are small enough for these estimators to be substantially more precise than unbiased estimators. The development of principal component regression is done by Massy [1965] to handle the problem of multicollinearity by eliminating model instability and reducing the variances of the regression coefficients. In principal component regression, the latent variables are chosen among the principal components of design matrix of covariates. Thus, principal component regression is a discrete shrinkage approach as the parameter of interest is the number of latent variables introduced in the regression, whereas the ridge regression is a continuous shrinkage method as it depends on a penalty parameter.

Despite the extensive discussion of the collinearity in linear regression, less has been said about how to deal with multicollinearity in linear mixed model. Several authors developed ridge regression in the context of linear mixed model [Eliot et al., 2011, Liu and Hu, 2013, Ozkale and Can, 2017]. Eliot et al. [2011] first integrated the ridge regression into the framework of linear mixed model, deriving the parameter estimates via the expectation-maximization (EM) algorithm. Liu and Hu [2013] discuss the conditions for

superiority of ridge estimator over maximum likelihood estimator (MLE) in linear mixed model. Ozkale and Can [2017] also showed the out-performance of ridge estimator over MLE under the mean squared error (MSE) criterion. Besides its advantage, the ridge regression always contain a data-driven determination of the shrinkage parameter, for example based on cross-validation [Hoerl and Kennard, 1970a,b]. In the context of linear mixed model which parameter estimations requires iterative optimization, including additional shrinkage parameters will further complicate the algorithm and increase the computational burden.

In the Bayesian framework, the classical ridge estimator coincides with a version of posterior Bayes regression estimate [Lindley and Smith, 1972]. Assume the prior distribution of linear regression coefficients β as $\beta|\tau \sim N(0, \tau^2 I)$, the posterior mean of β would be same with the ridge estimator. Such a normal prior distribution on coefficient is a special form of Gaussian regularization priors for shrinking coefficient vectors. Generally, by specifying appropriate informative priors on regression coefficient $p(\beta|\tau)$, inference on β would be regularized, where τ represent shrinkage parameters.

Compared with traditional penalty estimator, such as ridge estimator, Bayesian formulation has several advantages. First, to derive the penalty estimator, one must decide the value of penalty parameter τ via some methods such as cross validation. Bayesian framework allows to circumvent such external penalty parameter selection mechanisms. Instead, hyperpriors can be assigned on parameters included in τ. This enables joint estimation of both penalty parameters and regression coefficients. Second, inference

for β would be based on marginal posterior $p(\beta|y)$ by integrating out τ from the joint distribution. But the classical penalty estimator rely on plug-in estimates $\hat{\beta} = \hat{\beta}(\hat{\tau})$. Third, the classical penalty estimator in the linear mixed effect model framework may report computational difficulties in attempting to minimize likelihood equation. Instead, conditionally Gaussian priors assigned on coefficients lead to Gibbs sampler, the posterior updating will have efficiency gain, therefore facilitating full Bayesian Markov chain Monte Carlo (MCMC) inference.

In addition to correlated covariates, incomplete data is another issue which deserves attention in longitudinal studies. Incomplete data have always been a problem as it may induce biases into the data analysis. Longitudinal research is likely to suffer higher rates of missing data, because the same people are followed over time, making dropout or losing follow-up more likely. Thus, statistical models of incomplete longitudinal data have been studied extensively in literature [Little and Schenker, 1995, Diggle et al., 2002, Daniels and Hogan, 2008, Fitzmaurice et al., 2008, Verbeke and Molenberghs, 2009, Fitzmaurice et al., 2012], which usually deal with incomplete response by assuming fully-observed covariates. However, missing time-varying covariates is a common phenomenon in longitudinal studies. For example, when a subject drop out from a longitudinal study prior to the completion of the study, time-varying covariates, along with the outcome variable, would not be observed past dropout. Hence, the assumption of completely observed covariates is often not realistic in the presence of time-varying covariates. Models that account for both missing response and (time-varying) covariates

are therefore necessary.

Some statistical methods for missing covariates in longitudinal data models have been developed by assuming the fully-observed outcome. A joint-modeling strategy that accounts for both intermittently missing and left-censored time-varying covariates in the Bayesian framework is given by Chen et al. [2014]. Wu and Wu [2001] developed a Gibbs sampler for estimating parameters in non-linear mixed effect models with missing time-independent covariates, and extended the algorithm to accounting missing time-dependent covariates [Wu and Wu, 2002]. The models for non-ignorable missing time-dependent covariates is later given by Wu [2007].

To accommodate missing response and missing covariates simultaneously in longitudinal data, Stubbendick and Ibrahim [2003] developed a maximum likelihood method for non-ignorable missing longitudinal responses and baseline covariates, who later presented the method for non-ignorable missing binary response and covariates [Stubbendick and Ibrahim, 2006]. Roy and Lin [2002] proposed the model for non-ignorable dropouts and dropout-related missing covariates, and extended it to the generalized linear mixed model [Roy and Lin, 2005]. A pseudo-likelihood approach is given by Parzen et al. for modeling non-ignorable missing binary outcomes and time-varying covariates over time [Parzen et al., 2006]. Schafer [1997b] proposed multivariate mixed effect model to treat all missing response and missing time-varying covariates as multiple outcomes. The detailed algorithms for implementation is described by Schafer and Yucel [2002].

Multiple imputation (MI) is a popular method for analyzing incomplete datasets

[Rubin, 1996, Schafer, 1997a, Rubin, 2004], which generates a number of "completed" datasets by filling in missing values from an appropriate predictive distribution, conditional on the observed data. The substantive model of interest is then fitted to each of imputed datasets, and the results are combined for final inference using Rubin's rule [Rubin, 2004].

Two-stage multiple imputation [Shen, 2000, Harel, 2009, Reiter and Raghunathan, 2007], an extension of multiple imputation, is an appealing alternative to tackle missing response and missing covariates. It is often the case that missing data are of two distinct types, such as unplanned and planned nonresponse, dropout and intermittent missing data in longitudinal data, or missing response and missing covariates. Two-stage multiple imputation is advantageous when imputation of one type of data would be substantially easier if the other type were known, and/or when different imputation models are desired for two types of missing data.

1.2 The Motivating Data

Uganda faces a dual burden of HIV and unhealthy alcohol use. Alcohol is currently the most widely distributed and commonly used recreational drug in the country. The prevalence of heavy drinking among drinkers is the highest worldwide [World Health Organization, 2018]. The country also has a high prevalence of HIV among the adults [Uganda AIDS Commission, 2017]. Such dual burden of heavy alcohol use and HIV may

present some challenges. This is because alcohol consumption likely has a large impact on the HIV epidemic via behavioral pathways such as sexual risk-taking behaviors, decreased self-care behaviors such as poor medication adherence, and biological pathways such as impaired immunity [Hahn et al., 2011]. Numerous studies have assessed the consequences of such impact of heavy alcohol use: it results in increased HIV incidence [Baliunas et al., 2010, Scott-Sheldon et al., 2016]; it has been a consistent risk factor for HIV care and treatment cascades [Azar et al., 2010, Vagenas et al., 2015], especially for the non-adherence to antiretroviral therapy (ART) [Samet et al., 2004, Hendershot et al., 2009] with a dose-response relationship [Braithwaite et al., 2005, Sileo et al., 2016].

Chronic alcohol use impacts both innate and adaptive immune functioning [Szabo and Saha, 2015], and chronic alcohol use and HIV independently damage the intestinal mucosa, enabling increased microbial translocation with subsequent increased inflammation [Bagby et al., 2015]. Experimental studies in which high doses of alcohol are administered to macaques before and after infection with simian immunodeficiency virus (SIV) found increased levels of SIV viremia and mortality compared to control macaques who are infected with SIV but who receive a sucrose control [Bagby et al., 2003, Kumar et al., 2005, Bagby et al., 2006, Poonia et al., 2006]. Thus alcohol use might be an important factor in HIV disease progression.

Despite the high biologic plausibility of an effect of alcohol use on HIV disease progression, the results of human observational studies have been mixed. No prospective study conducted in the period before the advent of ART find an association between

alcohol consumption and the onset of AIDS [Hahn and Samet, 2010], and a retrospective analysis of persons not yet on ART participating in a large clinical HIV cohort find no association between risky alcohol use and CD4 cell count [Conen et al., 2013]. However, two studies conduct since the advent of ART suggest a detrimental effect of alcohol use prior to ART use, with one study reporting a difference in mean CD4 cell count of 49 cells/mm^3 among those reporting heavy drinking compared to those abstaining [Samet et al., 2007], and another reporting a strong association between frequent alcohol use ($\geq$ 2 drinks daily) and time to CD4 cell count below 200 cells/mm^3 [Baum et al., 2010, Monroe et al., 2016].

Among longitudinal studies of persons on ART, the findings have been mixed as well. Several studies among persons on ART have found no association between high levels of alcohol use (various defined as heavy, hazardous, problem, or severe risk alcohol use) and CD4 cell count and/or HIV viral load after controlling for ART adherence [Samet et al., 2007, Kowalski et al., 2012, Conen et al., 2013, Cagle et al., 2017]. Two recent studies conduct mediation analyses to separate out effects of alcohol use on CD4 cell counts due to reduced adherence versus other pathways. One finds direct effects of heavy alcohol use on CD4 cell counts [Kahler et al., 2017], while the other finds only indirect effects of alcohol use via adherence [Wandera et al., 2017].

Several methodological considerations might explain these inconsistent findings. First, inability to accurately measure alcohol use may impact the results [Baliunas et al., 2010]. Inclusion the biomarkers of alcohol use can provide an objective measurement. Second,

in some populations, alcohol use may be associated with illicit drug use, which may be associated with more rapid HIV progression (i.e. illicit stimulant use [Cook et al., 2008, Carrico et al., 2014]), and thus a spurious association of alcohol use with HIV progression may occur. A solution to this is to exclude other substance use, or conduct studies in settings with very little substance use, such as in Uganda [John-Langba et al., 2006]. Third, studies of persons on ART may be susceptible to residual confounding due to imperfect measurement of adherence, such as in the case of exaggerated self-reported adherence [Stirratt et al., 2015]. Thus, restricting the sample to those who are not yet on ART avoids this potential pitfall. Lastly, the relationship between alcohol use and HIV disease progression may be confounded over time, if individuals who engage in heavy drinking experience declines in their health, and thus reduce their subsequent drinking. This circumstance may spuriously reduce the apparent relationship between heavy drinking and HIV disease progression [Robins et al., 2000], previous analyses of this issue have not accounted for this possibility.

Thus, to clarify the previous inconsistent results of biological impact of unhealthy alcohol consumption on HIV disease progression, the reliable and accurate measurement of alcohol consumption is fundamentally important. Multimodal alcohol assessment is usually recommended, including both self-report and biological measures, because each method of measurement has strengths and weaknesses. Self-report measures is the most common way to assess alcohol intake, because they are quick, straightforward, and inexpensive. But they are easily limited by factors such as recall bias toward under-report

of alcohol use. Social desirability may cause under-reporting if there is the perception of negative consequences associated with reported use. Persons infected with HIV who are receiving or hoping to receive HIV ART may under-report their alcohol consumption if they fear that they will be denied ART if they report alcohol consumption. The accuracy of self report is also limited by significant difficulties in precisely measuring content and volume of alcohol consumed. For example, many drinkers consume locally-made alcohols with variable ethanol contents.

To address the biases and inaccuracy associated with self-report, several alcohol biomarkers have been developed to objectively assess alcohol use [Wurst et al., 2005, Hannuksela et al., 2007, Litten et al., 2010]. One such alcohol biomarker is phosphatidylethanol (PEth). Studies indicate the PEth is a valid marker of assessing alcohol consumption among individuals who are HIV-infected [Hahn et al., 2012]. However, the accuracy of PEth can vary between persons consuming the same amount of alcohol, likely due to differences in alcohol metabolism [Javors et al., 2016]. Also the biomarker is more expensive, and may be unavailable in resource-limited setting.

Several research groups have compared PEth to self-reported assessments of alcohol use. Some studies indicate that PEth and self-reported alcohol consumption can show strong concordance [Hahn et al., 2012, Stewart et al., 2014]. This conclusion is made based on the assumption that self-reports are valid. Given the possible bias of self-reported alcohol use, other studies have used PEth as the "gold standard" to validate the self report. They've found considerable disagreement between self-reported alcohol

consumption and PEth status [Kip et al., 2008, Bajunirwe et al., 2014].

Consequently, self report and biological marker together may provide the basis for a more accurate estimate of participants' past alcohol consumption than either does alone, given the lack of complete agreement between self-report and biological measurement. This is the reason why simultaneous analysis of self report and biomarker is appealing. Joint modeling of associated self-reported responses and biological responses is common, facilitated by development of various statistical methods on joint modeling of associated outcomes [Catalano, 1997, Li et al., 2016]. On the other hand, less is discussed about the case where researchers intend to combine biological marker and self report as explanatory variables in the study. For example, When alcohol intake is considered as the explanatory variable, the common act is to categorize alcohol exposure according to some cutoffs values of self report and biomarkers [Eyawo et al., 2018, Hahn et al., 2018]. Such categorization, however, depends on the researchers' choice of cutoff for two measurement. Different cutoffs and different levels of categories may affect the study result, which could be subjective and even misleading.

Beginning in 2011, 447 participants were recruited from the Immune Suppression Syndrome (ISS) Clinic of the Mbarara Regional Referral Hospital of the Mbarara University of Science and Technology. Study enrollment was conducted from September 2011 to August 2014. Eligibility criteria are adult patient of the Mbarara ISS Clinic not yet meeting eligibility criteria for ART (i.e., CD4 cell count <350 cells/mm^3, World Health Organization disease stage III or IV, or AIDS defining illness). At baseline and

follow-up visits every six months, study assessments include alcohol consumption and CD4 cell count. Demographic information and HIV viral load are also measured at baseline. Alcohol consumption is assessed using the Alcohol Use Disorders Identification Test - Consumption (AUDIT-C) [Bradley et al., 2007]. Due to the inability to accurately measure alcohol use of self report, a biomarker of alcohol use, phosphatidylethanol (PEth) is also measured in the study as a direct metabolite of alcohol use that is highly specific and reasonably sensitive for measuring prior 2-3 weeks alcohol use [Wurst et al., 2015].

The analysis of interest is the association between alcohol consumption and HIV disease progression measured by CD4 cell count over time, estimated using a linear mixed effect model adjusted for baseline covariates including demographic variables and HIV viral load. The summary of variables used in the model is given in Table 1. Instead of categorizing the alcohol intake based on AUDIT-C or PEth, we will include both two measurements directly as two time-dependent covariates in the linear mixed model.

There are two major research issues associated with our proposal. First, the correlation between self-report measures and biological marker would induce the multicollinearity in design matrix of linear mixed model. Second, after enrollment and baseline testing, some patients lost to follow-up or withdraw from the study. Figure 1 reveals the patterns of missing observations in CD4 cell count, AUDIT-C and PEth over the course the study. First, when a patient miss one visit, measurements of CD4 cell count and alcohol use are all missing. That is the response and time-varying covariates in our analysis model have same missing percentages at each visit. Second, patients were lost to follow

up starting at first follow-up. At first follow-up (6 month), the proportion of missing value is around 25%, however, this proportion increases as time goes by. But at the end of the study, some people show up again where the missing proportion is just slightly higher than that in the first follow-up. This is an indication of the hard work the data collection team were doing in order to reduce missingness, and due to the importance of missingness in the last time point. Finally, none of patients attend all follow-up visits. 78 patients attend the baseline visit and final follow-up visit, 77 patients additionally attend the first follow-up visit, 69 patients also attend the second follow-up.

In sum, models that account for multicollinearity of incomplete longitudinal where both outcome and time varying covariates are incomplete would be necessary. This is the motivation and objective of our work.

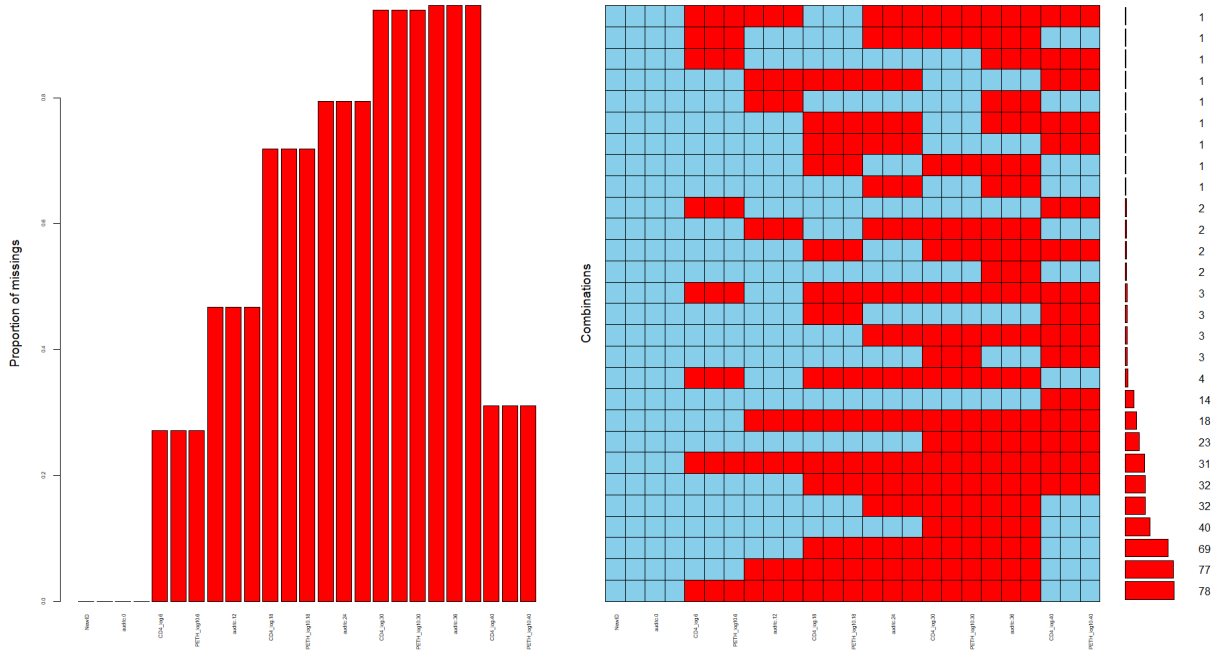
1.3 Dissertation Narratives

The first difficulty arise in our work is the high correlation between self-report and biological marker when we include two measurements of alcohol use simultaneously in the liner mixed model. At baseline, the correlation between AUDIT-C and PEth is 0.68, which suggests the issue of multicollinearity in the linear mixed model. To settle this issue, in Chapter 2, we develop a Bayesian shrinkage models for linear mixed model. We assume a normal-gamma prior on the fixed effects to induce the shrinkage of coefficients estimation as a remedy to the multicollinearity. We carry out a detailed analysis of

Table 1: HIV-infected persons in southwestern Uganda: participants characteristics at baseline (N=447)

Time-varying Variables	Statistics	Values
CD4 cell count (log10)	Mean(Std)	6.28 (0.37)
	Median	6.31
AUDIT-C Score	Mean(Std)	2.11(2.8)
	Median	1
PEth (log10)	Mean(Std)	1.06(1.11)
	Median	0.93
	Correlation with AUDIT-C	0.68
Baseline Covariates	Statistics	Values
Age	Mean(Std)	34.42(10.07)
	Median	32
Religion	Catholic	157
	Protestant/Anglican	220
	Moslem	41
	Saved/Other	29
Sex	Female	303
	Male	144
HIV Virus load (log10)	Mean(Std)	3.66(1.04)
	Median	3.74

Figure 1: Aggregation plot of CD4 cell count in log scale, AUDIT-C and PEth over the course the study



the HIV data using the proposed methodology and our results are generally consistent with those reported in the literature. In consideration of a large amount of patients fail to follow up all visit, we restrict the analysis to $N = 145$ patients who all show up at 6-months, 12-months, and 40-months follow-up.

Second, in view of large amount of missing observations, in Chapter 3, we propose a two-stage-multiple imputation to account for incomplete responses and incomplete time-varying covariates in longitudinal or clustered data. We develop two versions of two-stage multiple imputation by shifting the order of imputing the missing response and missing time-varying covariates in two stages respectively. We compare the two proposed two-stage multiple imputation approaches to the existing multiple imputation approaches

for missing time-varying covariates in a variety of simulation scenarios. To avoid the multicollinearity problem, we apply the proposed two-stage multiple imputation to the data example, while only including the PEth as the covariates. A simulation study is set to evaluate the performance of proposed two-stage multiple imputation approach across different missingness scenarios.

In Chapter 4, we extend the two-stage multiple imputation to further account the case where partly-observed covariates are highly correlated. We incorporate the Bayesian shrinkage prior developed in Chapter 2 into the imputation for the missing highly-correlated covariates. Simulations of various scenarios are carried out to evaluate the performance of proposed approach. We illustrate the proposed approach by applying to the data example.

Finally, in Chapter 5, we conclude with a brief discussion, and point out some interesting extensions for future research. These are the topics that have not been covered in this dissertation, but raised our attention during our exploration of the fields. One potential extension focuses on accommodating missing not at random mechanism.

Chapter 2

Application of Bayesian Shrinkage Priors in Linear Mixed Effect Model

2.1 Introduction

In this chapter, we propose Bayesian shrinkage priors as remedies to multicollinearity in linear mixed effect model. Kreft and De Leeuw [1998] showed that multicollinearity can make the interpretation of model coefficients difficult, especially when dealing with cross-level interactions. Slight changes in the model led to very different results, with coefficients for correlated variables changing over models and standard errors changing even more. Therefore, multicollinearity in linear mixed model must be proceed with caution.

Remedies to multicollinearity have been discussed extensively in the context of linear regression, yet less has been said about how to deal with multicollinearity in linear mixed model. Several authors have discussed the frequentist approach by developing ridge regression in the framework of linear mixed model. Ridge regression [Hoerl and Kennard,

1970a,b] is a classical shrinkage approach to overcome the multicollinearity. Eliot et al. [2011] first integrated ridge regression into the framework of linear mixed model, deriving the parameter estimates via the expectation-maximization (EM) algorithm. Liu and Hu [2013] discuss the superiority conditions of ridge estimator over maximum likelihood estimator (MLE) in linear mixed model. Ozkale and Can [2017] also showed the outperformance of ridge estimator over MLE under the mean squared error (MSE) criterion. However, ridge regression always contain a data-driven determination of the shrinkage parameter, for example based on cross-validation [Hoerl and Kennard, 1970a,b]. In linear mixed models, which parameter estimations requires iterative optimization, including additional shrinkage parameters will further complicate the algorithm and increase the computational burden.

Motivated by this, we consider a Bayesian approach to model the correlated covariates in linear mixed model. We propose a Bayesian shrinkage estimator for fixed effects in linear mixed model to provide a possibility to treat the multicollinearity in linear mixed effect framework. We assign the normal-gamma prior proposed by Griffin et al. [2010], which has shown the computational efficiency in linear regression with correlated design matrix, as well as ability to avoid overshrinkage of large coefficient. Simulation studies will be conducted to compare the performance of proposed method with that of MLE under MSE criterion.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the background needed for the proposed model. In Section 2.3, we fully develop the

Bayesian shrinkage mixed estimator by outlining the detailed computational algorithm. In Section 2.4, a simulation study is carried out to evaluate the performance of the proposed approach. In Section 2.5, we illustrate the proposed approach using the motivating data. We conclude this chapter with some discussion in Section 2.6.

2.2 Methodology

2.2.1 Multicollinearity in linear regression

In linear regression, $y = X\beta + \epsilon$, for p unknown parameters $\beta = (\beta_1, \cdots, \beta_p)'$, the ordinary least square (OLS) estimation β has the form $\hat{\beta} = (X'X)^{-1}X'y$. This estimator, however, is only well-defined when $X'X$ is nonsingular. When some columns in the design matrix X are highly correlated or even perfectly correlated, the matrix $X'X$ become singular. In the view of such multicollinearity issue, Hoerl and Kennard [1970a] proposed ridge regression as an ad-hoc fix: replace $X'X$ by $X'X + kI_p$, where k is a penalty parameter. The ridge estimator in linear regression thus is given by

$$\hat{\beta}_r = (X'X + kI_p)^{-1}X'y. \tag{2.1}$$

The ridge estimator is a member of penalized least squares (PLS), which is derived from minimization of a penalized sum of squares [Frank and Friedman, 1993, Fan and

Li, 2001]:

$$\min||y - X\beta||^2 + k||\beta||^2 \tag{2.2}$$

where $||\beta||^2$ is a shrinkage penalty, and the penalty parameter k has the effect of shrinking the estimates of β towards zero.

By adopting a Bayesian perspective, we interpret the linear regression $y = X\beta + \epsilon$ as a conditional distribution of y, such as $y|\beta, \sigma \sim N(X\beta, \sigma^2 I)$, if we assume $\epsilon \sim N(0, \sigma^2 I)$. Additionally, suitable priors would be assumed for β and σ [Gelman et al., 2013]. For example, by assuming $\beta \sim N(0, \tau^2 I)$, we can derive the posterior distribution of β by Bayes theorem as $\beta|y, \sigma \sim N\left((X'X + kI)^{-1}X'y, \sigma^2(X'X + kI)^{-1}\right)$ with $k = \frac{\sigma^2}{\tau^2}$. Hence, the posterior mean coincide with the ridge estimator derived in equation (2.1) [Lindley and Smith, 1972].

In a generic form, inference on β can be regularized by specifying appropriate informative priors $p(\beta|\tau)$, where the hyperparameters τ includes parameters controlling shrinkage. $p(\tau)$ may be defined hierarchically with the aim to encourage shrinkage on β. The relationship to frequentist regularization via penalized inference is seen if τ is fixed with some plug-in values.

2.2.2 Bayesian Gaussian Shrinkage Priors

To induce the shrinkage on coefficients in the Bayesian framework, conditional Gaussian priors $\beta_i \sim N(0, \tau_i), i = 1, \cdots, p$ for each element in β is a common choice. Suitable

hyperpriors on τ_i would induce non-Gaussian marginal priors of β_i, enforcing desired shrinkage properties. The framework of conditionally Gaussian priors facilitates full Bayesian Markov chain Monte Carlo (MCMC) inference: for Gaussian regression models, full conditional for all regression coefficients are Gaussian, leading to the Gibbs sampler.

Different prior assumptions for τ_i^2 would induce different marginal distribution of β_i, leading to different shrinkage effect. For example, if τ_i^2 follows an inverse gamma distribution, β_i marginally follows a scaled t-distribution [Fahrmeir et al., 2010]; if τ_i^2 follows an exponential distribution, β_i instead marginally follows a double exponential distribution [Andrews and Mallows, 1974]. Griffin et al. [2010] developed a generalization of the double exponential priors by proposing the normal-gamma priors. They showed that the normal-gamma prior is a natural extension of double exponential prior, inducing a wide range of shrinkage behavior. The normal gamma prior is given by:

$$\beta_i|\Phi_i \sim N(0, \Phi_i), \quad \Phi_i \sim \Gamma(\lambda, \frac{1}{2\gamma^2}) \tag{2.3}$$

where λ is the shape parameter, and $\frac{1}{2\gamma^2}$ is the rate parameter. The choice of λ and ϕ plays an information role in estimation. Griffin et al. [2010] assume $\lambda \sim exp(1)$ to induce a Bayesian Lasso. And the prior for γ conditional on λ is given by $2\lambda\gamma^2 \sim \text{IG}(2, M)$, where IG denotes the inverse gamma distribution, $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^{\,2}$.

2.2.3 Linear mixed model

Linear mixed model is an extension of linear regression, by including both fixed and random effects as predictor variables [Laird and Ware, 1982, Jennrich and Schluchter, 1986, Laird et al., 1987, Lindstrom and Bates, 1988, Searle et al., 2009]. Let y_i denote an n_i vector of responses for sample unit i, $i = 1, 2, \ldots, n$. The model for the y_i is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{2.4}$$

where $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ are known covariate matrices, β is a p-dimensional vector of regression coefficients common to all units, and b_i is a q-dimensional vector of coefficients specific to unit i. β and b_i are called "fixed effects" and "random effects" respectively. We assume that random effects are distributed as $b_i \sim N(0, \psi)$ independently for $i = 1, \ldots, n$, and n_i rows of ϵ_i are independently distributed as $N(0, \sigma^2)$.

let Y, X and Z be appropriately defined matrices representing the concatenation of the corresponding variables over all individual i, and $V = \text{var}(y) = Z\psi Z + \sigma^2 I$. The maximum likelihood estimator of β can be represented as $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$, with conditional distribution as $g(\hat{\beta}|\beta) = N(\beta, (X'V^{-1}X)^{-1})$.

2.3 Bayesian Shrinkage Mixed Estimator

In view of flexible shrinkage effect of normal-gamma prior, we extend it to the framework of linear mixed model. To our knowledge, so far there has been no discussion about the

extension of normal-gamma prior to linear mixed model specifically with multi-collinear design matrix. We'll first adapt the algorithms to the mixed effect structure, and use the simulations to assess the performance of this prior on multicollinearity in next section.

2.3.1 Normal-gamma prior for fixed effect

In the framework of linear mixed model, we assume a normal-gamma prior for fixed effect β. To facilitate the analysis, we assume the covariance matrix of random effect b_i is $\psi = \sigma_b^2 I_q$. Prior distributions on all unknown parameters as follows induced by Bayes theorem:

$$\beta_i|\Phi_i \sim N(0, \Phi_i), \quad \Phi_i|\lambda, \gamma \sim \Gamma(\lambda, \frac{1}{2\gamma^2}), \tag{2.5}$$

$$\lambda \sim \exp(a_\lambda), \quad 2\lambda\gamma^2 \sim \text{IG}(2, M), \tag{2.6}$$

$$\sigma \sim \text{flat prior}, \quad \sigma_u \sim \text{flat prior}. \tag{2.7}$$

The flat prior for variance components in linear mixed model is assumed to be a uniform distribution (0, 100) within scope of this analysis. The posterior distribution of the parameters can be simulated using a Gibbs sampler with an additional Metropolis - Hastings update. The full conditionals used in the updating steps are given below:

Updating β The full conditional distribution of β follow a normal distribution with mean $(X'V^{-1}X + \sigma^2\Lambda)^{-1}X'V^{-1}y$ and variance $\sigma^2(X'V^{-1}X + \sigma^2\Lambda)^{-1}$, where

$$\Lambda = \text{diag}(\frac{1}{\Phi_0}, \frac{1}{\Phi_1}, \cdots, \frac{1}{\Phi_p})$$

Updating σ^2 and σ_b^2

The full conditional distribution of σ^2 is $IG(c, d)$, with $c = \frac{n}{2}$ and $d = (y - X\beta - Zb)'(y - X\beta - Zb)/2$, while the full conditional distribution of σ_b^2 is $IG(c^*, d^*)$, with $c^* = \frac{n}{2}$ and $d^* = \sum_{i=1}^{n} b_i b_i'/2$

Updating Φ_i

The full conditional distribution of $\Phi_i, i = 1, \cdots, p$ is Generalized Inverse Gaussian distribution $\text{GIG}(\lambda - \frac{1}{2}, \frac{1}{\gamma^2}, \beta_i^2)$.

Updating λ and γ

The full conditional distribution of γ^{-2} is $\Gamma(e, f)$, with $e = 2 + p\lambda$, $f = \frac{M}{2} + \frac{1}{2}\sum_{i=1}^{p} \Phi_i$. In terms of shape parameter, let $\pi(\lambda)$ be the density function of $\exp(a_\lambda)$, then the full conditional distribution of λ would be

$$\pi(\lambda)\frac{1}{(2\gamma^2)^{p\lambda}(\Gamma(\lambda))^p}\left(\prod_{i=1}^{p}\Phi_i\right)^{\lambda}$$

which need to be updated using a Metropolis-Hastings random walk update on logλ.

Updating b_i

The full conditional distribution of random effect b_i is a normal distribution $(\tilde{b_i}, U_i)$, with $U_i = (\sigma_b^{-2} I + \frac{Z_i^T Z_i}{\sigma^2})^{-1}$, $\tilde{b_i} = U_i Z_i^T (y_i - X_i \beta)/\sigma^2$.

The choice of λ plays an important role in estimation. It is showed that, in the setting of (multiple) linear regression, smaller λ will lead to larger shrinkage [Griffin et al., 2010], with the amount of shrinkage is also related to size of β and value of σ.

2.3.2 Shrinkage for multicollinearity

The motivation of imposing a normal-gamma prior distribution is that such Bayesian estimation can lead to shrinkage of regression estimator, where shrinkage is an efficient way to deal with multicollinearity. The multicollinearity would cause the variance inflation, while the appropriate shrinkage would be able to decrease the variance while increase the bias at the same time. Suppose that the error variance σ^2 and random variance matrix ψ are known in model (2.4), by extending the result of Griffin et al. [2010], we are able to express the posterior expectation and variance for fixed effect coefficients β given $\hat{\beta}$ as

$$E[\beta|\hat{\beta}] = \left(I - S(\hat{\beta})\right) \hat{\beta}, \tag{2.8}$$

$$V[\beta|\hat{\beta}] = \sigma^2 (X'V^{-1}X)^{-1} - \sigma^4 (X'V^{-1}X)^{-1} W(\hat{\beta}) (X'V^{-1}X)^{-1}, \tag{2.9}$$

where the marginal distribution of $\hat{\beta}$ is $h(\hat{\beta}) = \int g(\hat{\beta}|\beta)\pi(\beta)d\beta$, $S(\beta) = \sigma^2(X'V^{-1}X)^{-1}R(\hat{\beta})$ where $R(x)$ is a diagonal matrix with $R_{ii}(x) = -\frac{1}{x_i}\frac{\partial}{\partial x_i}\log h(x)$ and $W(x) = -\frac{\partial}{\partial x}\frac{\partial}{\partial x'}\log h(x)$. Equations (2.8) indicates the posterior expectation and variance of fixed effect coefficient β are matrix-shrunken versions of those of the maximum likelihood estimator. Henceforth, by imposing a normal-gamma prior on β, called as " Bayesian shrinkage mixed estimator (BSME)", we expect a smaller mean squared error (MSE) for this newly-developed estimator compared with that of maximum likelihood estimator. Simulation studies would be conducted to examine the performance for newly proposed method.

2.4 Simulations

The goals of the simulation are twofold. First, we aim to compare the effect of a_λ in equation (2.6) on the performance of Bayesian shrinkage mixed estimator (BSME). More importantly, we wish to compare the performance of maximum likelihood estimator (MLE), ridge regression in linear mixed effect model (RR), and Bayesian shrinkage mixed estimator (BSME) across various scenarios.

We generate two correlated covariate using the following device which is able to achieve different collinearity [McDonald and Galarneau, 1975]:

$$x_{ij} = (1 - \varphi^2)^{1/2} z_{ij} + \varphi z_{i3}, i = 1, 2, \cdots, n, j = 1, 2 \tag{2.10}$$

where z_{i1}, z_{i2}, z_{i3} are independent standard normal pseudo-random numbers, and φ

is specified so that the correlation between any two explanatory variables is given by φ^2. The response is generated through:

$$y_{ij} = (\beta_0 + b_i) + \beta_1 x_{i1} + \beta_2 x_{i2} + e_{ij} \tag{2.11}$$

where b_i and e_{ij} are independent normal random numbers with mean being 0 and variance being 1 respectively. For simplicity, $\beta_0, \beta_1, \beta_2$ are set as $(1, 1, 1)$. We perform a set of simulations for different combinations of data generation models. First, we vary φ to alter the degree of correlation. We consider scenarios with slightly high ($\varphi^2 = 0.6$), high ($\varphi^2 = 0.8$), and extremely high ($\varphi^2 = 0.99$) correlations. Second, we vary the sample size to small (n=100) and medium (n=500) samples. In each scenario, 1000 datasets are generated. The target parameters $\beta_0, \beta_1, \beta_2$ are estimated from each dataset. For each parameter, $\beta_i, i = 0, 1, 2$, its estimator $\hat{\beta}_i$ and estimated standard error $\text{se}(\hat{\beta}_i)$, we calculate the percentage bias: $(\frac{\hat{\beta}_i - \beta_i}{\beta_i}) * 100\% = \frac{\text{bias}_i}{\beta_i} * 100\%$; the mean squared error (MSE): $\text{bias}_i^2 + \text{se}(\hat{\beta}_i)^2$, and 95% credible interval for Bayesian approach, and confident interval for maximum likelihood approach. Then the mean of percentage bias, mean of MSE, mean of credible (confidence) interval width, and coverage rate are reported, with coverage rate is given by the percentage of the 1000 datasets for which the credible (confidence) interval contained the true value of β_i.

2.4.1 Simulation Results

Table 2 shows the performance of BSME for different mean of prior exponential assigned on λ with different level of correlation. Generally, for every value of a_λ, estimates obtained across all three levels correlation have relatively small bias and credible intervals achieve nominal coverage. Furthermore, the magnitude of MSE and credible interval increase as the correlation increase for coefficient β_1 and β_2. The MSE and credible interval width are worst in scenario with extremely high correlation 0.99 across all values of a_λ (MSE between 0.912 and 1.271 compared with 0.099 to 0.124 with slightly high correlation 0.6, credible interval width between 3.133 and 3.563 compared with 0.944 and 0.963). On the other hand, correlation levels do not have significant effect on β_0 in terms of MSE and credible interval width. This is because β_0 represents the intercept, which can be ignored if all covariates standardized. In a summary, a small value of a_λ such as 0.1 would help to achieve relatively small MSE and credible interval width.

Table 3 shows detailed results for the BSME, MLE and RR across every scenarios with a_λ being 1. The performances for other values of a_λ present similar patterns and thus are not present here. Full simulation results could be found in Tables 11, 12 and 13 in appendix. First, the performance of BSME and MLE have similar pattern across all scenarios. With different levels of correlations and different samples sizes, the estimated percentage bias is small, between 0.4% and 3.1% and credible (confidence) interval achieve nominal coverage between 92% and 98%. The magnitude of the MSE and width of credible (confidence) intervals both increase as the level of correlation

increase. The larger sample size lead to smaller MSE and shorter credible interval width. However, the ridge regression (RR) produces larger percentage bias (as high as 13.99%) especially in the case of small sample size. Such large bias lead to poor confidence interval coverage rate (much lower than nominal coverage rate), although the MSE and confidence interval width calculated from RR are smaller than those from other two approaches. This suggests that, the ridge regression is an efficient way to reduce the variance of regression coefficients estimates, but fails to produce reasonable bias.

Notably, the traditional MLE approach yield large MSE with extremely high correlation. The MSE of MLE is worst in scenarios with high correlation 0.99 and small sample size 100 (MSE are 3.280 and 3.284 for β_1 and β_2). Wider credible interval are also observed for MLE approach with high correlation (as high as 4.9 with sample size 100, and 2.2 with sample size 500). Hence, solely as a result of the behavior of the regression coefficient estimator, the traditional MLE approach, yield poor MSE and credible interval width when the multicollinearity is severe. Furthermore, the ridge regression in linear mixed mode is able to yield small MSE by reducing the variance of coefficients estimates. However, in views of large bias and lower-than-nominal-rate confidence interval coverage, the ridge regression still remain questionable as an efficient remedy to multicollinearity. The proposed BMSE has notably reduced MSE compared with MLE across three levels of correlations. Especially with sample size 100 and correlation 0.99, the MSE produced by BSME is around 0.92, compared with 3.28 produced by MLE.

The proposed approach also result in shorter credible intervals.

In a conclusion, the proposed BMSE is able to to reduce the MSE caused by the multicollinearity, at the same time yielding small bias. The outperformance of BMSE is notable when the sample size is small or the correlation is high.

Table 2: Results from simulation study. Percentage bias (Bias %), MSE, 95% credible interval coverage (Cov), and 95% credible interval width (CI width) from different values of expected value of λ, a_λ

a_λ		Bias (%)			MSE			Cov			CI Width		
	φ^2	0.6	0.8	0.99	0.6	0.8	0.99	0.6	0.8	0.99	0.6	0.8	0.99
0.1	β_0	-3.1	-0.8	-1.2	0.051	0.069	0.060	0.99	0.92	0.96	0.695	0.694	0.691
	β_1	1.4	0.4	-1.1	0.109	0.161	0.923	0.96	0.96	0.97	0.960	1.129	3.137
	β_2	-4.0	-0.7	-0.9	0.102	0.158	0.923	0.98	0.97	0.97	0.957	1.133	3.136
0.2	β_0	-2.3	-0.5	-2.4	0.057	0.075	0.062	0.95	0.93	0.95	0.691	0.696	0.691
	β_1	-0.6	-0.2	1.8	0.106	0.155	1.253	0.96	0.94	0.95	0.953	1.129	3.563
	β_2	1.6	-3.0	-3.4	0.113	0.150	1.271	0.94	0.96	0.95	0.957	1.133	3.565
0.3	β_0	-3.3	-5.2	-2.8	0.063	0.059	0.064	0.95	0.97	0.95	0.698	0.696	0.693
	β_1	0.6	0.6	-1.8	0.121	0.162	1.167	0.96	0.97	0.97	0.958	1.130	3.476
	β_2	0.1	-2.2	1.0	0.113	0.152	1.169	0.97	0.95	0.97	0.962	1.130	3.477
0.4	β_0	-2.8	2.8	-1.9	0.064	0.065	0.066	0.97	0.95	0.94	0.687	0.686	0.692
	β_1	-0.3	0.6	-2.2	0.124	0.144	1.104	0.93	0.99	0.98	0.946	1.128	3.402
	β_2	-2.5	-3.9	0.7	0.113	0.152	1.100	0.96	0.96	0.97	0.948	1.130	3.403
0.5	β_0	-3.5	-2.6	-1.9	0.061	0.065	0.063	0.96	0.93	0.95	0.697	0.691	0.692
	β_1	1.5	-3.6	-2.5	0.118	0.154	1.020	0.98	0.96	0.97	0.960	1.118	3.334
	β_2	-3.0	-0.7	1.6	0.105	0.144	1.023	0.97	0.96	0.97	0.963	1.120	3.333
0.6	β_0	-2.5	-2.7	-3.8	0.063	0.056	0.062	0.94	0.99	0.95	0.691	0.694	0.689
	β_1	-1.5	-1.8	-3.0	0.106	0.167	0.995	0.99	0.97	0.96	0.952	1.121	3.268
	β_2	-5.7	1.3	0.3	0.123	0.155	0.983	0.94	0.97	0.96	0.954	1.124	3.270
0.7	β_0	0.9	-0.2	-3.2	0.057	0.062	0.064	0.97	0.96	0.94	0.696	0.695	0.689
	β_1	-0.7	-2.8	1.1	0.113	0.182	0.967	0.98	0.92	0.98	0.944	1.121	3.232
	β_2	-3.2	1.3	-2.9	0.112	0.163	0.964	0.96	0.96	0.98	0.944	1.125	3.233
0.8	β_0	0.0	-0.2	-1.9	0.065	0.058	0.061	0.95	0.99	0.95	0.700	0.693	0.690
	β_1	4.8	-0.9	-7.2	0.123	0.158	0.924	0.92	0.96	0.97	0.948	1.114	3.191
	β_2	-6.5	2.1	4.9	0.107	0.155	0.925	0.98	0.94	0.96	0.952	1.113	3.192
0.9	β_0	-3.9	-2.9	-3.2	0.059	0.066	0.061	0.95	0.92	0.95	0.689	0.686	0.693
	β_1	0.2	-0.9	-1.0	0.115	0.162	0.912	0.96	0.96	0.94	0.948	1.112	3.159
	β_2	-2.1	-0.7	-0.9	0.111	0.151	0.923	0.95	0.98	0.96	0.948	1.113	3.160
1	β_0	-3.5	-0.6	-3.1	0.072	0.060	0.060	0.91	0.96	0.97	0.690	0.688	0.690
	β_1	-0.3	-0.9	0.3	0.120	0.153	0.924	0.95	0.95	0.97	0.950	1.111	3.134
	β_2	-2.2	-1.2	-3.1	0.099	0.162	0.917	0.97	0.94	0.97	0.946	1.112	3.133

Table 3: Results from simulation study with expected value of λ, a_λ being 0.1. Percentage bias (Bias %), MSE, 95% credible interval coverage (Cov), and 95% credible interval width (CI Width) from the proposed method BSME, along with traditional approaches MLE and RR. Results are based on 1000 replications each

n	ϕ^2		Bias	(%)		MSE			Cov			CI	Width	
			BSME	MLE	RR	BSME	MLE	RR	BSME	MLE	RR	BSME	MLE	RR
100	0.6	β_0	-3.098	-1.494	-13.993	0.051	0.053	0.069	0.99	0.99	0.97	0.695	0.716	0.719
		β_1	1.354	2.468	0.757	0.109	0.113	0.061	0.96	0.95	0.74	0.960	0.972	0.493
		β_2	-3.989	-3.105	-5.138	0.102	0.105	0.055	0.98	0.98	0.72	0.957	0.970	0.493
	0.8	β_0	-0.831	0.948	-11.419	0.069	0.071	0.078	0.92	0.93	0.91	0.694	0.708	0.714
		β_1	0.406	1.041	-2.281	0.161	0.172	0.100	0.96	0.96	0.71	1.129	1.152	0.574
		β_2	-0.668	0.576	-0.975	0.158	0.168	0.094	0.97	0.97	0.67	1.133	1.155	0.574
	0.9	β_0	-1.172	1.553	-10.480	0.060	0.063	0.069	0.956	0.954	0.964	0.691	0.712	0.715
		β_1	-1.071	-0.926	-2.297	0.923	3.280	0.242	0.99	0.936	0.702	3.137	4.915	1.330
		β_2	-0.885	0.976	-0.732	0.923	3.284	0.243	0.994	0.932	0.69	3.136	4.916	1.330
500	0.6	β_0	-1.230	-1.047	-2.925	0.012	0.013	0.013	0.95	0.97	0.95	0.310	0.313	0.313
		β_1	0.708	0.932	0.536	0.024	0.024	0.017	0.94	0.94	0.82	0.433	0.434	0.284
		β_2	-0.406	-0.193	-0.488	0.022	0.022	0.016	0.96	0.96	0.83	0.432	0.433	0.284
	0.8	β_0	-0.810	-0.296	-2.143	0.011	0.011	0.012	0.98	0.98	0.98	0.309	0.313	0.313
		β_1	-0.443	-0.273	-0.696	0.035	0.035	0.025	0.98	0.98	0.78	0.515	0.516	0.336
		β_2	1.057	1.276	0.671	0.034	0.035	0.025	0.95	0.95	0.79	0.515	0.516	0.335
	0.9	β_0	-1.774	-1.350	-3.170	0.014	0.014	0.015	0.96	0.96	0.96	0.314	0.315	0.315
		β_1	-2.344	-2.924	-1.388	0.447	0.574	0.174	0.96	0.95	0.84	2.020	2.201	1.112
		β_2	1.357	2.208	-0.058	0.446	0.573	0.174	0.98	0.95	0.86	2.020	2.201	1.113

2.5 Data Application

The purpose of the HIV data study was to evaluate the relationship between CD4 cell count and alcohol use. Over the course of study, about two-third (67%) participants graduate from the cohort due to ART initiation, or lost to follow-up, or withdraw from the study. The treatment of incomplete data in these set-up would be the topic of Chapter 4. Thus we will restrict the analysis in this chapter to $n = 145$ patients who all show up at 6-months, 12-months, and 40-months follow-up. At baseline, the average CD4 cell count in log scale is 6.370, the average AUDIT-C score is 1.731, the average PEth level is 0.89, and the correlation between AUDIT-C and PEth level is 0.64.

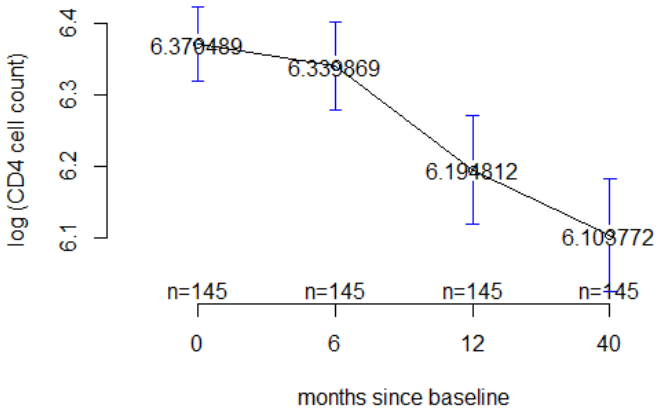
Figure 2 suggest a decline in CD4 cell count during the study. To examine whether alcohol use is associated with such decline, we fit a linear mixed model by including time-varying AUDIT-C and PEth as the main explanatory variables, in addition to some time-independent baseline covariates such as age, sex, religion, and HIV viral load. In view of the correlation between self-report measures and biological marker, we apply the propose Bayesian shrinkage mixed estimation approach, and compare the results with those derived from maximum likelihood estimation approach. We also rerun the analysis by using alternative measures of alcohol consumption, such as AUDIT-C alone and PEth alone.

The resulting point estimates, together with the corresponding standard deviations and 95% credible interval for the regression coefficients are displayed in the Table 5. All

Table 4: HIV-infected persons in southwestern Uganda: participants characteristics at baseline (N=145)

Variable	Statistics	Value
Age	Mean (Std)	35.46 (9.46)
	Median	35.00
Religion	Catholic	54
	Moslem	64
	Saved/Other	14
	Protestant/Anglican	13
Sex	Female	94
	Male	51
Viral Load (log10)	Mean (Std)	3.422 (1.07)
	Median	3.619
CD4 cell count (log10)	Mean (Std)	6.370 (0.31)
	Median	6.407
AUDIT-C score	Mean (Std)	1.731 (2.72)
	Median	0.000
PEth level (log10)	Mean (Std)	0.892 (1.09)
	Median	0.041
	Correlation with AUDIT-C	0.64

Figure 2: Average CD4 cell count in log scale for available cases



models show decline in CD4 cell count over time. The BSME approach shows a negative relationship between AUDIT-C and CD4 cell count, as well as between PEth and CD4 cell count. This suggests that higher alcohol consumption will lead the decline in CD4 cell count, although this relationship is not significant (95% credible interval of AUDIT-C : -0.016 to 0.024; 95% credible interval of PEth : -0.045 to 0.051). The separate models also suggest a insignificant negative association between alcohol consumption and CD4 cell count (95% confidence interval of AUDIT-C : -0.016 to 0.013; 95% confidence interval of PEth : -0.046 to 0.029). However, the MLE approach shows that the discordant impacts of PEth and AUDIT-C (Estimated coefficient of AUDIT-C is 0.0001, estimated coefficient of PEth is -0.009). This disagreement in the relationship between alcohol consumption and CD4 cell count may be just caused by high correlation between AUDIT-C and PEth, which distort the sign of estimated coefficient. For other covariates, all

Table 5: Application to the HIV data: Estimated coefficients obtained by fitting linear mixed model, using maximum likelihood estimation and Bayesian shrinkage mixed estimation to incorporate related AUDIT-C and PEth

	Est	Std	95% CI	Est	Std	95% CI
Joint Analysis		MLE			BSME	
Intercept	6.798	0.125	(6.555, 7.041)	6.718	0.138	(6.441, 6.980)
time	-0.007	0.001	(-0.009, -0.005)	-0.006	0.008	(-0.023, 0.010)
AUDIT-C	0.0001	0.009	(-0.016, 0.017)	-0.004	0.010	(-0.016, 0.024)
PEth	-0.009	0.021	(-0.051, 0.033)	-0.003	0.025	(-0.045, 0.051)
age	-0.002	0.002	(-0.006, 0.002)	-0.002	0.002	(-0.007, 0.002)
religion	-0.039	0.023	(-0.084, 0.007)	-0.035	0.025	(-0.084, 0.014)
sex	-0.035	0.046	(-0.124, 0.054)	-0.041	0.048	(-0.133, 0.053)
viral load	-0.074	0.021	(-0.115, -0.034)	-0.052	0.022	(-0.094, -0.009)
Separate Analysis		MLE			MLE	
Intercept	6.800	0.125	(6.558, 7.043)	6.798	0.124	(6.557, 7.041)
time	-0.007	0.001	(-0.008, -0.005)	-0.007	0.001	(-0.009, -0.005)
AUDIT-C	-0.002	0.008	(-0.016, 0.013)	-	-	-
PEth	-	-	-	-0.009	0.019	(-0.046, 0.029)
age	-0.002	0.002	(-0.006, 0.002)	-0.002	0.002	(-0.006, 0.002)
religion	-0.037	0.023	(-0.082, 0.008)	-0.039	0.024	(-0.084, 0.007)
sex	-0.038	0.046	(-0.126, 0.051)	-0.035	0.046	(-0.124, 0.054)
viral load	-0.075	0.020	(-0.115, -0.034)	-0.074	0.020	(-0.115, -0.034)

approaches yield similar results, all suggesting a statistically significant relationship between HIV viral load and CD4 cell count, whereas no significant difference in age, religion, and gender. It must be noted, that the small sample size may have an impact on the confidence intervals and efficiency of the estimation.

In sum, we find that the alcohol consumption has a negative impact on CD4 cell count, where such impact is not significantly significant. This finding is consistent with several studies [Weiser et al., 2014, Cagle et al., 2017, Hahn et al., 2018].

2.6 Discussion

In this chapter, we consider longitudinal data with highly-related covariates. This problem will arise when multiple measurements of same construct are simultaneously included in the analysis as explanatory variables, such as self-reports and biomarkers. The existing statistical models of longitudinal data such as linear mixed model generally require independent or approximately-independent covariates. The traditional maximum likelihood estimation (MLE) of coefficients may be misleading when covariates are related. In this chapter, we consider a Bayesian perspective, proposing the Bayesian shrinkage mixed estimation (BSME) to deal with the highly-correlated covariates in linear mixed model.

Simulations demonstrate that the proposed approach perform well with small percentage bias, small MSE, and confidence intervals that achieve nominal coverage across all scenarios studied. In scenarios with high correlation and/or small sample size, the proposed approach is clearly preferable to traditional maximum likelihood estimation approach.

The proposed Bayesian shrinkage mixed estimation approach facilitates the analysis intending to include both self-reported and biomarker measurement directly as explanatory variables. We apply the proposed approach to data example which examines the impact of alcohol use on CD4 cell count, where alcohol consumption is measured by both self-report and biomarker. It is shown that both self-report and biological marker

of alcohol intake have a negative impact on CD4 cell count, although such impacts are not statistically significant.

While the simulation studies report excellent performance of proposed approach, we omit subjects with missing observations when we apply the proposed method to the data example. This is called complete case analysis (CCA), which may yield biased estimates because the complete case can be unrepresentative of the full population. Also, it can result in a very substantial loss of information by deleting all case with missing value, and this gives an impact on reduced statistical precision and power. In view of this, we are going to develop models which can also account for missing observations in following chapters.

The proposed approach can be also applied to a broad variety of studies where both self-reported instruments and biological markers of the behavior of interest are recorded, such as smoking, dietary intake, and sexual activity. In addition, the proposed approach can be applied in studies where multiple biological markers are collected and the impacts of those biomarkers are of interest.

Future work could include consideration and comparisons of more shrinkage priors for Bayesian shrinkage estimation and inferences in linear mixed models. Additionally, the assessment of proposed approach in the generalized linear mixed model rather than linear mixed model would also be of interest, when the primary outcome in the analysis is binary or categorical variable.

Last, the proposed approach in this chapter assume the data is fully observed. However, this assumption is usually not satisfied in the reality. Henceforth, in next two chapters, we will take the incomplete data into accounts.

Chapter 3

Application of Two-stage Multiple Imputation for Missing Response and Missing Time-varying Covariates

3.1 Introduction

In this chapter, we propose two-stage multiple imputation to accommodate missing response and missing time-varying covariates in longitudinal or clustered data. In practice, missing covariates arise in longitudinal studies when patients fail to show up during the follow-up. In such cases, all time-dependent variables including time-dependent response and time-dependent covariates would be missing.

Multiple imputation (MI) is a popular method for analyzing incomplete datasets [Rubin, 1996, Schafer, 1997a, Rubin, 2004], which generates a number of "completed"

datasets by filling in missing values from an appropriate predictive distribution, conditional on the observed data. The substantive model of interest is then fitted to each of imputed datasets, and the results are combined for a final inference using Rubin's rule [Rubin, 2004].

In light of missing covariates in longitudinal study, Schafer [1997b] and Schafer and Yucel [2002] proposed the multivariate mixed effect model as the imputation model for both missing response and missing time-varying covariates. This "multivariate multiple imputation " provides a convenient way to impute any missing time-varying variables in longitudinal data, regardless of being a response or covariates.

The drawbacks of "multivariate multiple imputation" is that it assumes a univariate linear mixed model for each incomplete variable given other incomplete variables. However, this may be undesirable because researchers may prefer different imputation models for different types of variables in the data. In such cases, two-stage multiple imputation give an excellent way to handle both missing response and missing covariates flexibly. To this end, we propose two-stage multiple imputation to accommodate the partly-observed response and partly-observed time-dependent covariates simultaneously.

The remainder of this chapter is organized as follows. In Section 3.2, we introduce the backgrounds needed for the proposed approach. In Section 3.3, we fully explained the imputation models and computational algorithms of proposed two-stage multiple imputation. In Section 3.4, a simulation study is conducted to compare the performance of the proposed approach with existing approach. In Section 3.5, we illustrate the

proposed approach using the motivating data. We conclude this chapter with some discussion in Section 3.6.

3.2 Methodology

3.2.1 Missing Data Mechanism

Denote the complete data as Y_{com}, which can be partitioned as Y_{obs} and Y_{mis}, representing the observed and missing parts of data respectively. The distribution of Y_{com} is characterized by parameters θ, such that $P(Y_{com}|\theta)$. In practice, θ is the parameter of interest, and in complete data one can estimate it using maximum likelihood or Bayesian procedures. Let R be an array of the same size as Y_{com}, where its value would be 0 for corresponding element of Y_{obs} and 1 for Y_{mis}. We refer to R as the missingness, and its distribution, $P(R|Y_{com}, \phi)$, as missingness mechanism, where ϕ are the parameters of the R distribution. Based on the work of Rubin [1976], Little and Rubin [2019], missingness mechanism can be categorized as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). When the missingness is not related to any other variables in the data, such that $P(R|Y_{com}, \phi) = P(R|\phi)$, the missingness is categorized as MCAR. One example of MCAR might occurs when the blood sample is damaged in the lab. When the missingness is related to other fully-observed variables measured in the data, $P(R|Y_{com}, \phi) = P(R|Y_{obs}, \phi)$, MAR holds. A simple example of MAR is a survey where subjects over a certain age refuse to answer

a particular survey question and age is an observed covariate. While if neither of the above two cases hold, the missingness is relevant to levels of the missing variables, the missingness is called MNAR. For example, when subjects who have higher incomes are more likely to refuse to report the income in the survey, the income variable is MNAR.

The missingness is considered as random variable, and therefore need to be jointly modeled with the observed data. The joint model for the complete data and missingness can be expressed as

$$P(Y_{com}, R, \theta, \phi) = P(Y_{com}|\theta)P(R|Y_{com}, \phi)P(\theta, \phi).$$

If prior distributions for θ and ϕ are independent, such as $P(\theta, \phi) = P(\theta)P(\phi)$, and the MAR assumption holds, we refer the missingness mechanism as ignorable since the distribution of R can be ignored when making likelihood-based or Bayesian inference about θ.

The predictive distribution of Y_{mis} given observed values Y_{obs} and R is given in general by

$$P(Y_{mis}|Y_{obs}, R) \propto \int\int P(Y_{com}|\theta)P(R|Y_{com}, \phi)P(\theta, \phi)d\theta d\phi.$$

Under ignorability assumption, the predictive distribution for Y_{mis} would be reduced to $P(Y_{mis}|Y_{obs})$, which acts as the imputation model in multiple imputation.

3.2.2 Two-stage Multiple Imputation

Multiple imputation (MI) [Rubin, 1996, Schafer, 1997a, Rubin, 2004, Harel and Zhou, 2007] is a widely-used approach to deal with incomplete data in practice. MI replaces each missing value by $M > 1$ plausible values, resulting in M complete datasets.

Two-stage multiple imputation [Harel, 2009, Rubin, 2003, Reiter and Raghunathan, 2007] involves generating imputations through a two-step process to account for two different types of missing values. For example, missing responses and missing covariates. If the missing data are of two different types, it may be beneficial to treat them differently.

The general idea of two-stage multiple imputation is to impute one type of missing values M times. Then for each of the M imputed datasets, impute the other type of missing values N times, treating the imputations for the first type of missing values as fixed. Henceforth, this yields MN complete imputed datasets.

Suppose the missing data can be partitioned into two parts such that $Y = (Y_{obs}, Y^A_{mis}, Y^B_{mis})$. One can imagine an extended indicator random variable R^+ such that R^+ is 0 for elements of Y_{obs}, 1 for Y^A_{mis}, and 2 for Y^B_{mis}. Then to carry out two-stage multiple imputation, we first draw M values of Y^A_{mis} from

$$Y^{A(j)}_{mis} \sim P(Y^A_{mis}|Y_{obs}, R^+),$$

and then given each $Y^{A(j)}_{mis}$ draw N values of Y^B_{mis} from

$$Y_{mis}^{B(j,k)} \sim P(Y_{mis}^{B}|Y_{obs}, Y_{mis}^{A(j)}, R^{+}).$$

Under some ignorability condition [Harel and Schafer, 2009], the predictive distributions can be reduced to $Y_{mis}^{A(j)} \sim P(Y_{mis}^{A}|Y_{obs})$ and $Y_{mis}^{B(j,k)} \sim P(Y_{mis}^{B(j,k)}|Y_{obs}, Y_{mis}^{A(j)})$.

After imputing the missing values and analyzing MN complete datasets, results are combined following Shen's rules [Shen, 2000]. Let Q be the measure of interest, and U its variance. For each of the MN complete datasets, we obtain estimate of Q and associated squared standard error, $(\hat{Q}^{(j,k)}, U^{(j,k)})$, $j = 1, \cdots, M, k = 1, \cdots, N$. The overall estimate is

$$\bar{Q} = \frac{1}{MN}\sum_{j=1}^{M}\sum_{k=1}^{N}\hat{Q}^{(j,k)} = \frac{1}{M}\sum_{j=1}^{M}\bar{Q}_{(j.)},$$

where $\bar{Q}_{j.} = \frac{1}{N}\sum_{k}\hat{Q}^{(j,k)}$. The uncertainty in Q arises from three components: (1) the estimated complete-data variance $\hat{U}_{..} = \frac{1}{MN}\sum_{j=1}^{M}\sum_{k=1}^{N}\hat{U}^{(j,k)}$; (2) the between-nest imputation variance $U_{\text{between}} = \frac{1}{M-1}\sum_{j=1}^{M}(\bar{Q}_{j.} - \bar{Q}_{..})^2$; and (3) the within-nest imputation variance $U_{within} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N-1}\sum_{k=1}^{N}(\hat{Q}^{(j,k)} - \bar{Q}_{j.})^2$. The total variance is calculated by :

$$U_{\text{total}} = \bar{U}_{..} + (1 - \frac{1}{N}U_{within}) + (1 + \frac{1}{M}U_{\text{between}}).$$

It has been shown that $\sqrt{U_{\text{total}}}(Q - \bar{Q}_{..})$ follows a t-distribution with degree freedom being $\frac{1}{M(N-1)}(\frac{(1-1/N)U_{\text{within}}}{U_{\text{total}}})^2 + \frac{1}{M-1}(\frac{(1+1/M)U_{\text{between}}}{U_{\text{total}}})^2$.

3.3 Multiple Imputation for Missing Covariates in Linear Mixed Model

Suppose we have a sample of l individuals, where each individual i have t measurements. This is a simple representation for longitudinal data where t represent time. However, this also can be set as clustered data where t represent the clusters. Here, we will consider t as time. For individual i, let $y_i = (y_{i1}, \cdots, y_{it})^T$ be a $(t \times 1)$ vector of response of interest. A total of $q + q^*$ time-varying covariates are measured for each of l individuals throughout the study. For the first q covariates, missing values occur for each covariates at some time points for some individual, whereas for remaining q^* covariates each covariate is fully observed for each individual at each time point. Let $Z_i = (z_{ijk})$ be a $t \times q$ covariate matrix containing q incompletely observed covariates (each column vector contains at least one missing value). Let $X_i = (x_{ijk})$ be a $t \times q^*$ covariate matrix containing q^* completely observed covariates. In addition, assume all time-independent covariates are completely-observed.

If all time-varying variables (y_i, Z_i, X_i) are complete, a linear mixed model is assumed for the analysis model as follows:

$$y_i = (1, Z_i, X_i)\beta_i + e_i, \tag{3.1}$$

$$\beta_i = (1, A_i)\beta + u_i.$$

Where e_i is a $(t \times 1)$ vector of random intra-individual errors, β_i is a $(s \times 1)$ vector with $s = 1 + q + q^*$. β is a $(k \times 1)$ vector of fixed effect with $k = p + 1$, A_i is an $(s \times p)$ matrix incorporating time-independent covariates, u_i is an $(s \times 1)$ vector of random effects associated with individual i. We assume that each element in e_i independently follows a normal distribution $N(0, \sigma^2)$, and u_i is distributed as $N_s(0, \Omega)$. Without missing observations, unknown parameters (β, σ, Ω) can be estimated via Maximum likelihood estimation (MLE).

When both y_i and Z_i are partly observed, we can consider multiple imputation for missing observations. Generally, multiple imputation imputes missing response and time-varying covariates based on appropriate imputation model, and then use the imputed datasets to fit the analysis model (equation 3.1). We first review the imputation model used in multivariate multiple imputation (multivariate MI) proposed by Schafer [1997b] in section 3.1. Then we discuss the imputation model for two-stage multiple imputation proposed in section 3.2. Within the scope of this work, we restrict the analysis with single incomplete response and multiple incomplete time-varying covariates.

3.3.1 Multivariate Multiple Imputation

Schafer proposed a multivariate mixed effects [Schafer, 1997b, Schafer and Yucel, 2002] model as the imputation model for missing response and time-dependent covariates in longitudinal setting. In principle, the multivariate mixed effects model is an extension of univariate mixed effect model, assuming multiple outcomes are of interest. When applied as the imputation model, the multivariate mixed effect model will include all time-variables which are partly-observed as outcomes needed of imputation. That is, both incomplete response and time-varying covariates would be on the left side of the equation. Specifically, under the same notation for analysis model (equation 3.1), the imputation model is given by:

$$(Y_i, Z_i) = (1, X_i)\alpha_i + c_i, \tag{3.2}$$

$$\alpha_i = (1, A_i)\alpha + d_i,$$

where (Y_i, Z_i) is a $(t \times r)$ matrix with $r = 1 + q$ of all variables for imputation, α_i is a $((1 + q^*) \times r)$ matrix, α is a $(k \times r)$ matrix of fixed effect, c_i is a $t \times r$ matrix of individual error, and d_i is a $((1 + q^*) \times r)$ random effect matrix. $\text{vec}(c_i)$ is assumed to distribute normally with mean being 0 and variance being Σ_c, $\text{vec}(d_i)$ is also assumed to follow a normal distribution with mean being 0 and variance being Σ_d.

It is noted that the model (equation 3.2) is merely used to impute missing values.

The interpretation of its parameters $(\alpha, \Sigma_c, \Sigma_d)$ is not of interest. The estimation of these parameters rely on MCMC algorithm which described in details in [Schafer and Yucel, 2002]. The R package "pan" [Zhao and Schafer, 2018] make it easy to implement multivariate multiple imputation.

Multivariate multiple imputation provides a convenient way to account for missing response and missing covariates simultaneously. It assumes a univariate linear mixed effect model for each column in the left side of equation (3.2) given other column. However, it is likely in practice that the relationship between some incomplete variables and fully-observed variables is nonlinear. For example, given a data of a binary outcome and continuous covariates, one would assume the nonlinear model for the incomplete outcome and linear model for incomplete covariates. In that case a simultaneous imputation model would be inappropriate. Further, the approach assumes the same missingness mechanism for all partly-observed variables, which may lead to misleading results when missing values in multiple variables are caused by different reasons.

3.3.2 Two-stage Multiple Imputation

The limitation of multivariate multiple imputation approach motivates us to propose two-stage multiple imputation approaches which are described in the following sections. In principle, the response y_i and time-varying covariates X_i in the analysis model (equation 3.1) are considered as two types of data. Thus we are going to impute them separately in two stages. The imputation model for X_i is referred to as covariate model,

whereas imputation model for y_i is referred to as the response model.

The benefit of two-stage multiple imputation over multivariate multiple imputation is that covariate model and response model could be tailored according to the research interest and data characteristic. For example, researchers can include different sets of predictors in the two imputation models, or can assume a linear model for covariate model while a nonlinear model for response imputation model. Further, it is often the case that missing proportions of response and covariates are different. The variables that have more observed values will contain more information which would be useful to impute the missing values. Thus two-stage multiple imputation in this case is beneficial by imputing such variables which has lower missing percentages in the first stage. Last, proposed approaches also facilitate imputation when covariates and response subject to different missing mechanisms. This can be done, for example, by simulating non-ignorable missing observations in the first stage and ignorable missing values in the second stage, which may help simplify the imputation model and thus release the computational complexity usually associated with non-ignorable imputation.

There are two alternatives ways to implement two-stage MI which the order of imputation differs: we denote the approach which imputes the covariates first as " two-stage-covariates-response", and which imputes the response first as " two-stage-response-covariates".

Two-Stage-Covariates-Response

Denote $y = (y_1, \cdots, y_l)$, $X = (X_1, \cdots, X_l)$, $Z = (Z_1, \cdots, Z_l)$. Suppose we impute the missing covariates in the first stage, and missing response in the second stage given the imputed value of missing covariates. That is, we first assume a conditional distribution for the incompletely observed covariates given the completely observed covariates $(Z|X)$, which is the covariate model. Then we assume a conditional distribution $(y|X, Z)$ as the response model.

In the longitudinal data setting, imputation models in each of two stages are given by:

- imputation model at first stage (covariate model):

$$Z_i = (1, X_i)B_i + \epsilon_i \tag{3.3}$$

$$B_i = (1, A_i^*)B + V_i$$

- imputation model at second stage (response model):

$$y_i = (1, Z_i, X_i)\beta_i + e_i$$

$$\beta_i = (1, A_i)\beta + u_i$$

where B_i is a $(q^* + 1) \times q$ matrix, A_i^* is a $(q^* + 1) \times p^*$ matrix incorporating time-independent covariates, B is a $p^* \times q$ matrix of fixed parameters, ϵ_i is a $t \times q$ matrix which each of m rows is independently distributed as $N_q(0, \Sigma_\epsilon)$ and V_i is a $q^* \times q$ random matrix, distributed as $\text{vec}(V_i) \sim N_{q^* \times q}(0, \Psi)$ independently for $i = 1, \cdots, l$.

Denote $y_i = (y_{i,obs}, y_{i,mis})$ and $Z_i = (Z_{i,obs}, Z_{i,mis})$. The implementation of two-stage-covariates-response is generally given by:

- imputation at first stage:

  drawing m independent samples $Z_{i,mis}^{(1)}, \cdots, Z_{i,mis}^{(M)}$ from $P(Z_{i,mis}|Z_{i,obs}, X_i)$

- imputation at second stage:

  given each $Z_{i,mis}^{(k)}, k = 1, \cdots, M$, drawing $y_{i,mis}^{(k,1)}, \cdots, y_{i,mis}^{(k,N)}$ from

  $P(y_{i,mis}|y_{i,obs}, Z_{i,obs}, Z_{i,mis}^{(k)}, X_i)$.

Gibbs sampler is used to draw values from the multidimensional distribution $P(Z_{i,mis}|Z_{i,obs}, X_i)$ and $P(y_{i,mis}|y_{i,obs}, Z_{i,obs}, Z_{i,mis}^{(j)}, X_i)$. Let $Z_{obs} = (Z_{1,obs}, \cdots, Z_{l,obs})$, $Z_{mis} = (Z_{1,mis}, \cdots, Z_{l,mis})$, $V = (vec(V_1), \cdots, vec(V_l))^T$. Denote the collection of all unknown parameters in covariate model as $\theta_z = (B, \Sigma_\epsilon, \Psi)$. The prior distributions for θ_z are assumed as:

$$P(B) \propto \text{constant},$$

$$\Sigma_\epsilon \sim W^{-1}(\gamma_1, \Lambda_1),$$

$$\Psi \sim W^{-1}(\gamma_2, \Lambda_2),$$

where $W^{-1}(\cdot)$ denotes the inverted Wishart distribution. In choosing the hyperparameters, we may consider $\gamma_1^{-1}\Lambda_1^{-1}$ as a prior guess for Σ_ϵ and $\gamma_2^{-1}\Lambda_2^{-1}$ as a prior guess for Ψ. Under these priors, we can apply the Bayes theorem to obtain the posterior distributions for $\theta_z, V_i, Z_{i,mis}$:

Updating B

The posterior distribution of B is a multivariate normal with mean being $\hat{B}$, and covariance matrix being $\Sigma_\epsilon \otimes \left(\sum_{i=1}^{l} W_i^T W_i\right)^{-1}$, where $\hat{B} = \left(\sum_{i=1}^{l} W_i^T W_i\right) \sum_{i=1}^{l} W_i^T (Z_i - (1, X_i)V_i), \quad W_i = (1, X_i)(1, A_i^*)$.

Updating Ψ and Σ_ϵ

The full conditional distribution of Ψ is an inverse Wishart distribution $(\gamma_2 + l, (\Lambda_2^{-1} + \sum_{i=1}^{l} V_i^T V_i)^{-1})$, and Σ_ϵ is also an inverse Wishart $(\gamma_1 + lt - (q^* + q), (\Lambda_1^{-1} + \sum_{i=1}^{l} \hat{\epsilon_i}^T \hat{\epsilon_i})^{-1})$, where $\hat{\epsilon_i} = Z_i - W_i\hat{B} - (1, X_i)^* V_i, \quad i = 1, \cdots, l$.

Updating $vec(V_i)$

The full conditional distribution of stacked columns $vec(V_i)$ follows a multivariate normal distribution $(vec(\hat{V_i}), D_i)$, where $vec(\hat{V_i}) = D_i(\Sigma_\epsilon^{-1} \otimes (1, X_i)^T) vec(Z_i - W_i B)$, $D_i = \{\Psi^{-1} + (\Sigma_\epsilon^{-1} \otimes (1, X_i)^T (1, X_i))\}^{-1}$

Updating $Z_{i,mis}$

The missing values in time-dependent covariates, $Z_{i,mis}$, can be drawn from its full

conditional distribution $N\{W_iB + (1, X_i)V_i, \Sigma_\epsilon \otimes I_t\}$

Repeating above steps from a starting value $(\theta_z^{(0)}, V^{(0)})$ would yield stochastic sequences $Z_{mis}^{(j)} : j = 1, 2, ...$ which converge in distribution to $P(Z_{mis}|Z_{obs}, X)$. Thus, for a sufficiently large value of j, we can regard $Z_{mis}^{(j)}$ as an approximate draw from $P(Z_{mis}|Z_{obs}, X)$.

In the second stage, given a version of $Z_{mis}^{(k)}$ from the marginal distribution for Z_{mis}, we are able to draw y_{mis} from conditional distribution $P\left(y_{mis}|y_{obs}, X, Z_{obs}, Z_{mis}^{(k)}\right)$. This process is equivalent to the imputation of single response assuming all covariates complete.

Two-Stage-Response-Covariates

We further propose to a two-stage multiple imputation which imputes the incomplete response in the first stage, that is, two-stage-response-covariates. Two imputation models are given by:

- imputation model at first stage (response model):

$$y_i = X_i\eta_i + e_i^* \tag{3.4}$$

$$\eta_i = C_i\eta + v_i$$

- imputation model at second stage (covariate model):

$$Z_i = (y_i, X_i)H_i + E_i, \tag{3.5}$$

$$H_i = C_i^* H + U_i,$$

where η_i is a $(q^* \times 1)$ vector, C_i is a $(q^* \times r)$ design matrix which include fully-observed time-independent covariates, η is a $(r \times 1)$ vector representing fixed parameters for imputation model of y_i, e_i^* is a $(l \times 1)$ vector each component independently distributed as $N(0, \sigma^{*2})$, and v_i is a $(q^* \times 1)$ vector where each row follows $N(0, \sigma_v^2)$. For the covariate model, H_i is a $((1 + q^*) \times q)$ matrix, C_i^* is a $((1 + q^*) \times r)$ design matrix, H is a $(r \times q)$ matrix containing fixed parameters for imputation of Z_i. E_i is a $(l \times q)$ matrix, where each row distributed as $N_q(0, \Sigma^*)$, and U_i is a $((1 + q^*) \times q)$ matrix, where $\text{vec}(U_i) \sim N_{1+q^*+q}(0, \Phi^*)$.

Instead of imputing covariates first, we will first draw M independent samples $y_{i,mis}^{(1)}, \cdots, y_{i,mis}^{(M)}$ from $P(y_{i,mis}|X_i, C_i)$. Then given $y_{i,mis}^{(k)}, k = 1, \cdots, M$, we further draw $Z_{i,mis}^{(k,1)}, \cdots, Z_{i,mis}^{(k,N)}$ from $P(Z_{i,mis}|X_{i,mis}, y_{i,mis}^{(k)}, C_i, C_i^*)$. Similarly, this implementation requires prior distribution for all unknown parameters. We will follow similar scheme as we did with two-stage-covariates-response to run the Gibbs sampler to gain the posterior distribution for missing values.

3.4 Simulation Studies

We aim to compare the performance across various scenarios of following approaches: multivariate multiple imputation method (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariates (TSMI-R).

We consider a sample size $l = 500$ and planned visit times $t = 4$. For each individual $i = 1, \cdots, 500$ at each visit $j = 1, \cdots, 4$, we generate three time-varying covariates: $(X_{1ij}, X_{2ij}, X_{3ij})$, which follow a multivariate distribution independently with mean being 0, variance being 1 . The correlation between X_{1ij}, X_{2ij} equals to 0.1, the correlation between X_{1ij} and X_{3ij}, and correlation between X_{2ij} and X_{3ij} are both equal to 0.7. The response is generated for such that

$$y_{ij} = \beta_0 + u_{0j} + \beta_1 * X_{1ij} + \beta_2 * X_{2ij} + \beta_3 * X_{3ij} + e_i \tag{3.6}$$

where u_{0i} is independently distributed as $N(0, \psi)$ for $i = 1, \cdots, 500$, and e_i independently distributed as $N(0, \sigma^2)$ across j and i. We set $\beta = (1, 1, 2, 0.5)$.

We perform a set of simulations for each combination of missing scenario and variation scenario. First, we consider the scenario with missing complete at random (MCAR). We randomly delete some observations in y_{ij} with the ratio p_y, and observations in X_{1ij} and X_{2ij} with same ratio p_x. We vary the p_x and p_y to alter the proportion of missing values to different missingness: $(p_y = 0.2, p_x = 0.4)$, $(p_y = 0.4, p_x = 0.2)$,$(p_y = 0.4, p_x = 0.4)$, $(p_y = 0.6, p_x = 0.6)$. Second, we consider the scenario with missing at random

(MAR). We assume the missing proportion depend on the visit time. The latter visit time, the larger missing proportion. Same combinations of p_x and p_y as with MCAR are used in case of MAR. Third, we consider small number of imputation, $M = 5, N = 2$, and large number of imputation $M = 25, N = 2$ for two-stage multiple imputation, where corresponding number of imputation for multivariate multiple imputation is $M * N$. Fourth, we consider small, and large variation by taking $\psi \in (1, 4)$ and $\sigma^2 \in (1, 4)$.

In each scenario, 500 datasets are generated. The target parameters, $\beta_0, \beta_1, \beta_2, \beta_3$ are estimated from each data set. For each kth component of β, given its estimate $\hat{\beta}_k$ and associated error s_k, we calculate the bias as $\hat{\beta}_k - \beta_k$, standardized bias $\frac{\text{bias}_k}{s_k} \times 100\%$, root mean squared error (RMSE) $\sqrt{\text{bias}_k^2 + s_k^2}$, 95% confidence interval coverage rate and confidence interval width. Then average of those statistics are taken across 500 generated datasets. Standardized bias and RMSE represent both accuracy and precision of an estimators, thus act as important measures of practical utility. One rule of thumb we used is that standardized bias exceeding 40% is considered to be severe [Collins et al., 2001]. Furthermore, the mean estimate of ψ and σ^2 are also calculated to assess the performance in terms of variance components.

The R package pan [Zhao and Schafer, 2018] is used to implement three multiple imputation approach, while analysis of each imputed dataset is fitted using the R package lme4 [Bates et al., 2015].

3.4.1 Simulation Results

Tables 6 (M=5, N=2) and 7 (M=50, N=2) present the detailed results in the setting of sample size 500, $\sigma = 1, \phi = 1$, which represents medium variation in random intercept and individual residual. Results in scenarios with sample size 100, small and large variations in random component present similar patterns and thus are not shown here. Those results could be found in Tables 5.3.1, 15 and 16 in the appendix.

Missing Completely at Random

Under missing completely at random (MCAR) with $p_y = 0.2, p_x = 0.4, M = 5, N = 2$, multivariate multiple imputation and two-stage-covariates-response both lead to large standardized bias. The standardized bias of two-stage-covariates-response are worst in terms of $\beta_1, \beta_2, \beta_3$. Instead, standardized bias and RMSE obtained from two-stage-response-covariates are both low (standardized bias $< 25\%$ and RMSE < 0.129 for all coefficients). In addition, the multivariate multiple imputation and two-stage-response-covariates achieve nominal coverage rate for 95% confidence interval, while two-stage-covariates-response shows lower coverage rate for β_1 and β_2 (78% for both). This implies that, when response suffers less than covariates with respect to missingness, it is beneficial to implement a two-stage-multiple imputation, first imputing parts with smaller missing percentage.

The results obtained from scenario with $p_y = 0.4, p_x = 0.2, M = 5, N = 2$ can lead to similar conclusions that two-stage multiple imputation outperforms by imputing the

parts of data with less missing values first. When covariates have less missing values compared with the response, the two-stage-covariates-response outperforms with smaller standardized bias, smaller RMSE, shorter confidence interval and nominal confidence interval coverage rate.

In the scenario with same missingness such as $p_y = 0.4, p_x = 0.4, M = 5, N = 2$, all three approaches achieve moderate bias with standardized bias all approximately less than 40%, except the standardized bias of $\beta_2 = 60.52\%$ by two-stage-covariates-response. The 95% confidence interval coverage rate calculated by three approaches are all around nominal level. Two-stage-response-covariates yield relatively smallest standardized bias (between -9.56 to 9.56), moderate confidence interval width, and moderate RMSE value, only RMSE of β_1 and β_4 are slightly larger than those obtained from other two approaches.

When the proportion of missingness increase from 0.4 to 0.6, standardized bias, RMSE and confidence interval width increase for all three approaches, whereas the confidence interval coverage rate decrease. This is expected because more missing observations in the data will lead to larger bias and larger estimated standard error for estimated coefficients. Notably, multivariate multiple imputation and two-stage-covariates-response lead to several severe standardized bias (as high as 84%). Again, the two-stage-response-covariates achieve acceptable standardized bias. Although the associated RMSE is slightly higher than those of other two approaches, the difference is very small, especially compared with pronounced improvement in terms of standardized

bias.

Last, the number of imputation also has effect on the results. As the number of imputation increase from $M = 5, N = 2$ (Table 6) to $M = 25, N = 2$ (Table 7), the magnitude of standardized bias, RMSE and confidence interval width decrease given same missingness pattern and missingness mechanism. The reduction is especially notable for two-stage multiple imputation with high proportion of missingness such as 60%: for instance, the magnitude of standardized bias for β_2 decreases from 84.33 to 57.90 under two-stage-covariates-response, and 7.44 to 1.17 under two-stage-response-covariates.

Missing at Random

In general, missingness at random results in more biased estimates compared with missingness completely at random, with larger standardized bias, RMSE and wider confidence intervals (Tables 6 and 7).

In the scenario with $p_y = 0.2, p_x = 0.4$, only two-stage-response-covariates achieve acceptable standardized bias for all coefficients (-8.07% to 41.51%), while other two approaches have several large standardized bias (as high as -492.80%). At the same time, two-stage-response-covariates also obtain smaller RMSE. The confidence interval coverage rate is around the nominal level 95%, and interval widths are also acceptable compared with other two approaches. Thus two-stage-response-covariates is again preferred when the response suffers smaller missing proportions.

In the scenario with covariates suffer smaller missing proportions, such as $p_y =$

$0.4, p_x = 0.2$, two-stage-covariate-response approach instead outperforms, with smallest standardized bias (-6.58% to 43.47%), smallest RMSE (0.062 to 0.142) and shortest confidence interval (0.226 to 0.398) . On the contrary, the two-stage-response-covariates performs poorly with extremely high bias and RMSE, as well as lower-than-nominal confidence interval coverage rate. We therefore arrive at similar conclusions we had with MCAR case, that two-stage-covariates-response is advantageous when time-varying covariates have smaller missing percentages.

When response and time-varying covariates suffer from same percentages missing values, the three approaches have similar performance. All three approaches achieve nominal confidence interval coverage rate and similar confidence interval width. However, two-stage-response-covariates still outperforms the other methods by obtaining smaller bias and smaller RMSE. It is also the only approach that show acceptable standardized bias which are all less than or around 40%.

Table 6: Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval width (CI Width) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with ($\sigma = 1, \psi = 1, M = 5, N = 2$)

Mechanism	Missing		Bias			RMSE			Cov			CI Width		
MCAR			MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R
	$p_y = 0.2$	β_0	8.18	-3.28	3.03	0.059	0.061	0.058	0.99	0.99	0.99	0.221	0.227	0.212
	$p_x = 0.4$	β_1	26.59	**-259.92**	9.31	0.095	0.265	0.096	0.99	0.78	0.99	0.298	0.373	0.308
		β_2	**94.29**	**-484.38**	24.8	0.108	0.476	0.092	0.95	0.78	0.98	0.282	0.384	0.292
		β_3	**-50.85**	**387.56**	-11	0.133	0.522	0.129	0.99	0.92	0.99	0.399	0.517	0.415
	$p_y = 0.4$	β_0	5.3	5.1	-5.67	0.063	0.063	0.068	0.99	0.99	0.99	0.228	0.222	0.239
	$p_x = 0.2$	β_1	13.87	12.71	**-309.72**	0.097	0.097	0.302	0.97	0.97	0.81	0.305	0.292	0.362
		β_2	**68.52**	42.78	**-612.03**	0.11	0.111	0.569	0.96	0.90	0.82	0.301	0.284	0.363
		β_3	-37.58	-36.35	**465.48**	0.14	0.141	0.609	0.98	0.95	0.89	0.420	0.398	0.505
	$p_y = 0.4$	β_0	7.92	6.71	1.73	0.064	0.064	0.061	0.99	0.99	0.99	0.228	0.222	0.215
	$p_x = 0.4$	β_1	15.45	21.19	9.18	0.098	0.096	0.101	0.99	0.98	0.99	0.309	0.292	0.322
		β_2	**48.17**	**60.52**	6.86	0.109	0.111	0.105	0.96	0.92	0.98	0.303	0.287	0.321
		β_3	-35.42	-42.86	-9.56	0.14	0.137	0.143	0.98	0.96	0.99	0.426	0.400	0.446
	$p_y = 0.6$	β_0	7.95	13.69	4.16	0.076	0.074	0.077	0.99	0.99	0.99	0.255	0.240	0.247
	$p_x = 0.6$	β_1	16.82	29.41	10.5	0.136	0.129	0.136	0.97	0.93	0.96	0.400	0.366	0.396
		β_2	**55.86**	**84.33**	-7.44	0.144	0.14	0.148	0.95	0.90	0.93	0.392	0.363	0.410
		β_3	-37.79	**-48.08**	-13.57	0.19	0.186	0.195	0.97	0.93	0.96	0.547	0.506	0.561
MAR	$p_y = 0.2$	β_0	-9.49	-12.71	-8.07	0.060	0.063	0.058	0.99	0.99	0.99	0.225	0.231	0.213
	$p_x = 0.4$	β_1	42.98	**-247.40**	39.94	0.095	0.256	0.094	0.99	0.73	0.99	0.297	0.377	0.307
		β_2	**96.90**	**-492.80**	41.51	0.107	0.476	0.094	0.94	0.76	0.98	0.278	0.376	0.286
		β_3	**-60.26**	**383.13**	-36.30	0.133	0.514	0.132	0.98	0.87	0.99	0.397	0.513	0.411
	$p_y = 0.4$	β_0	-8.81	-6.58	-7.64	0.063	0.062	0.070	0.99	0.99	0.99	0.232	0.226	0.247
	$p_x = 0.2$	β_1	32.81	43.47	**-288.38**	0.098	0.098	0.282	0.98	0.97	0.75	0.304	0.291	0.360
		β_2	**75.22**	36.22	**-591.76**	0.110	0.110	0.553	0.93	0.91	0.73	0.299	0.284	0.365
		β_3	**-48.85**	-40.93	**447.35**	0.142	0.142	0.584	0.97	0.94	0.82	0.419	0.398	0.502
	$p_y = 0.4$	β_0	-8.37	-6.68	-4.98	0.063	0.063	0.062	0.99	0.99	0.99	0.232	0.227	0.218
	$p_x = 0.4$	β_1	34.60	**48.69**	32.99	0.098	0.100	0.098	0.99	0.97	0.99	0.305	0.291	0.319
		β_2	**73.45**	**89.23**	35.60	0.109	0.113	0.104	0.95	0.91	0.97	0.302	0.286	0.314
		β_3	**-47.93**	**-60.48**	-34.66	0.140	0.143	0.140	0.98	0.95	0.98	0.423	0.400	0.440
	$p_y = 0.6$	β_0	-21.49	-16.56	-15.36	0.066	0.066	0.065	0.99	0.99	0.99	0.241	0.236	0.231
	$p_x = 0.6$	β_1	**51.07**	**65.05**	41.88	0.124	0.121	0.120	0.97	0.93	0.94	0.371	0.340	0.363
		β_2	**69.42**	**84.99**	26.24	0.131	0.129	0.120	0.93	0.90	0.95	0.362	0.326	0.356
		β_3	**-57.50**	**-72.29**	-44.11	0.175	0.174	0.172	0.97	0.93	0.97	0.509	0.464	0.508

Table 7: Simulation Results - comparisons of standardized bias (%) and RMSE of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with ($\sigma = 1, \psi = 1, M = 50, N = 2$)

Mechanism	Missing		Bias			RMSE			Cov			CI	Length	
MCAR			MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R
	$p_y = 0.2$	β_0	7.54	-3.09	3.01	0.059	0.061	0.057	0.99	0.99	0.99	0.220	0.226	0.210
	$p_x = 0.4$	β_1	25.31	**-258.75**	9.25	0.093	0.262	0.094	0.99	0.81	0.99	0.295	0.366	0.301
		β_2	**93.91**	**-480.28**	22.64	0.107	0.475	0.092	0.95	0.80	0.99	0.278	0.371	0.289
		β_3	**-49.98**	**385.52**	-9.47	0.131	0.52	0.127	0.99	0.92	0.99	0.395	0.503	0.407
	$p_y = 0.4$	β_0	4.82	2.96	-2.92	0.062	0.062	0.067	1.00	1.00	1.00	0.226	0.221	0.237
	$p_x = 0.2$	β_1	12.28	11.22	**-306.26**	0.095	0.095	0.301	1.00	0.97	0.85	0.300	0.285	0.357
		β_2	**66.58**	38.38	**-610.78**	0.106	0.106	0.568	0.96	0.92	0.87	0.296	0.281	0.358
		β_3	-34.57	-35.89	**461.68**	0.135	0.135	0.606	0.99	0.96	0.90	0.414	0.392	0.499
	$p_y = 0.4$	β_0	6.20	5.98	1.45	0.062	0.062	0.059	0.99	0.99	0.99	0.227	0.221	0.214
	$p_x = 0.4$	β_1	13.74	20.56	8.14	0.095	0.096	0.1	0.99	0.98	0.99	0.302	0.293	0.312
		β_2	46.14	**58.09**	5.5	0.106	0.109	0.102	0.97	0.93	0.99	0.299	0.287	0.312
		β_3	-33.04	-40.33	-8.35	0.135	0.137	0.139	0.99	0.97	0.99	0.417	0.399	0.434
	$p_y = 0.6$	β_0	5.06	3.85	3.48	0.074	0.074	0.072	0.99	0.99	0.99	0.252	0.239	0.238
	$p_x = 0.6$	β_1	15.06	17.54	9.45	0.133	0.129	0.135	0.98	0.94	0.97	0.393	0.364	0.391
		β_2	**55.88**	**57.90**	1.17	0.141	0.136	0.144	0.95	0.91	0.95	0.388	0.355	0.409
		β_3	-37.34	-37.74	-7.01	0.185	0.183	0.192	0.98	0.94	0.97	0.539	0.500	0.556
MAR	$p_y = 0.2$	β_0	-9.39	-12.33	-7.61	0.060	0.062	0.058	0.99	0.99	0.99	0.224	0.229	0.211
	$p_x = 0.4$	β_1	42.90	**-246.19**	39.32	0.094	0.248	0.094	0.99	0.78	0.99	0.292	0.362	0.298
		β_2	**96.63**	**-484.97**	40.38	0.107	0.466	0.092	0.93	0.78	0.98	0.274	0.367	0.282
		β_3	**-60.06**	**379.40**	-36.16	0.133	0.502	0.129	0.98	0.89	0.99	0.391	0.498	0.403
	$p_y = 0.4$	β_0	-8.15	-6.06	-5.84	0.062	0.062	0.068	0.99	0.99	0.99	0.231	0.224	0.244
	$p_x = 0.2$	β_1	27.40	41.40	**-281.61**	0.093	0.095	0.282	0.99	0.98	0.76	0.299	0.284	0.353
		β_2	**67.49**	31.10	**-591.28**	0.104	0.107	0.552	0.96	0.92	0.75	0.295	0.279	0.355
		β_3	-39.04	-31.63	**445.42**	0.132	0.135	0.584	0.99	0.96	0.85	0.413	0.389	0.496
	$p_y = 0.4$	β_0	-8.40	-6.67	-4.07	0.062	0.062	0.060	0.99	0.99	0.99	0.230	0.226	0.217
	$p_x = 0.4$	β_1	32.26	44.78	31.41	0.094	0.096	0.098	0.99	0.98	0.99	0.298	0.291	0.307
		β_2	**72.62**	**83.74**	34.10	0.106	0.108	0.099	0.97	0.94	0.99	0.296	0.286	0.304
		β_3	**-47.34**	**-58.19**	-34.10	0.135	0.138	0.138	0.99	0.97	0.99	0.413	0.397	0.427
	$p_y = 0.6$	β_0	-17.57	-11.68	-10.16	0.064	0.065	0.062	0.99	0.99	0.99	0.238	0.233	0.224
	$p_x = 0.6$	β_1	**50.59**	**64.40**	40.04	0.121	0.121	0.120	0.97	0.94	0.96	0.357	0.326	0.356
		β_2	**67.51**	**84.41**	23.02	0.128	0.128	0.119	0.94	0.92	0.96	0.353	0.322	0.355
		β_3	**-56.64**	**-71.87**	-43.04	0.171	0.171	0.168	0.97	0.94	0.97	0.494	0.451	0.497

3.5 Data Application

One of the exploratory objectives of the study is to evaluate the relationship between alcohol intake and the HIV disease progression. For analyses, we fit the linear mixed model for CD4 cell count. In light of correlation between self report and biological marker, we only include PEth as the measurement of alcohol intake in the model to avoid the problem of multicollinearity. The treatment of multicollinearity in the context of complete data has already been discussed in Chapter 3, while treatment of multicollinearity in the context of incomplete data would be discussed in the next chapter. Other explanatory variables include age, gender, religion, and HIV virus load. As the data suffers from a considerable amount of missing data that occurred after the baseline, we apply the proposed two-stage multiple imputation as well as the multivariate multiple imputation to handle those missing values.

The resulting point estimates, together with the corresponding standard deviations (std) and p-values for the regression coefficients are displayed in the Table 8. Multivariate multiple imputation approach shows an insignificant negative relationship between PEth and CD4 cell count (estimated coefficient of PEth is -0.011, associated p-value is 0.668); two-stage-covariates-response approach arrives the similar conclusion with estimated coefficient of PEth being -0.007 and p-value being 0.590; and estimated coefficient of PEth produced by two-stage-response-covariates is -0.016, with p-value 0.380. In sum, all approaches suggest evidence of a negative association between PEth and CD4 cell

count, however those associations are not significant. Thus we argue that heavy alcohol use has an impact on CD4 cell count decline, although such impact is not significant. This finding is consistent with current studies results [Hahn et al., 2018, Cagle et al., 2017, Weiser et al., 2014]. It is noted that the size of estimation and associated standard deviation differs a lot across three approaches. The size of point estimation with respect to PEth produced by two-stage-covariate-response is approximately half as much as those produced by other two approaches (0.007 compared with 0.011 and 0.016), where the estimated standard deviation for two-stage-covariates-response is also smaller than those for other two approaches (0.013 compared with 0.026 and 0.018).

Table 8: Application to the HIV study: estimated coefficients by fitting linear mixed model, using various approach to handle missing response and missing time-varying covariates

	MULT			TSMI-C			TSMI-R		
	Estimate	Std	p-value	Estimate	Std	p-value	Estimate	Std	p-value
(Intercept)	6.684	0.093	0.000	6.627	0.091	0.000	6.665	0.099	0.000
PEth	-0.011	0.026	0.668	-0.007	0.013	0.590	-0.016	0.018	0.380
time	-0.005	0.002	0.014	-0.005	0.002	0.001	-0.004	0.002	0.024
age	-0.000	0.002	0.809	-0.000	0.001	0.929	-0.0001	0.001	0.701
religion	-0.026	0.018	0.135	-0.018	0.017	0.320	-0.016	0.018	0.375
sex	-0.024	0.036	0.511	-0.026	0.033	0.433	-0.027	0.040	0.504
virus load	-0.082	0.015	0.000	-0.072	0.018	0.000	-0.079	0.016	0.000

3.6 Discussions

In this chapter, we proposed a two-stage multiple imputation method for incomplete time-dependent covariates in linear mixed effects models, which imputes the missing response and missing time-varying covariates at two stages. We presented the simulation studies where we compare the performance yielded by two-stage multiple imputation and multivariate multiple imputation. We further consider some modifications of multivariate multiple imputation: given multivariate imputation of response and covariates, we impose the same missingness on the response and implement the multiple imputation as well as expectation maximum algorithm to estimate the coefficients in the linear mixed model. It turned out that such modified multivariate multiple imputation fail to yield better performance in terms of bias and mean squared error. In sum, our simulation studies show that the proposed method is more reliable than other methods in the sense that it provides the smallest bias and the smallest mean-squared errors in the estimates of covariate coefficients.

When we applied the three missing covariate methods to our HIV data, consistent results were obtained from different methods. Thus, we conclude that alcohol use are negatively related to the CD4 cell count (response). This result seems to be consistent with the findings in [Weiser et al., 2014, Cagle et al., 2017, Hahn et al., 2018].

The proposed two-stage multiple imputation offers a flexible method to tackle incomplete response and covariates. Outcome models and covariates models could be

cautiously chosen separately in a way of accounting for types of outcome variables and covariates variables, as well as assumptions of missingness mechanism.

Several extension of proposed two-stage multiple imputation may be of interest. First, linear mixed models have been developed for continuous outcomes and covariates in this chapter, generalized linear mixed model can be proposed as a natural way to handle categorical variables. Second, the assumptions on missing response and missing covariates are restricted to case of missing completely at random (MCAR) and missing at random (MAR) in this chapter. In many analyses, however, such assumptions may fail to hold. Missing data can often arise from a more complicated mechanism such as missing not at random (MNAR). In these cases, the use of non-ignorable imputation methods such as selection models [Diggle and Kenward, 1994, Molenberghs et al., 1997] or pattern-mixture models [Little, 1993, 1994, 1995, Daniels and Hogan, 2000, Roy, 2003] should be considered. Extensions to these scenarios will be the topic of future research.

Finally, it is assumed in this chapter all covariates are independent. In case of partly-observed correlated covariates, we will consider the incorporation of Bayesian shrinkage priors we propose in last section into the two-stage multiple imputation. This will done in next section.

Chapter 4

Application of Two-stage Multiple Imputation for Missing Correlated Time-varying Covariates

4.1 Introduction

Despite the current increased attention devoted to the modeling of incomplete longitudinal data, limited attention has been directed toward handling the issues of incomplete correlated time-varying covariates in the context of linear mixed model. Several authors proposed the model accounting for incomplete outcome and time-varying covariates [Roy and Lin, 2002, Stubbendick and Ibrahim, 2003, Roy and Lin, 2005, Parzen et al., 2006, Stubbendick and Ibrahim, 2006]. However, none of those approaches take into count the correlation between incomplete covariates. A notable exception is Schafer and Yucel's work where they proposed the multivariate linear mixed effect model [Schafer, 1997a] to impute all missing variables jointly [Schafer and Yucel, 2002]. They suggested a

ridge-like prior to handle the correlated incomplete covariates.

The use of ridge-like prior for incomplete data is described in [Schafer, 1997a], which is a diagonal prior distribution on random components for linear mixed model, functioning similar with ridge regression in the frequentest framework.

The purpose of this chapter is to propose a two-stage multiple imputation for jointly modeling time-varying covariates subject to high correlation and missingness at random. In doing so, we incorporate the Bayesian shrinkage approach for multicollinearity into the two-stage multiple imputation. In the first stage, the missing covariates are imputed with the multivariate mixed model; in the second stage, based on those imputed covariates, the missing response is imputed through a linear mixed model where the correlated covariates are proceeded with Bayesian shrinkage method.

The remainder of this chapter is organized in the following way. In the next section 4.2, we describe the proposed approach that can be used to model missing correlated time-varying covariates in the framework of linear mixed model. In section 4.3, simulation studies are carried out to evaluate the performance of proposed approach. An application of the proposed approach to real data is provided in section 4.4. Finally, comments and discussion are presented in Section 4.5.

4.2 Two-Stage Multiple Imputation for Incomplete Correlated Time-varying Covariates

In this chapter, we extend the work we develop in Chapter 3 for partly-observed response and time-varying covariates. Henceforth, we continue to use the data setting described in section 3.3, and linear mixed model (equation 3.1) as the analysis model. In light of related covariates, in the imputation process, we introduce the Bayesian shrinkage priors which we developed in Chapter 2.

4.2.1 Specification of Imputation Models

Missing values in time-varying outcome and correlated time-varying covariates present statistical methodological challenges. We seek a flexible imputation engine under multiple imputation to handle the missing outcome and missing correlated time-dependent covariates. To this end, we propose a two-stage multiple imputation by introducing the Bayesian shrinkage priors. In the first stage, we use a covariate model as the imputation model to fill in missing values in time-varying covariates; in the second stage, we predict the missing values in response through a response imputation model by introducing the Bayesian shrinkage priors to account for correlated covariates. Within the scope of this chapter, the model we develop can handle missing at random (MAR) missingness.

Covariate Model in First Stage

During the first stage, we use the multivariate mixed effect model proposed by Schafer [1997b] and Schafer and Yucel [2002] as the imputation model. In principle, the multivariate mixed effects model is an extension of univariate mixed effect model, assuming multiple outcomes are of interest. When applied as the covariate imputation model, the multivariate mixed effect model will include all time-varying covariates which are partly-observed as outcomes need imputation. The covariate model is given by:

$$Z_i = (1, X_i)B_i + \epsilon_i, \tag{4.1}$$

$$B_i = (1, A_i^*)B + U_i,$$

where B_i is a $(q^* + 1) \times q$ matrix, A_i^* is a $(q^* + 1) \times p^*$ matrix incorporating time-independent covariates, B is a $p^* \times q$ matrix of fixed parameters, ϵ_i is a $m \times q$ matrix which each of t rows is independently distributed as $N_q(0, \Sigma_\epsilon)$ and U_i is a $q^* \times q$ random matrix, distributed as $\text{vec}(U_i) \sim N_{q^* \times q}(0, \Psi)$ independently for $i = 1, \cdots, l$. Last, denote A^* and U as appropriately defined matrices representing the concatenation of the corresponding variables over every individual i

Denote the collection of all unknown parameters in covariate model as $\theta_z = (B, \Sigma_\epsilon, \Psi)$. The prior distributions for θ_z are assumed to be:

$$B \sim \text{flat prior},$$

$$\Sigma_\epsilon \sim W^{-1}(\gamma_1, \Lambda_1),$$

$$\Psi \sim W^{-1}(\gamma_2, \Lambda_2),$$

where $W^{-1}(\cdot)$ denotes the inverted Wishart distribution. Given above priors, we are able to simulate the posterior distributions for $\theta_z, U_i, Z_{i,mis}$ using the Gibbs sampler.

Updating B

The fixed effect B follows a multivariate normal distribution with mean being $\hat{B}$, and variance being $\Sigma_\epsilon \otimes \left((A^*)^T X^T X A^*\right)^{-1}$, where $\hat{B} = \left((A^*)^T X^T X A^*\right)(A^*)^T X^T (Z - XA^*U)$.

Updating Ψ and Σ_ϵ

The posterior distribution of Ψ is an inverse Wishart: $W^{-1}\{\gamma_2 + l, (\Lambda_2^{-1} + U^T U)^{-1}\}$, and Σ_ϵ is also an inverse Wishart: $W^{-1}\{\gamma_1 + lt - (q^* + q), (\Lambda_1^{-1} + \sum_{i=1}^{l} \hat{\epsilon_i}^T \hat{\epsilon_i})^{-1}\}$, where $\hat{\epsilon_i} = Z_i - X_i A_i^* \hat{B} - X_i A_i^* U_i, i = 1, \cdots, l$.

Updating U_i

The stacked U_i follows a multivariate normal distribution, such as $vec(U_i) \sim N(vec(\hat{U_i}), D_i)$ where $vec(\hat{U_i}) = A_i(\Sigma_\epsilon^{-1} \otimes W_i^T) vec(Z_i - W_i B), W_i = (1, X_i), D_i = \{\Psi^{-1} + (\Sigma_\epsilon^{-1} \otimes W_i^T W_i)\}^{-1}$.

Updating $Z_{i,mis}$

The missing values contained in Z_i can be simulated via its posterior distribution,

which is a multivariate normal distribution, with mean being $X_i A_i^*(B+U_i)$, and variance being $\Sigma_\epsilon \otimes I_t$.

Response Model in Second Stage

In the second stage, given the filled-in values in Z_i, we impute the response using the response imputation model which is equivalent to the analysis model, equation (3.1).

$$y_i = (1, Z_i, X_i)\beta_i + e_i,$$

$$\beta_i = (1, A_i)\beta + u_i.$$

The unknown parameters in response imputation model can be denoted as $\theta_y = (\beta, \sigma, \Omega)$. In light of high correlation in covariates, Z_i and X_i, we introduce the Bayesian shrinkage priors by assigning the normal-gamma priors on β.

$$\beta_i|\Phi_i \sim N(0, \Phi_i), \quad \Phi_i|\lambda, \gamma \sim \Gamma(\lambda, \frac{1}{2\gamma^2}),$$

$$\lambda \sim \exp(0.1),$$

$$2\lambda\gamma^2 \sim \mathrm{IG}(2, O),$$

$$\sigma \sim \text{flat prior},$$

$$\sigma_u \sim \text{flat prior},$$

where $O = \frac{1}{1+q+q^*} \sum_{k=0}^{q+q^*} \hat{\beta}_k{}^2$. Let $G_i = (1, Z_i, X_i)$, $H_i = G_i(1, A_i)$, G, H be defined matrices representing the concatenation of the corresponding variables over all individual i, $V = \text{var}(y) = \sigma_u GG^T + \sigma^2 I$. By applying the Bayes theorem, we are able to derive the full conditional for all unknown parameters, θ_y, u_i and $y_{i,mis}$.

Updating β

The full conditional distribution of β follow a normal distribution with mean $(H^T V^{-1} H + \sigma^2 \Lambda)^{-1} H^T V^{-1} y$ and variance $\sigma^2 (H^T V^{-1} H + \sigma^2 \Lambda)^{-1}$, where

$$\Lambda = \text{diag}(\frac{1}{\Phi_0}, \frac{1}{\Phi_1}, \cdots, \frac{1}{\Phi_{q+q^*}}).$$

Updating σ^2 and σ_b^2

The full conditional distribution of σ^2 is $IG(c, d)$, with $c = \frac{l}{2}$ and $d = (y - H\beta - Zb)'(y - X\beta - Gu)/2$, while the full conditional distribution of σ_b^2 is $IG(c^*, d^*)$, with $c^* = \frac{l}{2}$ and $d^* = \sum_{i=1}^{l} u_i u_i'/2$.

Updating Φ_i

The full conditional distribution of $\Phi_k, k = 0, \cdots, q+q^*$ is Generalized Inverse Gaussian distribution GIG($\lambda - \frac{1}{2}, \frac{1}{\gamma^2}, \beta_k^2$), where β_k is the k^{th} component of fixed effect β.

Updating λ and γ

The full conditional distribution of γ^{-2} is $\Gamma(e, f)$, with $e = 2 + (q + q^*)\lambda$, $f = \frac{O}{2} + \frac{1}{2}\sum_{i=1}^{q+q^*} \Phi_i$. In terms of shape parameter, let $\pi(\lambda)$ be the density function of exp(0.1), then the full conditional distribution of λ would be

$$\pi(\lambda)\frac{1}{(2\gamma^2)^{(q+q^*)\lambda}(\Gamma(\lambda))^{(q+q^*)}}\left(\prod_{i=1}^{q+q^*}\Phi_i\right)^{\lambda}$$

which need to be updated using a Metropolis-Hastings random walk update on logλ.

Updating u_i

The full conditional distribution of random effect u_i is a normal distribution $(\tilde{u_i}, F_i)$, with $F_i = (\sigma_b^{-2}I + \frac{G_i^T G_i}{\sigma^2})^{-1}$, $\tilde{u_i} = F_i G_i^T (y_i - H_i\beta)/\sigma^2$.

Updating $y_{i,mis}$

The missing values contained in y_i also distribute normally mean being $H_i\beta + G_i u_i$, and variance being $\sigma^2 I_t$.

Gibbs sampler, together with a Metropolis-Hastings random walk allows us to obtain the posterior distribution for Z_{mis} and y_{mis}.

4.3 Simulation studies

The simulation aims to evaluate and compare the performances across various scenarios of the multivariate multiple imputation and proposed two-stage multiple imputation.

The data generation model closely follows the data structure of motivating example, with the sample size $l = 500$, and schedule of planned visit times as $t = 8$. For each individual $i = 1, \cdots, l$, we generate three time-varying covariates, $(X_{1ij}, X_{2ij}, X_{3ij})$. X_{2ij} and X_{3ij} are both normally distributed as mean being 0 and standard deviation being 1, while $X_{1ij} = j$ representing the time of visit. The correlation between X_{2ij} and X_{3ij} is set as ρ, where $\rho \in (0.6, 0.8, 0.99)$ to represent different degree of correlation. The correlation between X_{1ij} and X_{2ij}, and correlation between X_{1ij} and X_{3ij} are both equal to 0.5. The response is generated for such that

$$y_{ij} = \beta_0 + u_{0j} + \beta_1 * X_{1ij} + \beta_2 * X_{2ij} + \beta_3 * X_{3ij} + e_i, \tag{4.2}$$

where u_{0i} is independently distributed as $N(0, \psi)$ for $i = 1, \cdots, 500$, and e_i independently distributed as $N(0, \sigma^2)$ across j and i. We set $\beta = (6, 0.05, -0.05, -0.01)$.

We perform a set of simulations for each combination of missing scenario and variation scenario. First, we consider the scenario with missing complete at random (MCAR). We randomly delete some observations in y_{ij}, X_{1ij}, X_{2ij} with the same ratio. We vary the ratio to alter the proportion of missing values to different missingness: 40% and 60%,

representing realistic and extreme scenarios respectively. Second, we consider the scenario with missing at random (MAR). We assume the missing proportion depend on the visit time. The latter visit time, the larger missing proportion. Third, we consider small number of imputation, $M = 5, N = 2$, and large number of imputation $M = 25, N = 2$ for two-stage multiple imputation, where corresponding number of imputation for multivariate multiple imputation is $M * N$. Fourth, we consider small, medium, and large variation by taking $\psi \in (0.01, 1, 4)$ and $\sigma^2 \in (0.01, 1, 4)$.

In each scenario, 500 datasets are generated. The target parameters, $\beta_0, \beta_1, \beta_2, \beta_3$ are estimated from each data set. For each kth component of β, given its estimate $\hat{\beta}_k$ and associated error s_k, we calculate the bias as $\hat{\beta}_k - \beta_k$, standardized bias $\frac{\text{bias}_k}{s_k} \times 100\%$, root mean squared error (RMSE) $\sqrt{\text{bias}_k^2 + s_k^2}$, and 95% confidence interval coverage rate. Then mean of those statistics are taken across 500 generated datasets. Standardized bias and RMSE represent both accuracy and precision of an estimators, thus act as important measures of practical utility. One rule of thumb is that the if standardized bias exceeds 40%, it is considered to be severe [Collins et al., 2001].

4.3.1 Simulation Results

Table 9 summarizes the simulation results across scenarios with different levels of correlation.

When correlation is moderate ($\rho = 0.6$) and data are missing with proportion of 40%,

we observe minimal bias under both multivariate multiple imputation (MULT) and two-stage multiple imputation (TSMI), with standardized bias not exceeding 40%. In both MCAR and MAR scenarios, the multivariate multiple imputation produce higher standardized bias and root mean squared error (RMSE) than two-stage multiple imputation. The confidence interval coverage rate achieve nominal rate under both approaches. However, when the proportion of missing values increase to 60%, the multivariate multiple imputation result in more biased estimator, particularly for coefficient β_2 (47.261% under MCAR and 54.712% under MAR). Instead, two-stage multiple imputation still performs better with smaller RMSE and smaller standardized bias, which remaining under 40% for all coefficients under both MCAR and MAR.

When the correlation level is 0.8, two-stage multiple imputation produce a substantially lower standardized bias and a slightly lower root mean squared error compared with multivariate multiple imputation. Across all missingness scenarios, standardized bias of β_2 given by multivariate multiple imputation are all exceeding acceptable level (51.288% under MCAR-40%, 58.633% under MCAR-60%, 57.903% under MAR-40%, and 59.822% under MAR-60%). Also, a slight under-coverage is reported by multivariate multiple imputation for β_2 under all missingness scenarios.

When the correlation level is high ($\rho = 0.99$), we observe a significant difference between the two approaches in standardized bias and root mean squared error. Two-stage multiple imputation results in standardized bias with maximum 40.654% in all scenarios, while multivariate multiple imputation produces high level of standardized

bias especially for both β_2 and β_3, the coefficients with two covariates are correlated and partly-observed (72.483% for β_2, -44.615 % for β_3 under MCAR-40%; 91.844 % for β_2, -48.004% for β_3 under MCAR-60%; 96.341% for β_2, -61.541% for β_3 under MAR-40%; 98.012% for β_2, -66.157% for β_3 under MAR-60%). We also observe larger root mean squared error with maximum of 0.466, and lower-than-nominal coverage ratio of confidence interval (88% with β_2 under MAR-40% and MAR-60%, 89% with β_3 under MAR-60%) for multivariate multiple imputation. However, the two-stage multiple imputation achieves nominal coverage rate of confidence interval in all missingness scenarios.

The results suggest that the correlation in the design matrix would affect the coefficients estimation: higher correlation level leads to more biased estimator (with larger standardized bias and root mean squared error). In addition, for both approaches, increase in the proportion of missing values is reflected in raise in standardized bias and root mean squared error. Higher standardized bias and higher root mean squared error are also observed for missing at random (MAR) scenario compared with missing completely at random (MCAR) scenario, given the same missing ratio and correlation level.

Table 9: Simulation Results - comparison standardized bias (%), root mean squared error (RMSE), and 95% confidence interval coverage rate (cov) of estimated coefficients between multivariate multiple imputation (MULT) and two-stage multiple imputation (TSMI)

ρ		0.6						0.8						0.99					
Missingness		Bias	%	MSE		Cov		Bias	%	MSE		Cov		Bias	%	MSE		Cov	
		MULT	TSMI	MULT	TSMI	MULT	TSMI	MULT	TSMI	MULT	TSMI	MULT	TSMI	MULT	TSMI	MULT	TSMI	MULT	TSMI
MCAR 40%	β_0	12.528	1.030	0.084	0.081	0.99	0.99	14.551	-3.083	0.084	0.084	0.99	0.99	15.259	-4.299	0.087	0.092	0.99	0.98
	β_1	4.031	2.203	0.147	0.138	0.96	0.97	5.185	2.886	0.149	0.140	0.96	0.93	7.101	3.330	0.175	0.162	0.97	0.98
	β_2	36.582	30.193	0.111	0.101	0.96	0.95	**51.288**	32.842	0.150	0.110	0.95	0.95	**72.483**	33.000	0.175	0.111	0.92	0.94
	β_3	-20.368	-10.798	0.140	0.140	0.96	0.95	-33.511	-20.960	0.204	0.209	0.95	0.92	**-44.615**	-25.101	0.330	0.221	0.93	0.94
MCAR 60%	β_0	15.128	6.677	0.085	0.081	0.99	0.99	16.079	9.381	0.087	0.083	0.99	0.99	18.087	10.236	0.088	0.085	0.99	0.99
	β_1	5.153	-2.500	0.151	0.139	0.97	0.97	6.998	-2.582	0.152	0.140	0.96	0.95	8.100	3.491	0.172	0.160	0.98	0.98
	β_2	**47.261**	33.990	0.152	0.147	0.94	0.94	**58.633**	34.195	0.159	0.149	0.93	0.95	**91.844**	36.657	0.268	0.156	0.89	0.94
	β_3	-22.769	-12.183	0.203	0.198	0.95	0.95	-34.975	-22.082	0.281	0.281	0.95	0.94	**-48.004**	-26.243	0.370	0.298	0.93	0.96
MAR 40%	β_0	16.094	14.666	0.095	0.093	1.00	1.00	17.047	14.900	0.101	0.100	0.98	0.99	18.068	15.36	0.106	0.102	0.99	0.99
	β_1	4.308	2.238	0.150	0.137	0.94	0.97	6.916	3.014	0.161	0.158	0.96	0.97	8.740	6.71	0.171	0.160	0.97	0.97
	β_2	37.488	32.268	0.121	0.125	0.93	0.93	37.903	34.123	0.157	0.126	0.95	0.95	**96.341**	39.940	0.189	0.134	0.88	0.93
	β_3	-29.613	-20.853	0.198	0.241	0.96	0.96	-33.857	-28.348	0.231	0.235	0.94	0.94	**-61.541**	-36.35	0.400	0.241	0.90	0.95
MAR 60%	β_0	18.838	15.646	0.110	0.093	0.99	0.99	19.878	16.000	0.121	0.115	0.98	0.99	20.043	16.56	0.121	0.116	0.99	0.99
	β_1	5.622	2.719	0.159	0.137	0.95	0.97	7.107	3.692	0.172	0.161	0.96	0.96	9.472	8.283	0.189	0.177	0.97	0.98
	β_2	**54.712**	34.913	0.158	0.150	0.94	0.92	**59.882**	35.921	0.168	0.160	0.92	0.95	**98.012**	40.653	0.196	0.165	0.88	0.93
	β_3	-29.641	20.985	0.241	0.246	0.94	0.94	-35.149	-29.872	0.292	0.290	0.95	0.95	**-66.157**	-37.58	0.466	0.291	0.89	0.95

4.4 Data Example

We consider linear mixed effect model accounting for the change in CD4 cell count. Time-varying variables include AUDIT-C and PEth, while baseline covariates include age, sex, religion, and HIV virus load. In view of missing observations in CD4 cell count, AUIDT-C and PEth, we apply both multivariate multiple imputation and two-stage multiple imputation to handle the missing data. The results are summarized in Table 10 .

Both approaches suggest negative associations between alcohol consumption and CD4 cell count, with estimated coefficients of AUDIT-C being -0.013, PEth being -0.014 under multivariate multiple imputation approach, while estimated coefficients of AUDIT-C being -0.014, PEth being -0.023 under two-stage multiple imputation approach. However, such associations are not statistically significant (none of p-values is smaller than 5%). It is also reported that HIV virus load is an important factor for change of CD4 cell count (estimated coefficient is -0.076, p-value less than 0.0001 under multivariate multiple imputation; estimated coefficient is -0.07, p-value is 0.0002 under two-stage multiple imputation). The results from two approaches confirm the finding in [Weiser et al., 2014, Cagle et al., 2017, Hahn et al., 2018].

Table 10: Application to the HIV data: Estimated coefficients obtained by fitting linear mixed model, using multivariate multiple imputation (MULT) and two-stage multiple imputation (TSMI) for missing values

	MULT			TSMI		
	Est	std	p-value	Est	std	p-value
Intercept	6.64	0.084	<0.0001	6.61	0.089	<0.0001
time	-0.005	0.002	0.0005	-0.005	0.002	0.001
AUDIT-C	-0.013	0.009	0.176	-0.014	0.005	0.100
PEth	-0.014	0.018	0.452	-0.023	0.014	0.113
age	0.00005	0.002	0.972	0.0001	0.001	0.943
religion	-0.018	0.02	0.378	-0.015	0.017	0.376
sex	-0.048	0.041	0.238	-0.036	0.033	0.272
viral load	-0.076	0.014	<0.0001	-0.07	0.018	0.0002

4.5 Discussion

We propose two-stage multiple imputation integrating Bayesian Gaussian shrinkage priors to handling incomplete CD4 cell count, incomplete AUDIT-C and incomplete PEth in the HIV data example. Bayesian Gaussian shrinkage priors are primarily used to account for the correlation between AUIDT-C and PEth. The results support that alcohol use has a negative effect on CD4 cell count, while such effect is not significant in our data example.

We compare the performance of the proposed two-stage multiple imputation and multivariate multiple imputation for handling missing data in response and correlated time-varying covariates in linear mixed effect model. We consider different missingness scenarios and different levels of correlation. Both approaches produce small bias under moderate correlation level ($\rho = 0.6$). When the correlation level is as high as 0.8 and/or

0.99, the multivariate multiple imputation produce biased estimator with some standardized bias exceeding 40%. In contrast, the proposed two-stage multiple imputation produce less biased estimates in all correlation level and all missingness scenarios investigated. This may be due to proposed two-stage multiple imputation accounting for the correlation between the covariates. In sum, the two-stage multiple imputation with Bayesian shrinkage priors has a robust performance even subject to multicollinearity.

Several extension of the proposed two-stage multiple imputation may be of interest. First, incorporation Bayesian Gaussian shrinkage priors other than normal-gamma prior used in this dissertation can be considered, and impact of different shrinkage priors on coefficients estimation needs further investigation. Second, we impute the missing values in response and covariates assuming they are missing at random. However, this may not be realistic in practice. For instance, patients may drop out of study or lost to follow-up due to heavier alcohol intake in our data example, thus the missingness would be subject to missing not at random (MNAR). Imputation models accounting for missing values and missingness indicators jointly would be necessary.

Chapter 5

Conclusion and Future Work

5.1 Overview

Incomplete data are prominent in longitudinal studies. Statistics literatures on handling incomplete outcome and covariates in longitudinal or clustered data are limited. This dissertation builds statistical methodology to address incomplete response and incomplete correlated covariates in the context of linear mixed model. In chapter 2 and 3, we develop two separate methodologies. First, we focus on multicollinearity in linear mixed model. We propose the Bayesian shrinkage mixed estimator which has reliable performance among extensive simulation studies. Second, we apply two-stage multiple imputation to tackle the incomplete response and time-varying covariates in linear mixed model. The proposed approach is shown to yield small bias of coefficient estimation. Finally, in Chapter 4, we bring together the findings of Chapter 2 and 3 to fully deal with incomplete response and correlated covariates.

The main contribution of this dissertations are (a) provides a framework to handle partly observed outcome and covariates via two-stage multiple imputation; (b) explore the benefits of introducing Bayesian shrinkage priors to deal with multicollinearity in

the context of linear mixed model (c) an application of proposed approaches to the HIV study.

The developments presented in this dissertation provides a foundation for extensions in many ways. We mainly discuss extensions to non-continuous data and missing not at random data.

5.2 Extension to Non-continuous Data

The ideas in this work can be extended to non-continuous data, such as categorical and count data. For example, if partly-observed outcome is a binary variable while partly-observed covariates are continuous data, one can implement two-stage multiple imputation by specifying multivariate linear mixed model as covariates model and logistic mixed effect model as outcome model. Future work will explore the performance of two-stage multiple imputation to tackle non-continuous data with comparison to other commonly-used missing data methods.

5.3 Extension to Non-ignorable Missing Data

While we assume missing response and missing time-varying covariates are missing at random (MAR) in this work, one cannot rule out the possibility that the process inducing missing variables is further related to unobserved variables. In particular, scientists may

be concerned that patients with heavier drinking behaviors and/or worst disease condition would not return for their next assessment (drop out). When the unobserved data itself provide information about its distribution, the missing data are said to be informative missingness, or non-ignorable, which belongs to missing not at random (MNAR), instead of MAR.

There've been much research on models for informative missing data, including selection models [Heckman, 1979, Diggle and Kenward, 1994, Molenberghs et al., 1997], pattern-mixture models [Little, 1993, 1994, 1995, Daniels and Hogan, 2000, Roy, 2003], and shared-parameter models [Wu and Carroll, 1988, Yuan and Little, 2009]. These models differ in the way the joint distribution of the outcome and missing data processes are factorized. In selection models, one specifies a marginal model for the outcome process and a conditional model for the distribution of the drop-out process given the outcome process; in pattern-mixture models, one specifies a conditional model for the outcome process given the drop-out time and the marginal distribution of the drop-out time ; and in shared-parameter models, the outcome and drop-out processes are assumed to be conditionally independent given shared random effects . Traditionally, these models have relied on very strong distributional assumptions in order to obtain model identifiability.

5.3.1 Extension of Two-Stage Multiple Imputation to Partial Ignorability

Harel and Schafer proposed the partial ignorability and latent ignorability [Harel and Schafer, 2009]. Partial ignorability assumption partitions the missing data and allow one (or more) of the partitions to be ignored given the other partition(s) and the observed data. Two stage multiple imputation can thus be extended to the situations where either missing covariates or missing outcome considered to be partial ignorable given the other one. For example, if it is assumed that individuals with higher alcohol intake would be more easily miss the follow-up assessment in the HIV data, then missing response can be ignored give the missing time-varying covariates and observed data. A two-stage-covariates-response, therefore, would be appropriate by assuming a joint distribution for covariates and missing data process of it in the first stage, and distribution for response proposed in this work in the second stage.

Appendix: Additional Tables

Table 11: Results from simulation study with different expected value of λ, a_λ for moderate correlation level 0.6. Percentage bias (Bias %), MSE, 95% confidence interval coverage (Cov), and 95% confidence interval length (CI Length) from the proposed method BSME along with traditional method MS. Results are based on 1000 replications each

a_λ		Bias	(%)	MSE		Cov		CI	Length
		BSME	MLE	BSME	MLE	BSME	MLE	BSME	MLE
0.2	β_0	-0.616	-2.255	0.059	0.057	0.96	0.95	0.711	0.691
	β_1	0.444	-0.578	0.111	0.106	0.96	0.96	0.970	0.953
	β_2	2.701	1.562	0.118	0.113	0.93	0.94	0.973	0.957
0.3	β_0	-1.790	-3.344	0.063	0.063	0.96	0.95	0.707	0.698
	β_1	1.750	0.569	0.128	0.121	0.95	0.96	0.975	0.958
	β_2	1.276	0.113	0.120	0.113	0.95	0.97	0.978	0.962
0.4	β_0	-0.331	-2.831	0.066	0.064	0.96	0.97	0.706	0.687
	β_1	0.957	-0.274	0.132	0.124	0.93	0.93	0.966	0.946
	β_2	-1.342	-2.543	0.119	0.113	0.95	0.96	0.967	0.948
0.5	β_0	-1.571	-3.532	0.063	0.061	0.97	0.96	0.715	0.697
	β_1	2.873	1.537	0.126	0.118	0.98	0.98	0.981	0.960
	β_2	-1.589	-2.963	0.111	0.105	0.96	0.97	0.983	0.963
0.6	β_0	0.064	-2.507	0.066	0.063	0.95	0.94	0.712	0.691
	β_1	-0.137	-1.549	0.113	0.106	0.97	0.99	0.973	0.952
	β_2	-4.477	-5.688	0.130	0.123	0.95	0.94	0.977	0.954
0.7	β_0	2.813	0.858	0.061	0.057	0.98	0.97	0.718	0.696
	β_1	0.510	-0.678	0.121	0.113	0.95	0.98	0.966	0.944
	β_2	-1.743	-3.227	0.120	0.112	0.94	0.96	0.966	0.944
0.8	β_0	1.917	0.048	0.067	0.065	0.93	0.95	0.713	0.700
	β_1	6.458	4.818	0.134	0.123	0.90	0.92	0.974	0.948
	β_2	-5.206	-6.457	0.113	0.107	0.97	0.98	0.977	0.952
0.9	β_0	-1.564	-3.931	0.061	0.059	0.97	0.95	0.707	0.689
	β_1	1.812	0.237	0.125	0.115	0.96	0.96	0.973	0.948
	β_2	-0.837	-2.145	0.118	0.111	0.95	0.95	0.974	0.948
1	β_0	-0.801	-3.482	0.075	0.072	0.93	0.91	0.711	0.690
	β_1	1.186	-0.285	0.130	0.120	0.93	0.95	0.976	0.950
	β_2	-0.714	-2.222	0.107	0.099	0.97	0.97	0.973	0.946

Table 12: Results from simulation study with different expected value of λ, a_λ for moderate correlation level 0.8. Percentage bias (Bias %), MSE, 95% confidence interval coverage (Cov), and 95% confidence interval length (CI Length) from the proposed method BSME along with traditional method MS. Results are based on 1000 replications each

a_λ		Bias	(%)	MSE		Cov		CI	Length
		MLE	BSME	MLE	BSME	MLE	BSME	MLE	BSME
0.2	β_0	1.309	-0.549	0.078	0.075	0.93	0.93	0.709	0.696
	β_1	0.700	-0.157	0.166	0.155	0.94	0.94	1.158	1.129
	β_2	-1.809	-3.004	0.160	0.150	0.95	0.96	1.161	1.133
0.3	β_0	-3.651	-5.170	0.060	0.059	0.97	0.97	0.712	0.696
	β_1	1.452	0.574	0.176	0.162	0.96	0.97	1.159	1.130
	β_2	-0.858	-2.212	0.165	0.152	0.95	0.95	1.159	1.130
0.4	β_0	4.851	2.817	0.069	0.065	0.96	0.95	0.710	0.686
	β_1	1.591	0.628	0.158	0.144	0.98	0.99	1.161	1.128
	β_2	-2.502	-3.886	0.164	0.152	0.96	0.96	1.162	1.130
0.5	β_0	0.000	-2.584	0.068	0.065	0.95	0.93	0.710	0.691
	β_1	-2.861	-3.562	0.168	0.154	0.94	0.96	1.155	1.118
	β_2	1.042	-0.680	0.157	0.144	0.96	0.96	1.155	1.120
0.6	β_0	-0.149	-2.741	0.057	0.056	0.99	0.99	0.708	0.694
	β_1	-1.260	-1.829	0.184	0.167	0.96	0.97	1.160	1.121
	β_2	3.366	1.327	0.172	0.155	0.97	0.97	1.162	1.124
0.7	β_0	2.279	-0.209	0.066	0.062	0.97	0.96	0.713	0.695
	β_1	-2.288	-2.838	0.202	0.182	0.90	0.92	1.160	1.121
	β_2	3.444	1.319	0.182	0.163	0.95	0.96	1.162	1.125
0.8	β_0	2.287	-0.214	0.061	0.058	0.97	0.99	0.711	0.693
	β_1	0.372	-0.885	0.176	0.158	0.92	0.96	1.154	1.114
	β_2	3.559	2.083	0.174	0.155	0.94	0.94	1.153	1.113
0.9	β_0	-0.370	-2.868	0.069	0.066	0.92	0.92	0.709	0.686
	β_1	0.428	-0.885	0.183	0.162	0.94	0.96	1.157	1.112
	β_2	0.697	-0.720	0.171	0.151	0.97	0.98	1.157	1.113
1	β_0	1.878	-0.552	0.064	0.060	0.95	0.96	0.710	0.688
	β_1	0.647	-0.935	0.173	0.153	0.94	0.95	1.155	1.111
	β_2	-0.031	-1.228	0.182	0.162	0.95	0.94	1.156	1.112

Table 13: Results from simulation study with different expected value of λ, a_λ for moderate correlation level 0.99. Percentage bias (Bias %), MSE, 95% confidence interval coverage (Cov), and 95% confidence interval length (CI Length) from the proposed method BSME along with traditional method MS. Results are based on 1000 replications each

a_λ		Bias	(%)	MSE		Cov		CI	Length
		MLE	BSME	MLE	BSME	MLE	BSME	MLE	BSME
0.2	β_0	-0.299	-2.372	0.064	0.062	0.95	0.95	0.713	0.691
	β_1	4.888	1.763	3.136	1.253	0.94	0.99	4.910	3.563
	β_2	-5.216	-3.389	3.171	1.271	0.94	0.99	4.912	3.565
0.3	β_0	-0.773	-2.815	0.065	0.064	0.95	0.95	0.711	0.693
	β_1	-2.196	-1.842	3.152	1.167	0.95	0.99	4.915	3.476
	β_2	2.756	1.011	3.154	1.169	0.95	0.99	4.915	3.477
0.4	β_0	0.225	-1.890	0.069	0.066	0.94	0.94	0.713	0.692
	β_1	-3.322	-2.206	3.171	1.104	0.95	0.99	4.928	3.402
	β_2	3.377	0.741	3.160	1.100	0.95	0.99	4.928	3.403
0.5	β_0	0.306	-1.935	0.065	0.063	0.95	0.95	0.714	0.692
	β_1	-4.665	-2.467	3.060	1.020	0.95	0.99	4.921	3.334
	β_2	5.570	1.644	3.067	1.023	0.96	0.99	4.921	3.333
0.6	β_0	-1.304	-3.759	0.064	0.062	0.95	0.95	0.711	0.689
	β_1	-4.848	-2.992	3.101	0.995	0.95	0.99	4.913	3.268
	β_2	3.785	0.297	3.077	0.983	0.95	0.99	4.913	3.270
0.7	β_0	-0.870	-3.215	0.066	0.064	0.95	0.94	0.712	0.689
	β_1	4.644	1.095	3.169	0.967	0.95	0.99	4.911	3.232
	β_2	-4.692	-2.919	3.162	0.964	0.95	0.99	4.910	3.233
0.8	β_0	0.513	-1.934	0.064	0.061	0.96	0.95	0.711	0.690
	β_1	-14.429	-7.164	3.118	0.924	0.94	0.99	4.920	3.191
	β_2	14.059	4.946	3.118	0.925	0.95	0.99	4.920	3.192
0.9	β_0	-0.554	-3.245	0.063	0.061	0.95	0.95	0.714	0.693
	β_1	0.058	-0.978	3.172	0.912	0.94	0.99	4.915	3.159
	β_2	0.046	-0.897	3.198	0.923	0.95	0.99	4.914	3.160
1	β_0	-0.387	-3.106	0.062	0.060	0.97	0.97	0.710	0.690
	β_1	3.821	0.318	3.323	0.924	0.94	0.99	4.919	3.134
	β_2	-4.576	-3.070	3.304	0.917	0.95	0.99	4.920	3.133

Table 14: Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval length (CI Length) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with sample size 100, $\sigma = 1, \psi = 1, M = 5, N = 2$

Missing		Bias (%)			RMSE			Cov			CI Length		
		MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R
$p_y = 0.2$	β_0	-14.05	-24.96	-10.18	0.145	0.157	0.141	0.99	0.99	0.99	0.461	0.472	0.442
$p_x = 0.4$	β_1	**-50.78**	**-255.96**	-42.52	0.252	0.468	0.251	0.92	**0.66**	0.95	0.629	0.796	0.641
	β_2	**-87.33**	**-321.5**	-41.68	0.231	0.691	0.22	0.94	**0.73**	0.94	0.593	0.815	0.609
	β_3	**65.95**	**271.5**	42.58	0.336	0.792	0.315	0.92	**0.84**	0.94	0.831	1.077	0.849
$p_y = 0.4$	β_0	-16.9	-10.61	-20.31	0.154	0.148	0.179	0.99	0.99	0.99	0.471	0.462	0.512
$p_x = 0.2$	β_1	-43.41	**-49.92**	**-294.63**	0.277	0.262	0.495	**0.87**	0.94	**0.69**	0.649	0.607	0.752
	β_2	**-80.28**	-37.9	**-573.05**	0.259	0.244	0.742	0.91	**0.89**	**0.71**	0.645	0.594	0.756
	β_3	**56.76**	**45.18**	**496.48**	0.367	0.347	0.83	0.9	0.93	**0.78**	0.887	0.818	1.041
$p_y = 0.4$	β_0	-11.95	-8.5	-7.59	0.151	0.146	0.142	0.99	0.99	0.99	0.473	0.461	0.458
$p_x = 0.4$	β_1	-36.43	**-54.34**	-39.29	0.268	0.257	0.253	0.91	**0.88**	0.94	0.655	0.623	0.66
	β_2	**-79.96**	**-93.94**	-43.86	0.253	0.243	0.246	0.9	0.9	0.93	0.631	0.614	0.671
	β_3	**53.64**	**62.13**	41.69	0.358	0.346	0.331	0.9	**0.88**	0.94	0.883	0.844	0.91
$p_y = 0.6$	β_0	-28	-21.54	-21.5	0.17	0.174	0.172	0.99	0.97	0.96	0.525	0.501	0.498
$p_x = 0.6$	β_1	**-59.07**	**-74.18**	**-46.83**	0.359	0.334	0.335	**0.86**	**0.87**	0.93	0.864	0.813	0.831
	β_2	**-73.84**	**-90.64**	-34.7	0.338	0.303	0.303	**0.89**	0.9	0.93	0.846	0.793	0.878
	β_3	**63.32**	**76.16**	**50.74**	0.47	0.427	0.418	**0.87**	**0.89**	0.92	1.152	1.084	1.183

Table 15: Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval length (CI Length) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with $\sigma = 2, \psi = 1, M = 5, N = 2$

Missing		Bias	(%)		RMSE			Cov			CI	Length	
		MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R
$p_y = 0.2$	β_0	-14.985	-21.831	-20.182	0.08	0.081	0.08	0.99	0.99	0.99	0.287	0.285	0.284
$p_x = 0.4$	β_1	21.054	**-161.367**	15.474	0.171	0.29	0.168	0.99	**0.74**	0.99	0.553	0.585	0.53
	β_2	**76.712**	**-302.238**	42.732	0.189	0.473	0.175	0.98	**0.04**	0.98	0.53	0.582	0.516
	β_3	-36.523	**248.093**	-23.36	0.234	0.549	0.231	0.99	**0.18**	0.99	0.745	0.798	0.722
$p_y = 0.4$	β_0	-11.244	-1.571	-20.068	0.09	0.09	0.095	0.99	0.99	0.99	0.31	0.301	0.319
$p_x = 0.2$	β_1	20.036	14.429	**-171.592**	0.176	0.174	0.302	0.99	0.99	**0.68**	0.562	0.538	0.579
	β_2	**69.963**	43.889	**-325.604**	0.2	0.195	0.5	0.96	0.94	**0.03**	0.561	0.53	0.573
	β_3	-39.905	-38.189	**257.74**	0.25	0.244	0.574	0.99	0.98	**0.19**	0.778	0.74	0.805
$p_y = 0.4$	β_0	-10.633	-13.192	-14.281	0.091	0.09	0.09	0.99	0.99	0.99	0.31	0.297	0.302
$p_x = 0.4$	β_1	5.748	10.229	-4.819	0.178	0.175	0.17	0.99	0.99	0.99	0.575	0.545	0.543
	β_2	**50.248**	**54.209**	21.84	0.192	0.193	0.182	0.99	0.96	0.98	0.56	0.528	0.545
	β_3	-20.192	-26.286	-1.214	0.249	0.243	0.24	0.99	0.99	0.99	0.789	0.744	0.759
$p_y = 0.6$	β_0	-9.911	-1.997	-8.425	0.113	0.113	0.117	0.99	0.97	0.99	0.363	0.339	0.371
$p_x = 0.6$	β_1	20.594	30.799	16.681	0.232	0.231	0.228	0.97	0.97	0.97	0.69	0.664	0.665
	β_2	**63.438**	**76.206**	42.024	0.248	0.246	0.234	0.94	0.94	0.95	0.682	0.652	0.666
	β_3	-31.579	-41.373	-20.259	0.325	0.33	0.319	0.97	0.95	0.96	0.948	0.92	0.921

Table 16: Simulation Results - comparisons of standardized bias (%), root mean squared error (RMSE), 95% confidence interval coverage rate (Cov) and confidence interval length (CI Length) of estimated coefficients between multivariate multiple imputation (MULT), two-stage-covariates-response (TSMI-C), two-stage-response-covariate (TSMI-R) in scenario with $\sigma = 1, \psi = 4, M = 5, N = 2$

Missing		Bias	(%)		RMSE			Cov			CI	Length	
		MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R	MULT	TSMI-C	TSMI-R
$p_y = 0.2$	β_0	16.291	8.088	12.757	0.077	0.078	0.074	0.99	0.99	0.99	0.29	0.294	0.278
$p_x = 0.4$	β_1	29.46	**-246.795**	-27.411	0.097	0.264	0.097	0.99	**0.2**	0.98	0.306	0.389	0.326
	β_2	**96.341**	**-468.783**	-44.017	0.111	0.474	0.101	0.92	**0.001**	0.94	0.286	0.393	0.32
	β_3	**-53.12**	**373.329**	37.903	0.137	0.521	0.107	0.98	**0.01**	0.96	0.409	0.534	0.445
$p_y = 0.4$	β_0	16.079	1.03	6.289	0.081	0.081	0.085	0.99	0.99	0.99	0.297	0.291	0.304
$p_x = 0.2$	β_1	9.472	2.238	**-330.588**	0.1	0.1	0.334	0.99	0.97	**0.03**	0.315	0.301	0.379
	β_2	**63.018**	36.582	**-653.66**	0.111	0.111	0.642	0.96	0.93	**0.001**	0.312	0.295	0.386
	β_3	-29.641	-20.853	**497.784**	0.141	0.141	0.682	0.99	0.97	**0.001**	0.435	0.41	0.532
$p_y = 0.4$	β_0	15.094	10.411	10.236	0.081	0.078	0.081	0.99	0.99	0.99	0.296	0.282	0.291
$p_x = 0.4$	β_1	4.031	-41.394	5.153	0.101	0.113	0.1	0.99	0.98	0.99	0.319	0.349	0.303
	β_2	47.261	**-94.83**	30.193	0.108	0.141	0.101	0.98	0.9	0.94	0.315	0.349	0.298
	β_3	-22.769	**72.483**	-10.798	0.141	0.175	0.14	0.99	0.95	0.98	0.44	0.48	0.414
$p_y = 0.6$	β_0	18.087	9.787	14.666	0.095	0.094	0.093	0.99	0.99	0.99	0.327	0.317	0.312
$p_x = 0.6$	β_1	6.916	**-49.752**	2.203	0.147	0.159	0.138	0.96	0.94	0.93	0.421	0.44	0.382
	β_2	**54.712**	**-132.911**	33.99	0.152	0.232	0.147	0.93	**0.74**	0.9	0.413	0.475	0.371
	β_3	-23	**91.844**	-12.183	0.203	0.268	0.198	0.95	**0.85**	0.91	0.572	0.641	0.525

Bibliography

David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974.

Marwan M Azar, Sandra A Springer, Jaimie P Meyer, and Frederick L Altice. A systematic review of the impact of alcohol use disorders on hiv treatment outcomes, adherence to antiretroviral therapy and health care utilization. *Drug and alcohol dependence*, 112(3):178–193, 2010.

Gregory J Bagby, David A Stoltz, Ping Zhang, Jay K Kolls, Julie Brown, Rudolf P Bohm, Richard Rockar, Jeanette Purcell, Michael Murphey-Corb, and Steve Nelson. The effect of chronic binge ethanol consumption on the primary stage of siv infection in rhesus macaques. *Alcoholism: Clinical and Experimental Research*, 27 (3):495–502, 2003.

Gregory J Bagby, Ping Zhang, Jeanette E Purcell, Peter J Didier, and Steve Nelson. Chronic binge ethanol consumption accelerates progression of simian immunodeficiency virus disease. *Alcoholism: Clinical and Experimental Research*, 30 (10):1781–1790, 2006.

Gregory J Bagby, Angela M Amedee, Robert W Siggins, Patricia E Molina, Steve Nelson, and Ronald S Veazey. Alcohol and hiv effects on the immune system. *Alcohol research: current reviews*, 37(2):287, 2015.

Francis Bajunirwe, Jessica E Haberer, Yap Boum II, Peter Hunt, Rain Mocello, Jeffrey N Martin, David R Bangsberg, and Judith A Hahn. Comparison of self-reported alcohol consumption to phosphatidylethanol measurement among hiv-infected patients initiating antiretroviral treatment in southwestern uganda. *PLoS One*, 9(12):e113152, 2014.

Dolly Baliunas, Jürgen Rehm, Hyacinth Irving, and Paul Shuper. Alcohol consumption and risk of incident human immunodeficiency virus infection: a meta-analysis. *International journal of public health*, 55(3):159–166, 2010.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Marianna K Baum, Carlin Rafie, Shenghan Lai, Sabrina Sales, John Bryan Page, and Adriana Campa. Alcohol use accelerates hiv disease progression. *AIDS research and human retroviruses*, 26(5):511–518, 2010.

Katharine A Bradley, Anna F DeBenedetti, Robert J Volk, Emily C Williams, Danielle Frank, and Daniel R Kivlahan. Audit-c as a brief screen for alcohol misuse in primary care. *Alcoholism: Clinical and Experimental Research*, 31(7): 1208–1217, 2007.

R Scott Braithwaite, Kathleen A McGinnis, Joseph Conigliaro, Stephen A Maisto, Stephen Crystal, Nancy Day, Robert L Cook, Adam Gordon, Michael W Bridges, Jason FS Seiler, et al. A temporal and dose-response association between alcohol consumption and medication adherence among veterans in care. *Alcoholism: Clinical and Experimental Research*, 29(7):1190–1197, 2005.

Anthony Cagle, Christine McGrath, Barbra A Richardson, Dennis Donovan, Sameh Sakr, Nelly Yatich, Richard Ngomoa, Agnes Chepngeno Langat, Grace John-Stewart, and Michael H Chung. Alcohol use and immune reconstitution among hiv-infected patients on antiretroviral therapy in nairobi, kenya. *AIDS care*, 29(9):1192–1197, 2017.

Adam W Carrico, Steven Shoptaw, Christopher Cox, Ronald Stall, Xiuhong Li, David G Ostrow, David Vlahov, and Michael W Plankey. Stimulant use and progression to aids or mortality after the initiation of highly active anti-retroviral therapy. *Journal of acquired immune deficiency syndromes*, 67(5):508, 2014.

Paul J Catalano. Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in medicine*, 16(8):883–900, 1997.

Qingxia Chen, Ryan C May, Joseph G Ibrahim, Haitao Chu, and Stephen R Cole. Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in medicine*, 33(26):4560–4576, 2014.

Linda M Collins, Joseph L Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4):330, 2001.

Anna Conen, Qing Wang, Tracy R Glass, Christoph A Fux, Maria C Thurnheer, Christina Orasch, Alexandra Calmy, Enos Bernasconi, Pietro Vernazza, Rainer Weber, et al. Association of alcohol consumption and hiv surrogate markers in participants of the swiss hiv cohort study. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 64(5):472–478, 2013.

Judith A Cook, Jane K Burke-Miller, Mardge H Cohen, Robert L Cook, David Vlahov, Tracey E Wilson, Elizabeth T Golub, Rebecca M Schwartz, Andrea A Howard, Claudia Ponath, et al. Crack cocaine, disease progression, and mortality in a multi-center cohort of hiv-1 positive women. *AIDS (London, England)*, 22 (11):1355, 2008.

Michael J Daniels and Joseph W Hogan. Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56(4):1241–1248, 2000.

Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC, 2008.

Peter Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43 (1):49–73, 1994.

Peter Diggle, Peter J Diggle, Patrick Heagerty, Patrick J Heagerty, Kung-Yee Liang, Scott Zeger, et al. *Analysis of longitudinal data*. Oxford University Press, 2002.

Melissa Eliot, Jane Ferguson, Muredach P Reilly, and Andrea S Foulkes. Ridge regression for longitudinal biomarker data. *The international journal of biostatistics*, 7(1):1–11, 2011.

Oghenowede Eyawo, Kathleen A McGinnis, Amy C Justice, David A Fiellin, Judith A Hahn, Emily C Williams, Adam J Gordon, Brandon DL Marshall, Kevin L Kraemer, Stephen Crystal, et al. Alcohol and mortality: Combining self-reported (audit-c) and biomarker detected (peth) alcohol measures among hiv infected and uninfected. *Journal of acquired immune deficiency syndromes (1999)*, 77(2):135–143, 2018.

Ludwig Fahrmeir, Thomas Kneib, and Susanne Konrath. Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20(2):203–219, 2010.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

Garrett Fitzmaurice, Geert Molenberghs, Marie Davidian, and Geert Verbeke. Generalized estimating equations for longitudinal data analysis. In *Longitudinal data analysis*, pages 51–86. Chapman and Hall/CRC, 2008.

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012.

LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Jim E Griffin, Philip J Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

Judith A Hahn and Jeffrey H Samet. Alcohol and hiv disease progression: weighing the evidence. *Current HIV/AIDS Reports*, 7(4):226–233, 2010.

Judith A Hahn, Sarah E Woolf-King, and Winnie Muyindike. Adding fuel to the fire: alcohol's effect on the hiv epidemic in sub-saharan africa. *Current HIV/AIDS Reports*, 8(3):172, 2011.

Judith A Hahn, Loren M Dobkin, Bernard Mayanja, Nneka I Emenyonu, Isaac M Kigozi, Stephen Shiboski, David R Bangsberg, Heike Gnann, Wolfgang Weinmann, and Friedrich M Wurst. Phosphatidylethanol (peth) as a biomarker of alcohol consumption in hiv-positive patients in sub-saharan africa. *Alcoholism: clinical and experimental research*, 36(5):854–862, 2012.

Judith A Hahn, Debbie M Cheng, Nneka I Emenyonu, Christine Lloyd-Travaglini, Robin Fatch, Starley B Shade, Christine Ngabirano, Julian Adong, Kendall Bryant, Winnie R Muyindike, et al. Alcohol use and hiv disease progression in an antiretroviral naive cohort. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 77(5):492–501, 2018.

Minna L Hannuksela, Marja K Liisanantti, Antti ET Nissinen, and Markku J Savolainen. Biochemical markers of alcoholism. *Clinical Chemical Laboratory Medicine*, 45(8):953–961, 2007.

James W Hardin and Joseph M Hilbe. *Generalized estimating equations*. Chapman and Hall/CRC, 2012.

Ofer Harel. *Strategies for data analysis with two types of missing values: from theory to application*. LAP Lambert Academic Publishing, 2009.

Ofer Harel and Joseph L Schafer. Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50, 2009.

Ofer Harel and Xiao-Hua Zhou. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077, 2007.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

Christian S Hendershot, Susan A Stoner, David W Pantalone, and Jane M Simoni. Alcohol use and antiretroviral adherence: review and meta-analysis. *Journal of acquired immune deficiency syndromes (1999)*, 52(2):180, 2009.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970a.

Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970b.

Martin A Javors, Nathalie Hill-Kapturczak, John D Roache, Tara E Karns-Wright, and Donald M Dougherty. Characterization of the pharmacokinetics of phosphatidylethanol 16: 0/18: 1 and 16: 0/18: 2 in human whole blood after alcohol consumption in a clinical laboratory study. *Alcoholism: Clinical and Experimental Research*, 40(6):1228–1234, 2016.

Robert I Jennrich and Mark D Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986.

J John-Langba, A Ezeh, G Guiella, A Kumi-Kyereme, and S Neema. Alcohol, drug use, and sexual-risk behaviors among adolescents in four sub-saharan african countries. *Ministry of Health*, 2006.

Christopher W Kahler, Tao Liu, Patricia A Cioe, Vaughn Bryant, Megan M Pinkston, Erna M Kojic, Nur Onen, Jason V Baker, John Hammer, John T Brooks, et al. Direct and indirect effects of heavy alcohol use on clinical outcomes in a longitudinal study of hiv patients on art. *AIDS and Behavior*, 21(7):1825–1835, 2017.

Miriam Julia Kip, Claudia Doris Spies, Tim Neumann, Yvonne Nachbar, Christer Alling, Steina Aradottir, Wolfgang Weinmann, and Friedrich Martin Wurst. The usefulness of direct ethanol metabolites in assessing alcohol intake in nonintoxicated male patients in an emergency room setting. *Alcoholism: Clinical and Experimental Research*, 32(7):1284–1291, 2008.

Stefan Kowalski, Elizabeth Colantuoni, Bryan Lau, Jeanne Keruly, Mary E McCaul, Heidi E Hutton, Richard D Moore, and Geetanjali Chander. Alcohol consumption and cd4 t-cell count response among persons initiating antiretroviral therapy. *Journal of acquired immune deficiency syndromes (1999)*, 61(4):455, 2012.

Ita GG Kreft and Jan De Leeuw. *Introducing multilevel modeling*. Sage, 1998.

Rakesh Kumar, Antonio E Perez-Casanova, Grissell Tirado, Richard J Noel, Cynthia Torres, Idia Rodriguez, Melween Martinez, Silvija Staprans, Edmundo Kraiselburd, Yasuhiro Yamamura, et al. Increased viral replication in simian immunodeficiency virus/simian-hiv-infected macaques with self-administering model of chronic alcohol consumption. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 39(4):386–390, 2005.

Nan Laird, Nicholas Lange, and Daniel Stram. Maximum likelihood computations with repeated measures: application of the em algorithm. *Journal of the American Statistical Association*, 82(397):97–105, 1987.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Qiuju Li, Jianxin Pan, and John Belcher. Bayesian inference for joint modelling of longitudinal continuous, binary and ordinal events. *Statistical methods in medical research*, 25(6):2521–2540, 2016.

Dennis V Lindley and Adrian FM Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972.

Mary J Lindstrom and Douglas M Bates. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

Raye Z Litten, Ann M Bradley, and Howard B Moss. Alcohol biomarkers in applied settings: recent advances and future research opportunities. *Alcoholism: Clinical and Experimental Research*, 34(6):955–967, 2010.

Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Roderick JA Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

Roderick JA Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the american statistical association*, 90(431):1112–1121, 1995.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. Wiley, 2019.

Roderick JA Little and Nathaniel Schenker. Missing data. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 39–75. Springer, 1995.

Xu-Qing Liu and Ping Hu. General ridge predictors in a mixed linear model. *Statistics*, 47(2):363–378, 2013.

William F Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

Gary C McDonald and Diane I Galarneau. A monte carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350): 407–416, 1975.

Geert Molenberghs, Michael G Kenward, and Emmanuel Lesaffre. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1):33–44, 1997.

Anne K Monroe, Bryan Lau, Michael J Mugavero, William C Mathews, Kenneth C Mayer, Sonia Napravnik, Heidi E Hutton, Hongseok S Kim, Sarah Jabour, Richard D Moore, et al. Heavy alcohol use is associated with worse retention in hiv care. *Journal of acquired immune deficiency syndromes (1999)*, 73(4):419, 2016.

M Revan Ozkale and Funda Can. An evaluation of ridge estimator in linear mixed models: an example from kidney failure data. *Journal of Applied Statistics*, 44 (12):2251–2269, 2017.

Michael Parzen, Stuart R Lipsitz, Garrett M Fitzmaurice, Joseph G Ibrahim, and Andrea Troxel. Pseudo-likelihood methods for longitudinal binary data with non-ignorable missing responses and covariates. *Statistics in medicine*, 25(16): 2784–2796, 2006.

Bhawna Poonia, Steve Nelson, Greg J Bagby, Ping Zhang, Lee Quinton, and Ronald S Veazey. Chronic alcohol consumption results in higher simian immunodeficiency virus replication in mucosally inoculated rhesus macaques. *AIDS research and human retroviruses*, 22(6):589–594, 2006.

Jerome P Reiter and Trivellore E Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480): 1462–1471, 2007.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Jason Roy. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, 59(4):829–836, 2003.

Jason Roy and Xihong Lin. Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association*, 97(457):40–52, 2002.

Jason Roy and Xihong Lin. Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. *Biometrics*, 61(3):837–846, 2005.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

Donald B Rubin. Nested multiple imputation of nmes via partially incompatible mcmc. *Statistica Neerlandica*, 57(1):3–18, 2003.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

Jeffrey H Samet, Nicholas J Horton, Seville Meli, Kenneth A Freedberg, and Anita Palepu. Alcohol consumption and antiretroviral adherence among hiv-infected persons with alcohol problems. *Alcoholism: Clinical and Experimental Research*, 28(4):572–577, 2004.

Jeffrey H Samet, Debbie M Cheng, Howard Libman, David P Nunes, Julie K Alperen, and Richard Saitz. Alcohol consumption and hiv disease progression. *Journal of acquired immune deficiency syndromes (1999)*, 46(2):194, 2007.

Joseph L Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997a.

Joseph L Schafer. Imputation of missing covariates under a multivariate linear mixed model, 1997b.

Joseph L Schafer and Recai M Yucel. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, 11(2):437–457, 2002.

Lori AJ Scott-Sheldon, Kate B Carey, Karlene Cunningham, Blair T Johnson, Michael P Carey, MASH Research Team, et al. Alcohol use predicts sexual decision-making: a systematic review and meta-analysis of the experimental literature. *AIDS and Behavior*, 20(1):19–39, 2016.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

Zijin Shen. *Nested multiple imputation*. PhD thesis, Harvard University, 2000.

Katelyn M Sileo, Leickness C Simbayi, Amber Abrams, Allanise Cloete, and Susan M Kiene. The role of alcohol use in antiretroviral adherence among individuals living with hiv in south africa: Event-level findings from a daily diary study. *Drug and alcohol dependence*, 167:103–111, 2016.

Scott H Stewart, David G Koch, Ira R Willner, Raymond F Anton, and Adrian Reuben. Validation of blood phosphatidylethanol as an alcohol consumption biomarker in patients with chronic liver disease. *Alcoholism: Clinical and Experimental Research*, 38(6):1706–1711, 2014.

Michael J Stirratt, Jacqueline Dunbar-Jacob, Heidi M Crane, Jane M Simoni, Susan Czajkowski, Marisa E Hilliard, James E Aikens, Christine M Hunter, Dawn I Velligan, Kristen Huntley, et al. Self-report measures of medication adherence behavior: recommendations on optimal use. *Translational behavioral medicine*, 5 (4):470–482, 2015.

Amy L Stubbendick and Joseph G Ibrahim. Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, 59(4):1140–1150, 2003.

Amy L Stubbendick and Joseph G Ibrahim. Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, 16(4):1143, 2006.

Gyongyi Szabo and Banishree Saha. Alcohol's effect on host defense. *Alcohol research: current reviews*, 37(2):159, 2015.

Uganda AIDS Commission. *UGANDA HIV/AIDS COUNTRY PROGRESS REPORT JULY 2016-JUNE 2017*. Uganda AIDS Commission, 2017.

Panagiotis Vagenas, Marwan M Azar, Michael M Copenhaver, Sandra A Springer, Patricia E Molina, and Frederick L Altice. The impact of alcohol use and related disorders on the hiv continuum of care: a systematic review. *Current HIV/AIDS Reports*, 12(4):421–436, 2015.

Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

Bonnie Wandera, Nazarius M Tumwesigye, Joaniter I Nankabirwa, Andrew D Kambugu, David K Mafigiri, Saidi Kapiga, and Ajay K Sethi. Hazardous alcohol consumption is not associated with cd4+ t-cell count decline among plhiv in kampala uganda: A prospective cohort study. *PloS one*, 12(6):e0180015, 2017.

Sheri D Weiser, Kartika Palar, Edward A Frongillo, Alexander C Tsai, Elias Kumbakumba, Saskia DePee, Peter W Hunt, Kathleen Ragland, Jeffrey Martin, and David R Bangsberg. Longitudinal assessment of associations between food insecurity, antiretroviral adherence and hiv treatment outcomes in rural uganda. *AIDS (London, England)*, 28(1):115, 2014.

World Health Organization. *Global status report on alcohol and health, 2018*. World Health Organization, 2018.

Hulin Wu and Lang Wu. A multiple imputation method for missing covariates in non-linear mixed-effects models with application to hiv dynamics. *Statistics in medicine*, 20(12):1755–1769, 2001.

Lang Wu. A computationally efficient method for nonlinear mixed-effects models with nonignorable missing data in time-varying covariates. *Computational statistics & data analysis*, 51(5):2410–2419, 2007.

Lang Wu and Hulin Wu. Missing time-dependent covariates in human immunodeficiency virus dynamic models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):297–318, 2002.

Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188, 1988.

Friedrich M Wurst, Christer Alling, Steina Aradottir, Fritz Pragst, John P Allen, Wolfgang Weinmann, Phillipe Marmillot, Pradeep Ghosh, Raj Lakshman, Gregory E Skipper, et al. Emerging biomarkers: new directions and clinical applications. *Alcoholism: Clinical and Experimental Research*, 29(3):465–473, 2005.

Friedrich M Wurst, Natasha Thon, Michel Yegles, Alexandra Schrück, Ulrich W Preuss, and Wolfgang Weinmann. Ethanol metabolites: their role in the assessment of alcohol intake. *Alcoholism: Clinical and Experimental Research*, 39(11): 2060–2072, 2015.

Han Yu, Shanhe Jiang, and Kenneth C Land. Multicollinearity in hierarchical linear models. *Social science research*, 53:118–136, 2015.

Ying Yuan and Roderick JA Little. Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, 65(2):478–486, 2009.

Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

Jing Hua Zhao and Joseph L. Schafer. *pan: Multiple imputation for multivariate panel or clustered data*, 2018. R package version 1.6.