STATISTICAL METHODS FOR CENSORED AND MISSING DATA IN SURVIVAL AND LONGITUDINAL ANALYSIS

Leah H. Suttner

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Sharon X. Xie, Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Sarah J. Ratcliffe, Professor of Biostatistics

Susan S. Ellenberg, Professor of Biostatistics

Daniel Weintraub, Professor of Psychiatry

STATISTICAL METHODS FOR CENSORED AND MISSING DATA IN SURVIVAL AND LONGITUDINAL ANALYSIS

ACKNOWLEDGEMENT

First and foremost, I would like to thank my dissertation advisor Dr. Sharon X. Xie. Your knowledge and guidance were invaluable throughout this work. Thank you for your patience and support in teaching me how to be a better statistician and conduct research. Thank you to my committee members, Dr. Sarah J. Ratcliffe, Dr. Susan S. Ellenberg, and Dr. Daniel Weintraub for sharing your time and expertise with me. I would also like to thank Dr. Russell T. Shinohara for his support in my first year of graduate school. Thank you for taking the time to teach me and answer my questions and for making me feel welcome in your lab when I was new.

I also have to thank all of the amazing gradute students, both past and present, that I have met here at Penn. You are an inspiration and I've truely valued your freindship. A special shout out to the Pooper Troopers for your support and always helping my to keep pedaling through.

Finally, I would not be where I am today without the love of support of my friends and family. I especially have to thank my parents Lisa and Leon Suttner, my brother Ethan, and my grandparents Lenard and Rita Millner. Thank you for the example you've set and for always believing in me.

ABSTRACT

STATISTICAL METHODS FOR CENSORED AND MISSING DATA IN SURVIVAL AND LONGITUDINAL ANALYSIS

Leah H. Suttner

Sharon X. Xie

Missing or incomplete data is a nearly ubiquitous problem in biomedical research studies. If the incomplete data are not appropriately addressed, it can lead to biased, inefficient estimation that can impact the conclusions of the study. Many methods for dealing with missing or incomplete data rely on parametric assumptions that can be difficult or impossible to verify. Here we propose semi-parametric and nonparametric methods to deal with data in longitudinal studies that are missing or incomplete by design of the study. We apply these methods to data from Parkinson's disease dementia studies. First, we propose a quantitative procedure for designing appropriate follow-up schedules in time-to-event studies to address the problem of interval-censored data at the study design stage. We propose a method for generating proportional hazards data with an unadjusted survival similar to that of historical data. Using this data generation process we conduct simulations to evaluate the bias in estimating hazard ratios using Cox regression models under various follow-up schedules to guide the selection of follow-up frequency. Second, we propose a nonparametric method for longitudinal data in which a covariate is only measured for a subset of study subjects, but an informative auxiliary variable is available for everyone. We use empirical and kernel density estimates to obtain nonparametric density estimates of the conditional distribution of the missing data given the observed. We derive the asymptotic distribution of the estimator for time-varying missing covariates as well as discrete or continuous auxiliary variables and show that it is consistent and asymptotically normally distributed. Through simulations we show that our estimator has good finite sample properties and is more efficient than the complete case estimator. Finally, we provide an **R** package to implement the method.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Incomplete data is a pervasive problem in biomedical research, giving rise to methodological challenges for accurate and efficient estimation. A variety of reasons and mechanisms can lead to incomplete data. Data can be missing if subjects drop out of a study or are lost to follow-up. In other cases, some data may not be collectible, observable, or available for some study subjects. Data can be missing by design of the study as a result of resource constraints. In this dissertation, we focus on data that are missing or incomplete by study design and propose methods to address the incomplete data at the design stage or at the time of analysis.

We consider methods to address two types of incomplete data situations motivated by studies of Parkinson's disease at the University of Pennsylvania Parkinson's disease Research Center.

First we consider interval-censored data, which is a type of incomplete data that is unique to survival analysis. Interval-censored data arises when the true time of an event is not known, but instead is observed to fall within a particular interval. For example, consider a clinical study of Parkinson's disease (PD) studying the time to progression from normal cognition to mild cognitive impairment (MCI). Cognitive function or status is measured by physician administered cognitive tests and neurological exams. Thus changes in cognitive status is only observed at patient follow-up visits when the cognitive assessments are conducted. The true time to MCI is therefore interval-censored between the follow-up visit at which the impairment was first measured and the previous visit.

Many methods have been developed to analyze interval-censored data. Some of these include nonparametric methods (Finkelstein, 1986), multiple-imputation algorithms (L Chen and J Sun, 2010; Pan, 2000), estimation-maximization (EM) algorithms (Goetghebeur and Ryan, 2000), and estimating equations methods (Heller, 2011). However, due to a lack of available software (Han, Andrei, and Tsui, 2014), none has been widely adopted. Instead, it is common to use methods for right-censored data, such as Cox proportional hazards (PH) models, with the right-endpoint of the censoring-interval as the event time. This 'right-endpoint imputation' approach is widely used despite having been shown to produce biased estimates of the hazard ratio (Law and Brookmeyer, 1992; Rücker and Messerer, 1988; X Sun and C Chen, 2010; Yu, 2012).

In Chapter 2 we address the problem of bias from right-endpoint imputation from a study design perspective. One of the main contributors to the estimation bias from right-endpoint imputation is the length of the censoring-interval, since longer intervals results in more overlap of those intervals, obscuring the true order of events. More frequent follow-up visits would shortened the censoring-intervals thereby reducing the bias; but increasing the frequency of visits may be constrained by funding resources and patient burden. Our goal then is to provide a method for designing follow-up schedules so that the frequency of follow-up visits reduces bias while conserving resources. Using what we already know about factors that contribute to bias from right-endpoint imputation in addition to new factors that we discovered through simulation studies, we develop a quantitative procedure for designing follow-up schedules. To implement our proposed procedure, we provide an easy to use Shiny (Chang et al., 2017) application.

In Chapters 3 and 4 we consider a second incomplete data situation in which a covariate is missing. When the measurement of a predictor is expensive, invasive, or otherwise unavailable, a surrogate or auxiliary variable may be measured in its place. Although prone to error, self-report data is relatively inexpensive and easy to collect; therefore self-report data, rather than more precise methods, are often used to obtain measures such as body mass index (Courtemanche, Pinkston, and Stewart, 2015; Xu, JK Kim, and Li, 2017) or nutritional intake (Yi et al., 2015). Similarly, fecal calprotectin measures are often used in place of the more accurate endoscopy, offering a less invasive, but generally adequate, predictor of inflammatory bowel disease activity and relapse (Røseth, Aadland, and Grzyb, 2004; Zhulina et al., 2016).

We focus on studies where the surrogate or auxiliary variable is measured for all study subjects while the expensive or invasive predictor is measured for only a subset. This subset of subjects makes up an ‘interval validation set’. For all other subjects, the covariate is missing by study design. For example, another study at Penn Parkinson’s disease Research Center is evaluating the relationship between the cerebral spinal fluid concentration of amyloid-β (CSF-aβ) and cognitive decline over time. Collection of this predictor from the cerebral spinal fluid requires a lumbar puncture which is both invasive and expensive. Therefore only a fraction of the study participants were assigned to undergo the procedure, while other participants are missing this covariate. Another biomarker, apolipoprotein E (APOE) genotype, which is measured through less invasive blood tests is available for most participants and has been shown to be associated with CSF-aβ (Tapiola et al., 2000), making APOE a good candidate for an auxiliary variable. However, the scientific question of interest still pertains to CSF-aβ. Since CSF-aβ is missing for a large sample of the study participants, the use of missing data methods is required.

The simplest approach to dealing with missing data is to use a complete case analysis, which would mean only using the subjects in the validation set who have no missing data. Since we assume that the validation set is a random subsample and therefore the data is missing completely at random (MCAR), the estimates would be unbiased but inefficient due to the drastically reduced sample size. We want to utilize the information in APOE to improve the efficiency of the analysis using more sophisticated methods. Some other popular missing covariate data methods include multiple imputation with chained equations (MICE) (Erler et al., 2016; Ibrahim et al., 2005), EM methods (Ibrahim, 1990), and Bayesian methods, all of which require some parametric assumptions about the distribution of the variables (Erler et al., 2016) that can be difficult to verify. Instead, we focus on semiparametric methods that do not require these distributional assumptions.

Pepe and Fleming (1991) and Carroll and Wand (1991) proposed semiparametric methods for missing covariate data that use empirical, nonparametric estimates of the densities of the missing (i.e. expensive) covariate and auxiliary or surrogate variable. Pepe and Fleming (1991) developed this method for linear regression where the auxiliary variable is discrete. Carroll and Wand (1991) assume a continuous surrogate variable but assume a logistic model (i.e discrete outcome). Xu, JK Kim, and Li (2017) use expected estimating equations to develop a general theory that is applicable to situations in which the auxiliary and outcome variables are continuous or discrete, but recommend bootstrapping the variance estimate. These three semiparametric methods assume a cross-sectional design and time-independent covariates.

In Chapter 3 we introduce our method that extends these semiparametric methods to longitudinal data with time-varying covariates. We derive the asymptotic distribution of the estimate and show that it is normally distributed. Through simulations we demonstrate that our estimator has good finite sample properties and estimates the variance well without the need to bootstrap. In Chapter 4 we describe the R package (R Core Team, 2018) that we wrote to perform this estimation.

In the final chapter we summarize our work and discuss how we could expand on these methods and topics moving forward.

CHAPTER 2

QUANTITATIVE METHOD FOR DESIGNING APPROPRIATE LONGITUDINAL FOLLOW-UP FREQUENCY

2.1. Introduction

Longitudinal studies with time-to-event as the main outcome are important in biomedical research because they enable us to study and understand the progression and risks associated with diseases. For example, they can be used to answer questions about time-to-clinical worsening or -relapse, progression-free survival, and time-to-death. Increasingly, the event of interest is not something that can be directly observed or measured by the patients, such as changes in disease status as indicated by biomarkers (Heller, 2011; Wellek, 2017) or other physician assessments administered at patient follow-up visits. A natural question then in designing longitudinal studies is how frequently participants should be followed-up.

With unlimited resources of time and money participants could be followed-up continuously. Of course continuous follow-up is not practical or feasible in real-world outpatient clinical trials or observational studies. Instead, follow-up schedules must be designed to balance resources and the precision of the collected data. This problem is uniquely challenging for time-to-event studies, where the true time of the event is not observed, but instead is interval-censored between follow-up visits.

The goal of this chapter is to provide a quantitative strategy for selecting follow-up schedules in order to avoid the problem of interval-censoring at the design stage. Methods to account for interval censoring after the data have been collected have some disadvantages that limit their use in practice. Prior to Wang et al. (2016), many of the interval censoring methods were too computationally intensive or complex for widespread use, and those that were available in software packages required parametric assumptions, provided biased parameter estimates, or did not provide variance estimates (Pan, 1999; Wang et al., 2016). As a result, this type of data is frequently analyzed with standard survival methods using right-endpoint imputation, i.e. the observed time of the event. This method is even recommended by the FDA for dealing with interval-censored progression-free survival in oncology clinical trials (FDA, 2015; Zeng et al., 2015). Multiple authors have shown that right-endpoint imputation may result in biased estimation (Rücker and Messerer, 1988; X Sun and C Chen, 2010; Yu, 2012; Zeng et al., 2015), yet there is limited advice for designing follow-up schedules to limit this bias. In this chapter, we propose a quantitative method to allow clinical investigators or statisticians to select an appropriate follow-up schedule so that the standard Cox model (i.e., right-end imputation approach) can generate reliable results with small bias. Thus, the impact of interval censoring on the bias can be reduced at the study design stage by applying our new procedure.

Current practice for designing longitudinal follow-up frequency in outpatient clinical research is mainly based on experience, tradition, and availability of resources. Below we describe several methods proposed to offer guidance in the design of follow-up frequencies in longitudinal studies where the outcome of interest is an interval-censored time-to-event. However, these methods generally rely on parametric assumptions about the underlying distribution of the rate of the event, fixed follow-up intervals, or complex programing.

Inoue and Parmigiani (2002) provides a method to choose the "optimal" follow-up times using a decision-theoretic approach in a Bayesian framework assuming a constant hazard rate (i.e. the time-to-event is exponential). Broad implementation of this method can be limited due to its computational complexity (Raab, Davies, and Salter, 2004). Raab, Davies, and Salter (2004) calculates the asymptotic loss of efficiency for interval-censored data given a parametric model and recommends interval lengths as a function of the median survival time. In addition to requiring parametric assumptions about the survival distribution, the method assumes fixed follow-up intervals over the duration of the study. Alexander (2008) provides an analytic expression for the Fisher information in terms of the interval length, also assuming a constant hazard rate, fixed follow-up intervals for the entire duration of the study, and no right-censoring. HY Kim, Williamson, and Lin (2016) calculates the power for different follow-up schedules assuming parametric distributions for the underlying survival and Wellek (2017) provides sample size formulas to calculate power for superiority and non-inferiority analyses under an accelerated failure time model.

In this chapter, we made several new contributions. First, we provide a novel method to design longitudinal follow-up schedules using semiparametric assumptions instead of parametric ones in contrast to existing literature which requires parametric assumptions. Moreover, our method allows for flexible lengths of follow-up, in contrast to existing methods which require fixed follow-up schedules. Second, we present novel discoveries on the factors that contribute to estimation bias from right-endpoint imputation. Third, we provide a user friendly web application to implement our procedure thereby enabling clinical researchers to apply this quantitative method to their follow-up frequency selection process.

The remainder of this chapter is organized as follows. Section 2.2 describes the Parkinson's disease (PD) cognition study that motivates this work. Section 2.3 describes simulations to explore the factors that influence estimation bias when using right-endpoint imputation. Section 2.4 introduces the novel quantitative procedure for evaluating follow-up schedules. The method is validated in Section 2.5. In Section 2.6 we apply the procedure for follow-up frequency selection in the PD setting. Finally, in Section 2.7 we discuss the implications and limitations of our procedure.

2.2. Motivating Study

This work is motivated by a University of Pennsylvania (Penn) Parkinson's Disease Center study (Pigott et al., 2015). Pigott et al. (2015) uses a convenience cohort of patients with PD and normal cognition (NC) at baseline to investigate potential baseline predictors of cognitive decline. The study included 141 patients and annual evaluations for four years, followed by biennial evaluations (i.e. participants were followed-up at years 1-4 and 6). The outcome of interest was the time-to-progression from NC to any cognitive impairment, and, for those who progressed to mild cognitive impairment (MCI), the time-to-progression from MCI to Parkinson's disease dementia (PDD). The effects of baseline predictors were estimated using standard Cox proportional hazards (PH) regression models. Due to the follow-up schedule, participants' true event times are interval-censored between subsequent visits. Therefore, right-endpoint imputation was used for the event times of participants who progressed to MCI or PDD. In other words, the event times were defined as the follow-up visit at which progression was first observed. Subjects who dropped out of the study or did not have the event by the end of the study were right censored. Investigators want to know if their follow-up schedules are appropriate for these research questions and if the schedules should be modified for future studies.

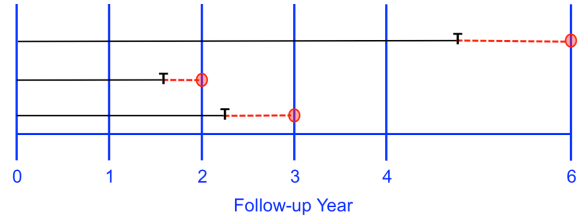


Figure 2.1: The true event time "T" is observed at time ("O"). If T falls before year 4, it would be interval-censored by one year, and if T falls between years 4 and 6, it would be interval-censored by two years.

2.3. Simulations to Study Factors Associated with Hazard Ratio Estimation Bias Due to Right-Endpoint Imputation

In this section we describe simulations used to explore how various factors impact the bias of hazard ratio (HR) estimates in PH models using right-endpoint imputation. First, we provide a new method for generating proportional hazards data that resembles a given Kaplan-Meier curve. Then we examine how the magnitude of the hazard ratio and the distribution of covariates effect the bias in the univariate setting. Finally, we compare the bias in univariate models to that of multivariate models.

2.3.1. Generating Proportional Hazards Data with Similar Unadjusted Kaplan-Meier Curves

It has been shown that the amount of ties impacts the HR estimation bias in Cox PH models, with more ties resulting in more biased estimates (Hertz-Picciotto and Rockhill, 1997). When participants are observed to fail in the same or overlapping follow-up intervals, the true order of events is lost, resulting in larger bias due to right-endpoint imputation, compared to when participants fail at more varied times. Therefore, we must control the number of ties in order to study how other factors impact the estimation bias. We can control the number of ties by controlling the shape of the survival curve. In this section, we describe a new method for generating PH data with similar unadjusted Kaplan-Meier curves.

The PH model can be written in terms of the survival distribution as

$$S(t|z) = S_0(t)^{\exp(\boldsymbol{Z}_i^T\boldsymbol{\beta})} \tag{2.1}$$

for i in 1 to N, where $S(t|z)$ is the survival at time t, $S_0(t)$ is the baseline survival distribution, $\boldsymbol{Z}_i$ is the vector of covariates for subject i, and $\boldsymbol{\beta}$=$\log(\boldsymbol{HR})$. Let $\hat{S}(t)$ be the target Kaplan-Meier curve. Our objective is to estimate an $S_0(t)$ given $\boldsymbol{\beta}$ and the distribution of $\boldsymbol{Z}$, so that the unadjusted Kaplan-Meier curve of the generated data, $\hat{S}_g(t)$, will be similar to $\hat{S}(t)$. If we obtain $\hat{S}(t)$ by generating survival times from a Weibull or Gompertz distribution, then $S_0(t)$ will also be Weibull or Gompertz, respectively, for some $\boldsymbol{\beta}$ and $\boldsymbol{Z}$, since these two distributions satisfy the PH assumption. Since both distributions can be expressed as a function of two parameters (scale=λ, shape=γ), we can estimate $S_0(t)$ by solving a system of two equations derived from Equation (2.1). The equations are defined by taking two points from $\hat{S}(t)$, so that we have (t_1,$\hat{S}(t_1)$) and (t_2,$\hat{S}(t_2)$). In addition, we define $Q_{Z\beta}(\hat{S}(t_1))$ and $Q_{Z\beta}(\hat{S}(t_2))$ to be the $\hat{S}(t_1)$ and $\hat{S}(t_2)$ quantiles of $\boldsymbol{Z}^T\boldsymbol{\beta}$. Now, for the Gompertz and Weibull distributions, we have the following systems of equations:

$$\begin{aligned} Gompertz: &\quad e^{Q_{Z\beta}(\hat{S}(t_k))}\lambda(1-e^{\gamma t_k})-\gamma\log\{\hat{S}(t_k)\} &= 0 \qquad (2.2)\\ Weibull: &\quad -e^{Q_{Z\beta}(\hat{S}(t_k))}\lambda t_k^{\gamma}-\log\{\hat{S}(t_k)\} &= 0 \qquad (2.3)\end{aligned}$$

where $k = 1, 2$. The Gompertz and Weibull distributions are supported only where $\lambda, \gamma > 0$. In order to satisfy the support of the distributions, we reparameterize the equations as

$$\begin{aligned} Gompertz: &\quad e^{Q_{Z\beta}(\hat{S}(t_k))}\lambda_{\dagger}^2(1-e^{\gamma_{\dagger}^2 t_k})-\gamma_{\dagger}^2\log\{\hat{S}(t_k)\} &= 0 \qquad (2.4)\\ Weibull: &\quad -e^{Q_{Z\beta}(\hat{S}(t_k))}\lambda_{\dagger}^2 t_k^{\gamma_{\dagger}^2}-\log\{\hat{S}(t_k)\} &= 0 \qquad (2.5)\end{aligned}$$

where $\lambda_{\dagger} = \sqrt{\lambda}$ and $\gamma_{\dagger} = \sqrt{\gamma}$. Then we can solve for the parameters in `R` (R Core Team, 2018) using the `multiroot` function from the `RootSolve` package (Soetaert, 2009; Soetaert and Herman, 2009). For starting values of the parameters, we use $\lambda_{\dagger 0} = \sqrt{\frac{\lambda^*}{e^{Q_{Z\beta}(\hat{S}(t_2))}}}$ and $\gamma_{\dagger 0} = \sqrt{\gamma^*}$, where λ^* and γ^* are the true parameters of $S(t)$ used to obtain $\hat{S}(t)$. In practice, we would not know the true distribution of $S(t)$ and instead can use maximum likelihood estimates for λ^* and γ^*, as we describe in Section 2.4.

2.3.2. Univariate Simulations

To investigate sources of bias in the univariate setting, we study a combination of different covariate distributions and hazard ratios. For each combination, we consider the % bias which is defined

$\frac{1}{SIM}\sum_{sim=1}^{SIM}\frac{\hat{HR}_{sim}-HR}{HR}\times 100\%$ where SIM is the number of datasets generated. In simulations (not shown), we found that in the univariate setting, the mean of a normally distributed covariate has no impact on the estimation bias. Therefore, we define the covariates in terms of skewness and variances. For zero skewness, we generate the covariate data from a normal distribution with a mean of 1 and the specified standard deviation (σ). For non-zero skewness, we generate data from a gamma distribution, since the distribution can be fully described by the skewness and standard deviation. Specifically, the shape parameter is calculated as $4/skew^2$ and the scale parameter is $\sigma \times skew/2$. We test skewness of 0, 1, 2, and 3 with standard deviations of 0.5, 1, 1.5, 2, and 2.5. For each combination we test HRs of 1.5, 2, and 2.5.

To investigate how the results are impacted when estimating the effect of multiple covariates, we compare the bias for individual predictors in a multivariate model to those in corresponding univariate models. We look at a total of five predictors, estimating each one in a univariate model and a sequence of multivariate models. The five covariates have skewness of 2.0, 1.0, 0, 0, and 1 and standard deviations of 1.5, 1.0, 1.0, 2.0, 0.5, respectively. The corresponding HRs are 1.5, 2.0, 2.5, 1.3, and 1.7. To look at how the number of predictors effects the bias in the estimated hazard ratios, we add one covariate to each subsequent model. We compare the resulting bias from the models with two, three, four, and five predictors to the respective univariate models.

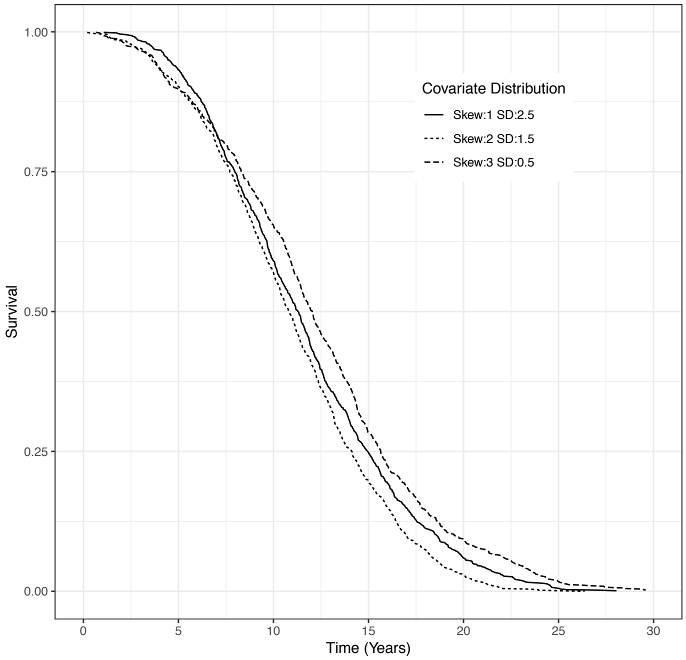
For all simulations, we use a sample size, N, of 200 and set $S(t)$ to be Weibull with shape=2 and scale=0.005. Figure 2.2 shows the Kaplan-Meier curves generated for HR=2.5 and a few of the tested covariate distributions. Censoring-intervals of 5 years were applied over 20 years.

2.3.3. Simulation Results

Table 2.1: HR estimation bias from right-endpoint imputation in univariate and multivariate models

					% Bias		
HR	**Skew**	**SD**	P1	P2	P3	P4	P5
1.5	2.0	1.5	-5.55	-7.34	-8.58	-9.00	-8.07
2.0	1.0	1.0	-8.58	-11.30	-13.87	-14.30	-12.92
2.5	0.0	1.0	-12.34		-16.40	-17.38	-15.60
1.3	0.0	2.0	-2.04			-5.25	-4.88
1.7	1.0	0.5	-1.92				-8.89

HR = Hazard ratio. Skew=Skewness. SD= Standard deviation.% Bias $= \frac{1}{2000}\sum_{sim=1}^{2000}\frac{\hat{HR}_{sim}-HR}{HR}\times 100\%$. P1 - P5 represents the number of predictors in each of the models. Proportional hazards models were used to estimate the HR associated with five predictors in univariate and multivariate models. The % bias from right-endpoint imputation is reported for each of the models. Column P1 shows the bias from the univariate models (i.e. 1 predictor). Columns P2 - P5 show the bias for each of the predictors in the multivariate models with 2, 3, 4, and 5 predictors.

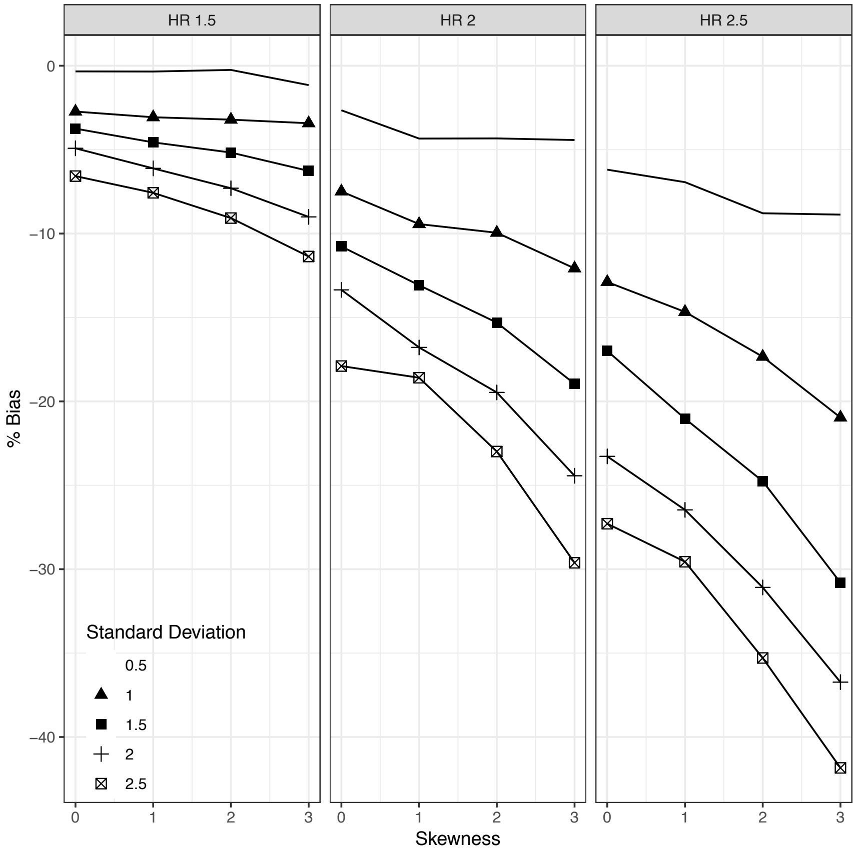


SD=Standard deviation

Figure 2.2: Kaplan-Meier curves for data generated from three different covariate distributions using our proposed method for controlling the unadjusted survival.

Figure 2.3 summarizes the bias in the univariate setting. As expected, the bias is negative, indicating that the estimated effect is attenuated by interval censoring. Interestingly, the magnitude of the bias increases with the standard deviation and with the skewness of the covariate. Additionally, the magnitude of the bias increases with hazard ratio. Zeng et al. (2015) found the magnitude of the HR to have little effect on the bias, but this is likely because their analyses were restricted to a single binary covariate. The results of the multivariate analyses are presented in Table 2.1. In general, when the number of predictors increases, the bias in each of the predictors also increases.

The significant impact of the covariate distribution on the resulting bias may be due to the way that ties are handled. The Efron method partially accounts for the fact that information on the true order of events is missing by approximating the average contribution to the likelihood for all possible orders of tied events. If the covariate values were the same for participants with tied event times, the contribution to the likelihood would be identical for all possible orders of events and thus those ties would not contribute to the bias. It is only when the covariate values differ that information is lost. When the standard deviation of the covariate distribution is larger, the values are likely to be



HR = Hazard ratio

$\% \text{ Bias} = \frac{1}{1000}\sum_{sim=1}^{1000}\frac{\hat{HR}_{sim}-HR}{HR}\times 100\%$

Figure 2.3: The bias from right-endpoint imputation increases with hazard ratio, predictor skewness, and predictor standard deviation.

more dissimilar at tied events, resulting in greater bias.

Moreover, the partial likelihood for the Efron method is a function of the mean weight $\frac{1}{d_j}\sum_{k\in D_j}\exp(\boldsymbol{Z}_k^T\boldsymbol{\beta})$, for the set of participants, D_j with failure time j, where d_j is the number of participants in D_j. When the distributions of the covariates are skewed instead of symmetric, using this mean may not be appropriate and may add to the estimation bias.

Similarly, the event rate impacts the bias by influencing the number of ties that are observed. When the rate of events is greater, more ties are observed resulting in a greater loss of information as the approximated likelihood is farther from the truth. Therefore, when the rate of change of survival is greater, more frequent follow-up is needed.

2.4. Proposed Longitudinal Follow-up Evaluation Procedure

Based on the results of the simulations in Section 2.3, we developed a novel procedure to evaluate the appropriateness of follow-up schedules. The procedure overview is as follows. First, the user specifies the HR and distributions for covariates based on pilot data. Then a baseline survival distribution is calculated such that the unadjusted Kaplan-Meier curve is similar to the historical data Kaplan-Meier curve provided by the user. Next, the user chooses potential follow-up schedules to investigate. Finally, the selected follow-up schedules are evaluated via simulations by generating data from the specified covariate and survival distributions. For each simulated dataset, Cox PH models are used with the observed (i.e. right-endpoint imputed) event times generated by each of the follow-up schemes. The bias for each of the respective models is averaged over all of the simulations. An appropriate follow-up frequency is chosen for which the bias is less than a pre-specified clinically significant threshold.

2.4.1. Covariate Definitions

The first step in our procedure is to define the covariate(s). As shown in Section 2.3, the shape of the covariate distributions can greatly impact the resulting HR estimation bias when using right-endpoint imputation. For each continuous covariate, the user can specify its skewness and standard deviation. Symmetric covariates are sampled from normal distributions and skewed covariates are sampled from gamma distributions. The user may also define discrete covariates by providing the possible values and their corresponding probabilities.

In addition, the user must specify HRs to test. Right-endpoint imputation attenuates the estimate of the effect, or shrinks the estimate towards 1 (Zeng et al., 2015). As a result, estimated HRs are more biased when the true HR is farther from 1. In practice, if no pilot data is available for the covariates, we recommend testing a range of plausible HRs and covariate distributions.

Finally, to allow the percent bias to be more comparable across HRs, all HRs should be defined as greater than or equal to 1. This could require re-defining a covariate if necessary. For example, if a binary covariate is defined as 1 for males with a HR=0.5, this should be redefined as 1 for females with a HR=2.

The covariates and HRs defined in this step are used to select the baseline survival distribution in

Step 2.

2.4.2. Distribution Selection

The second step of our new procedure is to calculate a survival distribution from which to generate data. This process is similar to that described in Section 2.3.1, with some additional steps to account for the fact that the true $S(t)$ is unknown. In place of $S(t)$, the user must provide the Kaplan-Meier curve from some historical data, which we define as $\hat{S}_h(t)$. Using $\hat{S}_h(t)$ we estimate $S_0(t)$ and check how similar the generated data is to the historical data.

In Section 2.5, we demonstrate that the procedure is not sensitive to the true distribution of $S_0(t)$. Specifically, we show that our procedure performs well even when the true baseline survival distribution is neither Weibull nor Gompertz. However, in order to automate our procedure, we let $S_0(t)$, and therefore $S(t)$, be from one of these two distributions.

To estimate $S_0(t)$, we first determine if $\hat{S}_h(t)$ more closely fits a Weibull or Gompertz distribution. We begin by simulating a large amount of "observed" data that is consistent with the historical data. To obtain this data, we use the inverse-cumulative distribution function (CDF) method, where $F(t) = 1 - \hat{S}_h(t)$. With the "observed" data, we calculate the maximum likelihood estimates (MLEs), λ^* and γ^*, of the shape and scale parameters, respectively, using `flexsurvreg` from the `flexsurv` package (Jackson, 2016) in `R` (R Core Team, 2018) assuming both a Weibull and a Gompertz distribution. We then select the distribution that better fits the data as defined by the Akaike information criterion (AIC).

Depending on the selected distribution, we can estimate $S_0(t)$ by solving either Equation 2.4 or 2.5, taking two points from the historical Kaplan-Meier curve, $\hat{S}_h(t)$. The MLEs, λ^* and γ^*, are used to calculate the starting values for the baseline parameters of $S_0(t)$. $Q_{Z\beta}(\hat{S}_h(t_1))$ and $Q_{Z\beta}(\hat{S}_h(t_2))$ are calculated as the $\hat{S}_h(t_1)$ and $\hat{S}_h(t_2)$ quantiles of $\boldsymbol{Z}^T\boldsymbol{\beta}$ as defined in Step 1.

After solving for the parameters of $S_0(t)$, we check that the chosen distribution is appropriate for the historical data. We first generate event times under the PH model using the calculated $S_0(t)$ along with the covariates and hazard ratios defined in Step 1. To make this data comparable to the historical data, we create censoring-intervals using the event times from the historical data. Then, using the right-endpoint of the censoring-interval for each event time, we calculate the Kaplan-

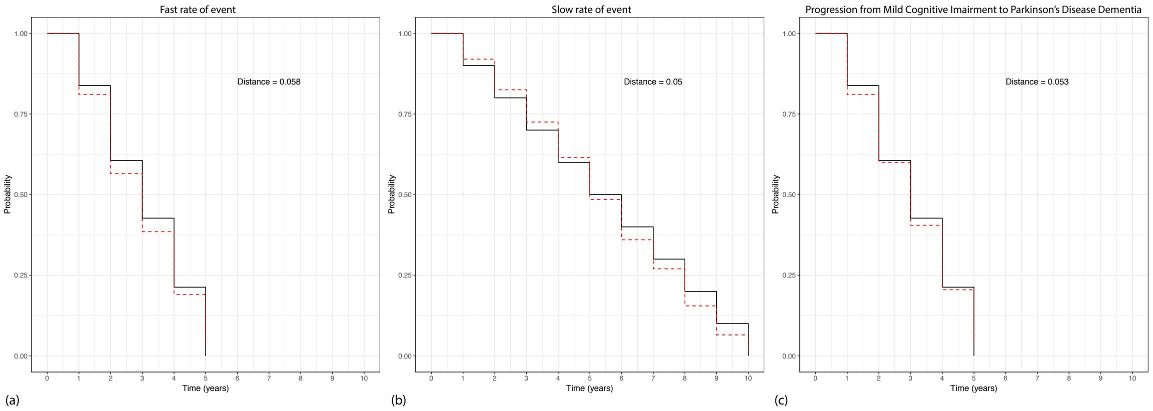


Figure 2.4: Comparison of generated data (dashed) and historic data (solid) Kaplan-Meier curves for a slow event rate (a), a fast event rate (b), and the progression from mild cognitive impairment to Parkinson's disease dementia. The 'dashed' curves in (a) and (b) are generated for a covariate with skewness of 1, a standard deviation of 1.5, and a hazard ratio of 2. The distance is calculated as the maximum difference in step-size between the two curves at any given point.

Meier curve, $\hat{S}_g(t)$, for the "generated" data. Finally, to compare the historical Kaplan-Meier curve to the "generated" Kaplan-Meier curve we define a similarity measure, or distance measure, as the maximum difference in step size between the two curves at any time t_j. That is, we calculate the distance as $\max_j \left| \{\hat{S}_g(t_{j+1}) - \hat{S}_g(t_j)\} - \{\hat{S}_h(t_{j+1}) - \hat{S}_h(t_j)\} \right|$ for $j \in 1$ to $D-1$, where D is the number of unique event times in the historical data. If this maximum difference is greater than a pre-specified desired threshold, we resample "observed" data using the inverse-CDF method and repeat the procedure until the difference is below the threshold.

Figure 2.4 shows examples of the Kaplan-Meier curves generated from the selected distributions (dashed) and how they compare to the historic Kaplan-Meier curves (solid). Figure 2.4a shows a rapid event rate, Figure 2.4b shows a slower event rate, and Figure 2.4c shows the observed historical data for the progression from MCI to PDD in Pigott et al. (2015). Both Figure 2.4a and 2.4b are generated for a covariate with skewness of 1, standard deviation of 1.5, and a HR of 2. Figure 2.4c is generated as described in Section 2.6.1.

2.4.3. Follow-up Schedules

In this step, the user selects the follow-up schedules under consideration. The schedules chosen for testing should include a range of follow-up frequencies in settings of both unlimited and limited resources. The frequency of follow-up may vary over the duration of the study, but since this procedure is for design purposes, we assume that all participants strictly adhere to the follow-up schedules as defined.

2.4.4. Simulations and Analysis

Simulations are performed by generating S datasets with the parameters defined in Step 1 and Step 2. For each dataset we apply the follow-up schedules of interest and use a standard Cox regression model to estimate the HRs. For tied event times, the Efron method (Efron, 1977) is used. For comparison, we also run the Cox regression using the true, unobserved event times.

The impact of right-endpoint imputation is evaluated using the percent bias of the HR estimates obtained from using the observed, right-endpoint of the censoring interval. The percent bias is defined for each hazard ratio estimate ($\hat{HR}$), as $\frac{1}{SIM}\sum_{sim=1}^{SIM}\frac{\hat{HR}_{sim}-HR}{HR}\times 100\%$.

2.5. Method Validation

Here we aim to demonstrate that our new method is not sensitive to misspecification of the underlying survival distribution. The purpose of our proposed procedure is to evaluate follow-up frequency by estimating the bias due to right-endpoint imputation. If the true baseline survival distribution were known, then a simple, standard simulation could be used to understand the impact of various follow-up schedules. Therefore, we validate our new procedure by comparing our estimates of the bias to those obtained given the true survival distribution. Moreover, we show that the results from our procedure are consistent with the truth even when $S_0(t)$ is neither Weibull nor Gompertz.

We define the true survival distribution to follow proportional hazards with a log-logistic baseline survival function. Given the HR, covariates, and observation times, we generate one set of "historical" data and calculate the Kaplan-Meier curve. Using this as the historical Kaplan-Meier curve, we apply our new simulation procedure using the same HR and covariate distributions. We also conduct the same simulation, but generate data from the true survival distribution and compare the resulting mean $\hat{HR}$s and mean % bias.

The true survival distribution has a log-logistic baseline survival, which is written as $\frac{1}{1+\lambda t^{\alpha}}$ where we let $\alpha = 1\times 10^{-5}$ and $\lambda = 4$. We define two covariates with HR=1.5 and 2.0, respectively. The distributions associated with these covariates have skewness of 2.0 and 1.0 and σ of 1.5 and 1.0, respectively. We consider follow-up schedules of 2 or 5 year intervals for 20 years.

Table 2.2: Comparison of simulation results using the true and procedure estimated survival distributions

Schedule	Skew	σ	HR	$\hat{S}_0(t)$ $\hat{HR}$(%Bias)	$S_0(t)$ $\hat{HR}$(%Bias)	Difference $\hat{HR}$(%Bias)
1	2.0	1.5	1.5	1.48 (-1.27)	1.48 (-1.43)	0.00 (0.16)
1	1.0	1.0	2.0	1.97 (-1.50)	1.96 (-1.91)	0.01 (0.41)
2	2.0	1.5	1.5	1.36 (-9.54)	1.36 (-9.12)	-0.01 (-0.42)
2	1.0	1.0	2.0	1.70 (-14.79)	1.72 (-13.93)	-0.02 (-0.86)

$\hat{S}_0(t)$ is the baseline survival distribution estimated by the proposed procedure and $S_0(t)$ is the true baseline survival distribution. For each baseline distribution, the mean hazard ratio (HR) estimate, $\hat{HR}$, is defined as $\frac{1}{2000}\sum_{sim=1}^{2000}\hat{HR}_{sim}$, and the mean percent bias, $\%$ Bias, is define as $\frac{1}{2000}\sum_{sim=1}^{2000}\frac{\hat{HR}_{sim}-HR}{HR}\times 100\%$, where $\hat{HR}_{sim}$ is the HR estimated by the proportional hazards (PH) model for the s^{th} generated dataset and HR is the true HR. The last column presents the difference in the mean HR estimate and mean percent bias for data generated from $\hat{S}_0(t)$ and $S_0(t)$. Schedule 1 involves follow up every 2 years for 20 years. Schedule 2 involves follow up every 5 years for 20 years. HR were estimated using bivariate PH models and the observed (right-endpoint imputed) event times. The two predictors were generated from gamma distributions with the specified skewness (skew) and standard deviation (σ).

Table 2.2 shows the results obtained using the baseline survival estimated by our proposed procedure, $\hat{S}_0(t)$, and the true baseline survival curve, $S_0(t)$. The maximum difference in the mean estimated HR is 0.02. For the mean % bias, the maximum difference is 0.86. These results demonstrate that the new procedure can estimate the bias in the $\log(\hat{HR})$ well. They also confirm that our procedure is not sensitive to the true distribution of the baseline hazard. In this case, the true $S_0(t)$ is log-logistic, while $\hat{S}_0(t)$ is Weibull; however, there is little difference in the mean estimates of the parameters.

2.6. Application to Parkinson's Disease Dementia Research

2.6.1. Implementation

We apply the follow-up frequency evaluation procedure to PD cognition research and evaluate follow-up schedules for the study of time-to-progression from MCI to PDD at the University of Pennsylvania Parkinson's Disease Center. The study involved PD patients who were already MCI. The study design was to follow-up the patients until they develop PDD or the study ended. The main scientific question was to examine factors on the risk of conversion to PDD. The investigators want to know the appropriate frequency to follow up these patients in order to produce unbiased analysis results without wasting resources and increasing patient burden. We consider three follow-up schedules. Schedule 1 requires annual follow-up for ten years, schedule 2 requires annual follow-up for four years followed by biennial follow-up for six years, and schedule 3 requires annual follow-up for five years followed by biennial follow-up for four years. The total duration of follow-up is ten years for schedules 1 and 2, and nine years for schedule 3. The maximum number of visits is ten for schedule 1, and seven for schedules 2 and 3. The observed event time for each subject is defined as the right-endpoint of the interval in which the true time-to-progression falls. For example, if a participant's progression occurs at 5.5 years, the observed event time would be 6 under schedules 1 and 2, and 7 under schedule 3. Subjects who progress after ten years are right-censored under schedules 1 and 2 and participants who progress after nine years are right-censored under schedule 3. In our simulations we generate SIM=1000 datasets with n=200 participants and use a similarity threshold of 0.06.

To select an appropriate follow-up schedule, we ran our procedure using the seven predictors Pigott et al. (2015) included in their model. These predictors are sex and age, disease duration, Hoehn & Yarh (H&Y) stage, Unified Parkinson's disease rating scale (UPDRS) motor score, geriatric depression score, and dementia rating score-2 at first study visit. To obtain HRs to test, we ran a standard Cox regression analysis with the time-to-progression from MCI to PDD as the outcome and the seven predictors as covariates. In the procedure, we use $\exp\left[\text{abs}\left(\log \hat{HR}\right)\right]$ as the HR for each predictor.

Finally, we had to define distributions for each of the covariates. Sex and H&Y stage are binary and categorical variables, respectively. Therefore we sampled these covariate values with probabilities equal to those seen in the data. For all other predictors, we calculated the standard deviation and skewness. If the skewness was less than 0.5, we sampled the predictor from a normal distribution using the sample mean and standard deviation. If the skewness was greater than 0.5, we sampled from a gamma distribution. The normally distributed predictors were age, UPDRS motor score, and dementia score. The skewed predictors were disease duration and depression score.

2.6.2. Results of Evaluation of Follow-up Schedules for PDD Patients

The results for the Parkinson's disease dementia study are shown in Table 2.3. Sex is the predictor with the largest HR and as a result the largest bias. Under schedules 1 and 3, the HR estimate for sex has approximately 5% bias, but under Schedule 2 the bias is about 7.5%. All other predictors have bias less than or equal to 3% in the HR estimates. Thus, all three schedules produce less than 10% bias. Therefore, we would recommend using either Schedule 2 or 3, which require 7 follow-up

Table 2.3: Right-endpoint imputation bias by follow-up schedule in the study of MCI to PDD

			% Bias	
Predictor	**HR**	**Schedule 1**	**Schedule 2**	**Schedule 3**
Sex	1.736	-5.09	-7.40	-5.49
Age	1.004	-0.05	-0.07	-0.06
Duration	1.094	-1.12	-1.46	-1.20
H&Y Stage	1.058	-0.61	-1.02	-0.68
UPDRS Motor	1.115	-1.30	-1.83	-1.43
Depression	1.161	-1.76	-2.35	-1.87
Dementia	1.211	-2.20	-3.12	-2.41

A multivariate proportional hazards (PH) model was run to estimate the hazard ratios (HR) of all 7 predictors. HRs estimated from a multivariate PH model using historical data were used for the simulations. $\%$ Bias $= \frac{1}{1000}\sum_{sim=1}^{1000}\frac{\hat{HR}_{sim}-HR}{HR}\times 100\%$. Schedule 1 involves annual follow-up for 10 years. Schedule 2 involves annual follow-up for 4 years followed by biennial follow-up for 6 years. Schedule 3 involves annual follow-up for 5 years followed by biennial follow-up for 4 years. In each setting the observed (right-endpoint imputed) event times were used.

visits, since both of these schedules provide sufficient follow-up while utilizing fewer resources than Schedule 1, which requires 10 visits.

2.7. Discussion

We have introduced a novel procedure to evaluate the appropriateness of follow-up frequencies and demonstrated its application for a Penn Parkinson's Disease Center study on dementia. This procedure provides a quantitative method to guide the design of time-to-event studies by utilizing historical data. Although we apply the method to the PD cognition setting, our procedure can be used in any research area that has sufficient historical data to enable the selection of appropriate survival and covariate distributions. Using our method to evaluate the bias for different follow-up designs can guide the selection of longitudinal follow-up frequency in a quantitative and robust way. Thus, it will help to save unnecessary study costs and reduce patient burden without sacrificing the accuracy in estimating the associations of interest.

Here we have focused on the estimation bias associated with right-endpoint imputation, unlike the methods discussed in the Introduction which consider the efficiency of estimation. As described above, those methods rely on parametric assumptions about the underlying hazards. When the assumptions hold, the parametric methods are unbiased, allowing for a meaningful assessment of efficiency. In contrast, we consider the popular and commonly used semiparametric Cox PH model. The advantage of the proportional hazards model is that it does not specify or estimate the baseline hazard. However, by imputing the event time as the right-endpoint of the censoring interval, bias is introduced into the estimation. Since we expect the estimate to be biased, we therefore focus on the magnitude of that bias.

An advantage of our procedure is that it does not require parametric assumptions about the survival distribution. We only require the distribution to satisfy the proportional hazards assumption of the Cox regression model. While the procedure uses Weibull or Gompertz to select a baseline survival distribution, our method is not sensitive to the selected distribution. The Cox regression model does not use or estimate the baseline hazard. Instead, it uses only the order of events to estimate the HR. Therefore, the chosen distribution only needs to have a similar shape to the historical data so that the relative number of events observed at each time-point is consistent with the historical data.

We consider the chosen distribution to be “similar” to the observed data if the distance measure is below a defined threshold. A smaller threshold would force the generated data to more strictly resemble the historical data and a larger value would allow for greater deviations. In choosing the threshold value, a user may want to consider the amount of information in the historical data, much like they would when defining a Bayesian prior. If there is a lot of prior information in the historical data, then a stricter threshold may be appropriate. However, if the historical data contains a small sample, then a higher threshold may be desired to allow for more uncertainty.

Although our method does not require parametric assumptions regarding the survival distribution, a limitation of our method is the need to correctly specify certain measures of the covariate distributions. As shown, the skewness and variance of the covariate distribution can greatly impact the bias in HR estimates due to right-endpoint imputation. If the distribution of the covariates in the target population is well known or can be estimated accurately, then this limitation is not a concern. However, if the distribution is not well known it would be necessary to do sensitivity analyses to test a variety of possible scenarios including highly skewed covariate distributions.

Longitudinal studies are increasingly important in early detection, prevention, and care management of neurodegenerative and other chronic diseases, but they are constrained by limitations on financial, administrative, and logistical resources. Therefore, we recommend carefully designing the follow-up frequency to minimize bias in statistical analyses without wasting resources or increasing patient burden.

CHAPTER 3

NONPARAMETRIC ESTIMATION FOR TIME-VARYING MISSING COVARIATES IN LONGITUDINAL MODELS

3.1. Introduction

Longitudinal studies are a useful tool in biomedical research, offering numerous advantages over cross-sectional studies. By following the same subjects over times, longitudinal studies can be used to investigate the effects of predictors on disease and disease progression. In addition, they are often more powerful than standard cross-sectional studies. However, longitudinal studies can suffer from missing data that is challenging to deal with.

The simplest method for dealing with missing covariate data is to use a complete case analysis, which is the default for many statistical programs. The complete case method drops from analysis all observations for which there is missing covariate data, resulting in less efficient estimates (Erler et al., 2016; Johansson and Karlsson, 2013). In addition, if the data are not missing completely at random (MCAR) the complete case estimates are known to be biased (Erler et al., 2016; Johansson and Karlsson, 2013).

Other common methods for dealing with missing covariate data, such as fully Bayesian methods, Estimation-Maximization methods (Ibrahim, 1990), and multiple imputation with chained equations (MICE) require parametric modeling of the covariate distributions (Erler et al., 2016; Ibrahim et al., 2005). MICE is particularly common and considered by some to be the current gold standard (Erler et al., 2016) for dealing with missing data. However for unbalanced longitudinal data it is unclear how the parametric models should be defined (Erler et al., 2016; Moons et al., 2006) and model misspecification can lead to bias (Little and Rubin, 2002). In recent years, there has been development in nonparametric multiple imputation, such as predictive-mean matching, which does not require the strong distributional assumptions. While these methods have been shown through simulation to work well, they lack statistical justification or theory for statistical inference (Bertsimas, Pawlowski, and Zhuo, 2018).

Estimating equations with inverse probability weights produce consistent and asymptotically normal estimates but are less efficient than the other methods and require that the data are missing at random (Ibrahim et al., 2005). If the data are MCAR these methods are not applicable.

In some cases, data can be missing by study design. If a predictor is expensive or invasive to measure, a study may be designed such that the predictor is only measured for a random sample of study participants. For example, a Parkinson's disease (PD) study at the University of Pennsylvania is interested in the relationship between biomarkers and changes in cognitive outcomes over time. One biomarker of interest is the cerebral spinal fluid concentration of amyloid-β (CSF-aβ). Whereas some biomarkers can be measured from saliva or blood, CSF-aβ requires a more invasive and costly lumbar puncture. As a result, only a subset of study participants is randomly assigned to undergo this additional procedure to measure their CSF-aβ. For all other study participants CSF-aβ is missing. Those subjects with non-missing data make up an 'internal validation set'.

A few nonparametric methods have been developed for missing covariate data with an internal validation set that do not require specifying parametric distributions for the covariates. Pepe and Fleming (1991) and Carroll and Wand (1991) developed similar methods for missing covariate data based on an estimated likelihood that employs a nonparametric estimate of the density of the missing covariate given an "auxiliary" or "surrogate" variable. However, Pepe and Fleming (1991) requires a discrete auxiliary variable and Carroll and Wand (1991) assumes a discrete outcome. Xu, JK Kim, and Li (2017) uses expected estimating equations to develop a general theory that is applicable to situations in which the auxiliary and outcome variables are continuous or discrete. Because their variance estimate underestimates the asymptotic variance, Xu, JK Kim, and Li (2017) recommends using a bootstrap method to calculate the variance of the parameter estimates. All of the above methods assume a cross-sectional design with time-independent covariates.

We propose to extend these nonparametric methods to longitudinal data. Our method offers two main contributions to the existing literature. First, our method can handle time-independent or time-varying missing covariates. For time-independent missing covariates, we allow for continuous or discrete auxiliary variables. In addition, if the missing covariate is time-varying we can allow for time-varying auxiliary variables. We allow for all of these situations without introducing additional parametric assumptions beyond those required for the standard linear mixed-effects model. Second, we derive the asymptotic distribution of the estimator and show through simulations that it has good finite sample properties. Therefore, the variance estimate of our estimator does not require bootstrapping.

The rest of the chapter is organized as follows. First we describe our proposed nonparametric estimator in Section 3.2 and discuss its asymptotic properties in Section 3.3. In Section 3.4 we demonstrate the performance of our method through the use of simulation studies. Then we apply our method to the Parkinson's disease data example in Section 3.5. Finally, we consider limitations of our method in Section 3.6.

3.2. Proposed Nonparametric Maximum Estimated Likelihood

Let Y be a continuous outcome with n repeated measures and let X and Z be time-varying or time-independent variables. Define X and Z as matrices with n rows and q_X and q_Z columns, respectively. We assume that Z is measured for all subjects but X is only available for a random subsample.

Those subjects for whom X is measured make up the validation set V. Subjects who are missing X make up the nonvalidation set $\bar{V}$. Here we note that for all subjects X is either fully observed or not observed at all. This means that for subjects in the validation set, each part of X is observed if $q_X > 1$ and measured at each observation time if X is time-varying.

In addition, we assume Z can be decomposed into (Z^*, A), where Z^* is the components of Z that are independent of the missing covariate and A is an auxiliary variable that contains some information about X. Thus, the observed data consists of (Y_i,X_i,Z_i^*,A_i) for $i \in V$ and (Y_j,Z_j^*,A_j) for $j \in \bar{V}$.

Now that we have defined A, we can explain what we mean by X is available for a random subsample. We assume that the missing mechanism is independent of the auxiliary variable, but not necessarily independent of the other covariates. So this assumption is less restrictive than the MCAR assumption.

We define a linear mixed-effects model for Y_i as

$$Y_i = X_i\beta_X + Z_i^*\beta_Z{}^* + \gamma_i b_i + \epsilon_i, \tag{3.1}$$

where γ_i is an $n_i \times q_\gamma$ known matrix of covariates for the $q_\gamma \times 1$ vector of random-effects b_i, ϵ_i is the

$n_i \times 1$ vector of random errors, and n_i is the number of observations for subject i. In addition, we assume as usual that ϵ_i's are independent and follow an n_i-variate normal distribution with mean 0 and variance $\sigma^2 \Lambda_i(\nu)$ where ν defines the parameters of Λ_i, b_i's are iid, independent of ϵ_i, and follow a q_γ-variate normal distribution with mean 0 and variance D. $P_\beta(Y_i|X_i, Z_i)$ then follows a multivariate normal distribution with mean $\mu_i = X_i\beta_X + Z_i^*\beta_Z{}^*$ and variance $\Sigma_i = \gamma_i D \gamma_i^T + \sigma^2\Lambda_i(\nu)$. Note that A_i is not used in the model to prevent problems due to the collinearity in A and X.

The full likelihood for the data in the validation and nonvalidation sets can be expressed as in Pepe and Fleming (1991) as

$$L = \prod_{i \in V} P_\beta(Y_i|X_i, Z_i) \prod_{j \in \bar{V}} P_\beta(Y_j|Z_j). \tag{3.2}$$

If the distribution of $P(X|A)$ were known, then $P_\beta(Y|Z)$ could be calculated as $\int P_\beta(Y|x, Z)P(X|A)dx$. However, $P(X|A)$ is not known and even if it were the calculation of $P_\beta(Y|Z)$ would likely require some form of numerical integration. Instead, following Pepe and Fleming (1991) and Carroll and Wand (1991), we obtain unbiased, nonparametric estimates of $P_\beta(Y|Z)$ using empirical estimates of $P(X|A)$ based on the random subsample that makes up the validation set. Since $P(X|A) = \frac{P(X,A)}{P(A)}$, we need estimates for $P(A)$. The empirical estimate for the distribution of discrete A is $\hat{f}_A(a_j) = \frac{1}{n^v}\sum_{i \in V} I(A_i = A_j)$, where n^v is the size of the validation set. For continuous A, kernel density estimates are used so that $\hat{f}_A(a_j) = \frac{1}{n^v h}\sum_{i \in V}\Phi(\frac{a_i - a_j}{h})$, where Φ is a symmetric density function and h is the bandwidth. Using these empirical estimates of $P(A)$, we can obtain unbiased estimates of $P(Y_j|Z_j)$ as defined below.

For brevity of notation, let $w_D = \frac{1}{n^v}$ and $w_C = \frac{1}{n^v h}$. Then, if X is time-independent, an unbiased estimate of $P(Y_j|Z_j)$ for subject j from the nonvalidation set can be written

$$\begin{aligned} \hat{P}(Y_j|Z_j) &= \frac{w_k \sum_{i \in V} P(Y_j|X_i, Z_j)K_k(A)}{w_k \sum_{i \in V} K_k(A)} \\ &= \frac{\sum_{i \in V} P(Y_j|X_i, Z_j)K_k(A)}{\sum_{i \in V} K_k(A)}, \end{aligned} \tag{3.3}$$

where $k = D, C$ for discrete or continuous A, respectively, and $K_D(A) = I(A_i = A_j)$ and $K_C(A) = \Phi\left(\frac{A_i - A_j}{h}\right)$. Note that $\hat{f}_a(A_j) = w_k \sum_{i \in V} K_k(A)$.

For time-varying covariates, we introduce the following notation. Let M_i be an $n_i \times q$ matrix representing X_i or A_i, where $q = q_X$ or q_A, respectively. Let t_i be an $n_i \times 1$ vector where t_i is time, which we assume is discrete. Then $M_i[t_j]$ is an $n_j \times q$ matrix with the rows of M_i that correspond to the positions where the elements of t_j are equal to the elements of t_i. For example, if $X_i = (1.2, 1.5, 1.3)'$, $t_i = (0, 1, 2)'$, and $t_j = (0, 2)'$, then $X_i[t_j] = (1.2, 1.3)'$. It is necessary to recognize that $M_i[t_j]$ will have n_j rows only if t_j is at least a subset of t_i. In other words, a subject i from the validation set can only contribute to the estimation of $\hat{P}(Y_j|Z_j)$ for a subject j from the nonvalidation set if $t_j \subseteq t_i$. We incorporate this condition into $K_k(A, t)$ which is the time-dependent version of $K_k(A)$ from Eq. 3.3.

Now we can define the estimate $\hat{P}(Y_j|Z_j)$ for a time-varying X as

$$\hat{P}(Y_j|Z_j) = \frac{\sum_{i \in V} P(Y_j|X_i[t_j], Z_j) K_k(A, t)}{\sum_{i \in V} K_k(A, t)}. \tag{3.4}$$

If A is time-independent, then $K_D(A, t) = I\left(A_i = A_j, t_j \subseteq t_i\right)$ and $K_C(A, t) = \Phi\left(\frac{A_i - A_j}{h}\right) I(t_j \subseteq t_i)$. If A is time-varying, $K_D(A, t) = I(A_i[t_j] = A_j, t_j \subseteq t_i)$ and $K_C(A, t) = \mathbf{\Phi}\left(\frac{A_i[t_j] - A_j}{h}\right) I(t_j \subseteq t_i)$, where $\mathbf{\Phi}\left(\frac{A_i[t_j] - A_j}{h}\right) = \Phi\left(\frac{A_i[t_{j1}] - A_j[t_{j1}]}{h}\right) \times \Phi\left(\frac{A_i[t_{j2}] - A_j[t_{j2}]}{h}\right) \cdots \times \Phi\left(\frac{A_i[t_{jn_j}] - A_j[t_{jn_j}]}{h}\right)$.

Then the estimated likelihood can be written as

$$\hat{L} = \prod_{i \in V} P_\beta(Y_i|X_i, Z_i) \prod_{j \in \bar{V}} \hat{P}_\beta(Y_j|Z_j). \tag{3.5}$$

We maximize this estimated likelihood using a pseudo Newton-Raphson algorithm and show that doing so yields consistent and asymptotically normal estimates for the unknown parameters.

3.2.1. Practical Considerations for Continuous Auxiliary Variables

Use of the kernel density estimator for continuous auxiliary variables introduces two important factors for consideration. First is the choice of bandwidth. Similar to Carroll and Wand (1991), we also use an *ad hoc* method to select the bandwidth based on the validation data. Specifically, we calculate the bandwidth based on the validation set auxiliary variable using the method of Sheather and Jones (1991), which is implemented as `bw.SJ` in `R` (R Core Team, 2018).

A second consideration for continuous auxiliary variables is how to handle nonvalidation data at or beyond the edge of the validation data. Consider the denominator of $\hat{P}(Y_j|Z_j)$ in Eq. 3.3 and Eq. 3.4 for a continuous auxiliary variable. If A_j, the auxiliary variable for the nonvalidation subject, is outside or near the edge of the range of A in the validation set, then $\Phi\left(\frac{A_i - A_j}{h}\right)$ will be small for all $i \in V$ and $\sum_{i \in V} K_C(A)$ or $\sum_{i \in V} K_C(A,t)$ will be close to zero. This can introduce bias and numerical instability to the estimate. Therefore, it is necessary to restrict the nonvalidation set to those subjects whose auxiliary values are interior to the auxiliary values in the validation set. How the 'interior' nonvalidation set is defined results in the common trade-off between bias and variance. More restrictive thresholds on the nonvalidation auxiliary variable result in smaller bias but reduce the size of the 'interior' nonvalidation set, thereby increasing the variance. For our simulations in Section 3.4, we use the second and third quartiles of the validation set auxiliary values as thresholds for inclusion of subjects in the 'interior' nonvalidation set.

3.3. Asymptotic Properties of the Maximum Estimated Likelihood Estimator

The asymptotic properties of the proposed estimator, i.e. the properties of the estimator as $N \to \infty$, are derived following similar arguments to those in Pepe and Fleming (1991) with modifications for continuous auxiliary variables and time-varying missing covariates and auxiliary variables. First, assume that the validation set V is a random subsample of the subjects from the data in the sense that the missing mechanism does not depend on A. Then let ρ^v be the proportion of subjects in the validation set. Assume that $\lim_{n\to\infty} \rho^v > 0$. For both discrete and continuous auxiliary variables, the maximum likelihood estimates $\hat{\beta}$ obtained by solving $\frac{\partial}{\partial\beta} \log \hat{L} = 0$ are asymptotically distributed

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \mathcal{I}^{-1} + \frac{(1-\rho^v)}{\rho^v}\mathcal{I}^{-1}\Sigma(\beta)\mathcal{I}^{-1}\right), \tag{3.6}$$

where $\mathcal{I}$ is the information matrix for the true likelihood $L(\beta)$ in Eq. 3.2 and $\Sigma(\beta) =$ var$\left\{\mathbb{E}\left[\frac{\partial \log P_\beta(Y|Z)}{\partial\beta}|X, A\right]\right\}$. The variance for $\hat{\beta}$ can be thought of in terms of two components that correspond respectively to the regular maximum likelihood estimate variance ($\mathcal{I}$) plus a penalty for estimating the probabilities for the nonvalidation set. Consistent estimates of $\mathcal{I}$ and Σ are given by

$$\hat{\mathcal{I}} = -\frac{\partial^2}{\partial\beta\beta^T} \log \hat{L} \tag{3.7}$$

and

$$\hat{\Sigma}(\hat{\beta}) = \text{vâr}\left\{\hat{\hat{W}}_{X_i,A_i}(\hat{\beta}), i \in V\right\}, \tag{3.8}$$

where vâr is the sample variance of $\bar{W}_{X_i,A_i}$, which is defined differently for discrete and continuous auxiliary variables. For a discrete auxiliary variable and time-independent X, $\hat{\bar{W}}_{X_i,A_i}$ is defined as in Pepe and Fleming (1991) as

$$\hat{\bar{W}}_{X_i,A_i} = \frac{\sum_{j\in\bar{V}} \left\{ \frac{\partial P(Y_j|X_i,Z_j)/\partial\beta}{\hat{P}(Y_j|Z_j)} - \frac{\partial\hat{P}(Y_j|Z_j)/\partial\beta}{[\hat{P}(Y_j|Z_j)]^2} P(Y_j|X_i,Z_j) \right\} K_D(A)}{\sum_{j\in\bar{V}} K_D(A)}.$$

If X is time-varying, $K_D(A)$ is replaced with $K_D(A,t)$.

For a continuous auxiliary variable and time-independent X, $\hat{\bar{W}}_{X_i,A_i}$ can be defined

$$\hat{\bar{W}}_{X_i,A_i} = \frac{n^V}{n^{\bar{V}}} \sum_{j\in\bar{V}} \left\{ \left[\frac{\partial P(Y_j|X_i,Z_j)/\partial\beta}{\hat{P}(Y_j|Z_j)} - \frac{\partial\hat{P}(Y_j|Z_j)/\partial\beta}{[\hat{P}(Y_j|Z_j)]^2} P(Y_j|X_i,Z_j) \right] \frac{K_C(A)}{\sum_{i\in V} K_C(A)} \right\}.$$

Again, if X is time-varying, $K_C(A)$ is replaced with $K_C(A,t)$.

The details for the derivation of the $\hat{\bar{W}}_{X_i,A_i}$ and the asymptotic properties of the estimator for a continuous auxiliary variable are provided in Appendix B.1.

3.4. Simulations

To evaluate the performance of our proposed estimator, we perform a series of simulations. We consider settings in which the missing and auxiliary variables are time-independent and time-varying. In addition, we let the auxiliary variable be continuous and discrete. Here we describe the simulations for auxiliary variables that are continuous and time-independent and discrete and time-varying. Simulations for a discrete time-independent auxiliary variable are described in Appendix B.2.

For each simulation setting, we generate longitudinal data based on the standard linear mixed-effects model in Eq. 3.1. The covariates included in the model are an intercept, time, and X, which will be missing for a subset of the sample. In addition, we include random intercepts and random slopes. Thus, $Z_i = \gamma_i = \begin{bmatrix} \underline{1} & t_i \end{bmatrix}$, where t_i is time in years.

In the simulations, we compare the performance of three estimators; the complete case estimator, the proposed estimator, and the oracle estimator. The complete case estimator drops all subjects with missing data from the analysis, the proposed estimator is implemented as described in Section 3.2, and the oracle estimator uses the unobservable full data. Using 1000 iterations of each simulation, we calculate the mean bias ($\hat{\beta} - \beta$), observed sample standard deviation (SD), mean estimated standard errors ($\hat{SE}$), mean relative efficiency (RE) compared to the oracle estimator where a lower RE is more efficient, and 95% coverage (Cov).

3.4.1. Time-Independent Missing Covariate with a Continuous Auxiliary Variable

First we describe the simulations involving a time-independent missing covariate X and a continuous auxiliary variable A. Since X is time-independent, this implies that A is also time-independent. To generate correlated X and A, we use a standard multivariate normal distribution where $\left(\begin{smallmatrix} X \\ A \end{smallmatrix}\right) \sim N\left[\left(\begin{smallmatrix} \mu_X \\ \mu_A \end{smallmatrix}\right), \left(\begin{smallmatrix} \sigma_X^2 & \rho\sigma_X\sigma_A \\ \rho\sigma_X\sigma_A & \sigma_A^2 \end{smallmatrix}\right)\right]$ and ρ is the correlation between X and A. We simulate data for ρ= 0.01, 0.25, .50, 0.75, and 1.0.

To create missing data, we first generate the full data, including the covariates, auxiliary variable, and outcome, for N=400 subjects. All subjects are considered to have observations at baseline and year one, but one-third of subjects are lost to follow-up at year two. Therefore, the data is balanced but incomplete. Then we randomly select subjects to be missing X to ensure that the missing mechanism does not depend on the auxiliary variable. We simulate data in which 25%, 50% and 75% of subjects are missing X.

Table 3.1 summarizes the results for 50% missing data. All three analyses (complete case, proposed, and oracle) have little bias and a good 95% coverage probability in this setting of moderate missingness. Similarly results are seen for 25% missing data (Appendix Table B.5). When the percent missing is high (75%) and the correlation between the missing and auxiliary variables is perfect (ρ=1.0), the proposed method is slightly more biased (Appendix Table B.6). Nevertheless, the mean $\hat{SE}$ estimates calculated based on the asymptotic theory for the proposed estimator is similar to the observed sample SD under all conditions. As a result, the coverage probability for 75% missing data is slightly low (92%) when the correlation is high.

The RE is calculated for each estimator as $\frac{1}{1000}\sum_{sim=1}^{1000}\frac{\hat{SE}_{m,sim}}{\hat{SE}_{oracle,sim}}$, where m = complete case, proposed, or oracle. The RE for the oracle estimator is 1 and larger values are less efficient. For the proposed estimator, the efficiency of the estimator increases with the correlation between X and A, but this result is more pronounced for high missingness. Under all conditions, including high missingness and low correlation, the proposed estimator is more efficient (smaller RE) than the complete case estimator. Moreover, even if the auxiliary variable provides little to no information about X, the proposed method is not necessarily equal to the complete case analysis with respect to relative efficiency. See Appendix B.1.2 for justification of this observation.

In these simulations, the RE of the proposed estimator is never 1, even for a perfectly correlated auxiliary variable (i.e. $\rho = 1$). This is because the proposed method does not use all of the subjects in the analysis. Instead, only those subjects in the validation set and 'interior' nonvalidation set are used in the analysis of the proposed method. Those subjects who are in the nonvalidation set but not the 'interior' nonvalidation set are not used. Thus the total sample size of the proposed method is smaller than that of the oracle method. However, in Appendix B.1.1 we show that the proposed estimator is fully efficient (i.e. RE = 1) for a perfectly correlated auxiliary variable if the sample used in the oracle analysis is restricted to be the same sample used in the proposed analysis. We should note that for a discrete auxiliary variable, the size of the oracle sample is equal to the size of the proposed method analysis sample since the 'interior' nonvalidation set is equal to the nonvalidation set.

3.4.2. Time-Varying Missing Covariate with a Time-Independent Continuous Auxiliary Variable

Next, we describe the simulations involving a time-varying covariate and a time-independent continuous auxiliary variable. When X is time-varying but A is time-independent, we generate X in two steps. First, we generate $\bar{X}_i$ and A_i, where $\bar{X}_i$ is the mean for X_i, from a multivariate normal distribution where $\begin{pmatrix} \bar{X}_i \\ A_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_A \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_A \\ \rho\sigma_X\sigma_A & \sigma_A^2 \end{pmatrix}\right]$. Second, we generate n_i observations of X_i from $N\left[\bar{X}_i, \left(\frac{\sigma_X}{2}\right)^2\right]$. Since X is time-varying but A is not, the correlation between X and A will never be exactly 1.0. Therefore, we only consider correlations below 1. Again we set the total sample size to be N=400 with one-third of subjects being lost to follow-up at year two. We also let the percent of subjects with missing X be 25%, 50%, and 75%. We evaluate the bias, efficiency, and 95% coverage probability for the three estimators for $\rho =$ 0.01, 0.25, 0.50, and 0.75. In addition, we compare the efficiency of the three methods for values of ρ at 0.05 increments between 0.2 and 0.75.

Table 3.2 shows the results for 50% of subjects with missing data. The results for 25% and 75% missing are provided in Appendix Tables B.7 and B.8. Again, the proposed method is unbiased and has good coverage for low to moderate missingness, but is slightly biased when the missingness and correlation are 75%. As in the previous section, the small bias results in a slightly lower coverage probability of 92%. Still, the proposed method is at least as efficient, if not more, than the complete case analysis for all scenarios.

Figure 3.1 plots the RE of the complete case and proposed estimators by the percent of subjects missing X and the correlation between X and A. When the percent missing is below 75%, the proposed method is as or more efficient than the complete case estimate for all three correlation conditions. As the percentage of missing data increases we see more efficiency gains in the proposed approach when the correlation is higher.

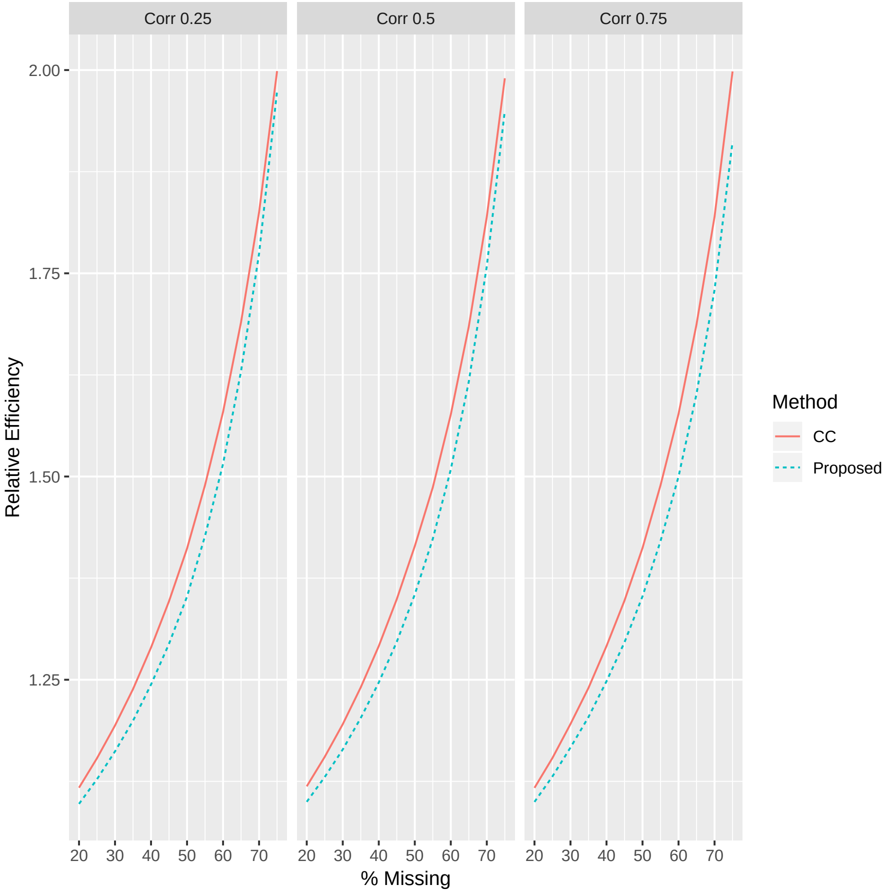


Figure 3.1: Relative efficiency vs % missing by correlation for a time-varying missing covariate and a continuous auxiliary variable. Relative efficiency is calculated as the mean of $\frac{\hat{SE}_m}{\hat{SE}_{oracle}}$ where m is ‘complete case’ (CC) or ‘proposed’. % missing is defined as the percentage of subjects in the nonvalidation set. ρ is the correlation between the missing covariate and the auxiliary variable.

3.4.3. Time-Varying Missing Covariate with a Time-Varying Discrete Auxiliary Variable

When A is continuous and time-varying, the proposed estimator can be biased and inefficient, as discussed in Section 3.6. Therefore, we only consider a discrete time-varying auxiliary here. For these simulations we generate balanced and complete data with three observations for each of the N=2000 subjects. The larger sample size is necessary in these situations to have a sufficient number of validation subjects who contribute to each estimate of $\hat{P}(Y_j|Z_j)$. For a validation subject to contribute to the estimate of $\hat{P}(Y_j|Z_j)$, the time-varying auxiliary variable must be matched at each timepoint (i.e. $I\left(A_i[t_j] = A_j, t_j \subseteq t_i\right)$) instead of just once (i.e. $I\left(A_i = A_j, t_j \subseteq t_i\right)$).

To generate data for a time-varying missing covariate and a time-varying discrete auxiliary variable, we first generate correlated continuous variables from the multivariate normal distribution described in Section 3.4.1. Since each subject has 3 observations, we generate $N \times 3$ draws from the specified distribution. Then we convert A to be a discrete variable by defining observations as 0, 1, or 2 based on the tertiles of A. We calculate the observed correlation between X and A using the Spearman correlation, which again will never be 1. In fact, the observed correlation is always smaller than the specified ρ. Therefore, in the simulations we set ρ = 0.01, 0.30, 0.57, and 0.95, to achieve the desired correlations of 0.01, 0.25, 0.5, and 0.75. We also let the percent missing be 25%, 30%, and 50%.

The results for these simulations are shown in Table 3.3 and Appendix Tables B.9 and B.10. For 25% and 30% missing data, the proposed method is unbiased and more efficient than the complete case analysis for all correlations. Table 3.3 also shows more pronounced gains in efficiency with increasing correlation, where the RE ranges from 1.18 for near zero correlation to 1.14 for a correlation of 0.75. When the percent missing is higher, such as 50%, and the correlation between X and A is low, the proposed method can be slightly biased and a little less efficient than the complete case analysis. However, when the correlation between the auxiliary variable and missing covariate is high, the proposed method performs well. For a correlation 0.75 between X and A and 50% missing data, the proposed method is unbiased, more efficient than the complete case analysis, and has good coverage. Therefore, when the missingness is greater than 30%, unless the correlation is very high between the missing and auxiliary variable, it may be beneficial to use a time-independent auxiliary variable instead of a time-varying one. For example, if scientifically reasonable, one may use the baseline value, the mean, or some other summary statistic of the auxiliary variable in place of its time-varying value.

3.5. Application to Parkinson's Disease Dementia Research

In this section, we apply our proposed method to Parkinson's disease (PD) using the Udall Intensive Cohort data. This data is from an ongoing study at the University of Pennsylvania Parkinson's Disease Center where 408 PD patients are followed longitudinally for clinical and cognitive assessments annually for the first four years and then biennially thereafter. Among these patients, only a fraction was randomly selected to receive extensive biomarker testing which included the measurement of CSF-aβ.

The purpose of this analysis is to understand how abnormal CSF-aβ, defined as values ≤ 192 ng/L (Shaw et al., 2009), affects the rate of change in the age adjusted Dementia Rating Scale total (DRStotalAge) in patients with PD. Since CSF-aβ values are only available for a subset of the study participants, we propose to use apolipoprotein E (APOE) genotype information as an auxiliary variable. APOE genotype information is available for most of the study participants and has been shown to be associated with CSF-aβ (Tapiola et al., 2000).

We consider the following mixed-effects model with subject-specific intercepts and slopes:

$$\begin{aligned} Y_i = b_{0i} + \beta_0 + \beta_1 \times \text{CSF-a}\beta_i + (\beta_2 + b_{1i}) \times \text{YEAR}_i \\ + \beta_3 \times \text{SEX}_i + \beta_4 \times \text{baseDRStotalAge}_i \\ + \beta_5 \times \text{CSF-a}\beta_i \times \text{YEAR}_i + \epsilon_i \end{aligned} \tag{3.9}$$

where Y_i=DRStotalAge, $\boldsymbol{b} \sim N\left[\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}D_{11} & D_{12}\\ D_{21} & D_{22}\end{smallmatrix}\right)\right]$, and $\epsilon_i \sim N(0, \sigma^2)$. CSF-a$\beta_i$ is the baseline CSF-aβ and defined as 0 (CSF-a$\beta >$ 192) or 1 (CSF-a$\beta \leq$ 192). YEAR$_i$ is the time from baseline defined starting from the first recorded DRStotalAge and rounded to the nearest 6 months. SEX$_i$ is 0 (female) or 1 (male) and baseDRStotalAge$_i$ is the baseline DRStotalAge.

We estimate the parameters of this model using our proposed method for time-independent missing values with a time-independent discrete auxiliary variable. Technically, we have two missing covariates; baseline CSF-aβ, which is time-independent, and its interaction with the time variable which creates a second, time-varying, missing covariate. However, since the interaction term can be created using CSF-aβ_i and the non-missing YEAR variable, we do not need to obtain $X_i[t_j]$. Therefore, we can use the time-independent version of our proposed estimator. We compare the results from our proposed method to those from a complete case analysis in which subjects with missing values are excluded from the study.

The raw data contained a total of 408 subjects. The validation set was defined as those subjects with non-missing APOE, non-missing CSF-aβ, a baseline DRStotalAge within 6 months of the CSF-aβ measure and at least two outcome observations, including baseline, over the course of eight years. In total, the validation set contained 134 subjects. The nonvalidation set consisted of 187 subjects (58% missing) with non-missing APOE genotype, missing CSF-aβ, and at least two observed outcomes within eight years.

The auxiliary variable used in our proposed method was APOE4, defined as 0, 1, or 2 corresponding to the number of APOE $\epsilon 4$ alleles. The Spearman correlation between the discrete CSF-aβ and APOE4 in the validation set was 0.18. To confirm that the auxiliary variable is independent of the probability of missing CSF-aβ, we performed a logistic regression and found that APOE4 is not a significant predictor of the probability of missing CSF-aβ (p-value = 0.43). Thus we do not have evidence against the assumption that the missing mechanism is independent of the auxiliary variable.

The results of the analyses for the complete case and proposed estimators are shown in Table 3.4. The proposed estimates are as or more efficient than the complete case estimates for all of the fixed-effects, with the non-missing covariates showing substantial gains in efficiency. As expected, since the complete case method is unbiased due to data missing by study design, the direction of association between the fixed-effects and cognitive decline is consistent between the two methods as well as with previously reported results (Pigott et al., 2015).

The primary scientific purpose of this analysis is to estimate the rate of change in DRStotalAge for PD patients with normal and abnormal CSF-aβ values. The rate of change is given by β_3 (YEAR) for patients with normal CSF-aβ values and $\beta_3 + \beta_5$ (YEAR + Interaction) for patients with abnormal CSF-aβ values. The estimated rate of change in DRStotalAge for PD patients with abnormal CSF-aβ is -0.645 ($\hat{SE} = 0.170$) in the complete case analysis and -0.699 ($\hat{SE} = 0.163$) for the proposed estimator, once again demonstrating that the proposed estimator is more efficient than the complete case analysis.

3.6. Discussion

We proposed a nonparametric estimator for longitudinal data with missing covariates and an internal validation subsample. By extending these nonparametric methods to longitudinal data we allow for time-varying missing covariates as well as time-varying auxiliary variables. In addition, we allow the auxiliary variable to be continuous or discrete. We derived the asymptotic distribution for our proposed estimator and proved that it is consistent and asymptotically normal. Through simulation studies, we showed that the estimator has good finite sample properties and is more efficient than the complete-case analysis even when the correlation between the missing covariate and auxiliary variable is small and there is moderate missing data. Finally, we applied our proposed estimator to a real data example for Parkinson's disease and showed that our method is more efficient than the complete case estimates in practice.

Our proposed estimator assumes that the missing mechanism is independent of the auxiliary variable. This assumption is valid for studies like the Parkinson's disease example, in which data is missing by study design. In the Parkinson's disease example, subjects were randomly assigned to a more invasive procedure. Other studies may randomly assign subjects to undergo more precise, but expensive testing, while everyone else has a cheaper, less accurate test. In these types of settings, the cheaper test could be used as an auxiliary variable for the more expensive covariate. If instead subjects were assigned to the expensive test based on their results of the cheaper test, our above missing data assumption would not hold and our proposed estimator would be biased. Therefore, it is important to consider how subjects will be selected for different types of tests or procedures at the design stage of the study.

When the missing covariate is time-varying, our proposed estimator has a couple of limitations. One limitation is that we assume discrete time. The discrete time assumption is necessary in order to define $X_i[t_j]$ and $A_i[t_j]$ in the estimation of $\hat{P}(Y_j|Z_j)$. Another limitation for time-varying missing covariates is the need for larger samples sizes. This requirement arises from the need to have a sufficient number of validation subjects with matching timepoints and matching auxiliary variables. When the data is not balanced and complete, there will be fewer subjects who "match on time" (i.e. $I(t_j \subseteq t_i)$). Therefore, a large sample size is needed to obtain a good estimate of $\hat{P}(Y_j|Z_j)$.

However, it should be noted that to “match on time”, a nonvalidation subject’s time only needs to be a subset of the validation subject’s time and not necessarily identical. Similarly, when the auxiliary variable is time-varying, larger sample sizes are needed since the auxiliary values must be matched at every timepoint. As such, our method may be applicable to electronic health records (EHR) data, which can contain large, longitudinal datasets.

Finally, we recognize that the proposed estimator does not perform well for continuous, time-varying auxiliary data. When the auxiliary variable is time-varying, multivariate kernel density estimation is needed to estimate $\hat{f}_A(a)$. Even for large sample sizes of N=5000, the proposed method is less efficient than the complete case estimator and even biased sometimes. Carroll and Wand (1991) acknowledge the same issue and note that high dimensional surrogates remains an open problem. Xu, JK Kim, and Li (2017) refers to this problem as the “curse of dimensionality”. However, this limitation can be overcome by converting a continuous time-varying auxiliary variable to a discrete time-varying auxiliary variable. Sometimes the auxiliary variable can be categorized using scientifically proven thresholds. If no proven threshold exists, the variable may be split into bins based on percentiles. For example, in our simulations we converted a continuous auxiliary variable into a discrete variable based on the tertiles of the continuous version of that variable. Although discretizing the auxiliary variable will reduces its correlation with the missing covariate, we will still see efficiency gains since the method performs well even for low correlations, as demonstrated in previous sections.

Table 3.1: Simulation results for 50% missing time-independent covariate with a continuous auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.0	β_X	-4.1E-03	0.097	0.095	1.41	0.948	1.9E-03	0.092	0.089	1.33	0.947	-3.6E-03	0.069	0.067	0.948
	β_0	1.6E-02	0.490	0.485	1.41	0.951	-8.8E-03	0.463	0.456	1.33	0.955	1.5E-02	0.352	0.343	0.950
	β_{Time}	-5.3E-03	0.117	0.118	1.41	0.953	-4.3E-03	0.097	0.097	1.16	0.951	-2.5E-03	0.085	0.084	0.938
0.2	β_X	-4.4E-03	0.097	0.095	1.41	0.943	-5.0E-03	0.092	0.089	1.33	0.946	-3.7E-03	0.068	0.067	0.953
	β_0	1.9E-02	0.493	0.485	1.41	0.950	2.0E-02	0.465	0.457	1.33	0.951	1.6E-02	0.345	0.343	0.952
	β_{Time}	-2.5E-03	0.116	0.118	1.41	0.954	-3.2E-03	0.097	0.097	1.16	0.951	-1.9E-03	0.085	0.084	0.938
0.5	β_X	-1.7E-03	0.099	0.095	1.41	0.940	-7.5E-03	0.094	0.090	1.33	0.947	-1.5E-03	0.068	0.067	0.951
	β_0	3.5E-03	0.500	0.485	1.41	0.939	2.7E-02	0.481	0.457	1.33	0.950	5.2E-03	0.350	0.343	0.946
	β_{Time}	-6.2E-03	0.117	0.118	1.41	0.949	-6.4E-03	0.099	0.097	1.16	0.945	-3.7E-03	0.086	0.084	0.935
0.8	β_X	-7.2E-04	0.098	0.095	1.42	0.946	-1.2E-02	0.095	0.090	1.34	0.944	-1.1E-03	0.067	0.067	0.950
	β_0	1.4E-04	0.495	0.485	1.42	0.941	5.6E-02	0.481	0.459	1.34	0.941	4.8E-03	0.345	0.343	0.945
	β_{Time}	-4.7E-03	0.117	0.119	1.42	0.950	-6.0E-03	0.098	0.097	1.16	0.943	-4.2E-03	0.086	0.084	0.941
1.0	β_X	-2.2E-04	0.100	0.095	1.42	0.932	-6.9E-03	0.096	0.090	1.33	0.930	1.0E-03	0.068	0.067	0.949
	β_0	1.5E-03	0.505	0.486	1.42	0.937	3.2E-02	0.482	0.455	1.33	0.929	-4.0E-03	0.346	0.343	0.947
	β_{Time}	-2.2E-03	0.121	0.119	1.42	0.951	-4.2E-04	0.098	0.097	1.16	0.949	-4.0E-03	0.088	0.084	0.935

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table 3.2: Simulation results for 50% missing time-varying covariate with a continuous auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	-1.6E-03	0.077	0.073	1.41	0.937	-2.1E-03	0.074	0.070	1.35	0.939	4.0E-04	0.053	0.052	0.948
	β_0	8.6E-03	0.400	0.378	1.41	0.932	1.1E-02	0.386	0.362	1.36	0.940	-2.8E-03	0.277	0.267	0.935
	β_{Time}	-2.2E-03	0.120	0.119	1.42	0.938	-4.1E-03	0.104	0.102	1.22	0.945	-1.8E-03	0.086	0.084	0.937
0.25	β_X	-1.9E-03	0.073	0.073	1.41	0.960	-4.5E-03	0.071	0.070	1.35	0.950	-1.1E-03	0.052	0.052	0.953
	β_0	1.1E-02	0.377	0.377	1.41	0.953	2.3E-02	0.370	0.360	1.35	0.944	6.3E-03	0.267	0.267	0.949
	β_{Time}	6.8E-03	0.119	0.119	1.41	0.943	8.9E-04	0.104	0.102	1.22	0.944	1.7E-03	0.086	0.084	0.943
0.50	β_X	-4.4E-03	0.075	0.072	1.42	0.946	-6.3E-03	0.072	0.069	1.35	0.949	-1.2E-03	0.052	0.051	0.953
	β_0	2.1E-02	0.387	0.374	1.42	0.940	2.9E-02	0.371	0.358	1.35	0.944	3.4E-03	0.265	0.264	0.952
	β_{Time}	-1.1E-02	0.121	0.119	1.42	0.955	-6.7E-03	0.105	0.102	1.22	0.940	-5.5E-03	0.082	0.084	0.947
0.75	β_X	4.0E-05	0.070	0.071	1.41	0.957	-9.5E-03	0.068	0.067	1.35	0.948	1.2E-03	0.048	0.050	0.955
	β_0	7.7E-03	0.367	0.366	1.41	0.958	5.5E-02	0.354	0.349	1.35	0.942	-8.6E-04	0.252	0.259	0.948
	β_{Time}	6.9E-03	0.119	0.119	1.41	0.949	6.2E-03	0.102	0.102	1.22	0.946	1.5E-03	0.084	0.084	0.950

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table 3.3: Simulation results for 30% missing time-varying covariate with a time-varying, discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	9.3E-04	0.020	0.020	1.19	0.949	3.3E-03	0.020	0.020	1.18	0.947	9.7E-04	0.016	0.017	0.950
	β_0	-4.8E-03	0.107	0.106	1.20	0.944	-1.7E-02	0.103	0.104	1.17	0.956	-6.0E-03	0.087	0.088	0.946
	β_{Time}	-1.7E-03	0.040	0.042	1.20	0.960	-1.1E-03	0.038	0.038	1.07	0.948	-1.1E-03	0.035	0.035	0.947
0.25	β_X	5.6E-05	0.020	0.020	1.20	0.941	1.6E-03	0.020	0.020	1.17	0.948	3.2E-04	0.016	0.017	0.950
	β_0	-1.1E-04	0.108	0.106	1.20	0.945	-8.3E-03	0.106	0.103	1.17	0.940	-2.0E-03	0.087	0.089	0.946
	β_{Time}	-6.8E-04	0.040	0.042	1.20	0.969	5.7E-05	0.036	0.038	1.07	0.966	7.1E-06	0.034	0.035	0.956
0.50	β_X	1.3E-04	0.021	0.020	1.19	0.943	-6.1E-05	0.020	0.019	1.16	0.948	5.9E-04	0.016	0.017	0.948
	β_0	-3.5E-04	0.111	0.106	1.19	0.937	-1.6E-04	0.105	0.102	1.16	0.944	-3.5E-03	0.088	0.089	0.948
	β_{Time}	-1.6E-05	0.040	0.042	1.19	0.970	-4.2E-04	0.036	0.038	1.06	0.961	-5.7E-04	0.035	0.035	0.950
0.75	β_X	2.8E-04	0.020	0.020	1.19	0.948	-9.0E-04	0.019	0.019	1.14	0.950	2.8E-04	0.016	0.017	0.956
	β_0	-1.7E-03	0.109	0.106	1.19	0.950	3.7E-03	0.102	0.101	1.14	0.942	-2.3E-03	0.088	0.089	0.958
	β_{Time}	3.3E-04	0.040	0.042	1.20	0.970	-4.8E-04	0.036	0.037	1.04	0.960	-3.0E-04	0.035	0.035	0.951

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table 3.4: Reslts of Parkinson's disease data example

	Complete Case			Proposed		
	$\hat{\beta}$	$\hat{SE}$	p-value	$\hat{\beta}$	$\hat{SE}$	p-value
(Intercept)	2.286	0.461	6.9E-07	1.539	0.298	2.4E-07
ABETA	-0.309	0.304	3.1E-01	-0.251	0.268	3.5E-01
Year	-0.197	0.077	1.1E-02	-0.254	0.059	1.7E-05
Sex	-0.340	0.228	1.4E-01	-0.320	0.164	5.1E-02
baseDRS	0.827	0.038	7.9E-106	0.886	0.025	7.9E-270
ABETA:Year	-0.448	0.187	1.7E-02	-0.445	0.187	1.7E-02

Complete Case = complete case estimator. Proposed = proposed estimator.

$\hat{\beta}$ = parameter estimate. $\hat{SE}$=estimated standard error

CHAPTER 4

LMEMVP - AN R PACKAGE FOR LINEAR MIXED EFFECTS MODELS WITH MISSING VALUES IN PREDICTORS.

4.1. Introduction

Scientific studies are often constrained by various factors, including funding, manpower, and time, that can impact the type and quality of the data that are collected. Consider a situation in which a covariate, or predictor, of interest requires an expensive, invasive, or time-consuming procedure such as a lumbar puncture, endoscopy, or echocardiogram. A study may not have the resources to allow for all study participants to undergo the procedure. Those who do not have the procedure will have missing data. However, there may exist a cheaper, faster procedure or test for an “auxiliary” or “surrogate” variable that can be measured for everyone and can provide some information about the missing predictor of interest. The goal then is to use the information provided by the auxiliary variable to better (i.e. more efficiently) estimate the effects of the predictor on the outcome being studied.

The default method for dealing with missing data for many programs in **R** (R Core Team, 2018), including `lm` and `glm,` is to drop observations with missing values from the analysis. This ‘complete case’ analysis yields inefficient estimates that can be biased if the data are not missing complete at random (MCAR) (Erler et al., 2016; Johansson and Karlsson, 2013).

Perhaps the most popular approach for handling missing data is to use multiple imputation methods. These methods are commonly implemented using the **R** packages `Amelia II` (Honaker, King, and Blackwell, 2011) or `mice` (van Buuren and Groothuis-Oudshoorn, 2011), which stands for multiple imputation by chained equations. However, these methods require parametric assumptions for the missing and non-missing data that can be difficult to verify. Moreover, Moons et al. (2006) found that imputation models for missing covariates should include both other covariates and the outcome. When the data are longitudinal and unbalanced it can be unclear exactly how these models should be constructed (Erler et al., 2016).

In this Chapter we describe our **R** package `lmeMVP` in which we implement our nonparametric method for missing covariates in longitudinal models that we describe in Chapter 3. First we review the data assumptions in Section 4.2. Then we discuss how we implement the method in Section 4.3 and provide details on how to use the package in Section 4.4. We provide a few examples for how to use the package in Section 4.5. In Section 4.6 we look at the performance of the package. And in Section 4.7 we discuss how the package could be expanded and improved in the future.

4.2. Data

Here we review the type of data applicable to the method implemented in `lmeMVP`. First, let time be discrete. Then let Y be a longitudinal outcome and X and Z be time-independent or time-varying covariates. While Z is observed for all study subjects, X is missing for a subset. Those subjects with observed X make up the validation set V, while those missing X are part of the nonvalidation set $\bar{V}$. We further assume that for each subject, X is either fully observed or not observed at all. This means that if X is time-varying, it is either measured at every observation or at none. Furthermore, we assume that Z can be decomposed into (Z^*, A), where Z^* contains the components of Z that are uninformative of X and A is an “auxiliary” variable that may provide some information about X. Finally, we assume that A and X are independent of the missing data mechanism.

The data is modeled as

$$Y_i = X\beta_X + Z^*\beta_{Z^*} + \gamma_i b_i + \epsilon_i$$

where b_i are the random-effects parameters for the random effects γ_i and ϵ_i is the random errors, which follows a multivariate normal distribution with mean 0 and variance $\sigma^2\Lambda_i(\nu)$. As we note in Section 4.4, we assume independence of the random errors and define $\Lambda_i(\nu) = \mathrm{I}_{n_i}$

4.3. Implementation

The primary utility of `lmeMVP` is to maximize the estimated likelihood

$$\hat{L}(\theta) = \prod_{i \in V} P(Y_i | X_i, Z_i) \prod_{j \in \bar{V}} \hat{P}(Y_j | Z_j)$$

where $P(Y_i|X_i, Z_i)$ has the multivariate normal distribution

$$P(Y_i|X_i, Z_i) = (2\pi)^{-\frac{n_i}{2}}|\boldsymbol{V}_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right]$$

where $\boldsymbol{\mu}_i = X_i'\beta_X + Z^*\beta_{Z^*}$, n_i is the number of observations for subject i, $\boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\rho})$ is the covariance matrix, and $\theta = (\boldsymbol{\beta}, \boldsymbol{\rho})$. The definition of $\hat{P}(Y|Z)$ depends on the type of the missing covariate and auxiliary variable. X can be a time-independent missing covariate (TIM) or a time-varying (TVM) one and A can be a discrete auxiliary variable (DAV) or a continuous (CAV) one. Unless otherwise noted, the auxiliary variable is time-independent, but it may be time-varying (TV). However, as we note in Section 5, the method does not perform well when the auxiliary variable is continuous and time-varying. Therefore, we do not allow for this type of auxiliary variable in the `lmeMVP` package. All other missing covariate - auxiliary variable combinations are implemented. Table 4.1 provides the definition of $\hat{P}(Y|Z)$ for each of these combinations. For the sake of notation, we write $\hat{P}(Y|Z)$ generically as

$$\hat{P}(Y|Z) = \frac{\sum_{i\in V} P(Y_j|x_i[t], Z_j)K(A,t)}{\sum_{i\in V} K(A,t)}$$

where $K(A,t)$ depends on the type of auxiliary variable(s) and $x_i[t]$ may denote a time-varying or time-independent missing covariate. As shown in Table 4.1, when A is a single discrete variable, $K(A,t)$ in an indicator function, and when A is a single continuous variable, $K(A,t)$ includes the kernel function $\Phi\left(\frac{A_i - A_j}{h}\right)$ where Φ is the cumulative distribution function of the standard normal distribution and h is the bandwidth. We calculate the bandwidth using the Sheather and Jones (1991) method, implemented as bw.SJ (R Core Team, 2018), on the validation set auxiliary variables.

Although not shown in Table 4.1, A can also represent multiple variables. In that case $K(A,t)$ will be a combination of kernel and/or indicator functions for each variable in A. However, we recommend using as few auxiliary variables as possible.

We maximize the likelihood using maxLik from the `maxLik` package (Henningsen and Toomet, 2011). We supply maxLik with the log-likelihood (Eq. 4.3) and the gradient (Eq. 4.4) for the estimated likelihood and use the “BFGS” (i.e. ”Broyden-Fletcher-Goldfarb-Shanno”) method, a quasi-

Table 4.1: $\hat{P}(Y|Z)$ for each combination of missing covariate X and auxiliary variable A

Variable Type	$\boldsymbol{\hat{P}(Y\|Z)}$
TIM-DAV	$\frac{\sum_{i\in V} P(Y_j\|x_i,Z_j)I(A_i=A_j)}{\sum_{i\in V} I(A_i=A_j)}$
TIM-CAV	$\frac{\sum_{i\in V} P(Y_j\|x_i,Z_j)\Phi\left(\frac{A_i-A_j}{h}\right)}{\sum_{i\in V} \Phi\left(\frac{A_i-A_j}{h}\right)}$
TVM-DAV	$\frac{\sum_{i\in V} P(Y_j\|x_i[t_j],Z_j)I(A_i=A_j)}{\sum_{i\in V} I(A_i=A_j)}$
TVM-CAV	$\frac{\sum_{i\in V} P(Y_j\|x_i[t_j],Z_j)\Phi\left(\frac{A_i-A_j}{h}\right)}{\sum_{i\in V} \Phi\left(\frac{A_i-A_j}{h}\right)}$
TVM-TVDAV	$\frac{\sum_{i\in V} P(Y_j\|x_i[t_j],Z_j)I(A_i[t_j]=A_j)}{\sum_{i\in V} I(A_i[t_j]=A_j)}$

TIM=time-independent missing covariate,TVM=time-varying missing covariate,
DAV=discrete auxilairy variable, CAV=continuous auxiliary variable,
TVDAV=time-varying discrete auxiliary variable
Φ=standard normal

Newton algorithm, for maximization. To ensure that the variance-covariance matrix $\boldsymbol{V}$ is positive-definite, we maximize the estimated likelihood with respect to the components of the cholesky decomposition of $\boldsymbol{V}$. Unless otherwise specified, the complete case analysis (CCA) estimates, obtained from the lme function provided in `nlme` (Pinheiro et al., 2018), are used as the algorithm's starting values.

From the output of maxLik, we get the parameter estimates for β and the Cholesky decomposition parameters of the random-effects covariance matrix. In addition, we get out the numerically calculated Hessian matrix. We confirmed that the numerical Hessian is equal to the analytical Hessian, but found maxLik to be more efficient when we did not supply the analytic Hessian. Therefore, we use the numerical Hessian, which we'll denote as $\mathcal{I}_n$, when calculating the variance of the parameter estimates. In Section 3.3, we derived the asymptotic variance as

$$\text{var}\left(\hat{\beta}\right) = \frac{\mathcal{I}^{-1}}{n} + \frac{\frac{(1-\rho^v)}{\rho^v}\mathcal{I}^{-1}\Sigma(\beta)\mathcal{I}^{-1}}{n}.$$

It's important to note that $\mathcal{I}_n = n\mathcal{I}$, where n is the number of subjects used in the analysis (i.e. the number of subjects in the validation set plus the number of subjects in the 'interior' nonvalidation set). So the variance of $\hat{\beta}$ calculated in `lmeMVP` is actually defined as

$$\hat{\text{var}}\left(\hat{\beta}\right) = \hat{\mathcal{I}}_n^{-1} + n\frac{(1-\hat{\rho}^v)}{\hat{\rho}^v}\hat{\mathcal{I}_n}^{-1}\hat{\Sigma}(\hat{\beta})\hat{\mathcal{I}}_n^{-1},$$

where $\hat{\Sigma}(\hat{\beta}) = \text{vâr}\left\{\hat{\hat{W}}_{X_i,A_i}(\hat{\beta}), i \in V\right\}$ and $\hat{\hat{W}}_{X_i,A_i}(\hat{\beta})$ is defined differently for continuous and discrete auxiliary variables. For a discrete auxiliary variable

$$\hat{\hat{W}}_{X_i,A_i} = \frac{\sum_{j\in\bar{V}}\left\{\frac{\partial P(Y_j|X_i,Z_j)/\partial\beta}{\hat{P}(Y_j|Z_j)} - \frac{\partial\hat{P}(Y_j|Z_j)/\partial\beta}{[\hat{P}(Y_j|Z_j)]^2}P(Y_j|X_i,Z_j)\right\}K(A,t)}{\sum_{j\in\bar{V}}K(A,t)} \tag{4.1}$$

whereas for a continuous auxiliary variable

$$\hat{\hat{W}}_{X_i,A_i} = \frac{n^V}{n^{\bar{V}}}\sum_{j\in\bar{V}}\left\{\left[\frac{\partial P(Y_j|X_i,Z_j)/\partial\beta}{\hat{P}(Y_j|Z_j)} - \frac{\partial\hat{P}(Y_j|Z_j)/\partial\beta}{[\hat{P}(Y_j|Z_j)]^2}P(Y_j|X_i,Z_j)\right]\frac{K(A,t)}{\sum_{i\in V}K(A,t)}\right\}. \tag{4.2}$$

Since A can represent multiple auxiliary variables, if any variable in A is continuous then Eq. 4.2 is used to calculate the variance of β. Eq. 4.1 is used only if all of the auxiliary variables are discrete.

Log-likelihood:

$$\log\hat{L} = \sum_{i\in V}\log P(Y_i|X_i,X_i^*,W_i) + \sum_{j\in\bar{V}}\left[\log\sum_{i\in V}P(Y_j|x_i,Z_j)K(A,t) - \log\sum_{i\in V}K(A,t)\right] \tag{4.3}$$

Gradient:

$$\begin{aligned}
\frac{\partial\log\hat{L}}{\partial\boldsymbol{\theta}} &= \sum_{i\in V}\frac{\partial}{\partial\boldsymbol{\theta}}\log P(Y_i|X_i,Z_i) + \sum_{j\in\bar{V}}\frac{\sum_{i\in V}\frac{\partial}{\partial\boldsymbol{\theta}}P(Y_j|x_i,Z_j)K(A,t)}{\sum_{i\in V}P(Y_j|x_i,Z_j)K(A,t)}\\
&\frac{\partial}{\partial\boldsymbol{\beta}}\log P(Y_i|X_i,Z_i) = \boldsymbol{X}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta})\\
&\frac{\partial}{\partial\rho_k}\log P(Y_i|X_i,Z_i) = -\frac{1}{2}\text{tr}\left(\boldsymbol{V}_i^{-1}\frac{\partial\boldsymbol{V}_i}{\partial\rho_k}\right) +\\
&\qquad\qquad\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta})'\boldsymbol{V}_i^{-1}\frac{\partial\boldsymbol{V}_i}{\partial\rho_k}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta})\\
&\frac{\partial}{\partial\boldsymbol{\beta}}P(Y_j|x_i,Z_j) = P(Y_j|x_i,Z_j)\boldsymbol{X}_j'\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j'\boldsymbol{\beta})\\
&\frac{\partial}{\partial\rho_k}P(Y_j|x_i,Z_j) = \frac{1}{2}P(Y_j|x_i,Z_j)\times\\
&\qquad\qquad\left[(\boldsymbol{y}_i - \boldsymbol{X}_j'\boldsymbol{\beta})'\boldsymbol{V}_j^{-1}\frac{\partial\boldsymbol{V}_j}{\partial\rho_k}\boldsymbol{V}_j^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j'\boldsymbol{\beta}) - \text{tr}\left(\boldsymbol{V}_j^{-1}\frac{\partial\boldsymbol{V}_j}{\partial\rho_k}\right)\right]
\end{aligned} \tag{4.4}$$

4.4. Usage

All of the estimators defined in Table 4.1 are implemented using lmeMVP, the main function of the `lmeMVP` package. The usage is as follows:

```
lmeMVP(fixed, data, random, auxiliary, time.varying.missing, time.name,
       auxBounds, start)
```

fixed Formula object for fixed effects that takes the form `response ~ covariates`. An intercept is included by default. To exclude an intercept add `-1` to the covariates.

data A data.frame with named columns containing multiple lines per subject. That is, the data should be in long form not wide form.

random One sided formula object for random effects of the form `~ random effects | grouping variable`. Random effects are separated by `+`. An intercept is included by default. To exclude a random intercept add `-1` to the random effect. Up to one random-intercept and/or one random-slope can be used. An independence structure is assumed for the random errors so the variance-covariance matrix of the response is assumed to have the form $\gamma^T D\gamma + \sigma^2 I$, were γ is the matrix of random effects, D is the variance-covariance matrix for the random-effects parameters, and σ^2 is the variance of the random errors. This is equivalent to the `random=reStruct(random, pdClass="pdSymm")` option in `nlme`.

auxiliary Formula object of the form `Missing covariate(s) ~ auxiliary variable(s)`. The missing and auxiliary variables are separated by `+`. The discrete or continuous nature of the auxiliary variables is detected automatically based on the variables' class. If an auxiliary variable is time-varying, this is specified by adding the time variable in `auxiliary variable(s)`. NOTE: The auxiliary variable can only be time-varying if the missing covariate is time-varying (i.e `time.varying.missing=TRUE`)

time.varying.missing Logical variable indicating if the missing covariate is time-varying. If there is more than one missing variable, time.varying.missing should indicate if any of the missing variables are time-varying.

time.name Character string that provides the name of the time variable.

auxBounds Vector of length 2 that determines the boundaries for inclusion of nonvalidation subjects in the 'interior' nonvalidation set. The values provided correspond to the percentiles of the validation set's continuous auxiliary variables. It is important to note that there is a bias-variance trade-off, where stricter bounds produce lower bias, but higher variance estimates. Less restrictive boundaries will reduce the variance by including more subjects from the nonvalidation set in the analysis, but can sometimes lead to biased estimates. The default is `c(0.25,0.75)` which corresponds to the 25^{th} and 75^{th} percentiles.

start Optional list of starting values to pass to `maxLik`. The names of the list elements must include each fixed effect (including the "intercept" if applicable) as well as the following:

`sigma` standard deviation of the random-error

`cov` positive definite covariance matrix for the random-effects parameters.

If `NULL` (default), the starting values are the complete case analysis estimates obtained from the `nlme` package.

4.4.1. lmeMVP output

The output of `lmeMVP` is an object of class lmeMVP that contains the following values:

coefficients Fixed-effect parameters.

vcov Variance-covariance matrix for the fixed-effect parameter estimates.

sigma Standard error estimates of the fixed-effect parameters, equal to the square root of the diagonal of vcov.

Code Return code from maxLik.

mle Object output from maxLik.

dataset Number of subjects from the validation set and nonvalidation set used in the analysis.

Type String describing the type of missing covariate(s) and auxiliary variable(s).

cca.nlme Object output by lme of the `nlme` package.

4.5. Examples

To demonstrate how to use the lmeMVP function, we created five datasets; one for each combination of missing covariate and auxiliary variable types. For each dataset, we generated 2000 subjects. For each subject, we generated the outcome `Y`, the missing covariate `X`, the auxiliary variable `S`, and the time variable `time`. In addition, we have `id`, the subject identification number, and `Xcomplete` which can be used for the 'oracle' calculation. However, in these examples we will not be using `Xcomplete`. A snippet from one of the datasets is shown below.

```
head(tim_cav)
```

```
##            Y        X time         S id Xcomplete
## 1 -7.0010400       NA    0 1.4234659  1  5.362817
## 2  0.6780589       NA    1 1.4234659  1  5.362817
## 3  2.5992257       NA    2 1.4234659  1  5.362817
## 4 -5.6299306 5.772728    0 0.5534901  2  5.772728
## 5 -4.5357426 5.772728    1 0.5534901  2  5.772728
## 6 -3.1026385 3.123607    0 0.4046146  3  3.123607
```

In each of the following examples, we use a total sample sizes of 500 subjects. First we show how to use lmeMVP for a time-independent missing covariate and a discrete auxiliary variable (TIM-DAV) with a random intercept and slope. In addition, we demonstrate how the supporting functions coef() and summary() can be used.

```
timdav.500<- filter(tim_dav,id %in% sample(unique(tim_dav$id),500,replace=F))
ex.timdav<-lmeMVP(fixed = Y ~ X + time, random = ~ time | id ,
                  auxiliary = X ~ S, data=timdav.500,
                  time.varying.missing = FALSE, time.name ="time")
```

```
## [1] "TIM.DAV"
```

```
ex.timdav
```

```
## Linear mixed-effects estimated likelihood for:
## Time independent missing variable(s) with 1 discrete auxiliary variable(s)
##
## BFGS maximization, 45 iterations
## Return code 0: successful convergence
## Log-likelihood: -2668.047
```

```
##
## Random effects: ~time | id
## Random effects variance covariance matrix
##             (Intercept)     time
## (Intercept)   0.7816617 1.263567
## time          1.2635666 2.042831
##
## Fixed effects: Y ~ X + time
##
## Coefficients:
##           X (Intercept)        time
##   -1.544683    2.205494    0.387927
##
## Subjects in validation set: 316
## Subjects used from nonvalidation set: 184
```

The lmeMVP object can be summarized similar to other mixed-effects models using summary(), which returns an object of class summary.lmeMVP.

```
summary.timdav<-summary(ex.timdav)
summary.timdav
```

```
## Linear mixed-effects estimated likelihood
## BFGS maximization, 45 iterations
## Return code 0: successful convergence
## Log-likelihood: -2668.047
##
## Random effects: Y ~ X + time
##             (Intercept)     time
## (Intercept)   0.7816617 1.263567
## time          1.2635666 2.042831
##
## Fixed effects: Y ~ X + time
##              Estimate    StdErr  z.value   p.value
## X           -1.544683  0.065907 -23.4374 < 2.2e-16 ***
## (Intercept)  2.205494  0.340362   6.4799 9.181e-11 ***
## time         0.387927  0.075815   5.1168 3.108e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In addition, coef() can be used on the lmeMVP object to get the parameter estimates of the fixed-effects or on the summary.lmeMVP object to extract the coefficient table with the standard errors and p-values.

```
coef(ex.timdav)
```

```
##           X (Intercept)        time
##   -1.544683    2.205494    0.387927
```

```
coef(summary.timdav)
```

```
##               Estimate     StdErr    z.value       p.value
## X            -1.544683 0.06590679 -23.437382 1.777869e-121
## (Intercept)   2.205494 0.34036177   6.479851  9.181322e-11
## time          0.387927 0.07581479   5.116772  3.108086e-07
```

For a time-independent missing covariate and continuous auxiliary variable (TIM-CAV), the same code is used as that for the TIM-DAV situation, since S is automatically detected to be continuous.

```
timcav.500<- filter(tim_cav,id %in% sample(unique(tim_cav$id),500,replace=F))
ex.timcav<-lmeMVP(fixed = Y ~ X + time, random = ~ time | id ,
                  auxiliary = X ~ S, data=timcav.500,
                  time.varying.missing = FALSE, time.name ="time")
```

```
## [1] "TIM.CAV"
```

```
ex.timcav
```

```
## Linear mixed-effects estimated likelihood for:
## Time independent missing variable(s) with 1 continuous auxiliary variable(s)
## and 0 discrete auxiliary variable(s)
##
## BFGS maximization, 42 iterations
## Return code 0: successful convergence
## Log-likelihood: -2095.292
##
## Random effects: ~time | id
## Random effects variance covariance matrix
##              (Intercept)     time
## (Intercept)    0.8874422 1.169191
## time           1.1691908 1.932225
##
## Fixed effects: Y ~ X + time
##
## Coefficients:
##           X (Intercept)        time
##  -1.4383493   1.5614308   0.3729133
##
## Subjects in validation set: 304
## Subjects used from nonvalidation set: 96
```

For a time-varying missing covariate, the argument `time.varying.missing=FALSE` must be changed to `TRUE`.

```
tvmcav.500<- filter(tvm_cav,id %in% sample(unique(tvm_cav$id),500,replace=F))
ex.tvmcav<-lmeMVP(fixed = Y ~ X + time, random = ~ time | id ,
                  auxiliary = X ~ S, data=tvmcav.500,
                  time.varying.missing = TRUE, time.name ="time")
```

```
## [1] "TVM.CAV"
```

```
ex.tvmcav
```

```
## Linear mixed-effects estimated likelihood for:
## Time-varying missing variable(s) with 1 continuous auxiliary variable(s) and
## 0 discrete auxiliary variable(s)
##
## BFGS maximization, 45 iterations
## Return code 0: successful convergence
## Log-likelihood: -2075.554
##
## Random effects: ~time | id
## Random effects variance covariance matrix
##             (Intercept)     time
## (Intercept)   0.9598867 1.081346
## time          1.0813463 2.080996
##
## Fixed effects: Y ~ X + time
##
## Coefficients:
##           X (Intercept)        time
##  -1.5897684   2.4438308   0.2982206
##
## Subjects in validation set: 296
## Subjects used from nonvalidation set: 92
```

Finally, for a time-varying auxiliary variable, `time` is added to the right-hand side of formula for the `auxiliary` input.

```
tvmtvdav.500<- filter(tvm_tvdav,id %in% sample(unique(tvm_tvdav$id),500,replace=F))
ex.tvtvmdav<-lmeMVP(fixed = Y ~ X + time,random = ~ time | id ,
                    auxiliary = X ~ S + time, data=tvmtvdav.500,
                    time.varying.missing = TRUE, time.name ="time")
```

```
## [1] "TVM.TVDAV"
```

```
ex.tvtvmdav
```

```
## Linear mixed-effects estimated likelihood for:
## Time-varying missing variable(s) with 1 time-varying discrete auxiliary
## variable(s)
##
## BFGS maximization, 46 iterations
## Return code 0: successful convergence
## Log-likelihood: -3174.399
##
## Random effects: ~time | id
## Random effects variance covariance matrix
##             (Intercept)     time
## (Intercept)   0.7210899 1.021634
## time          1.0216340 1.707517
##
## Fixed effects: Y ~ X + time
##
## Coefficients:
##           X (Intercept)        time
##  -1.4555027   1.8988722   0.3653801
##
## Subjects in validation set: 291
## Subjects used from nonvalidation set: 209
```

4.6. Package Performance

In Section 3.4, we demonstrated the performance of the method through a series of simulations. In this section we look at the performance of the `lmeMVP` package. Specifically, we compare the run-time of the lmeMVP function for each of the five situations presented in the Table 4.1. Again we assume a random-intercept and random-slope. For each missing covariate−auxiliary variable combination, we test four samples sizes (N): 200, 400, 500, and 1000. We compare the run-time of lmeMVP for each situation using `rbenchmark` (Kusnierczyk, 2012) with one iteration. The benchmark was performed on a Windows machine with the Intel(R) Core(TM)i7-6700 (3.40 GHz) processor and 16.0GB of RAM.

Figure 4.1 shows the run-time for each situation. For all situations, the run-time increases rapidly with sample size. In general, the function is faster for discrete auxiliary variables compared to continuous auxiliary variables for the same type of missing covariate despite maxLik using a similar number of iterations to maximize each likelihood. This is due in part to the form of $K(A,t)$ in the calculation of $\hat{P}(Y_j|Z_j)$. The numerator of $\hat{P}(Y_j|Z_j) = \sum_{i \in V} P(Y_j|x_i, Z_j)K(A,t)$. For discrete A,

$K(A,t)$ is an indicator function, only taking values of zero or 1. Therefore, for a given nonvalidation subject j, we do not need to calculate $P(Y_j|x_i, Z_j)$ to include in the summation any subjects from the validation set for whom $K(A,t) = 0$. For example, in the TIM-DAV situation $K(A,t) = I(A_i = A_j)$, so we only sum over the validation set subjects whose auxiliary variable matches that of the nonvalidation set subject. Contrarily, for a continuous A, $K(A,t)$ will not be zero and therefore we must sum over every validation set subject.

The results of the analyses are shown in Tables 4.2 and 4.3. Table 4.2 presents the results for the discrete auxiliary variables and Table 4.3 presents the results for the continuous auxiliary variables. For each situation, the results from lmeMVP are compared to the complete case analysis estimates. In most situations, the lmeMVP estimates are more efficient than the complete case analysis estimates. The only exception is for the time-varying missing covariate and time-varying discrete auxiliary variable (TVM-TVDAV) when the sample size is 200. This is because for each nonvalidation set subject there are not enough subjects in the validation set with matching auxiliary variables. For all other situations, the lmeMVP estimates are more efficient.

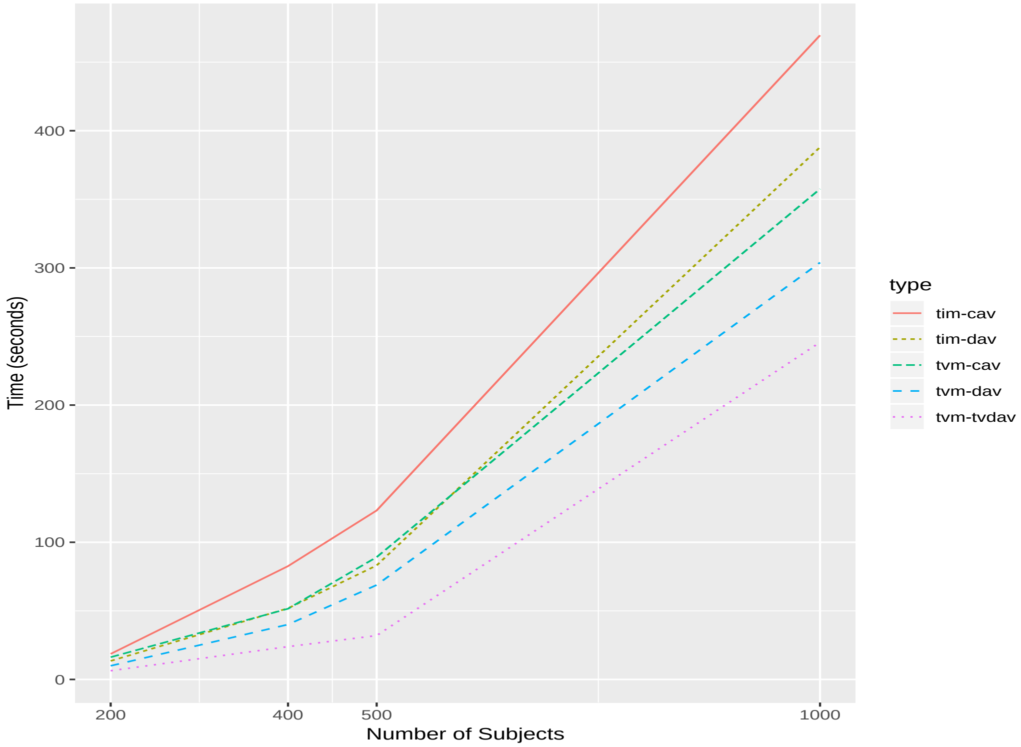


Figure 4.1: Runtime of `lmeMVP` for each missing covariate-auxiliary variable combination

4.7. Discussion

We introduced our package `lmeMVP` that implements the method described in Chapter 3 for linear mixed-effects models with missing values in the predictors. The package is flexible to allow each of the five combinations of missing covariate and auxiliary variable types: TIM-DAV, TIM-CAV, TVM-CAV, TVM-DAV, and TVM-TVDAV. In addition, our package allows for multiple missing covariates as well as multiple auxiliary variables; although we recommend using as few auxiliary variables as possible. If multiple auxiliary variables are needed, a larger sample size is required.

In the examples that we provide, all of the covariates, including the missing covariate, are treated as continuous. However, our function can also handle categorical predictors. To test if a categorical covariate is significant across all of its levels, a likelihood ratio test can be conducted using the log-likelihood that is reported in the output of lmeMVP.

In writing this package, we sought to be as consistent as possible with the commands for existing packages for mixed-effects models. Specifying the model using formula objects for the fixed and random effects as well as the auxiliary variable should be familiar to those who already use `lme` to run their mixed-effect models. In addition, the output and summary of `lmeMVP` is similar to that of `lme` and `maxLik` and therefore should be easy to read and interpret.

There are two areas in which we may improve our current package. The first would be to allow more flexible random-effects by allowing for both more and nested random-effects. Currently, `lmeMVP` allows for up to two random-effects, which may only be one of each type (slope and intercept). Furthermore, the variance structure cannot be specified. Instead, we assume independence in the random errors and correlated random-slopes and -intercepts. In the future, we may add more flexibility to specify different variance structures. Second, may improve the efficiency of our program. As shown in Section 4.6, the function can be slow to run, especially for larger sample sizes. The slowness of the program is mainly due to the need to repeatedly iterate through large loops. We use the family of `apply` functions to improve the efficiency of the summation calculations in Eq. 4.3 and Eq. 4.4, but there may be additional ways to reduce the run-time.

Table 4.2: Parameter estimates in the presence of discrete auxiliary variables

		TIM-DAV				TVM-DAV				TVM-TVDAV			
		CCA		lmeMVP		CCA		lmeMVP		CCA		lmeMVP	
N	Parm	est	SE	est	SE	est	SE	est	SE	est	SE	est	SE
	X	-1.56	0.116	-1.51	0.103	-1.61	0.085	-1.58	0.077	-1.61	0.061	-1.41	0.070
200	Intercept	2.33	0.598	2.02	0.533	2.55	0.432	2.37	0.392	2.59	0.326	1.62	0.362
	Time	0.40	0.162	0.30	0.118	0.42	0.138	0.51	0.114	0.43	0.132	0.39	0.119
	X	-1.44	0.092	-1.43	0.083	-1.54	0.071	-1.50	0.063	-1.53	0.048	-1.46	0.047
400	Intercept	1.66	0.480	1.71	0.432	2.23	0.367	2.07	0.329	2.26	0.261	1.93	0.249
	Time	0.25	0.102	0.34	0.080	0.24	0.101	0.36	0.084	0.23	0.103	0.28	0.083
	X	-1.60	0.078	-1.65	0.070	-1.49	0.057	-1.48	0.051	-1.52	0.045	-1.42	0.048
500	Intercept	2.54	0.403	2.76	0.359	1.80	0.296	1.82	0.267	2.13	0.239	1.60	0.252
	Time	0.38	0.093	0.32	0.072	0.31	0.106	0.35	0.082	0.20	0.090	0.24	0.078
	X	-1.57	0.052	-1.57	0.047	-1.48	0.043	-1.47	0.039	-1.52	0.029	-1.49	0.028
1000	Intercept	2.35	0.270	2.36	0.244	1.90	0.224	1.92	0.201	2.02	0.155	1.92	0.149
	Time	0.40	0.069	0.38	0.053	0.35	0.069	0.38	0.055	0.24	0.060	0.23	0.052

[*] CCA = complete case analysis performed using lme from the nlme package

[†] SE = standard error estimate

Table 4.3: Parameter estimates in the presence of continuous auxiliary variables

		TIM-CAV				TVM-CAV			
		CCA		lmeMVP		CCA		lmeMVP	
N	Parm	est	SE	est	SE	est	SE	est	SE
200	X	-1.76	0.128	-1.68	0.117	-1.47	0.083	-1.46	0.080
	Intercept	3.13	0.664	2.73	0.608	1.78	0.432	1.76	0.414
	Time	0.19	0.157	0.21	0.129	0.32	0.151	0.37	0.132
400	X	-1.58	0.085	-1.57	0.080	-1.49	0.071	-1.48	0.068
	Intercept	2.37	0.440	2.32	0.412	1.91	0.351	1.87	0.337
	Time	0.20	0.101	0.20	0.089	0.27	0.112	0.26	0.098
500	X	-1.42	0.078	-1.41	0.073	-1.51	0.061	-1.53	0.059
	Intercept	1.45	0.405	1.44	0.373	2.01	0.322	2.16	0.306
	Time	0.35	0.097	0.32	0.081	0.38	0.104	0.47	0.093
1000	X	-1.56	0.055	-1.56	0.051	-1.54	0.043	-1.52	0.041
	Intercept	2.17	0.277	2.18	0.259	2.15	0.225	2.06	0.215
	Time	0.33	0.072	0.29	0.062	0.32	0.072	0.36	0.064

[*] CCA = complete case analysis performed using lme from the nlme package

[†] SE = standard error estimate

CHAPTER 5

DISCUSSION

In this dissertation we proposed methods and provided tools for designing and analyzing longitudinal studies subject to missing information. We then applied these tools and methods to data from studies at the University of Pennsylvania Parkinson's Disease Center.

In Chapter 2 we proposed a method for designing follow-up schedules in longitudinal time-to-event studies in order to reduce the bias from right-endpoint imputation. We provide a tool to implement our method online at `https://lsapps.shinyapps.io/FollowUpDesign/`. The quantitative method evaluates the bias in the estimated hazard ratios from Cox regression models for simulated data. The simulated data is generated to resemble the survival curve of historical data, such as that from pilot data or previous studies. Although we used a Weibull or Gompertz distribution to generate the survival data, an advantage of our method is that it does not require parametric assumptions about the survival distribution. The simulated data does not need to be sampled from the true survival distribution, but only have a similar unadjusted survival curve. We defined the similarity between the generated and historical data based on the maximum difference in step-size at any point on their Kaplan-Meier curves. We chose this measure to ensure that the generated data has a similar amount of tied event times as the historical data. In the future we may compare other methods for measuring the similarity between two curves and evaluate what, if any, impact they have on the results.

Although our method does not rely on parametric assumptions, we do assume proportional hazards as required by the Cox regression model. A direction for future work would be to extend this method to situations with non-proportional hazards. In addition, we may consider more flexible or nonparametric data generation methods instead of relying on the Weibull and Gompertz distributions.

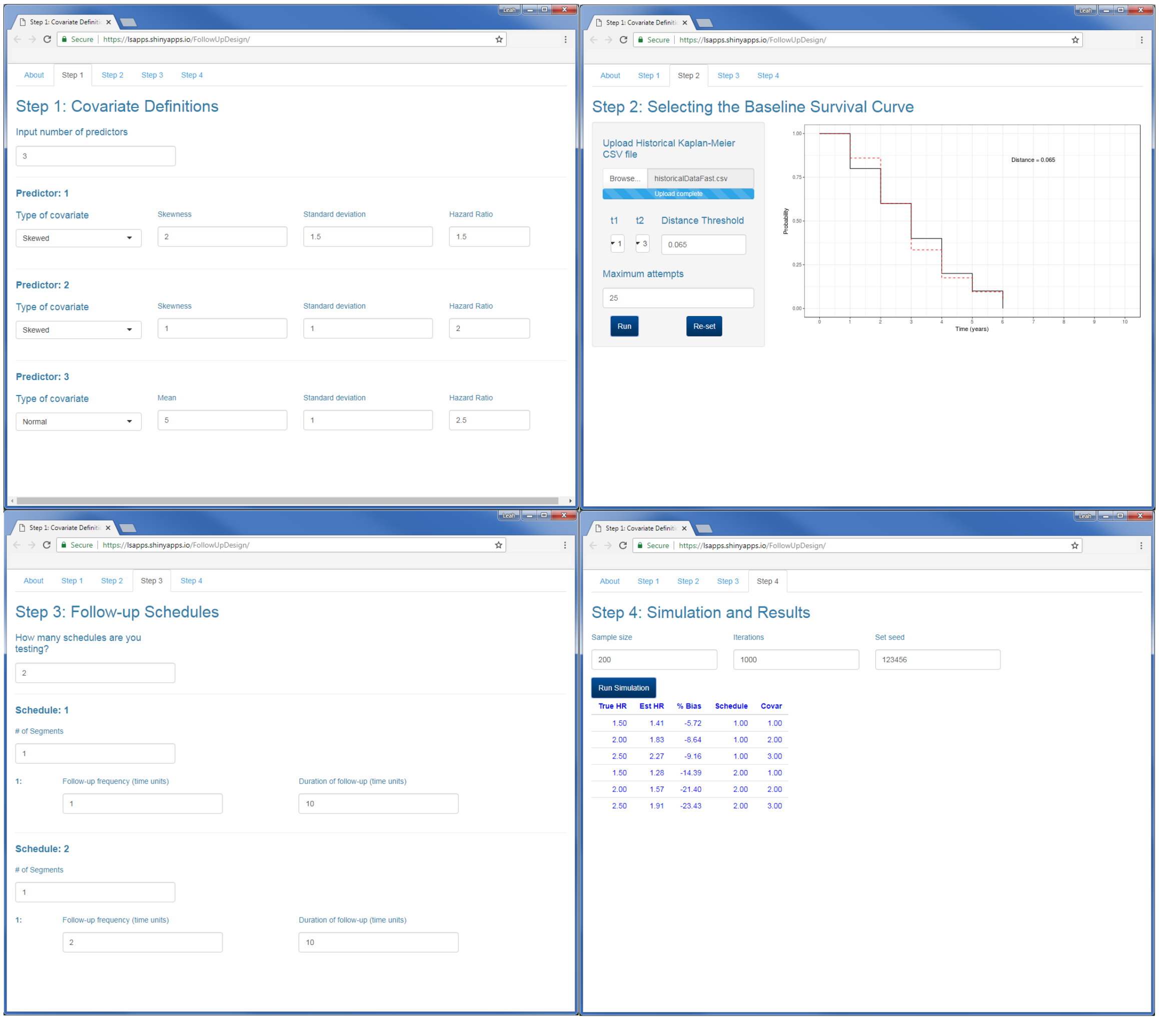
In Chapter 3 we proposed a nonparametric method for longitudinal data in which one or more covariates is missing for a subsample of subjects, but an auxiliary variable is available for everyone. We used empirical and kernel density estimates similar to Pepe and Fleming (1991) and Carroll and Wand (1991) to obtain nonparametric density estimates instead of relying on parametric assumptions about the conditional distribution of the missing data given the observed. Our proposed method can handle time-independent or time-varying missing covariates and auxiliary variables. In addition, the auxiliary variable can be discrete or continuous. We derived the asymptotic distribution of the estimator and showed that it is consistent and asymptotically normally distributed. Through simulations we showed that our estimator has good finite sample properties and is more efficient than the complete case estimator. We showed that the variance is well estimated using the asymptotic theory and therefore does not require bootstrapping. In Chapter 4 we provided the **R** package `lmeMVP` to implement the method.

A limitation of this method assumes discrete time. Extending the method to continuous time could be a direction for future work and would require overcoming two main challenges. First, time could be included as a continuous auxiliary variable by replacing the indicator function with the kernel function. However, this would give rise to the same dimensionality issues of other continuous, time-varying auxiliary variables. Second, for discrete time we “impute” values of X_i as $X_i[t_j]$, but the question remains how to “impute” X_i in the estimation $\hat{P}(Y_j|Z_j)$ if time is continuous. One possible solution might be to estimate X using splines which would avoid making parametric assumptions about X. However, more work is needed to understand the implications of this additional estimation on the proposed estimator.

APPENDIX A

FOLLOWUPDESIN: A SHINY APPLICATION

A.1. Shiny App Interface



Step 1: Covariate Definitions

Input number of predictors: 3

Predictor: 1

Type of covariate: Skewed; Skewness: 2; Standard deviation: 1.5; Hazard Ratio: 1.5

Predictor: 2

Type of covariate: Skewed; Skewness: 1; Standard deviation: 1; Hazard Ratio: 2

Predictor: 3

Type of covariate: Normal; Mean: 5; Standard deviation: 1; Hazard Ratio: 2.5

Step 2: Selecting the Baseline Survival Curve

Upload Historical Kaplan-Meier CSV file: historicalDataFast.csv

t1: 1; t2: 3; Distance Threshold: 0.065

Maximum attempts: 25

Step 3: Follow-up Schedules

How many schedules are you testing?: 2

Schedule: 1

# of Segments: 1

1: Follow-up frequency (time units): 1; Duration of follow-up (time units): 10

Schedule: 2

# of Segments: 1

1: Follow-up frequency (time units): 2; Duration of follow-up (time units): 10

Step 4: Simulation and Results

Sample size: 200; Iterations: 1000; Set seed: 123456

True HR	Est HR	% Bias	Schedule	Covar
1.50	1.41	-5.72	1.00	1.00
2.00	1.83	-8.64	1.00	2.00
2.50	2.27	-9.16	1.00	3.00
1.50	1.28	-14.39	2.00	1.00
2.00	1.57	-21.40	2.00	2.00
2.50	1.91	-23.43	2.00	3.00

Figure A.1: Screenshots of the FollowUpDesign Shiny application from https://lsapps.shinyapps.io/FollowUpDesign/

A.2. Shiny App Analysis Code

In this section we provide the **R** code used for the follow-up desgin analysis. We do not include the code for creating the Shiny App interface.

```r
library(shiny)
library(shinyjs)
library(ggplot2)
library(survival)
library(rootSolve)
library(flexsurv)
library(dplyr)
library(xtable)

# Function to apply the follow-up schedule to the generated data
regularFollowUp<-function(trueTime,nSeg,IL,dur) {
  obsTime<-numeric(length(trueTime))
  censorTime<-sum(dur)

  events<-which(trueTime < dur[1])
  obsTime[events]<-ceiling(trueTime[events]/IL[1])*IL[1]

  if (nSeg > 1) {
    for (i in 2:nSeg) {
      c0<-sum(dur[1:(i-1)])
      c1<-sum(dur[1:i])
      events<-which(c0 < trueTime & trueTime < c1 )
      obsTime[events]<-ceiling((trueTime[events]-c0)/IL[i])*IL[i] + c0
    }
  }

  # Censoring variable: subjects who fail after end of study are administratively
  # right censored (delta=0)
  delta<- trueTime < censorTime
  trueTime[!delta]<-censorTime
  obsTime[!delta] <- censorTime

  return(list(obsTime=obsTime,trueTime=trueTime,trueDelta=delta, obsDelta=delta))
}

# Function to generate the covariates
genCovars<-function(n,type, sd=NULL, values=NULL, prob=NULL, skew=NULL,mean=NULL) {
  if (type=="normal") {
    return(rnorm(n,mean,sd))
  } else if (type=="binary") {
    return(rbinom(n,1,prob))
  } else if (type=="categorical") {
    return(sample(x=values,size=n,prob=prob,replace = T))
  } else if (type=="skewed") {
    return(rgamma(n,shape=4/skew^2,scale=sd*skew/2))
  }
}

# Function to generate the survival time. Can be weibul or Gompertz
generateSurvival<-function(scale,shape,dist,covars,bet) {
  ut<-runif(nrow(covars),0,1)
  if (dist=="weibullPH") {
    genTime<-(-log(ut)/(scale*exp(covars%*%bet)))^(1/shape)
  } else if (dist=="gompertz") {
    genTime<-1/shape*(log(1-(shape*log(ut)/(scale*exp(covars%*%bet)))))
  }
  return(genTime)
}

# Funtion to choose the distribution family of the unadjusted survival
fitPH<-function(times,events=NULL) {

  gompFit<-flexsurvreg(Surv(times,events)~1,dist="gompertz")
  weibFit<-flexsurvreg(Surv(times,events)~1,dist="weibullPH")

  if (gompFit$AIC < weibFit$AIC) {
    dist<-"gompertz"
    fit<-gompFit
  } else {
    dist<-"weibullPH"
    fit<-weibFit
  }

  # Reparamaterize to match shape and scale of parameters as used in
  # the function generateSurvival

  if (dist=="weibullPH") {
    shape<-fit$res["shape","est"]
    scale<-fit$res["scale","est"]
  } else if (dist=="gompertz") {
    shape<-fit$res["shape","est"]
    scale<-fit$res["rate","est"]
  }

  return(data.frame(shape,scale,dist))
}


simulation<-function(n,S,km.hist,HR,
                     type, covarMean=NULL, covarSD=NULL, values=NULL,
                     probs=NULL,covarSkew=NULL,
                     nSched,nSeg,IL,dur,t12=NULL,distThres=0.05) {

  # select two points from historic data KM curve
  km.rm<-km.hist[km.hist$surv!=1 & km.hist$surv!=0,]
  if (is.null(t12)) {
    t12<-km.rm$time[c(1,nrow(km.rm))]
  }

  #t<-c(1,3)
  S1<-km.hist$surv[km.hist$time==t12[1]]
  S2<-km.hist$surv[km.hist$time==t12[2]]
```

```r
obs.out<-cox.out<-NULL

# Generate "observed" data using inverse-CDF method
m<-10000
bet<-log(HR)
nparms<-length(HR)
Ft<-1-km.hist$surv  # obtain estimated CDF
simMeas1<-1
diff.Surv.hist<-diff(-km.hist$surv,1)

while(simMeas1>distThres) {
  ut<-runif(m)

  # d = number of subjects who fail at each timepoint
  d<-numeric(length(Ft)-1)
  d[1]<-sum(ut<Ft[2])
  for (ts in 2:length(d)) {
    d[ts]<-sum(Ft[ts]<= ut & ut <Ft[(ts+1)])
  }
  c<-m-sum(d) # number of subjects with adminstrative right censoring

  # create dataframe of event times based on d
  times<-c(rep(km.hist$time[-1],d),rep(km.hist$time[length(km.hist$time)],c))
  events<-c(rep(1,m-c),rep(0,c))

  # Obtain PH distribtuion fit for the overall survival function
  fit<-fitPH(times,events=events)

  ## Calculate baseline survival parameters ##

  # obtain covariate values for corresponding points and generate one sample of
  # covars
  covars1<-numeric()
  for (i in 1:nparms) {
    covars1<-cbind(covars1,
                   genCovars(m,type[i],sd=covarSD[i],
                             skew = covarSkew[i],mean=covarMean[i],
                             values = values[[i]],prob = probs[[i]]))
  }
  # this is Q_{z\beta1} and Q_{z\beta2}
  gamma1<-quantile(exp(covars1%*%bet),S1)
  gamma2<-quantile(exp(covars1%*%bet),S2)

  # Gompertz function system of equations to solve
  fgomp<-function(x) {
    F1<-gamma1*x[1]^2*(1-exp(x[2]^2*t12[1])) - x[2]^2*log(S1)
    F2<-gamma2*x[1]^2*(1-exp(x[2]^2*t12[2])) - x[2]^2*log(S2)
    c(F1=F1,F2=F2)
  }
```

```r
# Weibull function to system of equations to solve
fweib<-function(x) {
  F1<-gamma1*(-x[1]^2*t12[1]^(x[2]^2)) - log(S1)
  F2<-gamma2*(-x[1]^2*t12[2]^(x[2]^2)) - log(S2)
  c(F1=F1,F2=F2)
}

if (fit$dist=="gompertz") {
  ss <- multiroot(f = fgomp,
                  start = c(sqrt(fit$scale/gamma2),sqrt(fit$shape)))
} else  ss <- multiroot(f = fweib,
                        start = c(sqrt(fit$scale/gamma2),sqrt(fit$shape)))
scale<-ss$root[1]^2
shape<-ss$root[2]^2

# Generate one sample of survival times
covars<-numeric()
for (i in 1:nparms) {
  covars<-cbind(covars,genCovars(n,type[i],sd=covarSD[i],
                                 skew = covarSkew[i],mean=covarMean[i],
                                 values = values[[i]],prob = probs[[i]]))
}
trueTime1<-generateSurvival(scale = scale,shape = shape,
                            dist=fit$dist,covars=covars,bet=bet)

# Obtain observed event times
hist.times<-km.hist$time
whichTime<-sapply(trueTime1, function(x) {
  which(order(c(hist.times,x),decreasing = F)==(length(hist.times)+1))
})

obsTime<-numeric(length(hist.times))
obsTime[whichTime<=length(hist.times)]<-hist.times[whichTime[
  whichTime<=length(hist.times)]
  ]
obsTime[whichTime>length(hist.times)]<-hist.times[length(hist.times)]
obsDelta<-numeric(length(hist.times))
obsDelta[whichTime<=length(hist.times)]<-1
obsDelta[whichTime>length(hist.times)]<-0

km.gen1<-survfit(Surv(obsTime,obsDelta)~1)
km.gen1<-data.frame(time=c(0,km.gen1$time),surv=c(1,km.gen1$surv))

diff.Surv.gen1<-diff(-km.gen1$surv,1)

if (length(diff.Surv.gen1)!=length(diff.Surv.hist)) {
  simMeas1<-1
} else {
  pdiffs<-abs(diff.Surv.hist-diff.Surv.gen1)
  simMeas1<-max(pdiffs)
}
```

```
  print(simMeas1)
}

simMaxes<-numeric(S)

# Generate and analyze S datasets
for (s in 1:S) {

  incProgress(1/S)

  # Generate survival time
  simMax<-1
  while (simMax ==1) {
    ut<-runif(n,0,1)
    covars<-numeric()
    for (i in 1:nparms) {
      covars<-cbind(covars,
                    genCovars(n,type[i],sd=covarSD[i],
                              skew = covarSkew[i],mean=covarMean[i],
                              values = values[[i]],prob = probs[[i]]))
    }

    trueTime<-generateSurvival(scale = scale,
                               shape = shape,dist=fit$dist,covars,bet=bet)

     # Obtain observed event times
    hist.times<-km.hist$time
    whichTime<-sapply(trueTime, function(x) {
      which(order(c(hist.times,x),decreasing = F)==(length(hist.times)+1))
    })
    obsTime<-numeric(length(hist.times))
    obsTime[whichTime<=length(hist.times)]<-hist.times[whichTime[
      whichTime<=length(hist.times)]
      ]
    obsTime[whichTime>length(hist.times)]<-hist.times[length(hist.times)]
    obsDelta<-numeric(length(hist.times))
    obsDelta[whichTime<=length(hist.times)]<-1
    obsDelta[whichTime>length(hist.times)]<-0
    km.gen1<-survfit(Surv(obsTime,obsDelta)~1)
    km.gen1<-data.frame(time=c(0,km.gen1$time),surv=c(1,km.gen1$surv))

    diff.Surv.gen1<-diff(-km.gen1$surv,1)

    if (length(diff.Surv.gen1)!=length(diff.Surv.hist)) {
      simMax<-1
    } else {
      pdiffs<-diff.Surv.hist-diff.Surv.gen1
      simMax<-pdiffs[which.max(abs(pdiffs))]
    }
  }
  simMaxes[s]<-max(pdiffs)
```

```r
    for (j in 1:nSched) {

      # Impose follow-up schedule
      obs<-regularFollowUp(trueTime,nSeg[[j]],IL[[j]],dur[[j]])

      # Run model
      m.obs<-coxph(Surv(obs$obsTime,obs$obsDelta)~covars)

      # Record Obs Estimates
      obsResults<-data.frame(summary(m.obs)$coefficients)
      obsResults$Bias<-obsResults$exp.coef.-HR
      obsResults$Perc<-obsResults$Bias/HR*100
      obsResults$Schedule<-j
      obsResults$Covar<-1:nparms
      obsResults<-cbind(HR,obsResults)
      obs.out<-rbind(obs.out,obsResults)
    }
  }

  obs.mean<-cox.mean<-data.frame(matrix(nrow=nSched*length(HR),ncol=5))
  r <- 0
  for (i in 1:nparms) {
    for (j in 1:nSched) {
      r<-r+1
      obs.mean[r,]<-apply(filter(obs.out,Schedule==j & Covar==i)[,
        c("HR","exp.coef.","Perc","Schedule","Covar")],2,mean)
    }
  }

  names(obs.mean)<-c("True HR","est HR","% Bias","Schedule","Covar")

  obs.mean<-obs.mean[order(obs.mean$Covar,obs.mean$Schedule),]

  return(list(Estimates=obs.mean,simMeas=mean(simMaxes)))
}
```

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1. Derivation of Estimated Likelihood Asymptotics

The proof for the proposed estimator with a time-independent or time-varying discrete auxiliary variables follows directly from the proof of Theorem 2 in Pepe and Fleming (1991). Therefore, here we focus on the derivation for the asymptotic distribution of the proposed estimator when the auxiliary variable is continuous and time-independent. The missing variable may be time-independent or time-varying.

Assume that V is the validation set which is a random subsample of size n^V from the total sample of size N and that $\lim_{n\to\infty} \frac{n^V}{N} = \rho^v > 0$. By random subsample, what we mean is that the missing mechanism does not depend on the auxiliary variables. Then $\hat{V}$ is the non-validation set of size $N - n^V$. For the estimated likelihood

$$\hat{L}(\theta) = \prod_{i \in V} P_\theta(Y_i|X_i, Z_i) \prod_{j \in \bar{V}} \hat{P}_\theta(Y_j|Z_j)$$

have the following result for the asymptotic normality from Theorem 2 in Pepe and Fleming (1991)

$$\frac{1}{n}\frac{\partial \log \hat{L}(\theta)}{\partial \theta} \xrightarrow{d} N\left(0, I(\theta) + \frac{(1-\rho^V)}{\rho^V}\Sigma\right)$$

where

$$I(\theta) = \rho^V \mathbb{E}\left[-\frac{\partial^2 \log P(Y|X,Z)}{\partial\theta\partial\theta'}\right] + (1-\rho^V)\mathbb{E}\left[-\frac{\partial^2 \log P(Y|Z)}{\partial\theta\partial\theta'}\right] = \mathbb{E}\left[-\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right]$$

and

$$\Sigma = \text{var}\left\{\mathbb{E}\left[-\frac{\partial \log P(Y|Z)}{\partial\theta}|X, A\right]\right\}.$$

Σ is estimated from the data as $\hat{\Sigma} = \hat{\text{var}}\{\hat{\hat{W}}_{x_i,s_i}(\hat{\theta}), i \in V\}$ where

$$\hat{\hat{W}}_{x_i,s_i}(\hat{\theta}) = \frac{\sum_{j\in\bar{V}} \frac{dP_\theta(Y_j|X_i,Z_j)/d\theta}{\hat{P}_\theta(Y_j|Z_j)} - \frac{d\hat{P}_\theta(Y_j|Z_j)/d\theta}{[\hat{P}_\theta(Y_j|Z_j)]^2} P_\theta(Y_j|X_i, Z_j) I(A_i = A_j)}{\sum_{j\in\bar{V}} I(A_i = A_j)}$$

We derive $\hat{\Sigma}$ for a continuous auxiliary variable and show that the asymptotic normally result holds.

Starting with the score equation, we have

$$\frac{1}{\sqrt{n}}\frac{\partial \log \hat{L}(\theta)}{\partial\theta} = \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{\partial\hat{P}_\theta(Y_j|Z_j)/\partial\theta}{\hat{P}_\theta(Y_j|Z_j)} + \frac{1}{\sqrt{n}}\sum_{i\in V}\frac{\partial P_\theta(Y_i|X_i,Z_i)/\partial\theta}{P_\theta(Y_i|X_i,Z_i)}$$

Let the partial derivatives of P with respect to θ be denoted by D. Then,

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial \log \hat{L}(\theta)}{\partial\theta} &= \frac{1}{\sqrt{n}}\sum_{i\in V}\frac{D(Y_i|X_i,Z_i)}{P_\theta(Y_i|X_i,Z_i)} + \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{\hat{D}(Y_j|Z_j)}{\hat{P}_\theta(Y_j|Z_j)}\\
&= \frac{1}{\sqrt{n}}\sum_{i\in V}\frac{D(Y_i|X_i,Z_i)}{P_\theta(Y_i|X_i,Z_i)} + \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{\hat{D}(Y_j|Z_j)}{\hat{P}_\theta(Y_j|Z_j)}+\\
&\qquad \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{P_\theta(Y_j|Z_j)} - \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{P_\theta(Y_j|Z_j)}\\
&= \frac{1}{\sqrt{n}}\sum_{i\in V}\frac{D(Y_i|X_i,Z_i)}{P_\theta(Y_i|X_i,Z_i)} + \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{P_\theta(Y_j|Z_j)}+\\
&\qquad \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\left\{\frac{\hat{D}(Y_j|Z_j)}{\hat{P}_\theta(Y_j|Z_j)} - \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{P_\theta(Y_j|Z_j)}\right\}\end{aligned}$$

The first two sums is the score function if $P(X|Z)$ were known. The third term is the penalty for estimating $\hat{P}_\theta(Y_j|Z_j)$.

To condense notation, let G represent A for a time-independent missing variable or S, t for a time-varying missing variable. Then $K_C(G)$ represent $K_C(A)$ or $K_C(S,t)$, where $K_C(A) = \Phi\left(\frac{A_i - A_j}{h}\right)$ and $K_C(S,t) = \Phi\left(\frac{A_i - A_j}{h}\right) I(t_j \subseteq t_i)$, where h is the bandwidth for the kernel function. Similarly, let $f(g)$ represent $f(s)$ or $f(s,t) = f(s)f(t)$. Assuming $nh \to \infty$ as $n \to \infty$, the third term above can be written

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\left\{\frac{\hat{D}(Y_j|Z_j)}{\hat{P}_\theta(Y_j|Z_j)} - \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{P_\theta(Y_j|Z_j)}\right\} \times \frac{\hat{P}_\theta(Y_j|Z_j)}{P_\theta(Y_j|Z_j)} + o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\left\{\frac{\hat{D}(Y_j|Z_j)}{P_\theta(Y_j|Z_j)} - \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{[P_\theta(Y_j|Z_j)]^2}\hat{P}_\theta(Y_j|Z_j)\right\} + o_p(1)\\
&\asymp \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\left\{\frac{\sum_{i\in V} D(Y_j|X_i,Z_j)K_C(G)}{P_\theta(Y_j|Z_j)\sum_{i\in V}K_C(G)} - \frac{1}{\sqrt{n}}\sum_{j\in\bar{V}}\frac{D(Y_j|Z_j)}{[P_\theta(Y_j|Z_j)]^2}\frac{\sum_{i\in V} P(Y_j|X_i,Z_j)K_C(G)}{\sum_{i\in V}K_C(G)}\right\}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}} \sum_{j \in \bar{V}} \sum_{i \in V} \left\{ \left[\frac{D(Y_j|X_i, Z_j)}{P(Y_j|Z_j)} - \frac{D(Y_j|Z_j)}{[P(Y_j|Z_j)]^2} P(Y_j|X_i, Z_j) \right] \frac{K_C(G)}{\sum_{i \in V} K_C(G)} \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i \in V} \sum_{j \in \bar{V}} \left\{ \left[\frac{D(Y_j|X_i, Z_j)}{P(Y_j|Z_j)} - \frac{D(Y_j|Z_j)}{[P(Y_j|Z_j)]^2} P(Y_j|X_i, Z_j) \right] \frac{K_C(G)}{\sum_{i \in V} K_C(G)} \right\}
\end{aligned}
$$

$$
\begin{aligned}
\text{let } W_{ij} &= \frac{D(Y_j|X_i, Z_j)}{P(Y_j|Z_j)} - \frac{D(Y_j|Z_j)}{[P(Y_j|Z_j)]^2} P(Y_j|X_i, Z_j) \\
&= \frac{1}{\sqrt{n}} \sum_{i \in V} \sum_{j \in \bar{V}} W_{ij} \frac{K_C(G)}{\sum_{i \in V} K_C(G)} \times \frac{n^{\bar{V}}}{n^{\bar{V}}} \times \frac{\frac{1}{n^V}}{\frac{1}{n^V}} \times \frac{\frac{1}{h}}{\frac{1}{h}} \\
&= \frac{1}{\sqrt{n}} \frac{n^{\bar{V}}}{n^V} \sum_{i \in V} \frac{1}{n^{\bar{V}}} \sum_{j \in \bar{V}} W_{ij} \frac{\frac{1}{h} K_C(G)}{\frac{1}{hn^V} \sum_{i \in V} K_C(G)} \\
&= \frac{1}{\sqrt{n}} \frac{n^{\bar{V}}}{n^V} \sum_{i \in V} \frac{1}{n^{\bar{V}}} \sum_{j \in \bar{V}} W_{ij} \frac{\frac{1}{h} K_C(G)}{\hat{f}(g_j)} \times \frac{\hat{f}(g_j)}{f(g_j)} + o_p(1) \\
&\asymp \frac{1}{\sqrt{n}} \frac{1-\rho^V}{\rho^V} \sum_{i \in V} \frac{1}{n^{\bar{V}}} \sum_{j \in \bar{V}} W_{ij} \frac{\frac{1}{h} K_C(G)}{f(s_j)} \times \left(\sum_{j \in \bar{V}} \frac{\frac{1}{h} K_C(G)}{f(s_j)} \right) \left(\frac{1}{\sum_{j \in \bar{V}} \frac{\frac{1}{h} K_C(G)}{f(s_j)}} \right)
\end{aligned}
$$

$$
\text{let } \bar{W}_{X_i, A_i} = \frac{1}{n^{\bar{V}}} \left(\sum_{j \in \bar{V}} \frac{\frac{1}{h} K_C(G)}{f(s_j)} \right) \left(\frac{1}{\sum_{j \in \bar{V}} \frac{\frac{1}{h} K_C(G)}{f(s_j)}} \right) \sum_{j \in \bar{V}} W_{ij} \frac{\frac{1}{h} K_C(G)}{f(s_j)}
$$

By the law of large numbers $\frac{1}{n^{\bar{V}}} \sum_{j \in \bar{V}} \frac{\frac{1}{h} K_C(G)}{f(g_j)} \xrightarrow{p} \mathbb{E}\left[\frac{\frac{1}{h} K_C(G)}{f(g_j)} \right]$. We will show that $\mathbb{E}\left[\frac{\frac{1}{h} K_C(G)}{f(g_j)} \right] = 1$ for $G = S, t$. That the results holds for $G = S$ is obvious.

$$
\begin{aligned}
\mathbb{E}\left[\frac{\frac{1}{h} \phi\left(\frac{A_j - A_i}{h} \right) I(t_j \subseteq t_i)}{f(s_j) f(t_j)} \right] &= \mathbb{E}\left[\frac{\frac{1}{h} \phi\left(\frac{A_j - A_i}{h} \right)}{f(s_j)} \right] \mathbb{E}\left[\frac{I(t_j \subseteq t_i)}{f(t_j)} \right] \\
&\left(\int_{-\infty}^{\infty} \frac{\frac{1}{h} \Phi\left(\frac{s - A_i}{h} \right)}{f(s)} f(s) ds \right) \times \frac{f(t)}{f(t)} \\
&= \int_{-\infty}^{\infty} \frac{1}{h} \Phi\left(\frac{s - A_i}{h} \right) ds = \int_{-\infty}^{\infty} \Phi(u) du \\
&= 1
\end{aligned}
$$

By Theorem 1 in Etemadi (2006)

$$
\frac{1}{\sum_{j \in \bar{V}} \frac{\frac{1}{h} K_C(G)}{f(g_j)}} \sum_{j \in \bar{V}} W_{ij} \frac{\frac{1}{h} K_C(G)}{f(g_j)} \xrightarrow{p} \mathbb{E}\left[W_{ij} | G \right]
$$

$$\begin{aligned}\mathbb{E}\left[W_{ij}|A,t\right] &= \mathbb{E}\left[\frac{D(Y|X,Z)}{P(Y|Z)} - \frac{D(Y|Z)}{[P(Y|Z)]^2}P(Y|X,Z)|G\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[\frac{D(Y|X,Z)}{P(Y|Z)} - \frac{D(Y|Z)}{[P(Y|Z)]^2}P(Y|X,Z)|Z\right]|G\right\} \\ &= \mathbb{E}\left\{\int \frac{D(Y|X,Z)}{P(Y|Z)}P(Y|Z)dY - \int \frac{D(Y|Z)}{[P(Y|Z)]^2}P(Y|X,Z)P(Y|Z)dY|G\right\} \\ &= \mathbb{E}\left[\int D(Y|X,Z)dY|A\right] - \mathbb{E}\left[\int \frac{D(Y|Z)}{P(Y|Z)}P(Y|X,Z)dY|G\right] \\ &= \mathbb{E}\left[\int \frac{D(Y|X,Z)}{P(Y|X,Z)}P(Y|X,Z)dY|A\right] - \mathbb{E}\left[\frac{D(Y|Z)}{P(Y|Z)}|X,G\right] \\ &= \mathbb{E}\left[\mathbb{E}\frac{\partial \log P(Y|X,Z)}{\partial\theta}|X,A\right] - \mathbb{E}\left[\frac{D(Y|Z)}{P(Y|Z)}|X,G\right] \\ &= -\mathbb{E}\left[\frac{D(Y|Z)}{P(Y|Z)}|X,G\right]\end{aligned}$$

By Slutsky's theorem $\bar{W}_{X_i,G_i} \overset{p}{\to} -\mathbb{E}\left[\frac{D(Y|Z)}{P(Y|Z)}|X,G\right]$ and the covariance converges to var$\left\{\mathbb{E}\left[\frac{D(Y|Z)}{P(Y|Z)}|X,G\right]\right\}$ = var$\left\{\mathbb{E}\left[\frac{\partial \log P(Y|Z)}{\partial\theta}|X,G\right]\right\}$. Finally, given $Y^{\bar{V}}$, $G^{\bar{V}}$, and $G^{\bar{V}}$, $\bar{W}_{X_i,G_i}$ are independent. Therefore, we can use the Lyapounov central limit theorem to show that $\sum_{i\in V}\bar{W}_{X_i,G_i}$ converges to a normal distribution.

B.1.1. Perfect Auxiliary Variable

Pepe (1992) describes a similar method for a surrogate outcome. Pepe explains that the estimated likelihood is fully efficient in the special case where the surrogate is perfect. Here we show that the same is true for our proposed estimator with a perfect auxiliary variable.

First, we show that for a discrete auxiliary variable, the estimated likelihood reduces to the full data likelihood when A is a perfect auxiliary variable for X.

$$\begin{aligned}\log \hat{L} =& \sum_{i\in V} P_\theta(Y_i|X_i,Z_i) + \sum_{j\in\bar{V}} \hat{P}_\theta(Y_j|X_j) \\ =& \sum_{i\in V} P_\theta(Y_i|X_i,Z_i) + \sum_{j\in\bar{V}} \frac{\sum_{i\in V} P_\theta(Y_j|X_i,Z_j)I(A_i - A_j)}{\sum_{i\in V} I(A_i - A_j)}\end{aligned}$$

If A is a perfect auxiliary variable, then given A the value of X is known. We can denote this as

$X(A)$. Then the estimated likelihood can be written

$$\log \hat{L} = \sum_{i \in V} P_\theta(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} \frac{\sum_{i \in V} P_\theta(Y_j|X_i(A_i), Z_j) I(A_i - A_j)}{\sum_{i \in V} I(A_i - A_j)}$$

Since $P_\theta(Y_j|X_i(A_i), Z_j)I(A_i - A_j)$ is zero except when $A_i = A_j$,

$$\begin{aligned}
\log \hat{L} &= \sum_{i \in V} P_\theta(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} \frac{\sum_{i \in V} P_\theta(Y_j|X_i(A_j), Z_j) I(A_i - A_j)}{\sum_{i \in V} I(A_i - A_j)} \\
\log \hat{L} &= \sum_{i \in V} P_\theta(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} \frac{P_\theta(Y_j|X(A_j), Z_j) \sum_{i \in V} I(A_i - A_j)}{\sum_{i \in V} I(A_i - A_j)} \\
\log \hat{L} &= \sum_{i \in V} P_\theta(Y_i|X_i, Z_i) + \sum_{j \in \bar{V}} P_\theta(Y_j|X(A_j), Z_j) \\
\log \hat{L} &= \log L
\end{aligned}$$

Since the estimated likelihood is equal to the full data likelihood when A is a perfect, discrete auxiliary variable, the proposed estimator is fully efficient even in small samples.

Next we show that the proposed estimator is fully efficient in large samples for both continuous and discrete perfect auxiliary variables. For a perfect auxiliary variable, $P(X|A)$ has unit mass. Therefore, $P_\theta(Y|Z) = \int P_\theta(Y|x, Z)P(x|A)dx = P_\theta(Y|X, Z)P(X|A)$. Then the score for $P(Y|Z)$

$$\frac{\partial}{\partial \theta} \log P_\theta(Y|Z) = \frac{D(Y|Z)}{P_\theta(Y|Z)} = \frac{\frac{\partial}{\partial \theta}\left[P_\theta(Y|X, Z)P(X|A)\right]}{P_\theta(Y|X, Z)P(X|A)} = \frac{D(Y|X, Z)}{P_\theta(Y|X, Z)} = \frac{\partial}{\partial \theta} \log P_\theta(Y|X, Z)$$

and the extra variance term $\Sigma = 0$ since

$$\Sigma = \text{var}\left\{\mathbb{E}\left[-\frac{\partial \log P_\theta(Y|X, Z)}{\partial \theta}|X, A\right]\right\} = 0.$$

We verify this result through large sample simulations using N=1000 subjects. To generate a perfect auxiliary variable, we require the correlation between the auxiliary variable and missing covariate to be 1. Therefore the auxiliary variable and missing covariate must either both be discrete or both be continuous. For the continuous auxiliary variable, we restrict the oracle analysis to only include those subjects from the validation set and the ’internal’ non-validation set so that the total sample is the same as that used in the proposed analysis.

Tables B.1 and B.2 summarize the results for the perfect discrete and continuous auxiliary variables, respectively. Table B.1 shows that for the discrete auxiliary variable the proposed method sample SD, $\hat{SE}$, and coverage are identical to those of the oracle (full data) method. In Table B.2 for the perfect continuous auxiliary variable, the proposed method sample SD is nearly identical to that of the oracle method, supporting the asymptotic theory. However, the $\hat{SE}$ is slightly underestimated resulting in lower coverage and REs less than 1.

Table B.1: Simulation results for a discrete missing covariate and a perfect, discrete auxiliary variable.

Missing		Proposed					Oracle			
(%)		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
	β_X	-2.3E-03	0.061	0.060	1.00	0.949	-3.5E-03	0.061	0.060	0.949
25	β_0	1.0E-02	0.277	0.274	1.00	0.953	1.7E-02	0.277	0.274	0.953
	β_T	-4.4E-03	0.054	0.053	1.00	0.946	-2.7E-03	0.054	0.053	0.946
	β_X	-2.3E-03	0.061	0.060	1.00	0.949	-3.5E-03	0.061	0.060	0.949
50	β_0	1.0E-02	0.277	0.274	1.00	0.953	1.7E-02	0.277	0.274	0.953
	β_T	-4.4E-03	0.054	0.053	1.00	0.946	-2.7E-03	0.054	0.053	0.946
	β_X	-2.3E-03	0.061	0.060	1.00	0.949	-3.5E-03	0.061	0.060	0.949
75	β_0	1.0E-02	0.277	0.274	1.00	0.953	1.7E-02	0.277	0.274	0.953
	β_T	-4.4E-03	0.054	0.053	1.00	0.946	-2.7E-03	0.054	0.053	0.946

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.
% Missing = percent of subjects with missing data. SD= standard deviation. $\hat{SE}$=estimated standard error.
RE = relative efficiency. Cov = coverage of 95% confidence interval. $\beta_T = \beta_{Time}$

B.1.2. Useless Auxiliary Variable

Next we discuss another special case in which the auxiliary variable is useless or independent of the missing covariate (i.e.$P(X|A) = P(X)$). The asymptotic variance for the score function of the estimated likelihood when A is independent of X, and therefore Y, is

$$\rho^V \mathbb{E}\left[-\frac{\partial^2 \log P(Y|X,Z)}{\partial\theta\partial\theta'}\right] + (1-\rho^V)\mathbb{E}\left[-\frac{\partial^2 \log P(Y|Z)}{\partial\theta\partial\theta'}\right] \\ + (1-\rho^V)\text{var}\left\{\mathbb{E}\left[-\frac{\partial \log P(Y|Z)}{\partial\theta}|X\right]\right\}$$

where $\rho^V \mathbb{E}\left[-\frac{\partial^2 \log P(Y|X,Z)}{\partial\theta\partial\theta'}\right]$ is the information from the validation set, $(1-\rho^V)\mathbb{E}\left[-\frac{\partial^2 \log P(Y|Z)}{\partial\theta\partial\theta'}\right]$ is the information from the non-validation set, and $\frac{(1-\rho^V)^2}{\rho^V}\text{var}\left\{\mathbb{E}\left[-\frac{\partial \log P(Y|Z)}{\partial\theta}|X\right]\right\}$ is the penalty

Table B.2: Simulation results for a continuous missing covariate and a perfect, continuous auxiliary variable.

Missing		Proposed					Oracle			
(%)		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
25	β_X	-1.9E-03	0.050	0.048	0.99	0.941	-2.9E-03	0.050	0.049	0.944
	β_0	1.1E-02	0.255	0.246	0.99	0.938	1.5E-02	0.255	0.247	0.937
	β_T	-3.4E-03	0.056	0.057	1.00	0.961	-3.0E-03	0.056	0.057	0.961
50	β_X	1.5E-04	0.058	0.057	0.98	0.942	-1.4E-03	0.059	0.058	0.943
	β_0	8.4E-03	0.299	0.290	0.98	0.935	8.3E-03	0.300	0.296	0.939
	β_T	-2.9E-03	0.059	0.061	1.00	0.961	-3.1E-03	0.059	0.061	0.959
75	β_X	7.3E-03	0.080	0.074	0.96	0.920	2.0E-03	0.080	0.078	0.942
	β_0	-4.3E-02	0.407	0.377	0.96	0.922	-9.0E-03	0.405	0.392	0.939
	β_T	-1.3E-03	0.064	0.067	1.00	0.961	-4.9E-04	0.065	0.067	0.962

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.
% Missing = percent of subjects with missing data. SD= standard deviation. $\hat{SE}$=estimated standard error.
RE = relative efficiency. Cov = coverage of 95% confidence interval.$\beta_T = \beta_{Time}$

for estimating $P(Y|Z)$. In order for the proposed method to have the same efficiency as the complete case method, both the second and third terms of the above equation would have to be zero. However, even when A, the auxiliary variable, is totally uninformative of X, there is still information about θ in $P(Y|Z)$, so $\mathbb{E}\left[-\frac{\partial^2 \log P(Y|Z)}{\partial\theta\partial\theta'}\right] \neq 0$. Therefore, even when A is useless, the efficiency of the proposed method will not necessarily be equal to that of the complete case analysis which uses only the validation data.

We demonstrate this using another set of large sample (N=1000) simulations, this time with useless discrete and continuous auxiliary variables. Web Table B.3 shows the results for a useless discrete auxiliary variables and Web Table B.4 shows the results for a useless continuous auxiliary variable. For both types of useless auxiliary variables, the proposed method is not equal to the complete case analysis in terms of efficiency.

Table B.3: Simulation results for a discrete missing covariate and a useless, discrete auxiliary variable.

Missing		Complete Case				Proposed			
(%)		Bias	SD	$\hat{SE}$	Cov	Bias	SD	$\hat{SE}$	Cov
25	β_X	-3.8E-04	0.035	0.035	0.944	8.4E-04	0.032	0.032	0.957
	β_0	6.4E-04	0.115	0.115	0.944	-6.3E-04	0.106	0.108	0.955
	β_T	-1.4E-03	0.059	0.061	0.960	1.4E-03	0.053	0.053	0.945
50	β_X	-3.7E-04	0.044	0.043	0.940	2.7E-04	0.037	0.036	0.947
	β_0	1.4E-03	0.145	0.141	0.941	-2.2E-03	0.124	0.123	0.950
	β_T	1.5E-03	0.074	0.075	0.957	1.0E-03	0.053	0.053	0.947
75	β_X	3.2E-03	0.062	0.060	0.935	-9.9E-03	0.053	0.048	0.910
	β_0	-6.2E-03	0.205	0.200	0.942	2.7E-02	0.178	0.167	0.931
	β_T	1.6E-03	0.106	0.106	0.934	1.1E-03	0.053	0.053	0.947

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.
% Missing = percent of subjects with missing data. SD= standard deviation. $\hat{SE}$=estimated standard error.
RE = relative efficiency.Cov = coverage of 95% confidence interval.$\beta_T = \beta_{Time}$

Table B.4: Simulation results for a continuous missing covariate and a useless, continuous auxiliary variable.

Missing		Complete Case				Proposed			
(%)		Bias	SD	$\hat{SE}$	Cov	Bias	SD	$\hat{SE}$	Cov
25	β_X	1.3E-03	0.048	0.049	0.957	2.8E-03	0.047	0.048	0.950
	β_0	1.1E-03	0.050	0.050	0.947	1.3E-03	0.049	0.049	0.947
	β_T	-1.7E-03	0.062	0.061	0.946	-8.6E-04	0.057	0.057	0.960
50	β_X	1.3E-04	0.059	0.060	0.949	6.6E-03	0.057	0.057	0.954
	β_0	3.0E-03	0.062	0.061	0.946	1.8E-03	0.060	0.059	0.943
	β_T	-2.3E-03	0.075	0.075	0.945	-2.9E-03	0.061	0.061	0.946
75	β_X	1.7E-03	0.085	0.085	0.953	8.2E-03	0.083	0.080	0.946
	β_0	6.2E-04	0.089	0.086	0.938	-4.5E-04	0.088	0.088	0.949
	β_T	-1.6E-03	0.105	0.106	0.951	-3.7E-03	0.068	0.067	0.950

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.
% Missing = percent of subjects with missing data. SD= standard deviation. $\hat{SE}$=estimated standard error.
RE = relative efficiency.Cov = coverage of 95% confidence interval.$\beta_T = \beta_{Time}$

B.2. Additional Simulation Results

To evaluate the performance of our proposed estimator, we perform a series of simulations. In the main text we describe the simulations for three settings:

1. Time-independent missing variable with a continuous auxiliary variable
2. Time-varying missing variable with a continuous auxiliary variable
3. Time-varying missing variable with a time-varying discrete auxiliary variable

Here we briefly reiterate the simulations described in the main text and present simulations analogous to settings 1 and 2, but with a discrete auxiliary variable. For each simulation setting we compare the performance of three estimators; the complete case estimator, the proposed estimators, and the oracle estimator. The complete case estimator drops all subjects with missing data from the analysis and the oracle estimator uses the unobservable full data. Using 1000 iterations of each simulation, we calculate the mean bias ($\hat{\theta} - \theta$), mean observed sample standard deviation (SD), mean estimated standard errors ($\hat{SE}$), mean relative efficiency (RE) compared to the oracle estimator where a lower RE is more efficient, and 95% coverage (Cov).

B.2.1. Time-independent Missing Covariate with a Continuous Auxiliary Variable

The missing covariate X and auxiliary variable A are generated using a standard multivariate normal distribution where $\left(\begin{smallmatrix} X \\ S \end{smallmatrix}\right) \sim N\left[\left(\begin{smallmatrix} \mu_X \\ \mu_A \end{smallmatrix}\right), \left(\begin{smallmatrix} \sigma_X^2 & \rho\sigma_X\sigma_A \\ \rho\sigma_X\sigma_A & \sigma_A^2 \end{smallmatrix}\right)\right]$ and ρ is the correlation between X and A. We simulate data for ρ= 0.01, 0.25, .50, 0.75, and 1.0. The full data is generated for all N=400 subjects where all subjects are considered to have observations at baseline and year one, but one-third of subjects are lost to follow-up at year two. Thus the data is balance but incomplete. X is then set to missing for 25%, 50% and 75% of subjects. This ensures that the data is MCAR.

The results are summarized in Table 3.1, Table B.5, and Table B.6. All three analyses (complete case, proposed, and oracle) have little bias and a good 95% coverage probability for low to moderate missing data. When the percent missing is high (75%) and the correlation between the missing and auxiliary variables is perfect (ρ=1.0), the proposed method is slightly more biased (Table B.6). Nevertheless, the mean $\hat{SE}$ estimates calculated based on the asymptotic theory for the proposed estimator is similar to the observed sample SD under all conditions. As a result, the coverage probability for 75% missing data is slightly low (92%) when the correlation is high.

The RE is calculated for each estimator as $\frac{1}{1000}\sum_{sim=1}^{1000}\frac{\hat{SE}_{m,sim}}{\hat{SE}_{oracle,sim}}$, where m = complete case, proposed, or oracle. The RE for the oracle estimator is 1 and larger values are less efficient. For the proposed estimator, the efficiency of the estimator increases with the correlation between X and A, but this result is more pronounced for high missingness. Under all conditions, including high missingness and low correlation, the proposed estimator is more efficient (smaller RE) than the complete case estimator. Therefore, even if the auxiliary variable provides little to no information about X, the proposed method is still at least as efficient as the complete case analysis.

Table B.5: Simulation results for 25% missing time-independent covariate with a continuous auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	-4.3E-03	0.080	0.077	1.15	0.944	-4.8E-03	0.078	0.075	1.12	0.942	-3.6E-03	0.069	0.067	0.948
	β_0	1.8E-02	0.407	0.395	1.15	0.947	2.6E-02	0.397	0.385	1.12	0.954	1.5E-02	0.352	0.343	0.950
	β_{Time}	-2.8E-03	0.097	0.097	1.15	0.941	-5.3E-05	0.090	0.090	1.07	0.941	-2.5E-03	0.085	0.084	0.938
0.25	β_X	-4.8E-03	0.079	0.078	1.15	0.951	-7.0E-03	0.077	0.075	1.12	0.948	-3.7E-03	0.068	0.067	0.951
	β_0	2.2E-02	0.402	0.396	1.15	0.950	4.0E-02	0.390	0.385	1.12	0.958	1.6E-02	0.345	0.343	0.952
	β_{Time}	-1.6E-03	0.097	0.097	1.15	0.940	-2.2E-03	0.090	0.090	1.07	0.939	-1.9E-03	0.085	0.084	0.938
0.50	β_X	-1.5E-03	0.080	0.077	1.15	0.951	-3.1E-03	0.078	0.075	1.13	0.950	-1.5E-03	0.068	0.067	0.950
	β_0	5.0E-03	0.410	0.395	1.15	0.949	2.8E-02	0.403	0.385	1.13	0.947	5.2E-03	0.350	0.342	0.946
	β_{Time}	-4.1E-03	0.097	0.096	1.15	0.941	-4.1E-03	0.091	0.090	1.07	0.942	-3.7E-03	0.086	0.084	0.935
0.75	β_X	-1.3E-04	0.078	0.077	1.15	0.953	-5.3E-03	0.077	0.076	1.13	0.947	-1.1E-03	0.067	0.067	0.950
	β_0	-1.8E-03	0.401	0.395	1.15	0.950	2.7E-02	0.396	0.387	1.13	0.944	4.8E-03	0.345	0.342	0.945
	β_{Time}	-3.4E-03	0.097	0.097	1.15	0.946	-3.8E-03	0.092	0.090	1.07	0.942	-4.2E-03	0.086	0.084	0.941
1.00	β_X	8.5E-04	0.081	0.078	1.15	0.938	-2.2E-03	0.079	0.076	1.13	0.937	1.0E-03	0.068	0.067	0.948
	β_0	-3.0E-03	0.410	0.396	1.15	0.937	5.8E-03	0.402	0.387	1.13	0.935	-4.0E-03	0.346	0.343	0.946
	β_{Time}	-3.6E-03	0.100	0.097	1.15	0.943	-5.1E-03	0.093	0.090	1.07	0.943	-4.0E-03	0.088	0.084	0.935

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.6: Simulation results for 75% missing time-independent covariate with a continuous auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	-1.9E-03	0.139	0.135	2.01	0.951	4.1E-03	0.131	0.127	1.89	0.937	-3.6E-03	0.069	0.067	0.948
	β_0	7.9E-03	0.703	0.688	2.01	0.952	-3.5E-02	0.665	0.650	1.90	0.941	1.5E-02	0.352	0.343	0.950
	β_{Time}	-1.9E-03	0.169	0.167	2.00	0.945	9.1E-04	0.109	0.107	1.27	0.947	-2.5E-03	0.085	0.084	0.938
0.25	β_X	-1.8E-04	0.140	0.135	2.01	0.950	1.4E-03	0.130	0.127	1.89	0.939	-3.7E-03	0.068	0.067	0.951
	β_0	-1.7E-03	0.708	0.687	2.00	0.953	-7.6E-03	0.663	0.648	1.89	0.941	1.6E-02	0.345	0.343	0.952
	β_{Time}	-3.3E-03	0.166	0.167	2.00	0.950	-1.7E-03	0.107	0.107	1.27	0.954	-1.9E-03	0.085	0.084	0.938
0.50	β_X	-1.3E-03	0.139	0.135	2.01	0.944	-8.0E-03	0.128	0.125	1.87	0.943	-1.5E-03	0.068	0.067	0.950
	β_0	4.5E-03	0.703	0.687	2.01	0.949	3.7E-02	0.652	0.640	1.87	0.944	5.2E-03	0.350	0.342	0.946
	β_{Time}	-2.9E-03	0.169	0.168	2.00	0.946	-7.0E-03	0.108	0.107	1.27	0.949	-3.7E-03	0.086	0.084	0.935
0.75	β_X	-3.7E-03	0.140	0.135	2.01	0.933	-2.8E-02	0.127	0.122	1.83	0.928	-1.1E-03	0.067	0.067	0.950
	β_0	2.3E-02	0.710	0.686	2.00	0.946	1.4E-01	0.647	0.624	1.82	0.927	4.8E-03	0.345	0.342	0.945
	β_{Time}	-6.2E-03	0.170	0.167	2.00	0.945	-4.2E-03	0.106	0.107	1.28	0.952	-4.2E-03	0.086	0.084	0.941
1.00	β_X	8.5E-04	0.141	0.135	2.00	0.939	-1.6E-02	0.127	0.116	1.73	0.923	1.0E-03	0.068	0.067	0.948
	β_0	-4.1E-03	0.715	0.686	2.00	0.936	7.6E-02	0.640	0.589	1.72	0.925	-4.0E-03	0.346	0.343	0.946
	β_{Time}	-5.4E-03	0.168	0.167	2.00	0.945	-4.8E-03	0.109	0.107	1.27	0.939	-4.0E-03	0.088	0.084	0.935

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

B.2.2. Time-varying Missing Covariate with a Continuous Auxiliary Variable

When X is time-varying but A is time-independent, we generate X in two steps. First, we generate $\bar{X}_i$ and A_i, where $\bar{X}_i$ is the mean for X_i, from a multivariate normal distribution where $\left(\begin{smallmatrix} \bar{X}_i \\ A_i \end{smallmatrix}\right) \sim N\left[\left(\begin{smallmatrix} \mu_X \\ \mu_A \end{smallmatrix}\right), \left(\begin{smallmatrix} \sigma_X^2 & \rho\sigma_X\sigma_A \\ \rho\sigma_X\sigma_A & \sigma_A^2 \end{smallmatrix}\right)\right]$. Second, we generate n_i observations of X_i from $N\left[\bar{X}_i, \left(\frac{\sigma_X}{2}\right)^2\right]$. Since X is time-varying but A is not, the correlation between X and A will never be exactly 1.0. Therefore, we only consider correlations of $\rho =$ 0.01, 0.25, 0.50, and 0.75. Again we set the total sample size to be N=400 with one-third of subjects being lost to follow-up at year two. We also let the percent of subjects with missing X be 25%, 50%, and 75%.

Table 3.2, Table B.7, and Table B.8 shows the results for 50%, 25%, and 75% of subjects with missing data, respectively. Again, the proposed method is unbiased and has good coverage for low to moderate missingness, but is slightly biased when the missingness and correlation are 75%. As in the previous section, the small bias results in a slightly lower coverage probability of approximately 92%. Still, the proposed method is at least as efficient, if not more, than the complete case analysis for all scenarios.

Table B.7: Simulation results for 25% missing time-varying covariate with a continuous auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	5.7E-04	0.062	0.060	1.15	0.937	7.1E-04	0.061	0.058	1.13	0.939	4.0E-04	0.053	0.052	0.948
	β_0	-4.5E-03	0.323	0.308	1.15	0.941	-4.9E-03	0.321	0.301	1.13	0.939	-2.8E-03	0.277	0.267	0.935
	β_{Time}	-2.8E-03	0.098	0.097	1.15	0.940	-3.5E-03	0.093	0.091	1.09	0.940	-1.8E-03	0.086	0.084	0.937
0.25	β_X	-5.1E-05	0.061	0.059	1.15	0.943	-1.3E-05	0.060	0.058	1.13	0.945	-1.1E-03	0.052	0.051	0.953
	β_0	2.5E-04	0.314	0.307	1.15	0.948	5.4E-05	0.311	0.300	1.13	0.946	6.3E-03	0.267	0.266	0.949
	β_{Time}	3.2E-03	0.098	0.097	1.15	0.940	1.6E-03	0.093	0.091	1.09	0.942	1.7E-03	0.086	0.084	0.943
0.50	β_X	-2.1E-03	0.059	0.059	1.16	0.938	-1.8E-03	0.058	0.058	1.13	0.936	-1.2E-03	0.052	0.051	0.952
	β_0	8.2E-03	0.305	0.305	1.16	0.943	7.3E-03	0.300	0.298	1.13	0.945	3.4E-03	0.265	0.264	0.952
	β_{Time}	-7.4E-03	0.097	0.097	1.15	0.941	-6.2E-03	0.092	0.091	1.09	0.939	-5.5E-03	0.082	0.084	0.947
0.75	β_X	4.5E-07	0.057	0.057	1.15	0.951	-2.5E-03	0.056	0.056	1.13	0.952	1.2E-03	0.048	0.050	0.955
	β_0	5.5E-03	0.299	0.298	1.15	0.948	1.8E-02	0.294	0.292	1.13	0.939	-8.6E-04	0.252	0.258	0.948
	β_{Time}	6.0E-03	0.095	0.097	1.16	0.953	4.3E-03	0.090	0.091	1.09	0.952	1.5E-03	0.084	0.084	0.950

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.8: Simulation results for 75% missing time-varying covariate with a continuous auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	2.8E-04	0.104	0.103	2.00	0.946	-3.0E-02	0.101	0.102	1.98	0.933	4.0E-04	0.053	0.052	0.948
	β_0	2.6E-04	0.536	0.533	2.00	0.946	1.5E-01	0.530	0.531	1.99	0.944	-2.8E-03	0.277	0.267	0.935
	β_{Time}	4.5E-04	0.165	0.167	1.99	0.953	3.9E-03	0.135	0.125	1.49	0.934	-1.8E-03	0.086	0.084	0.937
0.25	β_X	-2.4E-03	0.104	0.103	2.00	0.945	-3.6E-02	0.101	0.102	1.97	0.930	-1.1E-03	0.052	0.051	0.953
	β_0	1.4E-02	0.536	0.532	2.00	0.943	1.8E-01	0.536	0.529	1.99	0.927	6.3E-03	0.267	0.266	0.949
	β_{Time}	-2.6E-03	0.172	0.167	2.00	0.945	2.4E-04	0.125	0.125	1.49	0.949	1.7E-03	0.086	0.084	0.943
0.50	β_X	8.1E-04	0.104	0.101	1.99	0.944	-4.1E-02	0.099	0.099	1.95	0.932	-1.2E-03	0.052	0.051	0.952
	β_0	-7.4E-03	0.539	0.526	1.99	0.950	2.0E-01	0.513	0.517	1.96	0.928	3.4E-03	0.265	0.264	0.952
	β_{Time}	1.4E-04	0.164	0.167	1.99	0.961	-8.3E-03	0.121	0.125	1.49	0.943	-5.5E-03	0.082	0.084	0.947
0.75	β_X	3.5E-03	0.097	0.100	2.00	0.955	-4.5E-02	0.093	0.095	1.91	0.922	1.2E-03	0.048	0.050	0.955
	β_0	-1.4E-02	0.506	0.516	2.00	0.952	2.3E-01	0.482	0.494	1.91	0.930	-8.6E-04	0.252	0.258	0.948
	β_{Time}	-1.2E-02	0.170	0.167	1.99	0.938	-4.5E-03	0.123	0.124	1.48	0.942	1.5E-03	0.084	0.084	0.950

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

B.2.3. Time-varying Missing Covariate with a Time-Varying Discrete Auxiliary Variable

For a time-varying discrete auxiliary variable we generate balanced and complete data with three observations for each of the N=2000 subjects. The larger sample size is necessary in these situations to have a sufficient number of validation subjects who contribute to each estimate of $\hat{P}(Y_j|Z_j)$. For a validation subject to contribute to the estimate of $\hat{P}(Y_j|Z_j)$, the time-varying auxiliary variable must be matched at each timepoint (i.e. $I\left(A_i[t_j] = A_j, t_j \subseteq t_i\right)$) instead of just once (i.e. $I\left(A_i = A_j, t_j \subseteq t_i\right)$).

To generate data for a time-varying missing covariate and a time-varying discrete auxiliary variable, we first generate N × 3 observations from the multivariate normal distribution described above. To make A discrete, we define each observation as 0, 1, or 2 based on the tertiles of A. The observed correlation between X and A is calculated using the Spearman correlation, which is smaller than the specified ρ and never 1. Therefore, in the simulations we set ρ = 0.01, 0.30, 0.57, and 0.95, to achieve the desired correlations of 0.01, 0.25, 0.5, and 0.75. Here we let the percent missing be 25%, 30%, and 50%.

The results for these simulations are shown in Table 3.3 and Tables B.9 and B.10. For 25% and 30% missing data, the proposed method is unbiased and more efficient than the complete case analysis for all correlations. When the percent missing is higher, such as 50%, and the correlation between X and A is low, the proposed method can be slightly biased and a little less efficient than the complete case analysis. However, when the correlation between the auxiliary and missing variable is high, the proposed method performs well. For a correlation 0.75 between X and A and 50% missing data, the proposed method is unbiased, more efficient than the complete case analysis, and has good coverage.

Table B.9: Simulation results for 25% missing time-varying covariate with a discrete, time-varying auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	9.1E-04	0.019	0.019	1.15	0.951	3.7E-03	0.019	0.019	1.14	0.951	9.7E-04	0.016	0.017	0.950
	β_0	-5.0E-03	0.103	0.102	1.15	0.944	-1.9E-02	0.100	0.100	1.13	0.947	-6.0E-03	0.087	0.088	0.946
	β_{Time}	-1.8E-03	0.039	0.041	1.15	0.953	-1.2E-03	0.037	0.037	1.06	0.946	-1.1E-03	0.035	0.035	0.947
0.25	β_X	3.3E-04	0.019	0.019	1.15	0.946	2.5E-03	0.019	0.019	1.13	0.949	3.2E-04	0.016	0.017	0.950
	β_0	-1.9E-03	0.104	0.102	1.15	0.946	-1.3E-02	0.102	0.100	1.13	0.940	-2.0E-03	0.087	0.089	0.946
	β_{Time}	-8.4E-04	0.039	0.041	1.15	0.958	1.7E-04	0.036	0.037	1.06	0.965	7.1E-06	0.034	0.035	0.956
0.50	β_X	3.0E-04	0.020	0.019	1.15	0.939	8.3E-04	0.019	0.019	1.12	0.947	5.9E-04	0.016	0.017	0.948
	β_0	-1.7E-03	0.107	0.102	1.15	0.940	-4.5E-03	0.101	0.099	1.12	0.940	-3.5E-03	0.088	0.089	0.948
	β_{Time}	-4.3E-04	0.039	0.041	1.15	0.959	-3.8E-04	0.035	0.037	1.05	0.960	-5.7E-04	0.035	0.035	0.950
0.75	β_X	4.5E-04	0.020	0.019	1.15	0.940	-1.6E-04	0.018	0.018	1.11	0.954	2.8E-04	0.016	0.017	0.956
	β_0	-2.9E-03	0.106	0.102	1.15	0.946	9.0E-05	0.098	0.098	1.11	0.955	-2.3E-03	0.088	0.089	0.958
	β_{Time}	-1.4E-04	0.039	0.041	1.15	0.958	-3.5E-04	0.035	0.037	1.03	0.957	-3.0E-04	0.035	0.035	0.951

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.10: Simulation results for 50% missing time-varying covariate with a discrete, time-varying auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	1.1E-03	0.024	0.024	1.41	0.950	-1.4E-02	0.024	0.024	1.43	0.902	9.7E-04	0.016	0.017	0.950
	β_0	-5.5E-03	0.127	0.125	1.41	0.949	6.9E-02	0.125	0.125	1.41	0.904	-6.0E-03	0.087	0.088	0.946
	β_{Time}	-2.8E-03	0.048	0.050	1.41	0.958	-1.4E-03	0.042	0.042	1.18	0.955	-1.1E-03	0.035	0.035	0.947
0.25	β_X	-4.0E-04	0.024	0.024	1.41	0.951	-1.5E-02	0.023	0.024	1.42	0.910	3.2E-04	0.016	0.017	0.950
	β_0	2.7E-03	0.127	0.125	1.41	0.946	7.5E-02	0.125	0.124	1.40	0.907	-2.0E-03	0.087	0.089	0.946
	β_{Time}	-1.9E-03	0.047	0.050	1.41	0.965	-4.9E-04	0.041	0.041	1.17	0.959	7.1E-06	0.034	0.035	0.956
0.50	β_X	-2.0E-04	0.024	0.024	1.41	0.947	-1.4E-02	0.023	0.023	1.38	0.914	5.9E-04	0.016	0.017	0.948
	β_0	1.8E-03	0.131	0.125	1.41	0.941	7.2E-02	0.122	0.121	1.37	0.914	-3.5E-03	0.088	0.089	0.948
	β_{Time}	-8.3E-04	0.048	0.050	1.41	0.963	-1.1E-03	0.039	0.041	1.15	0.956	-5.7E-04	0.035	0.035	0.950
0.75	β_X	1.1E-04	0.024	0.024	1.41	0.948	-9.6E-03	0.022	0.022	1.33	0.936	2.8E-04	0.016	0.017	0.956
	β_0	-3.4E-04	0.129	0.125	1.41	0.943	4.8E-02	0.117	0.116	1.32	0.933	-2.3E-03	0.088	0.089	0.958
	β_{Time}	5.8E-05	0.048	0.050	1.41	0.964	-1.1E-03	0.038	0.039	1.10	0.947	-3.0E-04	0.035	0.035	0.951

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

B.2.4. Time-independent missing covariate with a discrete auxiliary variable

To generate the time-independent missing covariate and discrete auxiliary variable we generate data similar to that described above for a time-varying discrete auxiliary variable. Since both X and A are time-independent, we draw $N = 400$ observations from the specified multivariate normal distribution and descritize A based on its tertiles. Again, the observed correlation will be less than the specified ρ. Therefore, we set ρ = 0.01, 0.3, 0.57, and 0.83, to achieve correlations of approximately 0.01, 0.25, 0.5, and 0.75. We set the percent of subjects with missing data to be 25%, 50%, and 75%.

The results for these simulations are summarized in Tables B.11, B.12, and B.13. For low to moderate missingness, the proposed method is unbiased, has good coverage, and shows large efficiency gains over the complete case analysis. Even when the correlation and amount of missingess is low, the proposed method is substantially more efficient. The RE is 1.09 for the proposed method compared to 1.15 for the complete case analysis when there is 25% missing data and a correlation of approximately 0.01. When there is 75% missing data, the proposed method does not perform as well. The estimates are slightly more biased and the $\hat{SE}$ under estimates the SD a little. Nevertheless, the proposed estimator is still more efficient based on the observe SD and the 95% coverage probability is still okay ($\sim$93%).

Table B.11: Simulation results for 25% missing time-independent covariate with a discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	3.4E-04	0.078	0.077	1.15	0.950	1.2E-03	0.073	0.073	1.09	0.946	7.5E-04	0.067	0.067	0.950
	β_0	-1.5E-03	0.400	0.394	1.15	0.946	1.7E-03	0.369	0.374	1.09	0.947	-5.0E-03	0.342	0.342	0.951
	β_{Time}	-2.0E-03	0.099	0.097	1.15	0.946	5.2E-04	0.087	0.084	1.00	0.944	-2.4E-03	0.087	0.084	0.945
0.25	β_X	1.4E-03	0.080	0.078	1.15	0.939	2.3E-03	0.076	0.073	1.09	0.940	6.7E-04	0.067	0.067	0.944
	β_0	-1.0E-02	0.409	0.396	1.15	0.940	-2.0E-02	0.388	0.375	1.09	0.939	-4.7E-03	0.345	0.343	0.944
	β_{Time}	-9.1E-04	0.099	0.097	1.15	0.951	-5.7E-03	0.087	0.084	1.00	0.940	-4.1E-03	0.087	0.084	0.939
0.50	β_X	-2.2E-04	0.080	0.078	1.15	0.946	1.7E-03	0.076	0.074	1.09	0.951	-5.3E-04	0.068	0.067	0.953
	β_0	-8.2E-04	0.407	0.398	1.16	0.944	-1.6E-02	0.390	0.376	1.09	0.946	1.6E-03	0.346	0.344	0.953
	β_{Time}	-4.1E-03	0.100	0.097	1.15	0.949	-5.6E-03	0.089	0.084	1.00	0.935	-4.7E-03	0.089	0.084	0.938
0.75	β_X	3.1E-03	0.081	0.078	1.16	0.941	4.2E-03	0.076	0.074	1.09	0.940	9.9E-04	0.067	0.068	0.947
	β_0	-1.7E-02	0.411	0.399	1.16	0.943	-2.4E-02	0.389	0.376	1.09	0.942	-6.8E-03	0.344	0.345	0.948
	β_{Time}	3.0E-04	0.098	0.097	1.15	0.947	-5.8E-04	0.087	0.084	1.00	0.940	-3.9E-04	0.087	0.084	0.943

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.12: Simulation results for 50% missing time-independent covariate with a discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	9.1E-04	0.096	0.095	1.41	0.946	-1.1E-03	0.086	0.086	1.28	0.949	7.5E-04	0.067	0.067	0.950
	β_0	-3.4E-03	0.488	0.484	1.41	0.948	1.5E-02	0.434	0.438	1.28	0.952	-5.0E-03	0.342	0.342	0.951
	β_{Time}	-1.9E-03	0.119	0.119	1.41	0.950	5.2E-04	0.087	0.084	1.00	0.948	-2.4E-03	0.087	0.084	0.945
0.25	β_X	2.8E-03	0.098	0.095	1.42	0.939	5.2E-03	0.090	0.086	1.28	0.942	6.7E-04	0.067	0.067	0.944
	β_0	-2.0E-02	0.502	0.486	1.42	0.934	-2.6E-02	0.461	0.439	1.28	0.942	-4.7E-03	0.345	0.343	0.944
	β_{Time}	-3.4E-03	0.120	0.119	1.41	0.954	-5.3E-03	0.088	0.084	1.00	0.942	-4.1E-03	0.087	0.084	0.939
0.50	β_X	9.2E-04	0.098	0.096	1.42	0.940	2.7E-03	0.090	0.086	1.27	0.939	-5.3E-04	0.068	0.067	0.953
	β_0	-1.1E-02	0.503	0.488	1.42	0.943	-2.9E-02	0.461	0.438	1.27	0.939	1.6E-03	0.346	0.344	0.953
	β_{Time}	-5.1E-03	0.121	0.118	1.41	0.945	-7.0E-03	0.089	0.084	1.00	0.937	-4.7E-03	0.089	0.084	0.938
0.75	β_X	1.2E-03	0.101	0.096	1.42	0.933	1.5E-03	0.088	0.085	1.26	0.942	9.9E-04	0.067	0.068	0.947
	β_0	-1.0E-02	0.513	0.489	1.42	0.932	-1.6E-02	0.453	0.435	1.26	0.938	-6.8E-03	0.344	0.345	0.948
	β_{Time}	-5.6E-04	0.118	0.119	1.42	0.951	-1.2E-03	0.087	0.084	1.00	0.944	-3.9E-04	0.087	0.084	0.943

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.13: Simulation results for 75% missing time-independent covariate with a discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	2.5E-03	0.139	0.135	2.02	0.948	-3.4E-02	0.134	0.124	1.85	0.931	7.5E-04	0.067	0.067	0.950
	β_0	-1.7E-02	0.711	0.691	2.02	0.950	1.6E-01	0.685	0.636	1.86	0.932	-5.0E-03	0.342	0.342	0.951
	β_{Time}	-3.7E-03	0.168	0.169	2.01	0.946	-1.7E-04	0.087	0.084	1.00	0.947	-2.4E-03	0.087	0.084	0.945
0.25	β_X	-1.3E-03	0.137	0.136	2.02	0.953	-3.1E-02	0.130	0.125	1.85	0.933	6.7E-04	0.067	0.067	0.944
	β_0	8.6E-03	0.697	0.693	2.02	0.957	1.8E-01	0.662	0.638	1.86	0.932	-4.7E-03	0.345	0.343	0.944
	β_{Time}	-1.4E-02	0.167	0.168	2.01	0.944	-6.2E-03	0.088	0.084	1.00	0.939	-4.1E-03	0.087	0.084	0.939
0.50	β_X	-1.5E-03	0.138	0.136	2.01	0.942	-3.2E-02	0.130	0.122	1.81	0.926	-5.3E-04	0.068	0.067	0.953
	β_0	8.7E-03	0.703	0.692	2.01	0.946	1.9E-01	0.668	0.623	1.81	0.921	1.6E-03	0.346	0.344	0.953
	β_{Time}	-6.5E-03	0.169	0.169	2.01	0.942	-6.7E-03	0.089	0.084	1.00	0.937	-4.7E-03	0.089	0.084	0.938
0.75	β_X	-5.6E-03	0.141	0.136	2.01	0.933	-2.4E-02	0.124	0.118	1.75	0.932	9.9E-04	0.067	0.068	0.947
	β_0	2.4E-02	0.713	0.692	2.00	0.946	1.2E-01	0.628	0.603	1.75	0.936	-6.8E-03	0.344	0.345	0.948
	β_{Time}	-3.5E-03	0.168	0.168	2.01	0.945	-2.4E-03	0.087	0.084	1.00	0.942	-3.9E-04	0.087	0.084	0.943

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

B.2.5. Time-varying Missing Variable with a Discrete Auxiliary Variable

To generate the missing and auxiliary variables here we first generate data for N=400 subjects as described in Section B.2 for a time-varying missing covariate and a time-independent continuous auxiliary variable. Then we again convert A to a discrete variable based on its tertiles. To achieve the desired observed correlation we set ρ = 0.01, 0.3, 0.57, and 0.95. We set the percent of subjects with missing data to be 25%, 50%, and 75%.

Tables B.14, B.15, B.16 summarize the results for these simulations. For low (Table B.14) to moderate (Table B.15) missing data, the results are similar to those in Tables B.11 and B.12 for a time-independent missing covariate and a discrete auxiliary variable. For 25% and 50% missing data, the proposed method is unbiased, efficient, and has good coverage. When the percent of missing data is high (Table B.16), the proposed method does not perform well. The method does not estimate the parameter or its variance well, resulting in poor coverage. Additionally, the proposed method is less efficient than the complete case estimator. These result, while more drastic, is consistent with the results for the simulations described above in which a high percent of missing data resulted in worse performance of the proposed estimator. This is likely due to a lack of sufficient information in the validation set to estimate $\hat{P}(Y_j|Z_j)$ well.

Table B.14: Simulation results for 25% missing time-varying covariate with a discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	-1.8E-03	0.060	0.060	1.15	0.940	-1.7E-03	0.058	0.057	1.10	0.938	-1.6E-03	0.052	0.052	0.948
	β_0	7.8E-03	0.311	0.308	1.15	0.950	9.3E-03	0.301	0.295	1.11	0.937	6.7E-03	0.272	0.267	0.934
	β_{Time}	-1.6E-03	0.099	0.097	1.15	0.942	-2.3E-03	0.085	0.087	1.03	0.956	-2.5E-03	0.082	0.084	0.959
0.25	β_X	3.8E-04	0.063	0.060	1.15	0.932	6.5E-04	0.061	0.057	1.10	0.931	7.8E-04	0.054	0.052	0.946
	β_0	-1.4E-03	0.324	0.308	1.15	0.941	-1.8E-03	0.315	0.295	1.10	0.936	-2.5E-03	0.275	0.267	0.948
	β_{Time}	-2.0E-03	0.097	0.097	1.15	0.947	-1.9E-03	0.086	0.086	1.03	0.958	-1.9E-03	0.082	0.084	0.956
0.50	β_X	1.0E-03	0.059	0.059	1.16	0.946	-1.6E-03	0.057	0.056	1.10	0.939	1.4E-04	0.052	0.051	0.945
	β_0	-6.2E-03	0.303	0.306	1.15	0.951	7.9E-03	0.291	0.292	1.10	0.949	-1.9E-03	0.265	0.265	0.949
	β_{Time}	-3.7E-03	0.100	0.097	1.15	0.939	-1.8E-03	0.088	0.087	1.03	0.941	-2.3E-03	0.085	0.084	0.940
0.75	β_X	-2.6E-03	0.062	0.060	1.15	0.938	-4.8E-03	0.059	0.057	1.09	0.938	-2.4E-03	0.052	0.052	0.951
	β_0	1.4E-02	0.316	0.308	1.15	0.943	2.5E-02	0.301	0.292	1.09	0.938	1.3E-02	0.266	0.267	0.954
	β_{Time}	-1.3E-03	0.095	0.097	1.15	0.950	-2.7E-03	0.086	0.086	1.03	0.954	-2.9E-03	0.083	0.084	0.952

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.15: Simulation results for 50% missing time-varying covariate with a discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	-5.0E-03	0.075	0.073	1.41	0.937	-1.7E-02	0.069	0.068	1.31	0.933	-1.6E-03	0.052	0.052	0.948
	β_0	2.2E-02	0.383	0.377	1.41	0.942	8.9E-02	0.357	0.351	1.31	0.931	6.7E-03	0.272	0.267	0.934
	β_{Time}	2.5E-04	0.125	0.119	1.42	0.936	-2.0E-03	0.091	0.092	1.09	0.957	-2.5E-03	0.082	0.084	0.959
0.25	β_X	-1.8E-03	0.077	0.073	1.41	0.943	-1.5E-02	0.072	0.067	1.31	0.928	7.8E-04	0.054	0.052	0.946
	β_0	1.2E-02	0.393	0.376	1.41	0.938	7.8E-02	0.370	0.349	1.31	0.931	-2.5E-03	0.275	0.267	0.948
	β_{Time}	-2.7E-03	0.121	0.118	1.41	0.941	-2.7E-03	0.091	0.092	1.09	0.950	-1.9E-03	0.082	0.084	0.956
0.50	β_X	-7.2E-04	0.073	0.072	1.41	0.949	-1.6E-02	0.067	0.067	1.30	0.939	1.4E-04	0.052	0.051	0.945
	β_0	1.8E-03	0.374	0.374	1.41	0.947	7.8E-02	0.344	0.345	1.30	0.944	-1.9E-03	0.265	0.265	0.949
	β_{Time}	-6.7E-03	0.120	0.119	1.42	0.942	-1.2E-03	0.093	0.092	1.09	0.940	-2.3E-03	0.085	0.084	0.940
0.75	β_X	-3.5E-03	0.077	0.073	1.42	0.936	-1.6E-02	0.069	0.066	1.27	0.928	-2.4E-03	0.052	0.052	0.951
	β_0	1.7E-02	0.400	0.378	1.42	0.939	8.2E-02	0.356	0.340	1.27	0.930	1.3E-02	0.266	0.267	0.954
	β_{Time}	-1.5E-03	0.119	0.119	1.42	0.951	-4.0E-03	0.091	0.091	1.09	0.951	-2.9E-03	0.083	0.084	0.952

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

Table B.16: Simulation results for 75% missing time-varying covariate with a discrete auxiliary variable

		Complete Case					Proposed					Oracle			
ρ		Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	RE	Cov	Bias	SD	$\hat{SE}$	Cov
0.01	β_X	-1.6E-03	0.110	0.103	2.00	0.933	-1.6E-01	0.113	0.099	1.92	0.638	-1.6E-03	0.052	0.052	0.948
	β_0	5.2E-03	0.568	0.535	2.00	0.936	7.7E-01	0.592	0.516	1.93	0.672	6.7E-03	0.272	0.267	0.934
	β_{Time}	-5.5E-03	0.169	0.168	2.00	0.951	-5.9E-04	0.107	0.103	1.23	0.951	-2.5E-03	0.082	0.084	0.959
0.25	β_X	2.8E-03	0.103	0.104	2.01	0.948	-1.4E-01	0.109	0.098	1.91	0.673	7.8E-04	0.054	0.052	0.946
	β_0	-1.0E-02	0.529	0.536	2.01	0.956	7.2E-01	0.565	0.513	1.92	0.685	-2.5E-03	0.275	0.267	0.948
	β_{Time}	-1.5E-03	0.165	0.168	2.01	0.946	-5.3E-04	0.105	0.103	1.23	0.948	-1.9E-03	0.082	0.084	0.956
0.50	β_X	-2.2E-03	0.103	0.102	2.00	0.942	-1.3E-01	0.110	0.096	1.87	0.699	1.4E-04	0.052	0.051	0.945
	β_0	9.9E-03	0.531	0.531	2.00	0.952	6.3E-01	0.571	0.499	1.88	0.719	-1.9E-03	0.265	0.265	0.949
	β_{Time}	2.0E-03	0.168	0.168	2.01	0.945	-4.0E-03	0.108	0.103	1.23	0.937	-2.3E-03	0.085	0.084	0.940
0.75	β_X	-9.6E-04	0.104	0.104	2.01	0.957	-7.6E-02	0.102	0.092	1.78	0.839	-2.4E-03	0.052	0.052	0.951
	β_0	6.1E-03	0.534	0.536	2.01	0.954	3.7E-01	0.531	0.474	1.77	0.840	1.3E-02	0.266	0.267	0.954
	β_{Time}	-7.0E-03	0.169	0.168	2.01	0.947	-4.8E-03	0.108	0.103	1.23	0.936	-2.9E-03	0.083	0.084	0.952

Complete case = complete case estimator. Proposed = proposed estimator. Oracle = oracle estimator.

ρ=correlation between missing and auxiliary variable. SD= standard deviation. $\hat{SE}$=estimated standard error. RE = relative efficiency.

Cov = coverage of 95% confidence interval.

BIBLIOGRAPHY

Alexander, N (2008). Precision of rate estimation under uniform interval censoring. en. *Statistics in Medicine* 27.17, 3442–3445. ISSN: 02776715, 10970258. DOI: `10.1002/sim.3199`. URL: `http://doi.wiley.com/10.1002/sim.3199` (visited on 10/24/2017).

Bertsimas, D, Pawlowski, C, and Zhuo, YD (2018). From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research* 18.196, 1–39. ISSN: 1533-7928. URL: `http://jmlr.org/papers/v18/17-073.html` (visited on 05/24/2019).

Carroll, RJ and Wand, MP (1991). Semiparametric Estimation in Logistic Measurement Error Models. *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3, 573–585. ISSN: 00359246. URL: `http://www.jstor.org/stable/2345587`.

Chang, W, Cheng, J, Allaire, J, Xie, Y, and McPherson, J (2017). *shiny: Web Application Framework for R*. R package version 1.0.3. URL: `https://CRAN.R-project.org/package=shiny`.

Chen, L and Sun, J (2010). A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. eng. *Computational Statistics & Data Analysis* 54.4, 1109–1116. ISSN: 0167-9473. DOI: `10.1016/j.csda.2009.10.022`.

Courtemanche, C, Pinkston, JC, and Stewart, J (2015). Adjusting body mass for measurement error with invalid validation data. en. *Economics & Human Biology* 19, 275–293. ISSN: 1570677X. DOI: `10.1016/j.ehb.2015.04.003`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S1570677X15000350` (visited on 05/02/2019).

Efron, B (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association* 72.359, 557. ISSN: 01621459. DOI: `10.2307/2286217`. URL: `http://www.jstor.org/stable/2286217?origin=crossref` (visited on 08/26/2017).

Erler, NS, Rizopoulos, D, Rosmalen, Jv, Jaddoe, VWV, Franco, OH, and Lesaffre, EMEH (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. eng. *Statistics in Medicine* 35.17, 2955–2974. ISSN: 1097-0258. DOI: `10.1002/sim.6944`.

Etemadi, N (2006). Convergence of Weighted Averages of Random Variables Revisited. *Proceedings of the American Mathematical Society* 134.9, 2739–2744. ISSN: 00029939, 10886826. URL: `http://www.jstor.org/stable/4098124`.

FDA (2015). *Clinical Trial Endpoints for the Approval of NonSmall Cell Lung Cancer Drugs and Biologics Guidance for Industry*. (Visited on 08/26/2017).

Finkelstein, DM (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* 42.4, 845–854. ISSN: 0006-341X.

Goetghebeur, E and Ryan, L (2000). Semiparametric regression analysis of interval-censored data. *Biometrics* 56.4, 1139–1144. ISSN: 0006-341X.

Han, S, Andrei, AC, and Tsui, KW (2014). A Semiparametric Regression Method for Interval-Censored Data. en. *Communications in Statistics - Simulation and Computation* 43.1, 18–30.

ISSN: 0361-0918, 1532-4141. DOI: `10.1080/03610918.2012.697962`. URL: `http://www.tandfonline.com/doi/abs/10.1080/03610918.2012.697962` (visited on 05/01/2019).

Heller, G (2011). Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime Data Analysis* 17.3, 373–385. ISSN: 1572-9249. DOI: `10.1007/s10985-010-9191-8`.

Henningsen, A and Toomet, O (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics* 26.3, 443–458. DOI: `10.1007/s00180-010-0217-1`. URL: `http://dx.doi.org/10.1007/s00180-010-0217-1`.

Hertz-Picciotto, I and Rockhill, B (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. eng. *Biometrics* 53.3, 1151–1156. ISSN: 0006-341X.

Honaker, J, King, G, and Blackwell, M (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software* 45.7, 1–47. URL: `http://www.jstatsoft.org/v45/i07/`.

Ibrahim, JG (1990). Incomplete Data in Generalized Linear Models. en. *Journal of the American Statistical Association* 85.411, 765–769. ISSN: 0162-1459, 1537-274X. DOI: `10.1080/01621459.1990.10474938`. URL: `http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474938` (visited on 01/11/2019).

Ibrahim, JG, Chen, MH, Lipsitz, SR, and Herring, AH (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. en. *Journal of the American Statistical Association* 100.469, 332–346. ISSN: 0162-1459, 1537-274X. DOI: `10.1198/016214504000001844`. URL: `http://www.tandfonline.com/doi/abs/10.1198/016214504000001844` (visited on 01/11/2019).

Inoue, LYT and Parmigiani, G (2002). Designing Follow-Up Times. en. *Journal of the American Statistical Association* 97.459, 847–858. ISSN: 0162-1459, 1537-274X. DOI: `10.1198/016214502388618645`. URL: `http://www.tandfonline.com/doi/abs/10.1198/016214502388618645` (visited on 08/29/2017).

Jackson, C (2016). flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software* 70.8, 1–33. DOI: `10.18637/jss.v070.i08`.

Johansson, ÅM and Karlsson, MO (2013). Comparison of Methods for Handling Missing Covariate Data. en. *The AAPS Journal* 15.4, 1232–1241. ISSN: 1550-7416. DOI: `10.1208/s12248-013-9526-y`. URL: `http://link.springer.com/10.1208/s12248-013-9526-y` (visited on 01/11/2019).

Kim, HY, Williamson, JM, and Lin, HM (2016). Power and sample size calculations for interval-censored survival analysis: Interval-Censored Power Calculation. en. *Statistics in Medicine* 35.8, 1390–1400. ISSN: 02776715. DOI: `10.1002/sim.6832`. URL: `http://doi.wiley.com/10.1002/sim.6832` (visited on 10/24/2017).

Kusnierczyk, W (2012). *rbenchmark: Benchmarking routine for R*. R package version 1.0.0. URL: `https://CRAN.R-project.org/package=rbenchmark`.

Law, CG and Brookmeyer, R (1992). Effects of mid-point imputation on the analysis of doubly censored data. en. *Statistics in Medicine* 11.12, 1569–1578. ISSN: 02776715, 10970258. DOI:

10.1002/sim.4780111204. URL: http://doi.wiley.com/10.1002/sim.4780111204 (visited on 08/26/2017).

Little, RJA and Rubin, DB (2002). *Statistical analysis with missing data*. 2nd ed. Wiley series in probability and statistics. Hoboken, N.J: Wiley. ISBN: 978-0-471-18386-0.

Moons, KGM, Donders, RART, Stijnen, T, and Harrell, FE (2006). Using the outcome for imputation of missing predictor values was preferred. eng. *Journal of Clinical Epidemiology* 59.10, 1092–1101. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2006.01.009.

Pan, W (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* 56.1, 199–203. ISSN: 0006-341X.

Pan, W (1999). Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data. en. *Journal of Computational and Graphical Statistics* 8.1, 109–120. ISSN: 1061-8600, 1537-2715. DOI: 10.1080/10618600.1999.10474804. URL: http://www.tandfonline.com/doi/abs/10.1080/10618600.1999.10474804 (visited on 12/07/2018).

Pepe, MS (1992). Inference using surrogate outcome data and a validation sample. en. *Biometrika* 79.2, 355–365. ISSN: 0006-3444. DOI: 10.1093/biomet/79.2.355. URL: https://academic.oup.com/biomet/article/79/2/355/226037 (visited on 05/07/2019).

Pepe, MS and Fleming, TR (1991). A Nonparametric Method for Dealing With Mismeasured Covariate Data. 86.413, 108. ISSN: 01621459. DOI: 10.2307/2289720. URL: http://www.jstor.org/stable/2289720?origin=crossref (visited on 08/26/2017).

Pigott, K, Rick, J, Xie, SX, Hurtig, H, Chen-Plotkin, A, Duda, JE, Morley, JF, Chahine, LM, Dahodwala, N, Akhtar, RS, Siderowf, A, Trojanowski, JQ, and Weintraub, D (2015). Longitudinal study of normal cognition in Parkinson disease. eng. *Neurology* 85.15, 1276–1282. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000002001.

Pinheiro, J, Bates, D, DebRoy, S, Sarkar, D, and R Core Team (2018). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-137. URL: https://CRAN.R-project.org/package=nlme.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Raab, GM, Davies, JA, and Salter, AB (2004). Designing follow-up intervals. en. *Statistics in Medicine* 23.20, 3125–3137. ISSN: 02776715, 10970258. DOI: 10.1002/sim.1882. URL: http://doi.wiley.com/10.1002/sim.1882 (visited on 08/29/2017).

Røseth, AG, Aadland, E, and Grzyb, K (2004). Normalization of faecal calprotectin: a predictor of mucosal healing in patients with inflammatory bowel disease. en. *Scandinavian Journal of Gastroenterology* 39.10, 1017–1020. ISSN: 0036-5521, 1502-7708. DOI: 10.1080/00365520410007971. URL: http://www.tandfonline.com/doi/full/10.1080/00365520410007971 (visited on 05/02/2019).

Rücker, G and Messerer, D (1988). Remission duration: An example of interval-censored observations. en. *Statistics in Medicine* 7.11, 1139–1145. ISSN: 02776715, 10970258. DOI: 10.

1002/sim.4780071106. URL: http://doi.wiley.com/10.1002/sim.4780071106 (visited on 08/26/2017).

Shaw, LM, Vanderstichele, H, Knapik-Czajka, M, Clark, CM, Aisen, PS, Petersen, RC, Blennow, K, Soares, H, Simon, A, Lewczuk, P, Dean, R, Siemers, E, Potter, W, Lee, VMY, Trojanowski, JQ, and Alzheimer's Disease Neuroimaging Initiative (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. eng. *Annals of Neurology* 65.4, 403–413. ISSN: 1531-8249. DOI: 10.1002/ana.21610.

Sheather, SJ and Jones, MC (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 53.3, 683–690. ISSN: 00359246. URL: http://www.jstor.org/stable/2345597.

Soetaert, K (2009). *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*.

Soetaert, K and Herman, P (2009). *A Practical Guide to Ecological Modelling. Using R as a Simulation Platform*. Springer, 372.

Sun, X and Chen, C (2010). Comparison of Finkelstein's Method With the Conventional Approach for Interval-Censored Data Analysis. en. *Statistics in Biopharmaceutical Research* 2.1, 97–108. ISSN: 1946-6315. DOI: 10.1198/sbr.2010.09013. URL: http://www.tandfonline.com/doi/abs/10.1198/sbr.2010.09013 (visited on 08/07/2017).

Tapiola, T, Pirttilä, T, Mehta, PD, Alafuzofff, I, Lehtovirta, M, and Soininen, H (2000). Relationship between apoE genotype and CSF beta-amyloid (1-42) and tau in patients with probable and definite Alzheimer's disease. eng. *Neurobiology of Aging* 21.5, 735–740. ISSN: 0197-4580.

van Buuren, S and Groothuis-Oudshoorn, K (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45.3, 1–67. URL: https://www.jstatsoft.org/v45/i03/.

Wang, L, McMahan, CS, Hudgens, MG, and Qureshi, ZP (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data: A Novel Method for Fitting the Proportional Hazards Model to Interval-Censored Data. en. *Biometrics* 72.1, 222–231. ISSN: 0006341X. DOI: 10.1111/biom.12389. URL: http://doi.wiley.com/10.1111/biom.12389 (visited on 12/07/2018).

Wellek, S (2017). Sample size planning of two-arm superiority and noninferiority survival studies with discrete follow-up. en. *Statistics in Medicine* 36.20, 3123–3136. ISSN: 02776715. DOI: 10.1002/sim.7360. URL: http://doi.wiley.com/10.1002/sim.7360 (visited on 10/24/2017).

Xu, Y, Kim, JK, and Li, Y (2017). Semiparametric estimation for measurement error models with validation data. en. *Canadian Journal of Statistics* 45.2, 185–201. ISSN: 03195724. DOI: 10.1002/cjs.11314. URL: http://doi.wiley.com/10.1002/cjs.11314 (visited on 01/11/2019).

Yi, GY, Ma, Y, Spiegelman, D, and Carroll, RJ (2015). Functional and Structural Methods With Mixed Measurement Error and Misclassification in Covariates. en. *Journal of the American Statistical Association* 110.510, 681–696. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2014.922777. URL: http://www.tandfonline.com/doi/full/10.1080/01621459.2014.922777 (visited on 05/02/2019).

Yu, L (2012). Bias and Its Remedy in Interval-Censored Time-to-Event Applications. en. In: *Interval-Censored Time-to-Event Data*. Ed. by DG Chen, J Sun, and K Peace. Vol. 52. Chapman and Hall/CRC. DOI: `10.1201/b12290-15`. URL: `http://www.crcnetbase.com/doi/abs/10.1201/b12290-15` (visited on 08/26/2017).

Zeng, L, Cook, RJ, Wen, L, and Boruvka, A (2015). Bias in progression-free survival analysis due to intermittent assessment of progression. eng. *Statistics in Medicine* 34.24, 3181–3193. ISSN: 1097-0258. DOI: `10.1002/sim.6529`.

Zhulina, Y, Cao, Y, Amcoff, K, Carlson, M, Tysk, C, and Halfvarson, J (2016). The prognostic significance of faecal calprotectin in patients with inactive inflammatory bowel disease. en. *Alimentary Pharmacology & Therapeutics* 44.5, 495–504. ISSN: 02692813. DOI: `10.1111/apt.13731`. URL: `http://doi.wiley.com/10.1111/apt.13731` (visited on 05/02/2019).