2016

MULTI-STATE MODELS WITH MISSING COVARIATES

Wenjie Lou
University of Kentucky, louwjapply@gmail.com
Digital Object Identifier: http://dx.doi.org/10.13023/ETD.2016.111

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Wenjie Lou, Student

Dr. Richard J. Kryscio, Major Professor

Dr. Constance Wood, Director of Graduate Studies

MULTI-STATE MODELS WITH MISSING COVARIATES

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Arts and Sciences
at the University of Kentucky

By
Wenjie Lou

Lexington, Kentucky

Director: Dr. Richard J. Kryscio, Professor of Statistics

Lexington, Kentucky

2016

ABSTRACT OF DISSERTATION

MULTI-STATE MODELS WITH MISSING COVARIATES

Multi-state models have been widely used to analyze longitudinal event history data obtained in medical studies. The tools and methods developed recently in this area require the complete observed datasets. While, in many applications measurements on certain components of the covariate vector are missing on some study subjects. In this dissertation, several likelihood-based methodologies were proposed to deal with datasets with different types of missing covariates efficiently when applying multi-state models.

Firstly, a maximum observed data likelihood method was proposed when the data has a univariate missing pattern and the missing covariate is a categorical variable. The construction of the observed data likelihood function is based on the model of a joint distribution of the response longitudinal event history data and the discrete covariate with missing values.

Secondly, we proposed a maximum simulated likelihood method to deal with the missing continuous covariate when applying multi-state models. The observed data likelihood function was approximated by using the Monte Carlo simulation method.

At last, an EM algorithm was used to deal with multiple missing covariates when estimating the parameters of multi-state model. The EM algorithm would be able to handle multiple missing discrete covariates in general missing pattern efficiently.

All the proposed methods are justified by simulation studies and applications to the datasets from the SMART project, a consortium of 11 different high-quality longitudinal studies of aging and cognition.

KEYWORDS: Longitudinal event history data, multi-state model, missing covariate data, EM algorithm, maximum simulated likelihood, SMART project.

Author's Signature: Wenjie Lou

Date: January 23, 2016

MULTI-STATE MODELS WITH MISSING COVARIATES

By

Wenjie Lou

Richard J. Kryscio, PhD

Director of Dissertation

Constance Wood, PhD

Director of Graduate Studies

January 23, 2016

Date

ACKNOWLEDGMENTS

I would like to express my appreciation to my advisor Dr. Richard Kryscio for his advice and support during my graduate studies. His wisdom, guidance and excellent suggestions helped me better understand and finish my dissertation.

I want to thank my committee members, Dr. Constance Wood, Dr. William Griffith, Dr. David Allen and Dr. Erin L Abner for serving on my Supervisory Committee. I am grateful to them for their time, and the careful and critical reading of this dissertation. I also thank Lijie Wan for valuable comments and discussion on the SMART project.

I feel a deep sense of gratitude to my family and friends for their constant love, dedication and support. They are the source of my happiness and motivation behind my achievements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Overview

Longitudinal event history[1] data commonly arise in chronic disease studies in which patients are observed over time and discrete states of the disease are recorded. A change of the disease states is called a transition or event. The outcome data often consists of longitudinal records of time to transition and the types of transitions that occur. Most often, patients are only observed at discrete time points (e.g., annually), which leads to interval-censored transition times and unobserved transitions.

Multi-state models (MSM) [1-4] have become powerful tools in analysis of longitudinal event history data in recent years. These are extensions to the widely used survival models. In survival models, there are just two possible states at any time point, either “alive” or “dead”, and two possible transitions, from “alive” to “alive” or from “alive” to “dead”. Multi-state models allow researchers to investigate a process containing any finite number of states and transitions at the same time.

One major use of multi-state models is to identify and quantify the effects of potential risk factors associated with the different transitions among several states over time. One limitation of the current developed methods and software packages in this area is that they require the covariate data to be completely observed. However, the problem of missing covariates is very common in practice. Indeed, the topics of this dissertation

were motivated by the missing covariates data from the Statistical Modeling of Aging and Risk of Transition (SMART) project established at Sanders-Brown Center on Aging (University of Kentucky).

We aim to develop methods that can handle different types of missing covariates data efficiently in the application of multi-state models. In the following sections, we will introduce the SMART project, a review of multi-state models, the problem of missing covariates data and the methodologies we proposed to address this problem.

1.2 Background of the SMART Project

The Statistical Modeling of Aging and Risk of Transition (SMART) project at University of Kentucky aggregates data from mature, extremely data-rich, and well-known longitudinal cohorts of older adults: the Memory and Aging Project at Washington University (MAPWU); the Oregon Brain Aging Study; Sanders-Brown Healthy Brain Aging Volunteers, also known as the Biologically Resilient Adults in Neurological Studies (BRAiNS) cohort; the Nun Study; the Honolulu Asia Aging Study; the Religious Orders Study; the Memory and Aging Project (Rush University); the African American Dementia and Aging Project; the Klamath Exceptional Aging Project (KEAP); and the Einstein Aging Study (EAS). Participants included in SMART were primarily cognitively intact at baseline and were subsequently assessed for transition to mild cognitive impairment (MCI) and dementia over many years. This combined cohort

presents a unique opportunity to study dementia in terms of the risk factors that lead to cognitive impairment or promote resistance to impairment. Abner, et al. [5] presented a detailed description of this project and its database.

One issue with this project is that we have a large portion of data with missing values on the risk factor covariates of interests due to the different designs of the studies contributing data. For example, the EAS dataset has about 45% subjects with missing APOE4 allele status values; the KEAP dataset contains about 20% subjects with missing APOE4 and about 50% subjects with missing baseline high blood pressure status; and in the MAPWU dataset, there are about 70% patients with missing BMI values. All of the above mentioned covariates are important potential risk factors for the transitions among different types of cognition function states. Thus omitting them from the model is not an appropriate way to analyze the data.

1.3 Multi-State Models

Multi-state models [1-4, 6, 7] are very useful to describe the progression of a disease with several possible states over time. Many applications of multi-state models can be found in the literature. Siannis et al. [8] proposed a multi-state model for joint modeling of terminal and non-terminal events with application to a study of serious coronary heart disease. Abner et al. [9] built a seven-state model to investigate the effects of two different types of mild cognitive impairment (MCI) in the development of dementia. Kryscio et

al.[10] provided a semi-Markov multi-state model to identify risk factors for transitions to MCI and dementia after accounting for the competing risk of mortality. Commenges et al. [11] used an illness-death model to study the incidence and the prevalence of Alzheimer's disease. Other applications of multi-state models can also be found in studies of dementia [9], breast cancer [12-14] , liver cirrhosis [15], AIDS [16, 17],bone marrow transplantation [18, 19],etc.

There are two types of multi-state models: a discrete-time version and a continuous-time version. A discrete-time multi-state model views the transition process as a discrete-time Markov chain assuming the process is observed at equally spaced time points. While, a continuous-time multi-state model models the transitions as a continuous-time Markov process. In this study, we focus on the continuous-time version multi-state models.

1.3.1 Multi-State Process

Continuous-time multi-state models are based on the theory of multi-state processes. A multi-state process is a stochastic process $(X(t), t \geq 0)$, with a finite state space $\boldsymbol{S} = \{1,2, \dots, K\}$. It can be fully characterized by either the transition probability matrix or the transition intensity matrix [1-3, 20].

Denote $\mathcal{F}(s-)$ the history before current time s, which is a σ-algebra generated by $\{X(u), u \in [0, s)\}$. The transition probability matrix is a K by K matrix with entries

$$p_{lm}(s,t) = P\big(X(t) = m|\, X(s) = l; \mathcal{F}(s-)\big), s < t\,.$$

The $(l,m)th$ entry of the transition intensity matrix is defined as:

$$\alpha_{lm}(s) = \begin{cases} \lim_{\Delta t \to 0} \mathrm{P}\big(X(s+\Delta t) = m | X(s) = l; \mathcal{F}(s-)\big)/\Delta t & m \neq l \\ -\sum_{k \neq l} \alpha_{lk}(s) & m = l \end{cases}.$$

The transition intensity α_{lm} measures the instantaneous hazard of the process transition from state l to state m at time s.

1.3.2 Markov Models

The process $(X(t), t \geq 0)$ is Markovian if the transition probabilities and transition intensities are independent of the past history, that is, for any s, t with $0 \leq s < t$, we have

$$P\big(X(t) = m|\, X(s) = l; \mathcal{F}(s-)\big) = P(X(t) = m|\, X(s) = l)$$

and

$$\alpha_{lm}(s) = \begin{cases} \lim_{\Delta t \to 0} \mathrm{P}(X(s+\Delta t) = m | X(s) = l)/\Delta t & m \neq l \\ -\sum_{k \neq l} \alpha_{lk}(s) & m = l \end{cases}$$

For a Markov process, the future of the process after time s depends only on the state occupied at time s. Under the Markov assumption, the transition probabilities can be calculated from the intensities by solving the forward Kolmogorov differential equation [4].

In this dissertation, we focus on time-homogenous Markov models. In time homogeneous Markov models, all transition intensities are assumed to be constant functions of time. Thus, we have $\alpha_{lm}(s) = \alpha_{lm}$. Let $\boldsymbol{P}(s,t)$ be the transition probability matrix with the $(l,m)th$ entry be $p_{lm}(s,t)$, and also let $\boldsymbol{Q}$ be the transition intensity matrix with the $(l,m)th$ entry be α_{lm}. In this case, the Kolmogorov differential equation has an explicit solution using the decomposition of the intensity matrix into eigenvalues and eigenvectors [21], which leads to

$$\boldsymbol{P}(s,t) = \boldsymbol{P}(t-s) = exp\big((t-s)\boldsymbol{Q}\big) = \sum_{r=0}^{\infty} \boldsymbol{Q}^r (t-s)^r / r!\,, s < t.$$

If $\boldsymbol{Q}$ has unique eigenvalues $v_1, \cdots, v_K$ and $\boldsymbol{A}$ is the $K \times K$ matrix whose jth column is a right eigenvector corresponding to v_j, then $\boldsymbol{Q} = \boldsymbol{AVA^{-1}}$, where $\boldsymbol{V} = diag(v_1, \cdots, v_K)$. Then

$$\boldsymbol{P}(t-s) = \boldsymbol{A}\, diag\big(e^{v_1(t-s)}, \cdots, e^{v_K(t-s)}\big)\boldsymbol{A^{-1}}$$

1.3.3 Modeling Intensities

Covariates in multi-state models are often incorporated through the transition intensity functions to explain differences among individuals in the course of the disease progression. One popular choice is the proportional hazards model [22]. Suppose we have a baseline covariate vector $\boldsymbol{Z} = \big(Z_1, Z_2, \ldots, Z_p\big)$, whose values do not change over time, a time-homogenous multi-state model with proportional intensities has the following form:

$$\alpha_{lm}(\boldsymbol{Z}|\boldsymbol{\beta}) = \alpha_{lm,0}\, exp(\boldsymbol{\beta}_{\boldsymbol{lm}}^{T}\boldsymbol{Z}) = exp\big(\beta_{lm,0} + \boldsymbol{\beta}_{\boldsymbol{lm}}^{T}\boldsymbol{Z}\big);\ m \neq l.$$

Here $\alpha_{lm,0} = exp(\beta_{lm,0})$ is called the baseline intensity from state l to state m, and $\boldsymbol{\beta} = \big(\beta_{lm,0}, \boldsymbol{\beta}_{lm}; l = 1, \cdots, K;\ m = 1, \cdots, K; m \neq l\big)$, which represents all the parameters associated with the multi-state model.

1.4 Missing Covariates Data

Even though multi-state models have been widely used in practice, in the current literature there is still no efficient method for handling missing covariates data. The complete case (CC) method is the common approach in most multi-state regression studies and existing software packages [23-27]. There are several limitations to the CC method. It results in biased estimates of the model parameters when covariates are not missing completely at random (MCAR). Even when covariates are MCAR, dropping subjects with incomplete covariate measurement can effectively result in loss of, oftentimes, expensive-to-collect data[28]. Sometimes, if there's a large portion of missing data, the CC method would fail due to convergence problems. Thus, it is urgent for us to identify efficient ways to deal with missing covariates data in the framework of multi-state models. In order to study missing covariates data, we will first introduce two basic concepts of missing data in multi-state model framework.

1.4.1 Missing Data Patterns

Some methods apply only to special patterns of missing data, whereas others apply to any pattern. The concept of missing data pattern [29, 30] describes which values are observed and which values are missing.

Denote $\boldsymbol{X} = (X_1, X_2, \dots, X_M)$ and $\boldsymbol{T} = (t_1, t_2, \dots, t_M)$, here M is a random variable indicating the number of observations, and X_j is the corresponding occupied state of the process $X(t)$ at jth observation at time t_j. Denote $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_p)$ a p-dimensional covariate vector. In our case, the longitudinal response data $\boldsymbol{X}$ is completely observed, while some components of the covariates $\boldsymbol{Z}$ are possibly missing.

Consider three examples of missing data patterns among the covariates in Figures 1-3. For univariate missing data (see *Figure 1.1*), missing values are confined to a single covariate, say Z_1. If there are two or more components of $\boldsymbol{Z}$ with missing values and these components can be rearranged so that all $Z_1, \dots, Z_{j-1}$ are missing for subjects wherever Z_j is missing for all $j = 2, \cdots, p$, then the data is said to have a monotone missing pattern (see *Figure 1.2*). In most cases, we would have general missing data. *Figure 1.3* represents a general pattern with no special structure.

1.4.2 Missing Data Mechanisms

In the previous section we considered various patterns of missing data. A different issue concerns the mechanisms that lead to missing data, and in particular the question of

whether the fact that variables are missing is related to the underlying values of the variables in the data set. Missing data mechanisms are crucial since the properties of missing data methods depend very strongly on these mechanisms. The role of the mechanism in the analysis of data with missing values were largely ignored until the concept was formalized in the theory of Rubin[31], through treating the missing data indicators as random variables and assigning them a distribution.

Rearrange the covariates vector such that $\boldsymbol{Z} = (\boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}})$, where components $\boldsymbol{Z_{obs}}$ are completely observed and components of $\boldsymbol{Z_{mis}}$ are components subject to be missing. Define an indicator vector $\boldsymbol{R} = (R_1, R_2, \cdots, R_{pmis})$ such that

$$R_j := \begin{cases} 1 & \text{if } Z_j \text{ is observed} \\ 0 & \text{if } Z_j \text{ is missing} \end{cases}.$$

Here, $pmis$ is the dimension of the missing components $\boldsymbol{Z_{mis}}$. The missing data mechanism can be characterized by the following conditional distribution $P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ denotes unknown parameters.

The missing data mechanism is called missing completely at random (MCAR) if the missingness of $\boldsymbol{Z_{mis}}$ does not depend on the values of $\boldsymbol{X}$ and $\boldsymbol{Z}$, that is

$$P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}) = P(\boldsymbol{R}|\boldsymbol{\phi}).$$

The missing data mechanism is called missing at random (MAR) if the missingness is independent of the underlying value of $\boldsymbol{Z_{mis}}$, but might be dependent on the values of $\boldsymbol{X}$ and $\boldsymbol{Z_{obs}}$, that is $P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z}) = P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi})$. The mechanism is called not

missing at random (NMAR) if the missingness is also dependent on the underlying value of $\boldsymbol{Z_{mis}}$.

1.4.3 A Review of Methods for dealing with Missing Covariates

Literature focusing on dealing with missing covariates can be found in areas of linear regressions with incomplete observed values in regressors [29, 32], logistic regression models with missing covariates [33], generalized linear models [34-36], survival models [28, 37, 38], etc.

Because standard techniques for most regression models require full covariates information, one simple way to avoid the problem of missing data is to analyze only those subjects who are completely observed. This method, known as the Complete Case (CC) method, is still the default in most software packages. It is well known that the CC analysis can be biased when the data are not MCAR. When the data are MCAR, the CC method is unbiased. However, as the fraction of missing data increases, the deletion of all subjects with missing data is unnecessarily wasteful and quite inefficient [28, 36]. Despite its limitations, the CC method is easy to implement and serves as a useful baseline method for comparisons.

Like the CC method, the available case (AC) method [30] is also widely used in practice. By using the AC method, one removes any covariate with missing values from

the regression model. This method is not helpful when the goal of the analysis to investigate the effects of covariates on the response data.

Mean substitution (MS) [30, 39]is another widely used method in practice to deal with missing covariates data. The MS method primarily works for missing continuous variables. It imputes the missing covariate by its sample mean or its conditional mean calculated from the observed data.

For missing binary covariate, Schemper and Smith [40] proposed the method of probability imputation technique (PIT). By using the PIT method, the missing covariate values are replaced by an estimate for the probability that the unobserved value is equal to 1 based on the complete cases.

Rubin proposed the method of multiple imputation (MI) [41, 42]. MI has become one of the standard methods for dealing with missing values. The method involves three steps, namely (1) creating multiple complete datasets, (2) analyzing each complete dataset using standard analysis, and then (3) combining the parameters estimated from these complete datasets. Issues with MI method are that it requires the data can be presented in a matrix with rows representing independent subjects and columns having fixed dimension representing different variables, and the model used to analyze the multiply imputed data is the same as the model used to impute missing values[43]. In most multi-state model applications, the observed data contain longitudinal response data

with random length and unequal spacing. Thus, it is difficult to place the data in a matrix with independent rows and fixed dimension columns, and it is also difficult to come up with an imputation model for the missing covariates. This might be the reason we cannot find literature of MI application in the framework of multi-state models. An alternative solution would be to impute the missing covariate values conditioning only on other observed covariates. This solution is not helpful when the missing covariates are independent with the completely observed covariates or in cases where all the covariates have some missing values. Little [29] and Lin and Ying [28] also pointed out that imputing the missing covariates based on only the observed covariates could lead to bias and is inappropriate. Schafer provided a comprehensive coverage on this topic [43].

Model-based procedures have been widely applied to deal with missing covariates in generalized linear models (GLMs) [28, 34, 36]. In applying these procedures, we first define a model for the variables with missing values and then make statistical inferences based on ML methods. Model-based methods are quite flexible and clearly set forth underlying model assumptions so that they can be evaluated. In addition, asymptotic variance estimates can be obtained based on second derivatives of the log-likelihood, which takes into account the missing data. Model-based algorithms and techniques include methods based on factoring the likelihood function of the observed data, Newton-Raphson or quasi-Newton algorithms for directly maximizing the likelihood of the

observed data, and the EM algorithm of Dempster, Laird, and Rubin [44] for obtaining ML estimates (MLEs) from the complete-data likelihood [30].

1.5 Estimation Methods

Estimation methods of statistical models with missing covariates are often more complicated than those with complete data, simply because we can often treat the covariate variables in the model as constant if the data is complete. While, this is not the case when the data under study contains covariates with missing values. When the data contains missing covariates values, two things have to be take into consideration in order to analyze the data correctly and efficiently. First, we have to consider the mechanism that leads to the missing covariate data. The missing data mechanism would be critical for a particular method to work properly in the presence of missing data. Second, we have to consider the distribution of the missing covariates. Most of the time, the observed part of the data contains a lot information on the missing part of the data. In the following, we will introduction three estimation methods for multi-state models with missing covariates.

1.5.1 Maximum Likelihood Estimation (MLE)

Maximum Likelihood is a straightforward and easy-to-implement method for dealing with missing covariate data. In the framework of multi-state models, we propose

to construct the likelihood function based on the conditional joint distribution of $(\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}})$**.** Using properties of conditional distributions, we have

$$P(\boldsymbol{R}, \boldsymbol{X}, \boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}}) = P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big)P(\boldsymbol{X}|\boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\beta})P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}}, \boldsymbol{\gamma}).$$

Here $P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big)$ defines the missing data mechanism, $P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big)$ is determined by the multi-state model and $P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}}, \boldsymbol{\gamma})$ is the distribution for the missing components of the covariates conditioned on the observed covariates.

With the data, the likelihood can be calculated a

$$L = \int P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big)P(\boldsymbol{X}|\boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\beta})P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}}, \boldsymbol{\gamma})\, d\boldsymbol{Z_{mis}}.$$

Here the integration is over all possible values of $\boldsymbol{Z_{mis}}$.

If the data is assumed to be MAR, thus we have $P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big) = P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big)$ and the likelihood can be rewritten as

$$L = P\big(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big) \int P(\boldsymbol{X}|\boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\beta})P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}}, \boldsymbol{\gamma})\, d\boldsymbol{Z_{mis}}$$

If the nuisance parameter $\boldsymbol{\phi}$ is distinct from $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$**,** the missing data mechanism can be ignored and we have

$$L \propto \int P(\boldsymbol{X}|\boldsymbol{Z_{mis}}, \boldsymbol{Z_{obs}}, \boldsymbol{\beta})P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}}, \boldsymbol{\gamma})\, d\boldsymbol{Z_{mis}} \qquad (\mathbf{1.1}).$$

When all components of $\boldsymbol{Z_{mis}}$ are discrete variables, the integration in the above formula is a summation.

1.5.2 Maximum Simulated Likelihood (MSL)

When $\boldsymbol{Z_{mis}}$ are continuous variables, calculation of the above likelihood function involves integration, and in most cases it does not have a closed form. Numerical simulation methods can be used to approximate the likelihood function (1.1) in this situation.

Suppose $\boldsymbol{Z_{mis}}$ has a density $g(\boldsymbol{z}|\boldsymbol{\gamma})$. First, we draw H independent random variables $\boldsymbol{Z_1^*}, \cdots, Z_H^*$ from $g(\boldsymbol{z}|\boldsymbol{\gamma})$, and next calculate

$$L^H = \frac{1}{H}\sum_{r=1}^{H} P(\boldsymbol{X}|\boldsymbol{Z_r^*}, \boldsymbol{Z_{obs}}, \boldsymbol{\beta})$$

By the law of large numbers, we have $L^H \xrightarrow{a.s.} L, as\ H \to \infty$. Thus L^H can be used for estimation of the model instead of L, which involves complex integration in the calculation.

1.5.3 EM Algorithm

Expectation Maximization (EM) algorithm [44] is a widely used method in the literature of missing data problem. EM is an iterative algorithm for calculating the MLE in the presence of missing data. Each iteration of EM algorithm consists an E (expectation) step and an M (maximization) step. The E step finds the conditional expectation of the complete-data log likelihood given the observed data and current estimated parameters.

Assume the data is MAR, then the complete data log likelihood has the following form

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = log\big(P(\boldsymbol{X}|\boldsymbol{Z_{mis}},\boldsymbol{Z_{obs}},\boldsymbol{\beta})P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}},\boldsymbol{\gamma})\big)$$
$$= log\big(P(\boldsymbol{X}|\boldsymbol{Z_{mis}},\boldsymbol{Z_{obs}},\boldsymbol{\beta})\big) + log\big(P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}},\boldsymbol{\gamma})\big).$$

Suppose $\big(\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\big)$ is the estimate of the parameter $(\boldsymbol{\beta},\boldsymbol{\gamma})$ from the previous iteration. For the current iteration, the E step can be written as

$$Q\big(\boldsymbol{\beta},\boldsymbol{\gamma}\big|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\big) = E\big(l(\boldsymbol{\beta},\boldsymbol{\gamma})|\boldsymbol{X},\boldsymbol{Z_{obs}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\big)$$
$$= \int log\big(P(\boldsymbol{X}|\boldsymbol{Z_{mis}},\boldsymbol{Z_{obs}},\boldsymbol{\beta})\big)P\big(\boldsymbol{Z_{mis}}\big|\boldsymbol{X},\boldsymbol{Z_{obs}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\big)\,d\boldsymbol{Z_{i,mis}}$$
$$+ \int log\big(P(\boldsymbol{Z_{mis}}|\boldsymbol{Z_{obs}},\boldsymbol{\gamma})\big)P\big(\boldsymbol{Z_{mis}}\big|\boldsymbol{X},\boldsymbol{Z_{obs}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\big)\,d\boldsymbol{Z_{i,mis}}\,.$$

The M step determines $\big(\boldsymbol{\beta}^{(s+1)},\boldsymbol{\gamma}^{(s+1)}\big)$ by maximizing the expected complete data log likelihood function $Q\big(\boldsymbol{\beta},\boldsymbol{\gamma}\big|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\big)$.

1.6 Outline of the Dissertation

The remainder of this dissertation is organized as follows.

In Chapter 2, we propose a likelihood-based method for estimation of multi-state models with univariate missing discrete covariate data. The calculation of log likelihood with missing covariate data is discussed in detail. And the performance of the method is assessed by numerical studies as well as a real data application.

In Chapter 3, we deal with missing continuous covariate in multi-state models. The method of MSL is used for the estimation. Robustness in the assumption of the missing covariate distribution is assessed by numerical studies.

In Chapter 4, EM algorithm is used for estimation in situations where the data under study has a general missing pattern. The method is limited to datasets with only discrete missing covariates. The performance of the method is compared to the widely used CC method by extensive simulation studies.

Finally in Chapter 5, we discuss some topics of the work and offer some potential areas for future study.

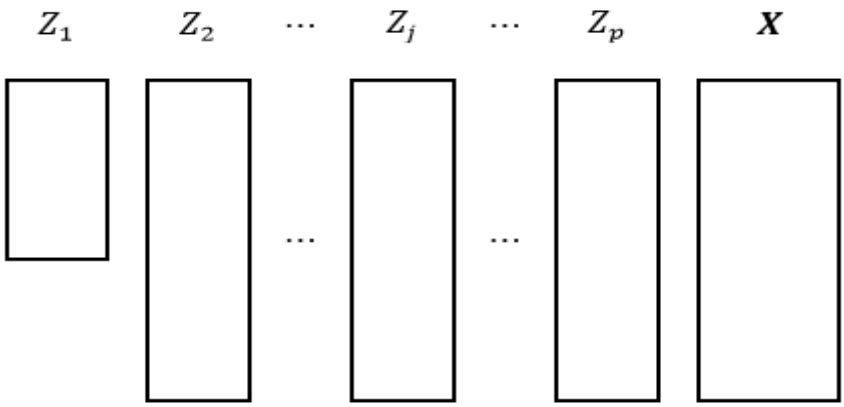

Figure 1.1: Pattern of Univariate Missing Data

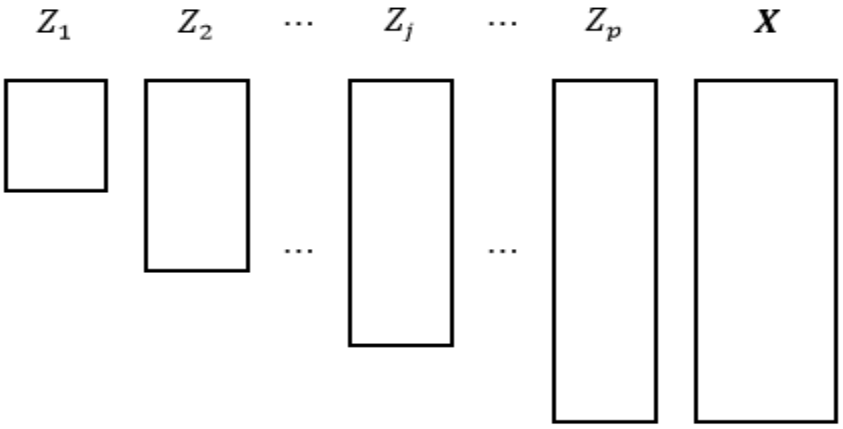

Figure 1.2: Pattern of Monotone Missing Data

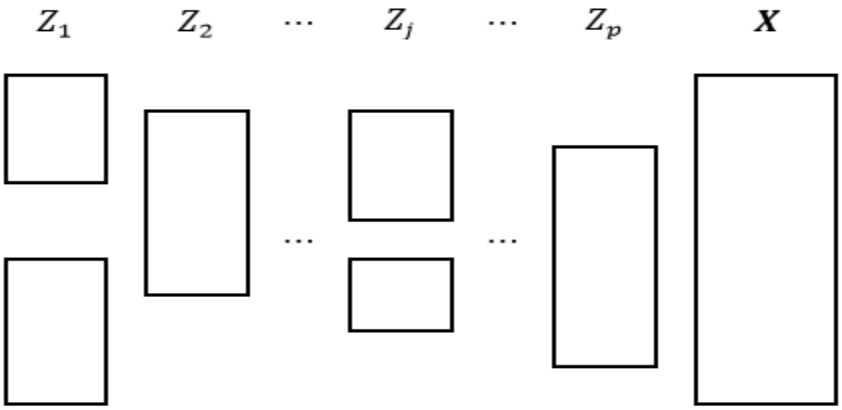

Figure 1.3: General Pattern of Missing Data

Chapter 2 Estimation of Multi-State Models with Missing Categorical Covariate based on Observed Data Likelihood

2.1 Introduction

Continuous-time multi-state models are commonly used to study diseases with multiple stages. In these models, potential risk factors associated with the disease are added to the transition intensities of the model as covariates. But, the problem of missing covariate values arises frequently in practice. In the current literature, estimation methods for multi-state models require complete covariate measurements. In the presence of missing covariate values, the complete case (CC) method is the default. The limitations of CC method are well known; it might produce biased estimates when missing data are not missing completely at random (MCAR), and even if the data are MCAR, dropping a large proportion of the data results in a substantial loss of information.

Other methods of handling missing covariate data in practice includes mean substitution[30, 39] and multiple imputation [30, 41]. The mean substitution method imputes the missing covariate by its sample mean or its conditional mean calculated from the observed data. Schemper and Smith [40] proposed the method of probability imputation technique (PIT) for missing binary covariate. By using the PIT method, the missing covariate values are replaced by an estimate for the probability that the unobserved value is equal to 1 based on the complete cases. Rubin [41, 42] proposed the method of multiple imputation (MI). The method involves three steps, namely (1)

creating multiple complete datasets, (2) analyzing each complete dataset using standard analysis, and then (3) combining the parameters estimated from these complete datasets. Issues with MI method are that it requires the data can be presented in a matrix with rows representing independent subjects and columns having fixed dimension representing different variables, and the model used to analyze the multiply imputed data is the same as the model used to impute missing values[43]. In most multi-state model applications, the observed data contain longitudinal response data with random length and unequal spacing. Thus, it is difficult to place the data in a matrix with independent rows and fixed dimension columns, and it is also difficult to come up with an imputation model for the missing covariate. An alternative is to impute the missing covariate values conditioning only on other observed covariates. Our simulation studies as well as real data application results showed that using multiple imputation method to impute the missing covariates based on only the observed covariates could lead to bias and is inappropriate.

In this chapter, we propose a maximum likelihood method to deal with the missing covariate data problem when estimating multi-state models. The method is based on the observed data log likelihood. The dataset can contain one partially missing categorical covariate as well as several other covariates with complete measurements. The proposed method works in situations even when the response event history data have mixed discrete-continuous pattern observations [20], in which the clinical status is

assessed at discrete visit times while the transition time to one of the absorption states (often death) is observed exactly. As long as the data is MAR, our simulation study shows that the proposed method works well. By adding the missing data mechanism to the model we also can extend this method to NMAR data.

The remainder of the chapter is organized as follows. In Section 2, we present the proposed method. In Section 3, simulation studies were carried out to compare the performance of our method to the widely used CC method, the probability imputation technique (PIT) method proposed by Schemper and Smith [40] and the multiple imputation (MI) method. We applied our method to the Einstein Aging Study (EAS) dataset in the SMART database in Section 4. In the concluding section, we discuss the advantages and possible extensions of our method.

2.2 Multi-State Models with Missing Covariate

In a time-homogenous Markov multi-state model, interesting covariates are often incorporated into the transition intensities using a Cox form regression model:

$$\alpha_{lm}(\boldsymbol{Z}|\boldsymbol{\beta}) = \alpha_{lm,0}\, exp(\boldsymbol{\beta}_{lm}^{T}\boldsymbol{Z}) = exp\big(\beta_{lm,0} + \boldsymbol{\beta}_{lm}^{T}\boldsymbol{Z}\big) \qquad m \neq l.$$

Here $\alpha_{lm,0} = exp(\beta_{lm,0})$ is called the baseline intensity from state l to state m, $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_p)$ is a vector of baseline covariates whose values do not change over time

and are completely observed, and $\boldsymbol{\beta} = \left(\beta_{lm,0}, \boldsymbol{\beta}_{lm}; l = 1, \cdots, K;\ m = 1, \cdots, K; m \neq l\right)$, which represents all the parameters associated with this multi-state model.

In real data applications, it is common that covariates might not be fully observed. In this section, we introduce a likelihood-based method for dealing with partially missing covariate data in the context of continuous-time multi-state models with a continuous-discrete mixed type of observation scheme [20]. Our method can handle one partially missing categorical covariate along with several other completely observed covariates in the model. The completely observed covariates can be either continuous or discrete.

2.2.1 Joint Modeling the Response Data and Missing Covariate Data

Denote $\boldsymbol{T} = (T_1, T_2, \dots, T_M)$ and $\boldsymbol{X} = (X_1, X_2, \dots, X_M)$, here M is a random variable indicating the number of observations, T_j is the time of jth observation and X_j is the corresponding occupied state of the process $X(t)$ at time T_j. To make it simple, we assume that Z_1 is a baseline binary covariate, and its value might be missing. Also denote $\boldsymbol{Z_{obs}} = (Z_2, \dots, Z_p)$**,** which is a vector of other baseline covariates whose values are all observed. We define an indicator R such that

$$R := \begin{cases} 1 & \text{if } Z_1 \text{ is observed} \\ 0 & \text{otherwise} \end{cases}.$$

Here we assume the observation process is ignorable [20], thus the time points of the observation process $\boldsymbol{T}$ can be viewed as fixed. The proposed method is based on the conditional joint distribution of $(R, \boldsymbol{X}, Z_1 | \boldsymbol{T}, \boldsymbol{Z_{obs}})$**.** In this dissertation, we assume the

missing data is MAR. Under the assumption of MAR, the joint distribution can be written as

$$P(R, \boldsymbol{X}, Z_1 | \boldsymbol{T}, \boldsymbol{Z_{obs}}) = P\big(R | \boldsymbol{X}, \boldsymbol{T}, \boldsymbol{Z_{obs}}, \boldsymbol{\phi}\big) P(\boldsymbol{X} | \boldsymbol{T}, Z_1, \boldsymbol{Z_{obs}}, \boldsymbol{\beta}) P(Z_1 | \boldsymbol{Z_{obs}}, \boldsymbol{\gamma}).$$

Here, $\boldsymbol{\beta}$ is the vector of parameters of interests and is associated with the multi-state model while $(\boldsymbol{\phi}, \boldsymbol{\gamma})$ are nuisance parameters.

Since Z_1 is binary, we use logistic regression to model the distribution of Z_1:

$$log\left(\frac{p_1}{1-p_1}\right) = \gamma_0 + \boldsymbol{\varphi}^T \boldsymbol{Z_{obs}}$$

Here $p_1 = P(Z_1 = 1 | \boldsymbol{Z_{obs}}, \boldsymbol{\gamma})$ and $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\varphi})$. In situations that Z_1 and $\boldsymbol{Z_{obs}}$ are independent or there's just one covariate Z_1 and no $\boldsymbol{Z_{obs}}$ variables, we have

$$log\left(\frac{p_1}{1-p_1}\right) = \gamma_0.$$

With the same method, we can generalize the model to the case where Z_1 has more than two levels.

2.2.2 The Likelihood Function

Assume that state 1 to state $K-1$ are all transient states, and the last state K represents death, an absorbing state. Since we cannot make observations in continuous time but only at a finite number of distinct times, the exact transition time to state $1, 2, \dots, K-1$ are unknown and are all interval-censored. The exact time of death can be retrieved, but the state just before death is unknown.

Denote $\boldsymbol{t_i} = (t_{i,1}, t_{i,2}, \dots, t_{i,m_i})$ and $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \dots, x_{i,m_i})$, where $t_{i,j}$ is the time point of the jth observation for subject i and $x_{i,j}$ is the corresponding occupied state at that time point. Define index of death $\delta_i = I(x_{i,m_i} = K)$. Thus we have $\delta_i = 1$ if subject i died at time t_{i,m_i}, and $\delta_i = 0$ otherwise. Write covariates $\boldsymbol{z_i} = (z_{i1}, \boldsymbol{z_{i,obs}})$, here $\boldsymbol{z_{i,obs}} = (z_{i2}, \dots, z_{ip})$. The likelihood contribution for the ith subject, L_i, can be calculated as

$$L_i = P(r_i|\boldsymbol{x_i}, \boldsymbol{t_i}, \boldsymbol{z_{i,obs}}, \boldsymbol{\phi}) \times \left(P(\boldsymbol{x_i}|\boldsymbol{t_i}, \boldsymbol{z_i}, \boldsymbol{\beta})P(z_{i,1}|\boldsymbol{z_{i,obs}}, \boldsymbol{\gamma})\right)^{r_i} \left(\sum_{z_{i,1}=0}^{1} P(\boldsymbol{x_i}|\boldsymbol{t_i}, \boldsymbol{z_i}, \boldsymbol{\beta})P(z_{i,1}|\boldsymbol{z_{i,obs}}, \boldsymbol{\gamma})\right)^{1-r_i}.$$

Here $r_i = 1$ if z_{i1} is observed and 0 otherwise, and the conditional distribution $P(\boldsymbol{x_i}|\boldsymbol{t_i}, \boldsymbol{z_i}, \boldsymbol{\beta})$, under the assumption that the process is Markov and given the baseline state, can be calculated as

$$P(\boldsymbol{x_i}|\boldsymbol{t_i}, \boldsymbol{z_i}, \boldsymbol{\beta}) = \prod_{j=2}^{m_i} P(x_{i,j}|x_{i,j-1}, t_{i,j-1}, t_{i,j}, \boldsymbol{z_i}, \boldsymbol{\beta}).$$

Here

$$P(x_{i,j}|x_{i,j-1}, t_{i,j-1}, t_{i,j}, \boldsymbol{z_i}, \boldsymbol{\beta}) = \begin{cases} p_{x_{i,j-1},x_{i,j}}(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta}) & \text{if } j \neq m_i \\ \left[p_{x_{i,j-1},x_{i,j}}(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta})\right]^{1-\delta_i} \left[\sum_{k\neq K} p_{x_{i,j-1},k}(t_{i,j}, t_{i,j}|\boldsymbol{z_i}, \boldsymbol{\beta})\alpha_{kK}(\boldsymbol{z_i}|\boldsymbol{\beta})\right]^{\delta_i} & \text{if } j = m_i \end{cases}$$

The log-likelihood function of all the subjects can be written as

$$l = l(\boldsymbol{\phi}) + l(\boldsymbol{\beta}, \boldsymbol{\gamma}).$$

Here

$$l(\boldsymbol{\phi}) = \sum_{i=1}^{n} \log\left(P\left(r_i|\boldsymbol{x_i}, \boldsymbol{t_i}, \boldsymbol{z_{i,obs}}, \boldsymbol{\phi}\right)\right)$$

and

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(r_i \log\left(P\left(\boldsymbol{x_i}|z_{i,1}, \boldsymbol{z_{i,obs}}, \boldsymbol{\beta}\right) P\left(z_{i,1}|\boldsymbol{z_{i,obs}}, \boldsymbol{\gamma}\right)\right) + (1 - r_i)\left(\sum_{z_{i,1}=0}^{1} P\left(\boldsymbol{x_i}|z_{i,1}, \boldsymbol{z_{i,obs}}, \boldsymbol{\beta}\right) P\left(z_{i,1}|\boldsymbol{z_{i,obs}}, \boldsymbol{\gamma}\right) \right) \right)$$

If the nuisance parameters $\boldsymbol{\phi}$ is not a function of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, then we have $l \propto l(\boldsymbol{\beta}, \boldsymbol{\gamma})$.The Newton-Raphson method can be used to maximize the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

2.3 Simulation Study

In this section, we evaluate the performance of the proposed method (referred to as OL throughout) through simulations. Both MCAR and MAR data will be considered here. We performed the full data method (referred to as FULL) on the original non-missing covariates datasets. This estimation acts like a benchmark, so we can evaluate the performance of our method in general. To show the advantage of the method, we also compared it to the CC method, the PIT method proposed by Schemper and Smith [40], and the MI method. In applying the MI method, we impute the missing covariate values

by a regression model only based on the other fully observed covariates and ignoring the longitudinal response data. The criteria used in the comparisons are percent bias (% bias), estimated standard error (SE) and 95% confidence interval coverage rate (95% CP). The results are based on 500 simulated datasets. All simulations were done in SAS 9.3 system[45]. PROC IML is used to calculate and maximize the likelihood functions. PROC MI and PROC MIANALYZE were used to impute datasets and to combine results for the MI method.

We consider two covariates (Z_1, Z_2) in this simulation study. The first covariate Z_1 is a baseline binary covariate, and its value would be missing for some subjects. The second covariate Z_2 is a continuous covariate, and its value will be completely observed. We generate Z_2 from a truncated normal distribution with $(\mu = 1, \sigma = 0.25)$ and truncated at 0 and 2. We generate Z_1 by a logistic regression model based on the value of Z_2, which has the following form

$$P(Z_1 = 1) = \frac{exp(\gamma_0 + \gamma_1 Z_2)}{\left(1 + exp(\gamma_0 + \gamma_1 Z_2)\right)}.$$

We set $(\gamma_0, \gamma_1) = (-0.2, 0.2)$ in the true model.

The longitudinal response data will be generated from a three-state model with the following transition intensity matrix:

$$\boldsymbol{Q} = \begin{Bmatrix} -\left(0.2 + \exp(-2.2 + \beta_1 Z_1)\right) & \exp(-2.2 + \beta_1 Z_1) & 0.2 \\ 0 & -\exp(-2.0 + \beta_2 Z_2) & \exp(-2.0 + \beta_2 Z_2) \\ 0 & 0 & 0 \end{Bmatrix}.$$

The true values of β_1 and β_2 are 0.8 and 0.4 respectively. Subjects began in state 1 at time 0. States 1 and 2 are transient states, while state 3 is an absorbing state. The states occupied were observed at 1-year intervals with a common censoring time 10 years if the subject was still in state 1 or state 2. The number of subjects (N) was 1000.

We consider both MCAR and MAR datasets in this study. For the MCAR datasets, we considered two cases. In the first case, we randomly set Z_1 missing for about 30% of the subjects. We denote this case as MCAR 1. In the second case, we randomly set Z_1 missing for about 70% subjects. Denote this type of dataset as MCAR 2. Both of these two types of missing data are MCAR data, since the missingness of the data is independent of underlying values of both the observed and unobserved data.

For the MAR datasets, we also considered two cases. In the first case, we set Z_1 missing if the first transition happens within the first year. In this situation, we have about 30% subjects with missingness in the covariate. We denote this type of dataset as MAR 1. In the second case, we set Z_1 missing if the first transition happens within the first 3 years. In this case, we had about 70% subjects with missingness in the covariate. We denote this dataset as MAR 2. Both of these two types of missing data are MAR, since whether the covariate is missing depends on the observed data, which is the transition history, but not on the unobserved data, which is the underlying value of Z_1.

The simulation results are listed in *Table 2.1*. When the datasets are MCAR, the CC, PIT and OL methods all worked well. The percent bias was small, and the real coverage rates for the nominal 95% confidence intervals were all near 95%. Both results were close to those provided by the full data estimates in which we had complete covariate values for all subjects. The PIT method has relative larger bias than the CC method and the proposed OL method for the coefficient estimate of the partially missing covariate Z_1, but the bias is in an acceptable range. The CC method is less efficient than the PIT method and the proposed OL method in terms of standard errors, especially for the datasets MCAR 2, which have a higher missing proportion. The MI method by imputing the missing data based on only the observed covariate performs the worst among the four methods considered. The results by MI method have large bias for the coefficient estimate of the partially missing covariate Z_1 even though the data is MCAR.

When the data were MAR, the proposed OL method still works well, while the other three methods considered here have relatively large biases. The CC method would have as large as 120% bias for the coefficient estimate of Z_1. The PIT and MI method perform better than the CC method, but both methods have at least 10% bias for the coefficient estimate of Z_1 no matter whether Z_1 and Z_2 are independent or not. However, our method (OL) still worked well for MAR data; the estimates provided by our method were unbiased. The simulation results also showed that our method did not lose much

efficiency because of the missing data. The increases of the standard errors were relatively small compared to those for the full data estimates (FULL), where we had complete covariate data for all subjects.

2.4 Application

In this application, we used the data of the Einstein Aging Study (EAS) from the SMART database. EAS, located at the Albert Einstein College of Medicine, Yeshiva University (New York City), is a population-based study of cognitive aging and dementia in a non-institutionalized, urban, and multi-ethnic community. Participants undergo annual in-person evaluations that include cognitive, neurological, functional, and physical assessments [46].

Three cognitive states are of interest, and they are: Cognitively Intact (state 1), MCI (state 2), and Dementia (state 3). Death (state 4) is also included in our model as a competing risk for MCI and Dementia. Being Cognitively Intact, MCI, and Dementia are transient states, while Death is an absorbing state. Participants with normal cognitive function may die or transition to MCI or dementia. Participants who transition to MCI may die, transition to Dementia, or reverse back to the Cognitively Intact state. Participants who were diagnosed as Demented cannot reverse back to being MCI or

Cognitively Intact; however, they can transition to Death. See *Figure 2.1* for the refined transition model structure.

The original dataset contains longitudinally observed data on 2,097 patients. In this application, we excluded participants if they had only one observation (n=500) or already entered into MCI or Dementia at baseline (n=395). For the 1,202 participants included in this study, all of these participants were Cognitively Intact at baseline, and there were 4,302 total observed transitions. This resulted in 4.6 ± 2.7 transitions per subject with a mean follow-up 5.7 ± 3.6 years.

One covariate of interest for our application is Apolipoprotein E4 allele (APOE4), a genetic marker of Alzheimer's risk; this covariate was not available for all EAS participants. There were 509 (42.4%) participants with missing APOE4. See *Table 2.2* for the observed transition frequencies for both the full data and the dataset with only the complete cases. Other risk factors of interest in this model are baseline age (Bage), female gender (Female) and education level (LowEdu). Education level was dichotomized into two groups, low education (≤ 12 years) and high education (> 12 years, above high school level). Baseline age, education level, and female gender were all fully observed. Summaries of these risk factors are listed in *Table 2.3*.

We used a time-homogenous Markov model to study the effects of these risk factors on each transition. This four-state model has the following transition intensity matrix:

$$\mathrm{Q}(\mathbf{Z}) = \begin{bmatrix} \alpha_{11}(\boldsymbol{Z}) & \alpha_{12}(\boldsymbol{Z}) & \alpha_{13}(\boldsymbol{Z}) & \alpha_{14}(\boldsymbol{Z}) \\ \alpha_{21}(\boldsymbol{Z}) & \alpha_{22}(\boldsymbol{Z}) & \alpha_{23}(\boldsymbol{Z}) & \alpha_{24}(\boldsymbol{Z}) \\ 0 & 0 & \alpha_{33}(\boldsymbol{Z}) & \alpha_{34}(\boldsymbol{Z}) \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Here $\boldsymbol{Z}$ is the vector of above risk factors, namely Bage, Female, LowEdu, and APOE4. Since all patients were over 60 years old at baseline, we centered the original baseline age at 60 to use as the age covariate in the model. Since the effects of risk factors on the backward transition from state MCI to state Cognitively Intact are not of interest in this study, we didn't add covariates on the transition intensity function for this path. Thus, the non-zero intensity functions in the transition intensity matrix $\mathrm{Q}(\mathbf{Z})$ have the following Cox-type regression form:

$$\alpha_{lm}(\boldsymbol{Z}) = \begin{cases} exp\left(\beta_{lm,0}\right) & if\ l = 2, m = 1 \\ exp\left(\beta_{lm,0} + \beta_{lm}^{T}\boldsymbol{Z}\right) & if\ l < m \\ -\sum_{h \neq l} \alpha_{lh}(\boldsymbol{Z}) & if\ l = m \end{cases}$$

We considered five methods to deal with the missing covariate data, the AC method, the CC method, the PIT method, the proposed OL method, and the MI method. For each method, a backwards algorithm was used for model selection. A basic model was fit to the data with all risk factors modeled on all possible transitions. At each step

the coefficient with the largest p value was eliminated from the model until all coefficients remaining in the model were significant at the 0.05 level.

First, we fitted the model without APOE4. With APOE4 excluded from the model, the data were completely observed. *Table 2.4* lists the hazard ratios and the corresponding 95% confidence intervals for all the significant risk factors (P<0.05). We use these results to see how the missing APOE4 data affected the results of our proposed OL method as well as the results of the CC, the PIT, and the conventional MI methods.

For the proposed OL method, the distribution of APOE4 is constructed by the logistic regression model based on the fully observed covariates, which are Bage, Female and LowEdu. The regression model has the following form:

$$log\left(\frac{P(APOE4=1)}{1-P(APOE4=1)}\right)=\varphi_0+\varphi_1 Bage+\varphi_2 Female+\varphi_3 LowEdu.$$

The results generated by the OL method are listed in *Table 2.5*.

Table 2.6 lists the results by using the PIT method. In the PIT method, missing APOE4 values are imputed by its percentage based the observed data, which is 21.6% in our data (see *Table 2.3*).

Since the likelihood function contained as many as 30 parameters in the initial model and 42.4% of the subjects had missing data, the model failed to converge using the

CC method. We refitted CC method on the reduced model based on the results of the PIT method and the proposed OL method. The results are listed in *Table 2.7*.

When applying the MI method, we imputed the missing APOE4 by logistic model using the fully observed covariates Bage, Female and LowEdu as predictors. The combined results showed that APOE4 is not significant on any of the transition paths. Thus the final model reduces to the one of the AC method. In order to compare to the proposed OL method, we refitted the MI method on the reduced model based on the results of the PIT method and the OL method. The results are listed in *Table 2.8*.

The PIT method and the proposed OL method provide very similar results. The effects of baseline age (Bage), female gender (Female), and low education (LowEdu) on each path by both methods are also very close to those by the AC method. In general, a 1-year increase in baseline age increased the transition intensities from Cognitively Intact to MCI and MCI to Dementia. Cognitively Intact women had a lower mortality rate than Cognitively Intact men. Cognitively Intact subjects with low education level had a higher risk for the transition to MCI. And both the PIT method and the OL method showed that APOE4 increased the transition intensities from Cognitively Intact to MCI. These findings are much in line with other similar studies [10, 47]. The CC method fails to indicate that Female gender has a significant effects on the transition intensity from Cognitively Intact state to Death, while the effects of baseline age, low education level

and APOE4 agree with the results by PIT method and OL method. The MI method, by ignoring the longitudinal outcome data, fails to indicate APOE4 is significant in the model.

We used the length of the 95% confidence intervals to compare the efficiency among the methods which are considered here. The results showed that the CC method is the least efficient method among the three. It has the largest length of each hazard ratio 95% confidence interval. The proposed OL method shows a small advantage over the PIT method. Its length of the 95% confidence interval of the hazard ratio for the missing covariate APOE4 on the path from state Cognitive Intact. to state MCI is less than the length provided by the PIT method (0.66 vs. 0.70). The MI method considered here has similar efficiency to the proposed OL method. However, MI method showed bias on the estimation of the effect of APOE4, which results in APOE4 not being significant in the model.

2.5 Discussion

Multi-state models have become a popular tool to study transitions in studies of chronic disease. But, most model estimation methods and applications require fully observed data. We proposed a likelihood-based method that would handle missing covariate data in continuous-time Markov multi-state models. The proposed method

works for dataset with one missing categorical covariate and several other completely observed covariates. The completely observed covariates are allowed to be either continuous or discrete.

Compared to the CC method, which is commonly used in multi-state models with missing covariate data, our method has three major advantages. First, when the data are MCAR our method was more efficient, especially when the data contained a large portion of cases with missing covariate data. Second, our method worked for both MCAR data and MAR data, while the CC method could provide biased estimates if the data are not MCAR. Third, when the dataset contains a large portion of missing covariate cases, the CC method might fail due to a convergence problem. In this case, the proposed method would be an alternative for the analysis.

From our simulation studies, the PIT method proposed by Schemper and Smith would also produce biased results when the data is not MCAR. The covariate distributions of the missing covariates for the missing cases and observed cases would be different when the missingness is dependent on other observed data. Thus the PIT method is not recommended when the data is MAR not MCAR in estimation of multi-state models with missing categorical covariate.

The MI method has become a widely used tool to deal with missing data. It is an efficient method and is also easy to carry out in many statistical modeling area with

missing data. However, in the area of multi-state models with missing covariates data, the proper MI method is often not easy to carry out. Our study found out that the conventional method of imputing the missing covariate by a model based only on the other observed covariates is biased even when the data is MCAR.

In our application to the EAS dataset, CC is inferior since it has the widest CIs and since it fails to find female to be protective for transitions from cognitive intact to death. Methods OL and PIT give the same results in the application, which were our expectation since the EAS dataset is more MCAR than MAR. The MI method based on observed covariates showed bias in the estimation of the effect of APOE4 in the model.

If the missing covariate is continuous, the observed data log-likelihood involves integration. In this circumstance calculation of the log-likelihood would be difficult, since there is never a closed form for the integration. Numerical integration methods, such as Gaussian quadrature, Monte Carlo simulations, quasi-Monte Carlo, etc., can be used to approximate observed data log-likelihood. We will explore this type of method in the next chapter.

The proposed method could also be extended to data with multiple categorical covariates with missing data. However, if the data contain too many covariates with missing values, this method might encounter programming difficulties or convergence problems when fitting the model. Modeling multiple covariates with missing data would

introduce many nuisance parameters into the likelihood function. With too many nuisance parameters in the likelihood function, there might be difficulties computing estimates using the usual maximization techniques, like the Newton-Raphson method. One possible solution to the problem discussed above is to use the EM algorithm as discussed by Ibrahim et al.[36]. By applying the EM algorithm, the nuisance parameters associated with the covariates model can be estimated separately from the model parameters associated with the multi-state model. In Chapter 4, we will have a detailed discussion about this method.

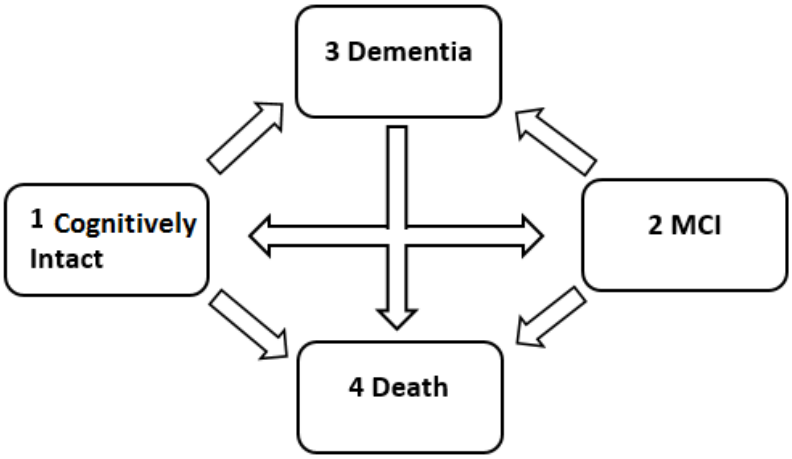

Figure 2.1: Multi-State Model Structure

Table 2.1: Simulation Results for four different types of missing covariate datasets

		β_1 (0.8)			β_2 (0.4)		
MissMech	Method	% bias	SE	95% CP	% bias	SE	95% CP
0	FULL	0.15%	0.064	94.6%	1.61%	0.056	94.4%
MCAR1	CC	0.27%	0.077	94.6%	2.05%	0.067	95.0%
	PIT	0.74%	0.071	93.8%	1.51%	0.056	94.0%
	OL	0.22%	0.071	94.2%	1.59%	0.056	94.0%
	MI	-14.37%	0.073	69.8%	-0.01%	0.056	94.8%
MCAR2	CC	0.59%	0.118	95.4%	2.54%	0.102	93.6%
	PIT	1.96%	0.084	92.2%	1.42%	0.056	94.2%
	OL	0.31%	0.085	94.4%	1.57%	0.056	94.4%
	MI	-33.78%	0.083	6.2%	-2.13%	0.056	93.8%
MAR1	CC	-51.56%	0.081	0.0%	-4.89%	0.071	93.4%
	PIT	-12.15%	0.069	73.8%	-0.80%	0.056	95.2%
	OL	0.23%	0.068	94.2%	1.65%	0.056	94.0%
	MI	-13.33%	0.074	75.4%	-0.43%	0.056	95.0%
MAR2	CC	-118.32%	0.128	0.0%	-11.41%	0.119	93.8%
	PIT	-18.36%	0.082	57.8%	-2.74%	0.056	94.0%
	OL	0.13%	0.072	94.8%	1.60%	0.056	94.4%
	MI	-30.65%	0.091	17.6%	-3.14%	0.056	93.6%

Note: MissMech=Missing Mechanism; FULL=Full Data Analysis; CC=Complete Case; PIT=Probability Imputation Technique; OL=Observed-data Likelihood; MI=Multiple Imputation; 95% CP=95% Confidence Interval Coverage Rate.

Table 2.2: Observed frequency of transitions.

From	To (Full Data)				To (Complete Case)			
	Co.I.	MCI	Dementia	Death	Co.I.	MCI	Dementia	Death
Co.I.	3027	328	37	454	2260	236	24	185
MCI	151	126	42	47	121	97	32	21
Dementia	.	.	48	42	.	.	30	27

Note: Co.I. =Cognitive Intact; Left panel: Full data; Right panel: Complete Case.

Table 2.3: Summary statistics of baseline risk factors (N=1,202).

Baseline Risk Factor	N Missing (%)	N Percent (%) or mean ± st. dev.
Baseline age	0 (0)	78.88 ± 5.34
Female	0 (0)	724 (60.2)
Low Education (≤12 years)	0 (0)	593 (49.3)
APOE4 (≥ 1 e4 allele)	509 (42.4)	150 (21.6)

Table 2.4: Hazard ratios (HR) for risk factors using Available Case method.

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	Co.I. to MCI	1.07	1.05	1.08	0.03
	MCI to Dementia	1.04	1.00	1.07	0.07
Female	Co.I. to Death	0.43	0.22	0.86	0.64
Low Education	Co.I. to MCI	1.54	1.30	1.82	0.52

Note: Co.I. =Cognitive Intact

Table 2.5: Hazard ratios (HR) for risk factors using Observed-data Likelihood method.

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	Co.I. to MCI	1.07	1.05	1.08	0.03
	MCI to Dementia	1.04	1.01	1.07	0.06
Female	Co.I. to Death	0.45	0.24	0.85	0.61
Low Education	Co.I. to MCI	1.54	1.30	1.83	0.53
APOE4	Co.I. to MCI	1.44	1.15	1.81	0.66

Note: Co.I. =Cognitive Intact

Table 2.6: Hazard ratios (HR) for risk factors using Probability Imputation Technique method.

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	Co.I. to MCI	1.07	1.05	1.09	0.03
	MCI to Dementia	1.04	1.01	1.07	0.06
Female	Co.I. to Death	0.44	0.23	0.84	0.61
Low Education	Co.I. to MCI	1.54	1.30	1.83	0.53
APOE4	Co.I. to MCI	1.47	1.16	1.86	0.70

Note: Co.I. =Cognitive Intact

Table 2.7: Hazard ratios (HR) for risk factors using Complete Case method on the reduced model

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	Co.I. to MCI	1.07	1.05	1.09	0.04
	MCI to Dementia	1.05	1.01	1.10	0.09
Female	Co.I. to Death	0.25	0.05	1.25	1.20
Low Education	Co.I. to MCI	1.59	1.29	1.97	0.68
APOE4	Co.I. to MCI	1.52	1.19	1.95	0.76

Note: Co.I. =Cognitive Intact

Table 2.8: Hazard ratios (HR) for risk factors using Multiple Imputation method on the reduced model

Risk Factor	Path	HR	95% Confidence Interval		
			L	U	Length
Baseline Age	Co.I. to MCI	1.07	1.05	1.09	0.03
	MCI to Dementia	1.04	1.01	1.07	0.07
Female	Co.I. to Death	0.43	0.22	0.85	0.62
Low Education	Co.I. to MCI	1.54	1.30	1.82	0.53
APOE4	Co.I. to MCI	1.25	0.98	1.56	0.58

Note: Co.I. =Cognitive Intact

Chapter 3 Estimation of Multi-State Models with Missing Continuous Covariate using Maximum Simulated Likelihood

3.1 Introduction

Multi-state models are very useful tools to model chronic disease processes in which patients might go through several different states. The common way to account for interpersonal difference in the disease process in multi-state models is to add covariates in the transition intensity functions. In the current literature, covariates have to be completely observed. However, missing covariate data is very common in practice.

In Chapter 2, we discussed how to deal with discrete missing covariate data in the framework of multi-state models. In this chapter, we propose a maximum simulated likelihood method for estimation of multi-state models with missing continuous covariate data. The method works for datasets with one missing continuous covariate. The dataset can also contain several other completely observed covariates and the completely observed covariates can be either discrete or continuous. The method is based on the assumption that the corresponding missing covariate follows a normal distribution. Through simulation studies, we showed that the proposed method is actually robust to the normal assumption. The method still works with moderate violation of the normal distribution assumption, and it works well for both MCAR and MAR type of missing data.

The rest of the chapter is organized as follows. In Section 2, we present the method of maximum simulated likelihood in the framework of multi-state models with missing continuous covariate data. In Section 3, the results of simulation studies are presented. In this section we check the robustness of the distribution assumption of the covariate. In Section 4, we apply the proposed method to a real data, the MAPWU dataset from the SMART database. In the discussion section, we discuss the advantages as well as limitations of the proposed method and lay out some possible future work.

3.2 The Maximum Simulated Likelihood Method

Covariates in multi-state models are often incorporated through the transition intensity functions to explain differences among individuals in the course of the disease progression. One popular choice is the proportional hazards model, which has the following form for a time-homogenous model:

$$\alpha_{lm}(\boldsymbol{Z}|\boldsymbol{\beta}) = \alpha_{lm,0}\, exp(\boldsymbol{\beta}_{\boldsymbol{lm}}^{T}\boldsymbol{Z}) = exp\big(\beta_{lm,0} + \boldsymbol{\beta}_{\boldsymbol{lm}}^{T}\boldsymbol{Z}\big).$$

Here $\alpha_{lm,0} = exp(\beta_{lm,0})$ is called the baseline intensity from state l to state m, $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_p)$ is a vector of baseline covariates whose values do not change over time, and $\boldsymbol{\beta} = \big(\beta_{lm,0}, \boldsymbol{\beta}_{lm}; l = 1, \cdots, K;\ m = 1, \cdots, K; m \neq l\big)$, which represents all the parameters associated with the multi-state model.

Let $\boldsymbol{t_i} = (t_{i,1}, t_{i,2}, \dots, t_{i,m_i})$ be the scheduled observation time points, and $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \dots, x_{i,m_i})$ be the corresponding occupied states for subject i. We view the observation time points $\boldsymbol{t_i}$ as given. Suppose z_{i1} is continuous and is missing for a subset of subjects, and covariates $\boldsymbol{z}_{i,obs} = (z_{i2}, \dots, z_{ip})$ are completely observed for all subjects. Components of $\boldsymbol{z}_{i,obs}$ can be either continuous or discrete. Define an indicator variable r_i such that

$$r_i := \begin{cases} 1 & \text{if } z_{i1} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

Assume the data is MAR, which assumes that $P(r_i|\boldsymbol{x_i}, z_{i1}, \boldsymbol{z}_{i,obs}) = P(r_i|\boldsymbol{x_i}, \boldsymbol{z}_{i,obs})$. The likelihood function is based on the conditional joint distribution of $(r_i, \boldsymbol{x_i}, z_{i1}|\, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}\,)$, which can be written as

$$P(r_i, \boldsymbol{x_i}, z_{i1}|\, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}\,) = P(r_i|\boldsymbol{x_i}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\phi})P(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\beta})P(z_{i1}|\boldsymbol{z}_{i,obs}; \boldsymbol{\gamma}).$$

In the above formula, we assume $P(z_{i1}|\boldsymbol{z}_{i,obs}; \boldsymbol{\gamma})$ has a normal distribution with mean $\mu = \mu_0 + \boldsymbol{\varphi}^T \boldsymbol{z}_{i,obs}$ and standard deviation σ. Here $\boldsymbol{\gamma} = (\mu_0, \boldsymbol{\varphi}, \sigma\,)$**.** The term $P(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\beta})$ is the same as in a multi-state model with completely observed covariates data. Give the baseline state $x_{i,1}$, we have

$$P(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\beta}) = \prod_{j=2}^{m_i} P(x_{i,j}|x_{i,j-1}, t_{i,j-1}, t_{i,j}, \boldsymbol{z_i}, \boldsymbol{\beta})$$

Assume the exact time of death is recorded but the state just before death is unknown. Define index of death $\delta_i = I(x_{i,m_i} = K)$, here state *K* means death. Thus we

have $\delta_i = 1$ if subject i died at time t_{i,m_i}, and $\delta_i = 0$ otherwise. In this type of observation scheme, we have

$$P\left(x_{i,j}|x_{i,j-1}, t_{i,j-1}, t_{i,j}, \boldsymbol{z}_i, \boldsymbol{\beta}\right)$$
$$= \begin{cases} p_{x_{i,j-1},x_{i,j}}\left(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta}\right) & \text{if } j \neq m_i \\ \left[p_{x_{i,j-1},x_{i,j}}\left(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta}\right)\right]^{1-\delta_i}\left[\sum_{k\neq K} p_{x_{i,j-1},k}\left(t_{i,j}, t_{i,j}|\boldsymbol{z_i}, \boldsymbol{\beta}\right)\alpha_{kK}(\boldsymbol{z_i}|\boldsymbol{\beta})\right]^{\delta_i} & \text{if } j = m_i \end{cases}$$

Here, $p_{x_{i,j-1},x_{i,j}}\left(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta}\right)$ is the one-step transition probabilities, which is defined as

$$p_{x_{i,j-1},x_{i,j}}\left(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta}\right) = P\left(X\left(t_{ij}\right) = x_{i,j}\middle|X\left(t_{ij-1}\right) = x_{i,j-1}; \boldsymbol{z_i}, \boldsymbol{\beta}\right).$$

It can be calculated from the intensities by solving the forward Kolmogorov differential equation [4].

The log-likelihood for all subjects can be written as $l = l(\boldsymbol{\phi}) + l(\boldsymbol{\beta}, \boldsymbol{\gamma})$. Here

$$l(\boldsymbol{\phi}) = \sum_{i=1}^{n} log\left(P\left(r_i|\boldsymbol{x_i}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\phi}\right)\right)$$

and

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} r_i\left(log\left(\left(P(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{iobs}, \boldsymbol{t_i}; \boldsymbol{\beta})\right)\right) + \log\left(P\left(z_{i1}|\boldsymbol{z}_{i,obs}; \boldsymbol{\gamma}\right)\right)\right)$$
$$+ (1 - r_i)\left(log\left(\int P\left(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\beta}\right)P\left(z_{i1}|\boldsymbol{z}_{i,obs}; \boldsymbol{\gamma}\right) dz_{i1}\right)\right)$$

Here $\boldsymbol{\beta}$ represents the parameters associated with the multi-state model and both $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are nuisance parameters whose estimations are not the main interest. Assume $\boldsymbol{\phi}$ and $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are distinct, thus we have $l \propto l(\boldsymbol{\beta}, \boldsymbol{\gamma})$. The likelihood $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ cannot be

calculated directly since it involves integrations which have no closed forms. We propose to approximate the log likelihood using Monte Carlo simulation method.

First, simulate H independent random variables from the standard normal distribution, $z_{i1,1}, z_{i1,2}, \cdots, z_{i1,H}$. Then calculate $z^*_{i1,r} = \mu_0 + \boldsymbol{\varphi}^T \boldsymbol{z}_{i,obs} + z_{i1,r}\sigma$, thus $z^*_{i1,r}$ follows a normal distribution with mean $\mu = \mu_0 + \boldsymbol{\varphi}^T \boldsymbol{z}_{i,obs}$ and standard deviation σ. Thus, we have

$$\int P(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{i,obs}, \boldsymbol{t_i}; \boldsymbol{\beta}) P(z_{i1}|\boldsymbol{z}_{i,obs}; \boldsymbol{\gamma})\, dz_{i1} \approx \frac{1}{H}\sum_{r=1}^{H} P(\boldsymbol{x_i}|z^*_{i1,r}, \boldsymbol{z}_{io}, \boldsymbol{t_i}; \boldsymbol{\beta})$$

Define the simulated log-likelihood function as

$$l^H(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} r_i \left(log\left(\left(P(\boldsymbol{x_i}|z_{i1}, \boldsymbol{z}_{iobs}, \boldsymbol{t_i}; \boldsymbol{\beta}) \right) \right) + \log\left(P(z_{i1}|\boldsymbol{z}_{i,obs}; \boldsymbol{\gamma}) \right) \right)$$
$$+ (1 - r_i) \left(log\left(\frac{1}{H}\sum_{r=1}^{H} P(\boldsymbol{x_i}|z^*_{i1,r}, \boldsymbol{z}_{io}, \boldsymbol{t_i}; \boldsymbol{\beta}) \right) \right)$$

We have $l^H(\boldsymbol{\beta}, \boldsymbol{\gamma}) \xrightarrow{a.s.} l(\boldsymbol{\beta}, \boldsymbol{\gamma})\ as\ H \to \infty$ by the law of large numbers when the normal distribution assumption of z_{i1} is correct. Estimates of the parameters can be achieved by maximizing the above simulated likelihood function $l^H(\boldsymbol{\beta}, \boldsymbol{\gamma})$ instead of the true log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by a numerical method such as the Newton-Raphson method.

3.3 Simulations

There are two purposes of the simulation study in this section. First, we will examine the sensitivity of the MLEs to the violations of the Normal distribution

assumption on the missing covariate. Second, we will compare the proposed method (referred to as MSL throughout) to the widely used complete case (CC) method and the method of mean substitution (MS). The criteria used for the comparison are the percent bias and the standard errors of the model parameter estimates. Estimations from the original complete datasets (without setting the covariate missing, marked as FULL in the tables) are also generated for comparisons.

Datasets are generated from the following true model

$$\boldsymbol{Q} = \begin{Bmatrix} \alpha_{11} & exp(\beta_{12,0}+, \beta_{12,1}Z_1) & exp(\beta_{13,0}) \\ 0 & \alpha_{22} & exp(\beta_{23,0} + \beta_{23,1}Z_1) \\ 0 & 0 & 0 \end{Bmatrix}.$$

Here, Z_1 is a continuous covariate and will be set to be missing for some subjects. The true value of $(\beta_{12,0}, \beta_{13,0}, \beta_{23,0}, \beta_{12,1}, \beta_{23,1})$ are set to $(-2.8, -2.7, -3, 0.8, -0.6)$ correspondingly. Observations are taken annually. The death times are recorded exactly but the state just before death is unknown. The observations will be right censored at time 25.

The values Z_1 are generated from one of the following distributions:

(1) Standard normal distribution: $Normal(1,1)$.

(2) Uniform distribution with lower bound -1.5 and upper bound 1.5:

$Uniform\ (-1.5,1.5)$.

(3) Weibull distribution with shape parameter 1.5 and scale parameter 1: $Weibull(1.5,1)$.

In case (1), the normal assumption on the distribution form of the covariate is correct. In case (2), the normal assumption on the covariate is violated but true distribution is still symmetric. In case (3), the normal assumption on the covariate is violated and true distribution is not symmetric.

Both MCAR and MAR datasets are studied in this simulation study. In the MCAR datasets, we randomly set about 30% or 60% subjects with missing Z_1 value. In the MAR datasets, we set Z_1 missing according to the time of its first transition. In Case (1), we set the covariate missing if the first transition happens within the first 5 years. In this setting, the datasets contain about 30% subjects with missing covariate values. In Case (2), we set the covariate missing if the first transition happens in 15 years. In this setting, the datasets contain about 60% subjects with missing covariate values

Simulations were set to have 500 iterations for each combination of the above mentioned cases, and each dataset contains 700 subjects. The sample size was chosen to correspond to the number of subjects we have in the MAPWU dataset. All the calculations were done using PROC IML in SAS 9.3 system [45]. We set $H = 50$ for the maximum simulated likelihood method. The results are listed in *Table 3.1* and *Table 3.2*.

From the results listed in *Table 3.1*, we can see that when the data is MCAR all three methods work relatively well. The CC method provides relatively larger standard errors for parameters estimates than the MS method and the proposed MSL method. The MS method is relatively sensitive to the datasets, and it has around 5% bias in certain cases. The proposed MSL method has the best performance among the three, the estimates provided by the MSL method is close to these we get from the original full dataset.

When the data is MAR, both the CC method and MS method have large biases. In the results listed in *Table 3.2*, the percentage bias (% Bias) can be as large as 90% for the CC method and 27% for the MS method. The proposed MSL method still performs well in our study.

The simulation studies also show that the estimations are relatively robust to the violation of the normal assumption on the missing covariate. The largest percentage bias in our simulation results was 6.2 % when the true covariate values were generated from a $Uniform\ (-1.5,1.5)$ distribution and the data is MAR with about 60% subjects with the missing covariate.

3.4 Application to the MAPWU Data from the SMART project

The Statistical Models of Aging and Risk of Transition (SMART) project is a multi-center dementia study. This project aggregates data from 11 mature, extremely data-rich, and well-known longitudinal cohorts of older adults with high autopsy rates [5]. In section, we apply the proposed method to the MAPWU data from the SMART project database. The MAPWU enrolls healthy volunteers from the community; exclusion criteria include existing neurological disorders (e.g., Parkinson's, Huntington's, or Alzheimer's disease) and psychiatric disorders (e.g., schizophrenia, substance abuse), as well as any active medical condition or treatment that impairs cognitive function [5, 48]. Cognitive function status of participants is assessed annually. Subjects included in the current study (n=732) were all cognitively normal at baseline and assessed at least two times.

Annual cognitive assessments are administered to each participant and used to classify them into either Cognitively Normal or Cognitive Impairment (including mild cognitive impairment and dementia). The state of death is also included in the model as a competing risk. A cognitively normal person would transition to either Cognitive Impairment state or die without Cognitive Impairment. A cognitive impairment person could recover to cognitively normal or die with cognitive impairment. *Figure 3.1* shows the model transition structure.

Each subject was scheduled to have cognitive assessments annually, while the true time intervals between two consecutive assessments vary among different patients and across time. The time intervals vary from 0.27 years to 12.08 years with mean ± SD 1.25 ± 0.68. The number of total longitudinal observations of each patient ranges from 2 to 31 with an average of 6.3 ± 4.7 observations.

Between two consecutive assessments, subjects in the Cognitive Normal state at prior assessment may remain at Cognitive Normal state or transition to Cognitive Impairment state at the next assessment. Subjects at Cognitive Impairment state at the prior assessment may remain at Cognitive Impairment state or reverse back to Cognitive Normal state at the next assessment. All subjects may die at any time with or without Cognitive Impairment. The exact death time could be retrieved, while the cognitive state just before death is unknown. *Table 3.3* lists the transition frequencies among these 3 states.

Our main goal was to study the effects of possible risk factors on the transitions among the three states. We focus on 4 risk factors in this study, namely baseline age (Bage), gender (Female), education level (Educ) and baseline BMI. Baseline age and BMI are continuous variables, and Female and Educ are binary variables. Educ is defined as less than 16 years of education, which is about below the college level. Bage, Female

and Educ are all completely observed. The values of Baseline BMI are missing for 521 (71.2%) participants. *Table 3.4* presents summary statistics for these baseline risk factors.

Figure 3.2 presents the histogram of the observed BMI values. The shape of this plot shows that the distribution of baseline BMI has a normal bell shape. Exploratory analyses based on the complete data show baseline BMI is independent with Female and Educ. Thus we assumes BMI follows a normal distribution with mean $\mu = \mu_0 + \varphi_1 Bage$ and standard deviation σ.

Since all participants were at least 60 years old at baseline, we centered the original baseline age at 60 in the model. We only considered patients with BMI>25 as overweight, so we defined a new variable $BMI^* = \max\{(BMI - 25)/5\,, 0\}$ as the risk factor. Effects of risk factors on the backward transition from cognitive impairment to cognitive normal was not of interest in this study, thus we have the following transition intensity functions for the multi-state model:

$$\alpha_{lm}(Bage, Female, Educ, BMI^*) = \begin{cases} exp(\beta_{lm,0}) & if\ l = 2, m = 1 \\ exp(\beta_{lm,0} + \beta_{lm,1} Bage + \beta_{lm,2} Female + \beta_{lm,3} Educ + \beta_{lm,4} BMI^*) & if\ m > l \end{cases}.$$

Table 3.5 lists the hazard ratio and the corresponding 95% confidence interval of each risk factor on each transition path using the proposed MSL method, the MS method and the CC method. The proposed MSL method and the MS method generated similar

results for the risk factors Bage, Female and Educ. Both methods show that female gender and education level have no significant effects on the transitions among these states. And both methods show baseline age is significant on all three transitions, which makes sense intuitively. Old people are likely to have higher risk of both cognitive impairment and death. The results of the two methods differ for BMI. The MSL method shows BMI is a significant risk factor of death for cognitively normal, and prevents cognitively normal people from transitioning to cognitive impairment. This finding indicates that the risk of death related to high body mass index (BMI) competes with the risk of progressing to cognitive impairment. The MS method failed to show BMI has significant effects on any of three possible transitions. The CC method showed less conclusive results. It only showed the significant effect of baseline age on transitions from Cognitively Normal to Cognitive Impairment, while failing to identify other significant effects.

3.5 Discussion

Multi-state models have been widely used in recent years, and several software packages have been available to fit various versions of multi-state models. Jackson[27] developed the R package “msm” for time homogenous and piece-wise time homogenous model. Meira-Machado et al. [4] developed an easy to use R library, call “tdc.msm”. The package fits 5 different multi-state models, including time homogenous Markov model

(THMM), non-homogenous Markov model (NHM), Cox Markov model (CMM), etc. Wu et al. [49]developed a SAS macro program for a non-homogeneous three-state progressive Markov multi-state model. Missing covariate data has been an issue in practice. No efficient methodology has been proposed to address the problem in this area so far. The maximum simulated likelihood method proposed in this manuscript provide a solution for data with missing continuous covariate data.

The proposed method requires H draws for each subject with missing covariate in construction of the simulated likelihood. In practice, the number of draws H has to be dependent on the sample size of the dataset, the proportion of subjects with incomplete covariate measurements, and also the complexity of the multi-state model being used. The way we select the number of draws in our study is that we try a sequence of numbers of H and see how the estimate convergences.

In this dissertation, we assume that the continuous partially missing covariate follows a normal distribution. Our simulation study showed that the proposed method is relatively robust to the violation of this assumption. Additional studies have to be conducted if we want to apply the method to a dataset in which the missing continuous covariate has a density much different from the normal distribution. In situations where the missing covariate does not have a normal shaped density, the method can be easily

modified by drawing values from the corresponding distribution. Again, we need to check the validity of the covariate distribution assumption.

The method is limited to data with univariate missing pattern, in which there's only one covariate with missing value. In a case where the data has a general missing data pattern with mixture of discrete and continuous covariates, we need to generate random values from the joint distribution of the covariates. Future works would focus on the construction of the joint distribution of covariates and the method that could draw random values from the joint distribution efficiently.

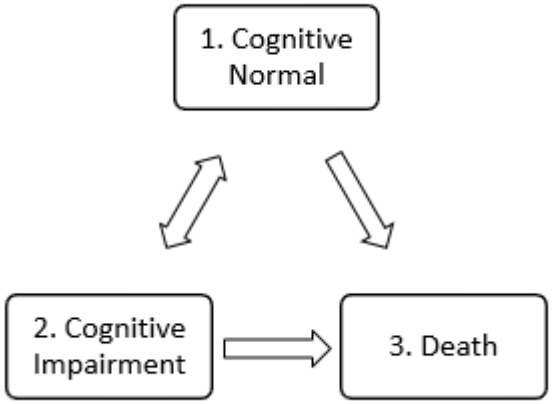

Figure 3.1: Three state model with backward transition

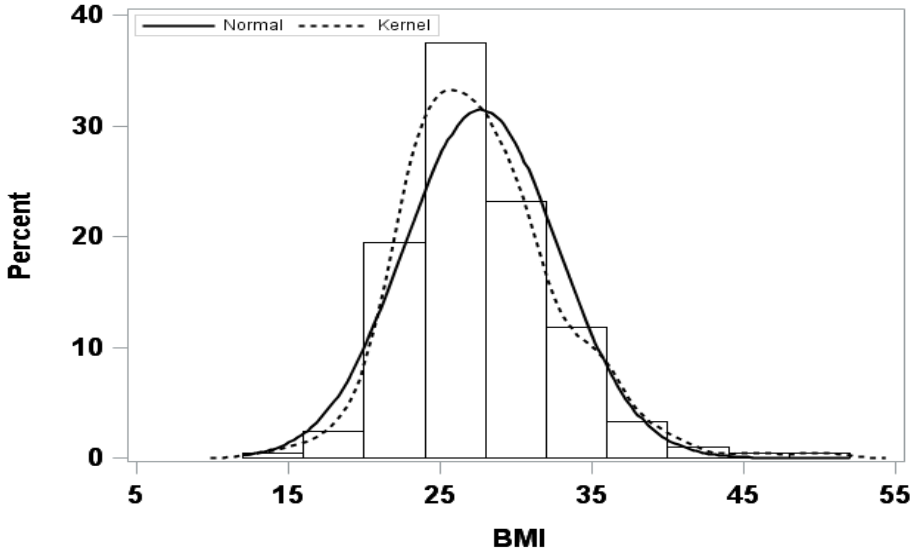

Figure 3.2: Histogram of the observed values of BMI from the MAPWU cohort with density curves. Solid curve is the fitted normal density curve and the dotted line is the kernel estimate

Table 3.1: Percent Bias and Standard Error of the model parameters for missing completely at random (MCAR) data

Distribution	Missing %	Method	$\beta_{12,0}$(-2.8)		$\beta_{13,0}$(-2.7)		$\beta_{23,0}$(-3.0)		$\beta_{12,1}$(0.8)		$\beta_{23,1}$(-0.6)	
			% Bias	SE	% Bias	SE	% Bias	SE	% Bias	SE	% Bias	SE
Normal	0	FULL	0.05%	0.07	0.04%	0.07	-0.05%	0.10	0.37%	0.07	0.17%	0.11
(0, 1)	Low	MSL	-0.10%	0.08	0.08%	0.07	0.30%	0.10	-0.02%	0.08	0.50%	0.12
		MS	-0.44%	0.07	0.04%	0.07	2.57%	0.10	0.33%	0.08	-0.41%	0.13
		CC	-0.23%	0.09	-0.03%	0.08	-0.24%	0.11	0.26%	0.09	2.10%	0.13
	High	MSL	-0.01%	0.09	0.07%	0.07	0.60%	0.11	-0.74%	0.10	0.23%	0.15
		MS	-0.91%	0.07	0.02%	0.07	5.34%	0.10	-0.13%	0.11	-0.63%	0.18
		CC	-0.06%	0.12	0.18%	0.10	-0.10%	0.15	1.10%	0.12	4.00%	0.17
Uniform	0	FULL	0.05%	0.07	-0.19%	0.07	0.37%	0.10	0.78%	0.08	0.54%	0.11
(-1.5, 1.5)	Low	MSL	0.01%	0.08	0.35%	0.07	0.21%	0.10	0.99%	0.10	0.07%	0.14
		MS	-0.33%	0.07	-0.02%	0.07	2.24%	0.09	-0.48%	0.10	-2.07%	0.14
		CC	-0.03%	0.09	-0.05%	0.08	0.08%	0.11	-0.40%	0.10	0.92%	0.14
	High	MSL	0.23%	0.08	0.17%	0.07	0.98%	0.11	2.90%	0.12	-0.54%	0.18
		MS	-0.66%	0.07	-0.13%	0.06	4.49%	0.10	-0.82%	0.13	-2.17%	0.19
		CC	-0.03%	0.12	0.00%	0.10	-0.18%	0.15	0.82%	0.13	2.10%	0.18
Weibull	0	FULL	0.12%	0.11	0.27%	0.08	-0.11%	0.16	0.52%	0.09	1.00%	0.16
(1.5, 1)	Low	MSL	-0.15%	0.12	0.56%	0.08	0.29%	0.19	-0.31%	0.10	-0.39%	0.18
		MS	0.74%	0.12	0.17%	0.08	0.44%	0.19	1.26%	0.11	0.50%	0.19
		CC	0.12%	0.13	-0.24%	0.09	-0.24%	0.19	0.36%	0.11	1.40%	0.19
	High	MSL	-0.06%	0.14	0.71%	0.08	0.12%	0.23	0.52%	0.13	0.25%	0.23
		MS	0.98%	0.15	-0.17%	0.08	0.06%	0.24	1.40%	0.14	5.55%	0.25
		CC	0.34%	0.17	-0.07%	0.12	-0.43%	0.26	1.50%	0.14	2.20%	0.25

Note: % Bias=Percent Bias. SE=standard error. Missing %: All subjects with covariate measures (0); about 30% subjects with missing covariate (Low); about 60% subjects with missing covariates (High);
Method: FULL=Full data analysis; MSL=maximum simulated likelihood; MS=method of mean substitution; CC=complete case method

Table 3.2: Percent Bias and Standard Error of the model parameters for missing at random (MAR) data

Distribution	Missing %	Method	$\beta_{12,0}$(-2.8)		$\beta_{13,0}$(-2.7)		$\beta_{23,0}$(-3.0)		$\beta_{12,1}$(0.8)		$\beta_{23,1}$(-0.6)	
			% Bias	SE	% Bias	SE	% Bias	SE	% Bias	SE	% Bias	SE
Normal	0	FULL	0.05%	0.07	0.04%	0.07	-0.05%	0.10	0.37%	0.07	0.17%	0.11
(0, 1)	Low	MSL	0.06%	0.07	-0.01%	0.07	0.00%	0.10	1.20%	0.08	1.50%	0.11
		MS	0.28%	0.07	0.07%	0.07	0.04%	0.10	5.96%	0.08	5.43%	0.11
		CC	-0.78%	0.08	23.00%	0.09	3.30%	0.10	4.90%	0.08	4.10%	0.11
	High	MSL	0.27%	0.08	-0.05%	0.07	0.25%	0.10	1.20%	0.09	3.40%	0.14
		MS	3.85%	0.08	0.24%	0.07	-0.86%	0.10	18.16%	0.09	7.59%	0.11
		CC	-2.10%	0.09	63.00%	0.20	23.00%	0.15	12.00%	0.09	15.00%	0.17
Uniform	0	FULL	0.05%	0.07	-0.19%	0.07	0.37%	0.10	0.78%	0.08	0.54%	0.11
(-1.5, 1.5)	Low	MSL	-0.21%	0.07	0.45%	0.07	-0.09%	0.09	1.00%	0.09	2.50%	0.12
		MS	0.17%	0.07	-0.15%	0.06	-0.28%	0.09	6.51%	0.09	3.21%	0.11
		CC	-1.00%	0.08	23.00%	0.09	3.30%	0.10	3.90%	0.09	2.60%	0.12
	High	MSL	0.46%	0.08	0.42%	0.07	-0.16%	0.10	6.60%	0.11	6.20%	0.15
		MS	3.65%	0.08	0.00%	0.06	-0.90%	0.10	23.91%	0.11	9.74%	0.12
		CC	-2.30%	0.09	63.00%	0.20	23.00%	0.15	15.00%	0.10	15.00%	0.18
Weibull	0	FULL	0.12%	0.11	0.27%	0.08	-0.11%	0.16	0.52%	0.09	1.00%	0.16
(1.5, 1)	Low	MSL	-0.49%	0.11	0.35%	0.08	0.05%	0.17	-1.10%	0.09	0.45%	0.17
		MS	1.49%	0.11	-0.01%	0.08	0.09%	0.16	5.48%	0.09	0.25%	0.15
		CC	-1.10%	0.11	32.00%	0.13	2.50%	0.17	4.00%	0.09	6.00%	0.16
	High	MSL	-1.40%	0.12	0.40%	0.08	0.72%	0.22	-4.20%	0.11	1.00%	0.24
		MS	3.29%	0.13	-0.07%	0.08	4.91%	0.17	3.66%	0.11	-27.27%	0.15
		CC	-6.10%	0.12	93.00%	0.37	22.00%	0.25	-0.40%	0.10	17.00%	0.24

Note: % Bias=Percent Bias. SE=standard error. Missing %: All subjects with covariate measures (0); about 30% subjects with missing covariate (Low); about 60% subjects with missing covariates (High);
Method: FULL=Full data analysis; MSL=maximum simulated likelihood; MS=method of mean substitution; CC=complete case method

Table 3.3: Numbers of transitions between each path at successive clinic visits

From	To		
	Cognitive Normal	Cognitive Impairment	Death
Cognitive Normal	3214	325	199
Cognitive Impairment	117	349	138

Table 3.4: Summary statistics of the baseline risk factors (n=732)

	N obs	Mean (Std) or %	N miss	Missing %
Baseline Age	732	76.07 (8.6)	.	.
BMI	211	27.68(5.07)	521	71.2
Female	732	60.66%	.	.
Educ	732	51.09%	.	.

Note: N obs=number of observed data, N miss=number of missing data

Table 3.5: Hazard ratios of the four risk factors by each transition path

Risk Factor	Path	MSL			MS			CC		
		HR	95% L	95% U	HR	95% L	95% U	HR	95% L	95% U
Baseline Age	C.N. to C.I.	1.06	1.04	1.08	1.07	1.06	1.08	1.14	1.07	1.21
	C.N. to Death	1.05	1.01	1.09	1.03	1.00	1.08	1.02	0.86	1.2
	C.I. to Death	1.05	1.03	1.07	1.05	1.03	1.07	0.98	0.83	1.15
Female	C.N. to C.I.	1.13	0.89	1.44	1.14	0.90	1.44	1.72	0.77	3.81
	C.N. to Death	0.64	0.34	1.20	0.62	0.31	1.23	1.50E-03	6.20E-10	3.62E+03
	C.I. to Death	1.04	0.74	1.47	1.05	0.74	1.50	1.29E+02	4.03E-03	4.14E+06
Educ	C.N. to C.I.	1.13	0.89	1.43	1.10	0.88	1.37	0.59	0.26	1.31
	C.N. to Death	1.54	0.79	3.03	1.74	0.85	3.56	5.65	0.57	56.26
	C.I. to Death	0.94	0.68	1.30	0.90	0.65	1.25	0.75	0.11	5.11
BMI*	C.N. to C.I.	0.53	0.34	0.83	0.71	0.47	1.08	0.65	0.35	1.23
	C.N. to Death	1.70	1.11	2.60	1.06	0.52	2.18	3.02	0.99	9.26
	C.I. to Death	1.18	0.75	1.87	1.23	0.30	5.10	0.27	0.02	3.44

Note: 95% L (U) =Low (Up) bound of the 95% confidence interval. $BMI^* = \max\{(BMI - 25)/5\,, 0\}$.

Chapter 4 Estimation of Multi-State Models with Missing Covariates by EM algorithm

4.1 Introduction

Multi-state models have been widely used to analyze longitudinal event history data obtained in medical and epidemiology studies. The tools and methods developed recently in this area require the dataset to be complete. However, missing covariates data is very common in practice, and it has been an issue in applications. In the last two chapters, we discussed how to deal with univariate discrete or continuous missing covariate data. In this chapter, we propose an Expectation-Maximization (EM) algorithm when applying multi-state models to datasets containing multiple missing categorical covariates. The missing data are allowed to have the general missing pattern [30]. Our simulation studies and real data application showed that the proposed EM algorithm performs well for both MCAR and MAR data.

The remainder of this chapter is organized as follows. In Section 2 we describe the EM method in detail. In Section 3, simulation studies are carried out to compare the performance of the proposed method with the widely used CC method. We applied our method to the Klamath Exceptional Aging Project (KEAP) cohort in Section 4. In the concluding section, we discuss the advantages and limitations of the EM Algorithm.

4.2 The Method

To detail the method, we will begin with time-homogenous multi-state model with compete covariates data. Then we will emphasize the EM algorithm in the case of

multiple missing categorical covariates. At the end of this section, we discuss the asymptotic variance-covariance matrix estimation.

4.2.1 Time-homogenous Multi-State Model

Suppose we have a baseline covariate vector $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_p)$, a time-homogenous multi-state model with proportional intensities has the following form:

$$\alpha_{lm}(\boldsymbol{Z}|\boldsymbol{\beta}) = \alpha_{lm,0}\, exp(\boldsymbol{\beta}_{\boldsymbol{lm}}^T \boldsymbol{Z}) = exp(\beta_{lm,0} + \boldsymbol{\beta}_{\boldsymbol{lm}}^T \boldsymbol{Z}).$$

Here $\alpha_{lm,0} = exp(\beta_{lm,0})$ is called the baseline intensity for the transition from state l to state m, and $\boldsymbol{\beta} = (\beta_{lm,0}, \boldsymbol{\beta}_{lm}; l = 1, \cdots, K;\ m = 1, \cdots, K; m \neq l)$, which represents all the parameters associated with the multi-state model.

4.2.2 Joint Modeling of the Response Data and the Partially Missing Covariates

Denote $\boldsymbol{T} = (T_1, T_2, \dots, T_M)$ and $\boldsymbol{X} = (X_1, X_2, \dots, X_M)$, here M is a random variable indicating the number of observations, T_j is the time of jth observation and X_j is the corresponding state of the process $X(t)$ at time T_j. Assume the observation process is ignorable [20], which means the observation time points T_j is determined by a process that is independent of the response $X(t)$. Thus, we view the observation time points $\boldsymbol{T}$ as fixed. Rearrange the vector of covariates such that we have $\boldsymbol{Z} = (\boldsymbol{Z_{pm}}, \boldsymbol{Z_{cc}})$. Here $\boldsymbol{Z_{pm}}$ is the components whose values might be partially missing for some subjects, and $\boldsymbol{Z_{cc}}$ is the observed components whose values are recorded for all subjects.

We assume that the covariates are MAR. If a covariate is MAR, it means that the probability of observing this covariate (conditional on the response and the other observed covariates) does not depend on the underlying value of the covariate, but may

depend on the response and the other observed covariates. Rubin [41] showed that if the data is MAR, likelihood-based inferences do not depend on the missing data mechanism. We also assume that each component of $\boldsymbol{Z_{pm}}$ is discrete. However, the completely observed covariates $\boldsymbol{Z_{cc}}$ are allowed to be a mixture of both continuous and discrete variables.

To deal with the missing covariates, we will view the partially missing covariates $\boldsymbol{Z_{pm}}$ as random variables. The likelihood will be based on the conditional joint distribution of $(\boldsymbol{X}, \boldsymbol{Z_{pm}})$ given $(\boldsymbol{T}, \boldsymbol{Z_{cc}})$, which can be modeled as

$$P(\boldsymbol{X}, \boldsymbol{Z_{pm}}|\boldsymbol{T}, \boldsymbol{Z_{cc}}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = P(\boldsymbol{X}|\boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{\beta})P(\boldsymbol{Z_{pm}}|\boldsymbol{Z_{cc}}, \boldsymbol{\gamma})$$

Here $\boldsymbol{\beta}$ is the vector of parameters associated with the multi-state model and $\boldsymbol{\gamma}$ is a vector of the nuisance parameters associated with the distribution of the partially missing covariates.

4.2.3 Likelihood under Interval-censored Data and Missing Covariates

Let $t_{i,j}$ be the time point of jth observation and $x_{i,j}$ be the corresponding observed state at $t_{i,j}$ point for subject i.Write $\boldsymbol{t_i} = (t_{i,1}, t_{i,2}, \dots, t_{i,m_i})$, and $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \dots, x_{i,m_i})$. Since we only make observations at a finite number of distinct time points, the transition times are interval-censored. Usually, the exact death time can be retrieved but the state just before death is unknown. We define $\delta_i = 1$ if the last observed state is death and $\delta_i = 0$ otherwise. Write $\boldsymbol{z_i} = (\boldsymbol{z_{i,pm}}, \boldsymbol{z_{i,cc}})$ and $\boldsymbol{z_{i,pm}} = (\boldsymbol{z_{i,mis}}, \boldsymbol{z_{i,obs}})$, where $\boldsymbol{z_{i,mis}}$ and $\boldsymbol{z_{i,obs}}$ are missing and observed components of $\boldsymbol{z_{i,pm}}$ respectively. If the data is MAR, then the likelihood for subject i based on the observed data of $\boldsymbol{x_i}$, $\boldsymbol{z_{i,cc}}$

and $\boldsymbol{z_{i.obs}}$ can be written as

$$L_i(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{x_i}, \boldsymbol{z_{i,cc}}, \boldsymbol{z_{i,obs}}) = \sum_{\boldsymbol{z_{i,mis}}} P(\boldsymbol{x_i}|\boldsymbol{z_i}, \boldsymbol{\beta}) P(\boldsymbol{z_{i,pm}}|\boldsymbol{z_{i,cc}}, \boldsymbol{\gamma}),$$

where the summation is over all possible underlying values of $\boldsymbol{z_{i,mis}}$.

Under the Markov assumption, we have

$$P(\boldsymbol{x_i}|\boldsymbol{z_i}, \boldsymbol{\beta}) = P(x_{i,1}|t_{i,1}) \prod_{j=2}^{m_i} P(x_{i,j}|x_{i,j-1}, t_{i,j-1}, t_{i,j}, \boldsymbol{z_i}, \boldsymbol{\beta}).$$

where, $P(x_{i,1}|t_{i,1})$ is the distribution for the baseline state and the transition probabilities have the following form

$P(x_{i,j}|x_{i,j-1}, t_{i,j-1}, t_{i,j}, \boldsymbol{z_i}, \boldsymbol{\beta}) =$

$$\begin{cases} p_{x_{i,j-1},x_{i,j}}(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta}) & \text{if } j \neq m_i \\ \left[p_{x_{i,j-1},x_{i,j}}(t_{ij-1}, t_{ij}|\boldsymbol{z_i}, \boldsymbol{\beta})\right]^{1-\delta_i} \left[\sum_{k \neq K} p_{x_{i,j-1},k}(t_{i,j}, t_{i,j}|\boldsymbol{z_i}, \boldsymbol{\beta}) \alpha_{kK}(\boldsymbol{z_i}|\boldsymbol{\beta})\right]^{\delta_i} & \text{if } j = m_i \end{cases}$$

Since all components of $\boldsymbol{z_{i,pm}}$ are discrete, we propose a multinomial logit model for $P(\boldsymbol{z_{i,pm}}|\boldsymbol{z_{i,cc}}, \boldsymbol{\gamma})$. Suppose $\boldsymbol{z_{i,pm}}$ has two components $(z_{i,p1}, z_{i,p2})$. Assume $z_{i,p1}$ has A possible categories and $z_{i,p2}$ has B possible categories, we have

$$log\left(\frac{\pi_{ab}}{\pi_{00}}\right) = \gamma_{ab,0} + \boldsymbol{\gamma}_{ab}^T \boldsymbol{z_{i,cc}},$$

$$a \in \{0,1,\dots,A-1\}, b \in \{0,1,\dots,B-1\}, (a,b) \neq (0,0)$$

Here

$$\pi_{ab} = P(\boldsymbol{z_{i,pm}} = (a,b)|\boldsymbol{z_{i,cc}}, \boldsymbol{\gamma}), \sum_{a=0}^{A-1} \sum_{b=0}^{B-1} \pi_{ab} = 1$$

and

$$\boldsymbol{\gamma} = \left(\gamma_{ab,0}, \boldsymbol{\gamma_{ab}}; a \in \{0,1,\dots,A-1\}, b \in \{0,1,\dots,B-1\} \text{ and } (a,b) \neq (0,0)\right).$$

The idea can be easily generalized to model $\boldsymbol{z_{i,pm}}$ with more than two components.

The log likelihood for all subjects is

$$l(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} log\left(L_i(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{x_i},\boldsymbol{z_{i,cc}},\boldsymbol{z_{i,obs}})\right).$$

In most cases, the log likelihood function $l(\boldsymbol{\beta},\boldsymbol{\gamma})$ does not have a closed form, so directly maximizing the log likelihood is not straightforward, especially when nuisance parameter $\boldsymbol{\gamma}$ has a high dimension. To facilitate the estimation, we propose using an EM algorithm to obtain MLEs.

4.2.4 The EM algorithm.

The E-step of EM would be

$$\begin{aligned} Q_i\left(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right) &= E\left[l_i(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{x_i},\boldsymbol{z_i})|\boldsymbol{x_i},\boldsymbol{z_{i,obs}},\boldsymbol{z_{i,cc}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right] \\ &= \sum_{\boldsymbol{z_{i,mis}}} l_i(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{x_i},\boldsymbol{z_i}) P\left(\boldsymbol{z_{i,mis}}\middle|\boldsymbol{x_i},\boldsymbol{z_{i,obs}},\boldsymbol{z_{i,cc}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right). \end{aligned}$$

Here $l_i(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{x_i},\boldsymbol{z_i})$ is the complete data log likelihood for subject i and it has the form:

$$l_i(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{x_i},\boldsymbol{z_i}) = log\left(P(\boldsymbol{x_i}|\boldsymbol{z_i},\boldsymbol{\beta})P\left(\boldsymbol{z_{i,pm}}\middle|\boldsymbol{z_{i,cc}},\boldsymbol{\gamma}\right)\right) = l_{\boldsymbol{x_i}|\boldsymbol{z_i}}(\boldsymbol{\beta}) + l_{\boldsymbol{z_i}|\boldsymbol{z_{i,cc}}}(\boldsymbol{\gamma})$$

Denote $w_{i,(s)} = P\left(\boldsymbol{z_{i,mis}}\middle|\boldsymbol{x_i},\boldsymbol{z_{i,obs}},\boldsymbol{z_{i,cc}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$, then we have

$$\begin{aligned} Q\left(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right) &= \sum_{i}^{n} Q_i\left(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right) \\ &= \sum_{i=1}^{n}\sum_{\boldsymbol{z_{i,mis}}} w_{i,(s)}\, l_{\boldsymbol{x_i}|\boldsymbol{z_i}}(\boldsymbol{\beta}) + \sum_{i=1}^{n}\sum_{\boldsymbol{z_{i,mis}}} w_{i,(s)}\, l_{\boldsymbol{z_i}|\boldsymbol{z_{i,cc}}}(\boldsymbol{\gamma}) \end{aligned}$$

Write

$$Q_{x|z}\left(\boldsymbol{\beta}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)=\sum_{i}\sum_{\boldsymbol{z_{i,mis}}} w_{i,(s)}\, l_{\boldsymbol{x_i}|\boldsymbol{z_i}}(\boldsymbol{\beta})$$

and

$$Q_{z}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)=\sum_{i}\sum_{\boldsymbol{z_{i,mis}}} w_{i,(s)}\, l_{\boldsymbol{z_i}|\boldsymbol{z_{i,cc}}}(\boldsymbol{\gamma}),$$

now the "Q function" presented above is separated into two parts. Note that $Q_{x|z}\left(\boldsymbol{\beta}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ is a function of the multi-state model parameter $\boldsymbol{\beta}$ and does not contain the nuisance parameter $\boldsymbol{\gamma}$. And $\mathrm{Q}_{z}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ is a function of only the nuisance parameter $\boldsymbol{\gamma}$. The calculation of weights $w_{i,(s)}$ can be done by using Bayes theorem. For subject i, we have

$$w_{i,(s)}=P\left(\boldsymbol{z_{i,mis}}\middle|\boldsymbol{x_i},\boldsymbol{z_{i,obs}},\boldsymbol{z_{i,cc}},\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)=\frac{P(\boldsymbol{x_i}|\boldsymbol{z_i},\boldsymbol{\beta}^{(s)})*P(\boldsymbol{z_{i,pm}}|\boldsymbol{z_{i,cc}},\boldsymbol{\gamma}^{(s)})}{\sum_{\boldsymbol{z_{i,mis}}}P(\boldsymbol{x_i}|\boldsymbol{z_i},\boldsymbol{\beta}^{(s)})*P(\boldsymbol{z_{i,pm}}|\boldsymbol{z_{i,cc}},\boldsymbol{\gamma}^{(s)})}.$$

For the M-step, maximization of the function $\mathrm{Q}\left(\boldsymbol{\beta},\boldsymbol{\gamma}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ can be achieved by maximizing $Q_{x|z}\left(\boldsymbol{\beta}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ and $Q_{z}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ separately. The maximization of $\mathrm{Q_z}\left(\boldsymbol{\gamma}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ and $Q_{x|z}\left(\boldsymbol{\beta}\middle|\boldsymbol{\beta}^{(s)},\boldsymbol{\gamma}^{(s)}\right)$ do not have closed forms. We use the Newton-Raphson method for the maximization.

4.2.5 Asymptotic Variance-Covariance Matrix Estimation

The EM algorithm does not provide the estimates of asymptotic variances as its byproduct. Here we obtain the variance estimates by finding the observed information matrix. One way to derive the observed information matrix is to take the derivatives directly from the observed data log likelihood. Note that the observed data log likelihood has the form

$$l(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} log\left(L_i\left(\boldsymbol{\theta}, \boldsymbol{\gamma} | \boldsymbol{x_i}, \boldsymbol{z_{i,cc}}, \boldsymbol{z_{i,obs}}\right)\right)$$

and it does not have a closed form. We use the forward difference method to compute the Hessian matrix of $l(\boldsymbol{\theta}, \boldsymbol{\gamma})$. Suppose $f(\boldsymbol{x})$ is a real function and twice differentiable regarding to a *p*-dimensional vector $\boldsymbol{x}$, then the Hessian matrix of $f(\boldsymbol{x})$ at $\boldsymbol{x_0}$ can be approximated as follow:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}|_{x=x_0} = \frac{f\left(\boldsymbol{x_0} + h_i \boldsymbol{e_i} + h_j \boldsymbol{e_j}\right) - f(\boldsymbol{x_0} + h_i \boldsymbol{e_i}) - f\left(\boldsymbol{x_0} + h_j \boldsymbol{e_j}\right) + f(\boldsymbol{x_0})}{h_i h_j}$$

Here $\boldsymbol{e_i}$ is the ith coordinate vector, a vector with its ith component equal to 1 and all others equal to 0, and h_i is the step size.

4.3 Simulations

In this section, we study the performance of the proposed EM method through simulation studies. Datasets were generated from an "illness-death" model with backward transitions. Four cases are considered here:

1) MCAR data with about 45% of subjects having missing covariates.
2) MCAR data with about 70% of subjects having missing covariates.
3) MAR data with about 45% of subjects having missing covariates.
4) MAR data with about 70% of subjects having missing covariates.

To see the advantages of the proposed method (EM), we compare it to the widely used CC method as well as to the full data analysis (FULL). In the full data analysis, all covariate values are preserved and there is no missing data. It is a benchmark for evaluating our method's performance in general. The comparisons were made through

percent bias (% Bias), empirical standard error (ESE), asymptotic standard error (SE) and 95% confidence interval coverage probability (95% CP). The total number of subjects in each simulated dataset is 500 or 1000. All results are based on 500 simulation datasets, and calculations are made by using PROC IML in SAS 9.3® [45].

4.3.1 Generating the Dataset

Datasets were generated from an "illness-death" model with backward transitions. The model has the following transition intensity matrix:

$$Q(Z_1, Z_2) = \begin{pmatrix} \alpha_{11} & exp(\beta_{12,0} + \beta_{12,1}Z_1) & exp(\beta_{13,0} + \beta_{13,2}Z_2) \\ exp(\beta_{21,0}) & \alpha_{22} & exp(\beta_{23,0} + \beta_{23,1}Z_1) \\ 0 & 0 & 0 \end{pmatrix}.$$

Here

$$\alpha_{11} = -\big(exp(\beta_{12,0} + \beta_{12,1}Z_1) + exp(\beta_{13,0} + \beta_{13,2}Z_2)\big)$$

and

$$\alpha_{22} = -\big(\exp(\beta_{21,0}) + exp(\beta_{23,0} + \beta_{23,1}Z_1)\big).$$

There are 7 parameters , $(\beta_{12,0}, \beta_{13,0}, \beta_{21,0}, \beta_{23,0}, \beta_{12,1}, \beta_{23,1}, \beta_{13,2})$, in this model. The first four parameters measure the baseline transition intensities, and the last three parameters measure the effects of covariates on the transition intensities. We set the true values of these parameters to be $(2, -2.5, -3, -2, 0.5, 0.3, 0.4)$ respectively. Covariates (Z_1, Z_2) are both binary variables with the following joint mass function：

$$P(z_1, z_2) = \begin{cases} 0.1 & if\ z_1 = 0, z_2 = 0 \\ 0.4 & if\ z_1 = 0, z_2 = 1 \\ 0.3 & if\ z_1 = 1, z_2 = 0 \\ 0.2 & if\ z_1 = 1, z_2 = 1 \end{cases}.$$

Observations of the process are taken annually. State 3 means death in this model, thus the transition time into state 3 is recorded exactly, while state just before death is unknown. A common censoring time of 25 years is used, which results in right censored transition time for those who remain in state 1 or state 2 at that time. The covariates are baseline covariates; their values do not change over time.

4.3.2 Estimations with MCAR data

We study two MCAR data cases. In the first case, we randomly set covariate Z_1 and Z_2 missing. Z_1 is missing with probability 0.2 and Z_2 is missing with probability 0.3. Thus, there are about 6% of subjects with both Z_1 and Z_2 missing, and about 45% of subjects with at least one covariate missing. We denote this type of missing data MCAR 1. In the second case, we set Z_1 missing with probability 0.4 and Z_2 missing with probability 0.5. In these data, there are about 70% subjects with at least one covariate value unobserved. We denote this type of data MCAR 2. Note that the data are MCAR data, since the probability of the covariates being missing is independent of both the observed data and the underlying values of missing covariates.

The results for the two MCAR cases are presented in *Table 4.1*. When the data are MCAR, it shows that both CC method and the proposed EM method work well. The percent bias (% Bias) is relatively small. The 95% confidence interval coverage probabilities (95% CP) hover around 95% and the estimated asymptotic standard errors are close to the empirical standard error for all 7 parameters. Moreover, the results are close to those provided by the FULL data analysis. We note that the proposed EM method is more sufficient than the CC method since estimates provided by the EM method have smaller standard errors than those provided by the CC method.

4.3.3 Estimations with MAR data

In the first case of MAR data, we set covariate Z_1 missing if the first transition happens in the first year, and set covariate Z_2 missing if the second transition happens after year 8. Thus, approximately 45% of subjects have at least one covariate missing. In the second case, we set covariate Z_1 missing if the first transition happens in the first 3 years, and set covariate Z_2 missing if the second transition happens after year 6. This results in a dataset with approximately 70% subjects having at least one covariate missing. Both of these two types of data are MAR but not MCAR, since the missingness of covariates is dependent on the observed data but independent of the underlying values of the missing data.

The results for the MAR data are presented in *Table 4.2*. When the data are MAR but not MCAR, in general the CC method fails. The bias of the estimates provided by the CC method can be very large, as large as 105% in our simulation studies. And the 95% confidence interval coverage probability (95% CP) could be far away from 95%. The proposed EM method still works well in both of these MAR cases.

4.4 Application

In this application, we used the KEAP cohort from the SMART database. KEAP is a population-based study of the oldest-old residents of the Klamath Basin, which is a rural area of Oregon. Subjects enrolled in this study were at least 80 years old. Subjects are visited in their homes by a geriatric research nurse every six months for neuropsychiatric testing and structured clinical interview [50].

The cognitive functions of each patient were classified into the following two states: No Dementia (ND) and Dementia. A third state Death is also added to the model to account the competing risk for Dementia. State ND includes both normal cognition and mild cognitive impairment (MCI). Subjects who were in state ND at the previous visit may die before the next scheduled assessment or may transition to Dementia at the next scheduled assessment. Subjects who were in state Dementia may die before the next assessment. There are no backward transitions from state Dementia to state ND. *Figure 4.1* presents the model state structure.

The dataset contains 419 subjects. At baseline, there were 351 (83.8%) subjects in state ND and 68 (16.2%) subjects in state Dementia. The number of observations for each subject ranges from 2 to 22 with mean$\pm$SD of 7.7 ± 4.9. Cognitive assessments are administered to each subject with mean time interval between consecutive assessments 0.58 ± 0.46 years.

Covariates to be examined as potential risk factors for transitions among the 3 states in *Figure 4.1* are: baseline age (Bage), female gender (Female), low education (LowEdu; defined as high school or less), APOE4 (with or without an ε4 allele), and baseline high blood pressure (HBP). Bage, Female and LowEdu are all fully observed. APOE4 and HBP are missing for some subjects. There are 80 (19%) subjects with missing APOE4 and 206 (49%) subjects with missing HBP. 34 (8%) subjects have both APOE4 and BHP missing, and 252 (60%) subjects have at least one of these missing. *Table 4.3* lists a summary of these 4 risk factors.

Table 4.4 lists the observed transition frequency and row percentage for the original 419 subjects as well as the 167 subjects with complete data. In the original data, 4.1% ND subjects transitioned to Dementia state, and 6.9% ND subjects died without Dementia, 26.8% Dementia subjects died at the end of the study. In contrast for the CC data, only 3.2% of ND subjects developed Dementia and 3.4% of them died without Dementia. Among Dementia subjects, 19.8% died at the end of the study.

We used a time-homogenous Cox Markov model to investigate the effects of potential risk factors on these transitions, which have the following forms:

$$\alpha_{lm}(\boldsymbol{Z}) = \begin{cases} exp\left(\beta_{lm,0} + \beta_{lm}^{T}\boldsymbol{Z}\right), if\ l \neq m \\ -\sum_{h \neq l} \alpha_{lh}(\boldsymbol{Z}), \qquad if\ l = m \end{cases}$$

.Here $\boldsymbol{Z} = (Bage, Female, LowEdu, APOE4, HBP)$. Bage were centered at 80 in the model.

First, we conducted an available-case (AC) analysis, in which we dropped the two covariates with missing values, APOE4 and HBP, out of the model. Without APOE4 and HBP in the model, the data is fully observed, thus standard estimation methods, for example the "msm" R-package [27], could be used to fit the model without dropping any subjects. Then, we fitted the data with the proposed EM method after adding APOE4 and HBP in the model. At last, we also conducted a CC analysis for the model.

Table 4.5 lists the hazard ratios (HR) and the corresponding 95% confidence intervals (95% CI) for risk factors on each transition path by these three method. The results obtained from the AC method show that the baseline age would increase the hazard ratio for all three paths, from ND to Dementia and Death, and from Dementia to

Death. Its effect on paths from ND to Dementia and from Dementia to Death are significant (P value<0.05). The effects of Female gender and low education level on all three paths are not significant. The estimated effects of baseline age, female gender and low education level by applying the proposed EM method are close to these of the AC analysis. And we also find out that APOE4 has significant effect (P Value<0.05) of increasing the HR for transition path from ND to Dementia. High blood pressure has significant effects on paths from ND to Dementia and to Death. Subjects at ND with baseline high blood pressure have lower hazard ratio of developing dementia, while they have higher mortality rate than those without high blood pressure at baseline. Comparing to the results of the AC analysis and the proposed EM method, the CC method is less efficient. The lengths of 95% confidence intervals of the hazard ratio are generally larger than those provided by the AC method and the EM method. Also, the CC method failed to indicate the significant effects of baseline age on transition path from Dementia to Death and of APOE4 on path from ND to Dementia. The significant effect on baseline age on path from Dementia to Death was indicated by both the AC method and the proposed EM method. And the effect of APOE4 on path from ND to Dementia was well studied in literature [10, 51, 52].

4.5 Discussion

Multi-state models are useful tools to analyze longitudinal event history data, and have been widely applied in medical studies. Missing covariates in data has been an issue in practice. Most of the currently available methods and software packages use the CC method in the case of missing covariates data. The problem associated with the CC method is that it will provide biased estimates if the data are not MCAR. Even if the data

are MCAR, dropping all cases with missing covariates is inefficient and might cause convergence problems in particular applications. In contrast, the proposed EM method worked well for both MCAR and MAR data. In the case of MCAR data, the proposed EM method was also more efficient than the CC method.

Standard multiple imputation methods are also very difficult to carry out in the estimation of multi-state model in analyzing longitudinally collected event history data with multiple missing baseline categorical covariates data. Constructing an appropriate imputation model for the missing categorical covariates data is difficult when the observed data contains longitudinal response data with random lengths and unequal spaces.

Likelihood-based methods are common approaches to the analysis of missing data [30]. The observed data likelihood contains both model parameters and nuisance parameters used to model the distribution of missing covariates. When data have the general missing pattern with multiple missing data covariates, we would need a relatively large number of nuisance parameters to model the missing covariates, and directly maximizing the observed data likelihood is difficult. By using the EM algorithm, we were able to separate the model parameters and the nuisance parameters and make the maximization of the likelihood possible.

EM algorithms have been used to deal with missing data problems in other areas. Ibrahim [34] provided an EM algorithm for generalized linear model with incomplete covariates measurements. Lin et al. [28] also used EM algorithm in the Cox regression model with missing covariates data. Applying EM algorithm in the multi-state model framework is more difficult than in the above mentioned areas. First, closed formulas

usually do not exist for both the expectation step and the maximization step when applying EM algorithm to multi-state models with missing covariates data. Numerical methods have to be used to get the expectation log-likelihood function and to maximize of the expected complete data log likelihood. Another issue encountered with missing covariates data in the multi-state model is the usually large number of unknown parameters. In applying maximum likelihood type methods to deal with missing covariates data, a probability model with nuisance parameters has to be constructed for the missing covariates. Plus, each covariate would have different coefficient parameters on different transition paths. These two factors lead to a relatively larger numbers of parameters compared to other situations. Our proposed EM algorithm would enable researchers to estimate the nuisance parameters and model covariates coefficients separately.

Our study showed that the proposed EM algorithm is efficient in dealing with missing covariates data with a general missing pattern. However, this method is limited to only missing discrete covariates. In the case of the missing continuous covariate, the Q function in the E-step cannot be written as a weighted sum of the complete data log likelihood. One possible solution to deal with missing continuous covariates in multi-state models is to approximate the E-step using Gaussian quadrature or Monte Carlo integration techniques.

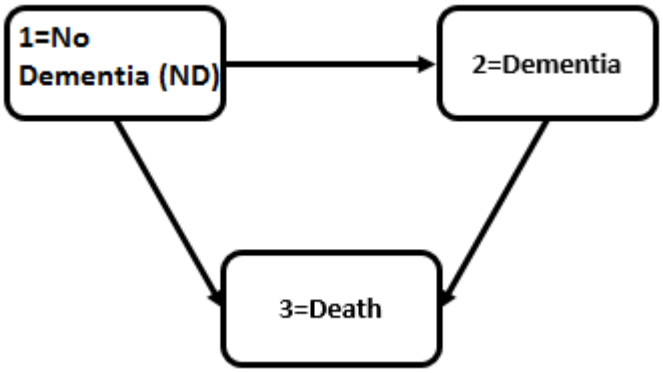

Figure 4.1: Transition flow diagram for the model

Table 4.1: Simulation results for MCAR data

			CC				EM				FULL			
Type	N	Parameters	%Bias	ESE	SE	95% CP	%Bias	ESE	SE	95% CP	%Bias	ESE	SE	95% CP
MCAR 1	500	$\beta_{12,0}$	0.15%	0.11	0.11	96%	-0.07%	0.08	0.09	97%	0.01%	0.08	0.09	96%
		$\beta_{13,0}$	0.62%	0.17	0.18	96%	1.00%	0.15	0.15	95%	0.74%	0.14	0.13	94%
		$\beta_{21,0}$	-0.10%	0.17	0.17	96%	-0.36%	0.13	0.13	95%	-0.35%	0.13	0.13	95%
		$\beta_{23,0}$	0.14%	0.15	0.14	95%	-0.15%	0.10	0.11	95%	0.08%	0.10	0.10	96%
		$\beta_{12,1}$	-0.61%	0.15	0.15	97%	-1.10%	0.13	0.12	96%	-0.71%	0.12	0.11	95%
		$\beta_{23,1}$	-1.20%	0.18	0.18	94%	-2.80%	0.14	0.14	96%	-0.02%	0.13	0.13	95%
		$\beta_{13,2}$	3.60%	0.23	0.21	93%	4.60%	0.19	0.19	95%	1.60%	0.17	0.16	93%
	1000	$\beta_{12,0}$	0.62%	0.08	0.08	94%	0.25%	0.07	0.06	93%	0.22%	0.06	0.06	94%
		$\beta_{13,0}$	-0.32%	0.14	0.13	93%	-0.48%	0.11	0.10	95%	-0.30%	0.10	0.09	94%
		$\beta_{21,0}$	0.57%	0.13	0.12	96%	0.35%	0.09	0.09	96%	0.35%	0.09	0.09	96%
		$\beta_{23,0}$	0.14%	0.10	0.10	94%	0.04%	0.08	0.08	93%	0.21%	0.07	0.07	97%
		$\beta_{12,1}$	3.30%	0.11	0.11	96%	1.90%	0.09	0.09	94%	1.60%	0.08	0.08	96%
		$\beta_{23,1}$	-0.43%	0.13	0.13	96%	-1.00%	0.11	0.10	95%	0.93%	0.09	0.09	95%
		$\beta_{13,2}$	-2.70%	0.16	0.15	95%	-4.60%	0.14	0.13	93%	-2.80%	0.12	0.11	94%
MCAR 2	500	$\beta_{12,0}$	0.36%	0.16	0.16	96%	0.03%	0.09	0.09	97%	0.01%	0.08	0.09	96%
		$\beta_{13,0}$	0.47%	0.26	0.25	95%	0.22%	0.18	0.17	93%	0.74%	0.14	0.13	94%
		$\beta_{21,0}$	0.91%	0.25	0.23	95%	-0.36%	0.13	0.13	95%	-0.35%	0.13	0.13	95%
		$\beta_{23,0}$	0.41%	0.19	0.19	96%	0.05%	0.11	0.11	95%	0.08%	0.10	0.10	96%
		$\beta_{12,1}$	0.99%	0.21	0.21	92%	-0.47%	0.14	0.14	93%	-0.71%	0.12	0.11	95%
		$\beta_{23,1}$	-0.54%	0.25	0.25	94%	-0.52%	0.16	0.16	94%	-0.02%	0.13	0.13	95%
		$\beta_{13,2}$	-1.30%	0.33	0.30	95%	-3.50%	0.24	0.21	93%	1.60%	0.17	0.16	93%
	1000	$\beta_{12,0}$	0.21%	0.10	0.11	98%	0.16%	0.07	0.07	93%	0.22%	0.06	0.06	94%
		$\beta_{13,0}$	-0.35%	0.17	0.17	96%	-0.37%	0.11	0.11	95%	-0.30%	0.10	0.09	94%
		$\beta_{21,0}$	0.47%	0.17	0.16	93%	0.35%	0.09	0.09	96%	0.35%	0.09	0.09	96%
		$\beta_{23,0}$	-0.34%	0.14	0.13	93%	0.19%	0.08	0.08	94%	0.21%	0.07	0.07	97%
		$\beta_{12,1}$	3.30%	0.15	0.15	94%	1.20%	0.10	0.10	95%	1.60%	0.08	0.08	96%
		$\beta_{23,1}$	-1.60%	0.18	0.17	94%	0.91%	0.12	0.12	95%	0.93%	0.09	0.09	95%
		$\beta_{13,2}$	-3.10%	0.21	0.20	94%	-3.30%	0.15	0.15	94%	-2.80%	0.12	0.11	94%

Note: ESE (empirical standard error) SE (estimated standard error) 95%CP (Coverage probability of the 95% confidence interval)

Table 4.2: Simulation results for MAR data

			CC				EM				FULL			
Type	N	Parameters	%Bias	ESE	SE	95% CP	%Bias	ESE	SE	95% CP	%Bias	ESE	SE	95% CP
MAR 1	500	$\beta_{12,0}$	22%	0.17	0.15	17%	0.20%	0.09	0.09	96%	0.01%	0.08	0.09	96%
		$\beta_{13,0}$	-13%	0.14	0.16	49%	0.69%	0.17	0.16	96%	0.74%	0.14	0.13	94%
		$\beta_{21,0}$	-23%	0.20	0.21	11%	-0.36%	0.13	0.13	95%	-0.35%	0.13	0.13	95%
		$\beta_{23,0}$	-39%	0.20	0.18	1.90%	0.06%	0.11	0.11	94%	0.08%	0.10	0.10	96%
		$\beta_{12,1}$	46%	0.20	0.19	80%	0.68%	0.13	0.13	95%	-0.71%	0.12	0.11	95%
		$\beta_{23,1}$	-54%	0.23	0.22	88%	-0.29%	0.15	0.15	95%	-0.02%	0.13	0.13	95%
		$\beta_{13,2}$	-32%	0.16	0.19	95%	1.20%	0.23	0.2	92%	1.60%	0.17	0.16	93%
	1000	$\beta_{12,0}$	23%	0.12	0.11	1.50%	0.28%	0.07	0.06	94%	0.22%	0.06	0.06	94%
		$\beta_{13,0}$	-13%	0.10	0.11	15%	-0.52%	0.12	0.11	95%	-0.30%	0.10	0.09	94%
		$\beta_{21,0}$	-23%	0.13	0.15	1.00%	0.35%	0.09	0.09	96%	0.35%	0.09	0.09	96%
		$\beta_{23,0}$	-40%	0.13	0.13	0.00%	0.07%	0.08	0.08	95%	0.21%	0.07	0.07	97%
		$\beta_{12,1}$	46%	0.14	0.14	61%	1.90%	0.09	0.09	96%	1.60%	0.08	0.08	96%
		$\beta_{23,1}$	-60%	0.17	0.16	75%	-0.71%	0.12	0.11	91%	0.93%	0.09	0.09	95%
		$\beta_{13,2}$	-36%	0.11	0.13	86%	-4.90%	0.15	0.14	95%	-2.80%	0.12	0.11	94%
MAR 2	500	$\beta_{12,0}$	92%	0.74	0.6	0.49%	0.05%	0.11	0.11	94%	0.01%	0.08	0.09	96%
		$\beta_{13,0}$	-14%	0.15	0.21	65%	0.59%	0.18	0.17	94%	0.74%	0.14	0.13	94%
		$\beta_{21,0}$	-6.40%	5.58	1.63	44%	-0.36%	0.13	0.13	95%	-0.35%	0.13	0.13	95%
		$\beta_{23,0}$	-72%	1.20	0.75	23%	-0.12%	0.13	0.12	98%	0.08%	0.10	0.10	96%
		$\beta_{12,1}$	114%	0.78	0.72	88%	-0.80%	0.16	0.16	95%	-0.71%	0.12	0.11	95%
		$\beta_{23,1}$	-69%	1.27	0.87	79%	-0.96%	0.19	0.19	93%	-0.02%	0.13	0.13	95%
		$\beta_{13,2}$	-70%	0.16	0.25	88%	0.85%	0.25	0.22	93%	1.60%	0.17	0.16	93%
	1000	$\beta_{12,0}$	93%	0.32	0.29	0.00%	0.40%	0.08	0.08	94%	0.22%	0.06	0.06	94%
		$\beta_{13,0}$	-15%	0.10	0.14	24%	-0.61%	0.12	0.12	94%	-0.30%	0.10	0.09	94%
		$\beta_{21,0}$	-36%	0.40	0.41	25%	0.34%	0.09	0.09	96%	0.35%	0.09	0.09	96%
		$\beta_{23,0}$	-64%	0.51	0.34	9.50%	-0.05%	0.09	0.09	94%	0.21%	0.07	0.07	97%
		$\beta_{12,1}$	105%	0.41	0.38	72%	2.40%	0.11	0.11	96%	1.60%	0.08	0.08	96%
		$\beta_{23,1}$	-52%	0.64	0.42	75%	-1.80%	0.15	0.13	91%	0.93%	0.09	0.09	95%
		$\beta_{13,2}$	-73%	0.11	0.17	61%	-5.60%	0.16	0.16	94%	-2.80%	0.12	0.11	94%

Note: ESE (empirical standard error) SE (estimated standard error) 95%CP (Coverage probability of the 95% confidence interval

Table 4.3: Summary statistics of the risk factors

Baseline Risk Factor	N Missing (%)	Mean (st. dev.) or percent
Baseline age	0 (0)	88.46 (4.01)
Female	0 (0)	66.11
Low Education (years <=12)	0 (0)	62.05
APOE4	80 (19.09)	15.99
High blood pressure	206 (49.16)	35.08

Table 4.4: Observed transition frequency (row %) for the original data

From	To					
	Original Data (N=419)			Complete Case (N=167)		
	ND	Dementia	Death	ND	Dementia	Death
ND	2037 (89.0%)	94 (4.1%)	157 (6.9%)	1401 (93.4%)	48 (3.2%)	51 (3.4%)
Dementia		378 (73.2%)	138 (26.8%)		166 (80.2%)	41 (19.8%)

Note: ND=No Dementia

Table 4.5: Hazard Ratio of each risk factor on each path by three methods

Risk factor	Path	AC method				EM method				CC method			
		HR	95 % CI			HR	95 % CI			HR	95 % CI		
			L	R	Length		L	R	Length		L	R	Length
Baseline	ND to Dementia	1.13	1.07	1.20	0.13	1.15	1.08	1.22	0.13	1.14	1.03	1.26	0.23
Age	ND to Death	1.07	0.99	1.15	0.15	1.05	0.98	1.13	0.15	0.95	0.82	1.11	0.29
	Dementia to Death	1.05	1.01	1.08	0.07	1.05	1.01	1.08	0.07	0.98	0.89	1.07	0.18
Female	ND to Dementia	1.11	0.72	1.71	1.00	1.37	0.88	2.15	1.27	0.88	0.47	1.64	1.17
	ND to Death	0.94	0.62	1.43	0.81	0.86	0.58	1.29	0.72	1.39	0.59	3.23	2.64
	Dementia to Death	0.93	0.65	1.34	0.69	0.94	0.65	1.36	0.70	0.64	0.30	1.36	1.06
Low Education	ND to Dementia	1.25	0.81	1.92	1.11	1.34	0.87	2.05	1.18	1.73	0.93	3.23	2.30
	ND to Death	0.80	0.53	1.19	0.66	0.75	0.51	1.12	0.61	0.55	0.26	1.14	0.88
	Dementia to Death	1.13	0.80	1.61	0.81	1.14	0.80	1.63	0.83	2.53	1.02	6.30	5.29
APOE4	ND to Dementia					1.84	1.15	2.96	1.82	1.44	0.74	2.83	2.10
	ND to Death					0.76	0.39	1.49	1.10	1.24	0.49	3.16	2.67
	Dementia to Death					1.07	0.73	1.57	0.84	0.67	0.31	1.48	1.17
High blood pressure	ND to Dementia					0.40	0.24	0.67	0.43	0.60	0.33	1.08	0.75
	ND to Death					3.27	1.11	9.63	8.52	4.24	1.06	16.92	15.86
	Dementia to Death					1.24	0.82	1.87	1.05	1.85	0.95	3.59	2.65

Note: L=lower bound of the 95% CI, U= upper bound of the 95% CI, Length=U-L

Chapter 5 Discussions and Future Research

In this dissertation, we proposed several methods for dealing with different types of missing covariates data in the framework of multi-state models. The methods discussed in the previous several chapters are all likelihood-based methods. The likelihood-based methods are one of most frequently used methods in the literature of missing data. Another popular method is imputation-based methods. However, the imputation-based methods are not feasible in the framework of multi-state model. The observed response data for most multi-state model analysis is longitudinal data with random length, thus it is a difficult to come up with an appropriate imputation model for the missing covariates conditioned on the observed covariates and the observed response longitudinal data. Our study also showed that the multiple imputation method conditioning only on the observed covariates is biased

Because of the unique features of multi-state models, calculating and maximizing the log likelihood function with missing covariates data directly would be very complicated and difficult. However, in cases that the multi-state models have relative simple state structures and the data has a univariate missing pattern with only one missing discrete covariate, the direct MLE method becomes feasible.

In situations where the data has a continuous missing covariate, the MSL would be an alternative for estimation. The MSL method replaces the true log likelihood function with a simulated one, thus we avoiding the integration in the calculation of the log likelihood, which often does not have a closed form for the integrand. One limitation

of the method is that a parametric model or a distribution form of the missing covariate has to be provided, which might be difficult to do in some applications.

When the data has a general missing pattern but the all missing covariates are discrete variables, the EM method would be a good choice for estimation of the multi-state model. The EM algorithm would allow us to estimate the model parameters associated with the multi-state model and the nuisance parameters associated with the covariates data separately in each iteration. And estimation of the nuisance parameters would be easier since there are closed form estimates in some circumstances. The limitation of the method lie in the slow convergence in some situations.

Despite the constraints and limitations of the methods we proposed in the previous several chapters, we are satisfied with the results to date. The method we proposed could help us deal with most cases of missing covariates data problems in practice. However, there are some potential extensions of future research in this area.

Throughout the methods discussed in this dissertation, we assume the data is either MCAR or MAR. Relaxing this assumption would help deal with a more general type of missing data, the NMAR data. Possible solution for this topic is to provide an appropriate model of missing mechanism for the data. For the EM algorithm, one could use more advanced maximization algorithms to speed up the convergence of estimation, and make this method feasible for more complicated multi-state models. More robustness studies on the assumption of the distribution form of the missing continuous covariate might be needed for more complicated models when using the MSL method for estimation. Another area of possible future research lies in mixture of missing continuous and discrete covariate variables. The likelihood-base method could still be feasible for the

mixture continuous and discrete missing covariates data, while more advanced methods have to be used to calculate and maximize the corresponding log likelihood function.

Appendices

A. SAS/IML modules for the Observed Data Likelihood Method

```
PROC IML; RESET storage= D.OLapp;
/*************************************************************************/
/*A.1: Calculate the transition probability p(xi|zi) for subject i*/
/*************************************************************************/
START ProbSubj(Z) global(Xi,GNZ,GNstate,Gmsmparms,Gdeath);
  n=nrow(Xi); logQ=J(GNstate,GNstate,0);
  do nc=1 to GNZ; logQ=logQ+Gmsmparms[,Gnstate*(nc-1)+1:Gnstate*nc]*Z[nc]; end;
  Q=exp(logQ); Q[loc(logQ=0)]=0;
  do ns=1 to GNstate; Q[ns,ns]=-Q[ns,+]; end;
  A=teigvec(Q); V=teigval(Q); L=1.0;
  do j=2 to n;
    from=Xi[j-1,2]; to=Xi[j,2]; ft=Xi[j-1,3]; tt=Xi[j,3]; timelag=tt-ft;
    D=diag(exp(V[,1]#timelag)); P=A*D*inv(A);
    if to=Gdeath then do;
      pdj=0.0; do h=1 to GNstate; if h^=Gdeath then pdj=pdj+P[from,h]*Q[h,Gdeath];end;
      L=L*pdj;
    end;
    else L=L*P[from,to];
  end; Return(L);
FINISH ProbSubj;

/*************************************************************************/
/*A.2: Calculate the Observed Data Log-likelihood*/
/*************************************************************************/
START llObs(parms) global(Gmsmdata, GNSubj, Xi, GZmis, GZful, GZmisCovs,
GNZmis, GNZful, GNZ, GNstate, Gmsmparms, Gdeath, GNp, Gnpmsm, Gnpz, Gprob);
  pmsm=parms[1:Gnpmsm]; Gmsmparms[loc(Gmsmparms^=0)]=pmsm;
  gamma=parms[Gnpmsm+1:Gnp]; logL=0.;
  do  id=1 to GNSubj;
    h=loc(Gmsmdata[,1]=id); Xi=Gmsmdata[h,1:3]; Zmis=Gmsmdata[h[1],GZmis];
    Zful=Gmsmdata[h[1],GZful]; ZmisCovs=Gmsmdata[h[1],GZmisCovs];
    exp1=exp(ZmisCovs*gamma); p1=exp1/(1+exp1);
    if Zmis=0 then Li=ProbSubj(Zful||0)*(1-p1);
    else if Zmis=1 then Li=ProbSubj(Zful||1)*p1;
    else Li=ProbSubj(Zful||0)*(1-p1)+ProbSubj(Zful||1)*p1;
    if Li>0 then logL=logL+log(Li); else Gprob=1+Gprob;
  end; return(logL);
FINISH llObs;

/*************************************************************************/
/*A.3: Estimation using Observed Data Log-likelihood*/
/*************************************************************************/
```

```
START Est_OL_App(Nsubj,parmsM0,gamma0,Zful,Zmis,ZmisCovs, death=0)
global(Gmsmdata,GNSubj,Xi,GZmis,GZful,GZmisCovs,GNZmis,GNZful,GNZ,GNstate,
Gmsmparms,Gdeath,GNp,Gnpmsm,Gnpz,Gprob);
  GNsubj=Nsubj;GZmis=Zmis;GZful=Zful; GZmisCovs=ZmisCovs;
  GNZmis=ncol(Zmis);GNZful=ncol(Zful); GNZ=GNZmis+GNZful;
  GNstate=nrow(parmsM0); Gmsmparms=parmsM0; Gdeath=death;  Gprob=0;
  parms_crt=parmsM0[loc(parmsM0^=0)]`; h0=parms_crt||gamma0; GNp=ncol(h0);
  Gnpmsm=ncol(parms_crt); optn={1 0 1 3};
  call NLPnra(rc,xres,"llObs",h0,optn); estimate=xres` ;
  call nlpfdd(f,g,hes,"llObs",estimate); cov=-ginv(hes); stderr=sqrt(vecdiag(cov));
  norqua=probit(1-0.05/2); low=estimate-norqua*stderr; up=estimate+norqua*stderr;
  z=abs(estimate/stderr); p=2*(1-probnorm(z));
  EstCoef=parmsM0; EstCoef[loc(parmsM0^=0)]=estimate[1:Gnpmsm];
  EstGamma=estimate[Gnpmsm+1:GNp];Pgamma=p[Gnpmsm+1:GNp];
  print EstCoef; print EstGamma pGamma;
  EstHR=parmsM0; EstHR[loc(parmsM0^=0)]=exp(estimate[1:Gnpmsm]);
  HR_low=parmsM0; HR_low[loc(parmsM0^=0)]=exp(low[1:Gnpmsm]);
  HR_up=parmsM0; HR_up[loc(parmsM0^=0)]=exp(up[1:Gnpmsm]);
  EstP=parmsM0; EstP[loc(parmsM0^=0)]=p[1:Gnpmsm]; Zs=Zful||Zmis;
  do iz=1 to GNZ;
    Covname=Zs[iz]; HRiz=EstHR[,(iz-1)*GNstate+1:iz*GNstate];
    Piz=EstP[,(iz-1)*GNstate+1:iz*GNstate];
    Lowiz=HR_low[,(iz-1)*GNstate+1:iz*GNstate];
    Upiz=HR_UP[,(iz-1)*GNstate+1:iz*GNstate];
    print Covname "From" "To" "HR" "Low" "UP" "Pvalue";
    do ir=1 to GNstate; do ic=1 to GNstate;
      HR=HRiz[ir,ic]; Low=Lowiz[ir,ic]; Up=Upiz[ir,ic]; pv=Piz[ir,ic];
      if HR^=0 then do; print ir ic HR Low Up pv; end;
    end;end;
  end; Ests=(estimate`)||(stderr`)||Gprob;
  create Results from Ests; Append from Ests; close Results;
FINISH Est_OL_App;  STORE module=_all_;
QUIT;
```

B. SAS/IML modules for the Maximum Simulated Likelihood method

```
PROC IML; RESET storage=D.SLapp ;
/*************************************************************************/
/*B.1: Calculate the transition probability p(xi|zi) for subject i*/
/*************************************************************************/
START ProbSubj(Z) global(Xi,GNZ,GNstate,Gmsmparms,Gdeath);
  n=nrow(Xi);  logQ=J(GNstate,GNstate,0);
  do nc=1 to GNZ; logQ=logQ+Gmsmparms[,Gnstate*(nc-1)+1:Gnstate*nc]*Z[nc]; end;
  Q=exp(logQ); Q[loc(Q=1)]=0; do ns=1 to GNstate; Q[ns,ns]=-Q[ns,+]; end;
  A=teigvec(Q); V=teigval(Q); L=1.0;
  do j=2 to n;
    from=Xi[j-1,2]; to=Xi[j,2]; ft=Xi[j-1,3]; tt=Xi[j,3]; timelag=tt-ft;
```

```
    D=diag(exp(V[,1]#timelag)); P=A*D*inv(A);
    if to=Gdeath then do; pdj=0.0;
      do h=1 to GNstate;
        if h^=Gdeath then pdj=pdj+P[from,h]*Q[h,Gdeath]; end; L=L*pdj;
      end;
    else L=L*P[from,to];
  end; Return(L);
FINISH ProbSubj;

/*****************************************************************/
/*B.2: Calculate the Simulated Log-likelihood*/
/*****************************************************************/
START llSL(parms) global(Gmsmdata, GNSubj, Xi, GZmis, GZful, GZmisCov,
GNZmis, GNZful, GNZ, GNstate, Gmsmparms, Gdeath, GNp, Gnpmsm, Gnpz, Gprob,
GNR);
  pmsm=parms[1:Gnpmsm]; Gmsmparms[loc(Gmsmparms^=0)]=pmsm;
  intcept=parms[Gnpmsm+1]; gamma=parms[Gnpmsm+2]; sigma=parms[Gnpmsm+3];
  logL=0.;
  do  id=1 to GNSubj;
    h=loc(Gmsmdata[,1]=id); Xi=Gmsmdata[h,1:3]; Zmis=Gmsmdata[h[1],GZmis];
    Zful=Gmsmdata[h[1],GZful]; ZmisC=Gmsmdata[h[1],GZmisCov];
    mu=intcept+gamma*ZmisC;
    if Zmis=. then do; simL=j(1,GNR,.);
      Zs = j(1,GNR,.); call randseed(id,1); call randgen(Zs,'NORMAL',mu,sigma);
      do mi=1 to GNR;
        if Zs[mi]<0 then Zsmi=0; else Zsmi=Zs[mi]; simL[mi]=ProbSubj(Zful||Zsmi);end;
        Li=simL[:]; if Li>0 then logL=logL+log(Li); else Gprob=1;
      end;
    else do;
      if Zmis<0 then Zmisabs=0; else Zmisabs=Zmis; Li=ProbSubj(Zful||Zmisabs);
      if Li>0 then logL=logL+log(Li)+logpdf('NORMAL',Zmis,mu,sigma); else Gprob=1;
    end;
  end; return(logL);
FINISH llSL;

/*****************************************************************/
/*B.3: Estimation using Simulated Log-likelihood*/
/*****************************************************************/
START EstSL(Nsubj,parmsM0,gamma0,Zmis,Zful,ZmisCov,death=0,NR=30)
global(Gmsmdata, GNSubj, Xi, GZmis, GZful, GZmisCov, GNZful, GNZ, GNstate,
Gmsmparms, Gdeath, GNp, Gnpmsm, Gnpz, Gprob, GNR);
  GNsubj=Nsubj;GZmis=Zmis;GZful=Zful; GZmisCov=ZmisCov; GNZmis=ncol(Zmis);
  GNZful=ncol(Zful); GNZ=GNZmis+GNZful; GNstate=nrow(parmsM0);
  Gmsmparms=parmsM0; Gdeath=death; Gprob=0;GNR=NR;
```

```
  parms_crt=parmsM0[loc(parmsM0^=0)]`; h0=parms_crt||gamma0; GNp=ncol(h0);
  Gnpz=ncol(gamma0); Gnpmsm=GNp-GNpz;con=J(2,Gnp,.); con[1,GNp]=0.01;
  optn={1 0 1 3}; call NLPnra(rc,xres,"llsl",h0,optn,con); estimate=xres` ;
  call nlpfdd(f,g,hes,"llsl",estimate); cov=inv(hes); stderr=sqrt(vecdiag(cov));
  norqua=probit(1-0.05/2); low=estimate-norqua*stderr; up=estimate+norqua*stderr;
  z=abs(estimate/stderr); p=2*(1-probnorm(z));
  EstHR=parmsM0; EstHR[loc(parmsM0^=0)]=exp(estimate[1:Gnpmsm]);
  HR_low=parmsM0; HR_low[loc(parmsM0^=0)]=exp(low[1:Gnpmsm]);
  HR_up=parmsM0; HR_up[loc(parmsM0^=0)]=exp(up[1:Gnpmsm]);
  EstP=parmsM0; EstP[loc(parmsM0^=0)]=p[1:Gnpmsm];
  Zs=Zful||Zmis;
  do iz=1 to GNZ;
    Covname=Zs[iz]; HRiz=EstHR[,(iz-1)*GNstate+1:iz*GNstate];
    Piz=EstP[,(iz-1)*GNstate+1:iz*GNstate];
    Lowiz=HR_low[,(iz-1)*GNstate+1:iz*GNstate];
    Upiz=HR_UP[,(iz-1)*GNstate+1:iz*GNstate];
    print Covname "From" "To" "HR" "Low" "UP" "Pvalue";
    do ir=1 to GNstate; do ic=1 to GNstate;
      HR=HRiz[ir,ic]; Low=Lowiz[ir,ic]; Up=Upiz[ir,ic]; pv=Piz[ir,ic];
      if HR^=0 then do; print ir ic HR Low Up pv; end;
    end; end;
  end; Ests=(estimate`)||(stderr`)||Gprob;
  Create Results from Ests; Append from Ests; Close Results;
FINISH EstSL;  Store module=_all_;
QUIT;
```

C. SAS/IML modules for the EM method

```
PROC IML; RESET storage=D.SLapp ;
/*************************************************************************/
/*C.1: Calculate the transition probability p(xi|zi) for subject i*/
/*************************************************************************/
START ProbSubj(Z) global(Xi,GNZ,GNstate,Gmsmparms,Gbsparms,Gdeath);
  n=nrow(Xi);  logQ=J(GNstate,GNstate,0);
  do nc=1 to GNZ; logQ=logQ+Gmsmparms[,Gnstate*(nc-1)+1:Gnstate*nc]*Z[nc]; end;
  Q=exp(logQ); Q[loc(Q=1)]=0; do ns=1 to GNstate; Q[ns,ns]=-Q[ns,+]; end;
  A=teigvec(Q); V=teigval(Q); L=1.0;
  do j=2 to n;
    from=Xi[j-1,2]; to=Xi[j,2]; ft=Xi[j-1,3]; tt=Xi[j,3]; timelag=tt-ft;
    D=diag(exp(V[,1]#timelag)); P=A*D*inv(A);
    if to=Gdeath then do; pdj=0.0;
      do h=1 to GNstate; if h^=Gdeath then pdj=pdj+P[from,h]*Q[h,Gdeath]; end;
L=L*pdj;
    end;
    else L=L*P[from,to];
  end; Return(L);
FINISH ProbSubj;
```

```
/*************************************************************************/
/*C.2: Calculate the Weights using Bayes Theorem*/
/*************************************************************************/
START MisDWeit(parms,pz) global(Gmsmdata, GNSubj, Xi, GZmis, GZful, GNZmis,
GNZful, GNZ, GNstate, Gmsmparms, Gdeath, Gnpmsm, Gnpz, Gprob);
  wet=J(GNSubj,Gnpz+1,0); Gmsmparms[loc(Gmsmparms^=0)]=parms;
Fpz=J(1,Gnpz+1,0); Fpz[1:Gnpz]=pz; Fpz[Gnpz+1]=1-Fpz[+];
  do id=1 to GNSubj;
    h=loc(Gmsmdata[,1]=id); Xi=Gmsmdata[h,1:3]; Zmis=Gmsmdata[h[1],GZmis];
    Zful=Gmsmdata[h[1],GZful]; Covs=Zmis;
    do zj=GNZmis to 1 by -1;
      if Zmis[,zj]=. then do;
        Cov0=Covs; Cov1=Covs; Cov0[,zj]=0; Cov1[,zj]=1; Covs=Cov0//Cov1;
      end;
    end;
    Nmis=nrow(Covs); *Number of mis patterns; Li=0; Weti=J(1,Gnpz+1,0);
    do j=1 to Nmis;
      Z=Covs[j,]; idxp=1; do jj=1 to GNZmis; idxp=idxp+Z[jj]*2**(GNZmis-jj); end;
      weti[idxp]=ProbSubj(Zful||Z)*Fpz[idxp];
    end;
    do wj=1 to Gnpz+1; wet[id,wj]=weti[wj]/weti[+]; end;
  end; return(wet);
FINISH MisDWeit;

/*************************************************************************/
/*C.3: Calculate the Expected Log Likelihood using Weights*/
/*************************************************************************/
START llWeit (parms) global (Gmsmdata, GNSubj, Xi, GZful, GZmis, GNZmis,
GNZful, GNZ, GNstate, Gmsmparms, Gnpmsm, Gdeath, Gnpz, Gwet, Gprob);
  logL=0.; Gmsmparms[loc(Gmsmparms^=0)]=parms;
  do id=1 to GNSubj;
    h=loc(Gmsmdata[,1]=id); Xi=Gmsmdata[h,1:3]; Zmis=Gmsmdata[h[1],GZmis];
    Zful=Gmsmdata[h[1],GZful];  Covs=Zmis;
    do zj=GNZmis to 1 by -1; if Zmis[,zj]=. then do;
       Cov0=Covs; Cov1=Covs; Cov0[,zj]=0; Cov1[,zj]=1; Covs=Cov0//Cov1;
    end; end;
    Nmis=nrow(Covs); *Number of mis patterns; Li=0;
    do j=1 to Nmis;
      Z=Covs[j,]; idxp=1; do jj=1 to GNZmis; idxp=idxp+Z[jj]*2**(GNZmis-jj); end;
      Li=ProbSubj(Zful||Z);
      if Li>0 then logL=logL+log(Li)*Gwet[id,idxp]; else Gprob=1+Gprob;
    end;
  end; return(logL);
FINISH llWeit;
```

```
/*************************************************************************/
/*C.4: Calculate the Observed Data Log Likelihood for Variance-Covariance Matrix*/
/*************************************************************************/
START llObs (parms) global(Gmsmdata, GNSubj, Xi, GZmis, GZful, GNZmis, GNZful,
GNZ, GNstate, Gmsmparms, Gdeath, GNp, Gnpmsm, Gnpz, Gprob);
  pz=j(1,GNpz+1,0); pz[1:Gnpz]=parms[Gnpmsm+1:Gnp]; pz[Gnpz+1]=1-pz[+];
  pmsm=parms[1:Gnpmsm]; Gmsmparms[loc(Gmsmparms^=0)]=pmsm; logL=0.;
  do id=1 to GNSubj;
    h=loc(Gmsmdata[,1]=id); Xi=Gmsmdata[h,1:3]; Zmis=Gmsmdata[h[1],GZmis];
    Zful=Gmsmdata[h[1],GZful]; Covs=Zmis;
    do zj=GNZmis to 1 by -1; if Zmis[,zj]=. then do;
      Cov0=Covs; Cov1=Covs; Cov0[,zj]=0; Cov1[,zj]=1; Covs=Cov0//Cov1;
    end; end;
    Nmis=nrow(Covs); *Number of mis patterns; Li=0;
    do j=1 to Nmis;
      Z=Covs[j,]; idxp=1; do jj=1 to GNZmis; idxp=idxp+Z[jj]*2**(GNZmis-jj); end;
      Li=Li+ProbSubj(Zful||Z)*pz[idxp];
    end;
    if Li>0 then logL=logL+log(Li); else Gprob=1+Gprob;
  end; return(logL);
finish llObs;

/*************************************************************************/
/*C.5: Estimation using Expectation Maximization (EM) method*/
/*************************************************************************/
START MisDEM(Nsubj,parmsM0,pz0,Zmis,Zful,death=0,Cvg=1.e-4) global(Gmsmdata,
GNSubj, Xi, GZful, GZmis, GNZmis, GNZful, GNZ, GNstate, Gmsmparms, Gnpmsm,
Gdeath, Gnpz, Gnp, Gwet, Gprob);
  Gprob=0; GNsubj=Nsubj; GZful=Zful;
GZmis=Zmis;GNZful=ncol(Zful);GNZmis=ncol(Zmis);   GNZ=GNZful+GNZmis;
GNstate=nrow(parmsM0); Gmsmparms=parmsM0; Gdeath=death;
  parms_crt=parmsM0[loc(parmsM0^=0)]`; GNpmsm=ncol(parms_crt);
  pz_crt=pz0; Gnpz=ncol(pz0); Gnp=Gnpmsm+Gnpz; Ntr=0; optn={1 0 1 3};
  do until(conv< cvg);
    Gwet=MisDWeit(parms_crt,pz_crt);
    call nlpnra(rc,curparms,"llweit",parms_crt,optn); curpz=Gwet[:,1:Gnpz];
    conv=sqrt((curparms-parms_crt)*(curparms-parms_crt)`
      +(curpz-pz_crt)*(curpz-pz_crt)`);
    parms_crt=curparms; pz_crt=curpz;
    file testfile;put Ntr;put conv;closefile testfile;
    Ntr=Ntr+1;
  end;
  estimate=curparms||curpz; call nlpfdd(f,g,hes,"llObs",estimate);
  cov=inv(hes); stderr=sqrt(vecdiag(cov)); norqua=probit(1-0.05/2);
  low=estimate`-norqua*stderr; up=estimate`+norqua*stderr;
  z=abs(estimate`/stderr); p=2*(1-probnorm(z));
```

```
  Est=parmsM0; Est[loc(parmsM0^=0)]=estimate[1:Gnpmsm];
  MStd=parmsM0; Mstd[loc(parmsM0^=0)]=stderr[1:Gnpmsm];
  MCI_low=parmsM0; MCI_low[loc(parmsM0^=0)]=low[1:Gnpmsm];
  MCI_up=parmsM0; MCI_up[loc(parmsM0^=0)]=up[1:Gnpmsm];
  Pvalue=parmsM0; Pvalue[loc(parmsM0^=0)]=p[1:Gnpmsm];
  Estp=curpz; Pp=p[Gnpmsm+1:GNpmsm+Gnpz];
  Stdp=stderr[Gnpmsm+1:GNpmsm+Gnpz];
  Lowp=low[Gnpmsm+1:GNpmsm+Gnpz];upp=up[Gnpmsm+1:GNpmsm+Gnpz];
  EstHR=parmsM0; EstHR[loc(parmsM0^=0)]=exp(estimate[1:Gnpmsm]);
  HR_low=parmsM0; HR_low[loc(parmsM0^=0)]=exp(low[1:Gnpmsm]);
  HR_up=parmsM0; HR_up[loc(parmsM0^=0)]=exp(up[1:Gnpmsm]);
  Zs=Zful||Zmis;
  do iz=1 to GNZ;
    Covname=Zs[iz]; HRiz=EstHR[,(iz-1)*GNstate+1:iz*GNstate];
    Piz=Pvalue[,(iz-1)*GNstate+1:iz*GNstate];
    Lowiz=HR_low[,(iz-1)*GNstate+1:iz*GNstate];
    Upiz=HR_UP[,(iz-1)*GNstate+1:iz*GNstate];
    print Covname "From" "To" "HR" "Low" "UP" "Pvalue";
    do ir=1 to GNstate; do ic=1 to GNstate;
      HR=HRiz[ir,ic]; Low=Lowiz[ir,ic]; Up=Upiz[ir,ic]; pv=Piz[ir,ic];
      if HR^=0 then do; print ir ic HR Low Up pv; end;
    end; end;
  end; Ests=(estimate`)||(stderr`)||Estp||stdp||Gprob;
  Create Results from Ests; Append from Ests; Close Results;
FINISH MisDEM;
```

References

1. Andersen, P.K. and N. Keiding, *Multi-state models for event history analysis.* Statistical Methods in Medical Research, 2002. **11**(2): p. 91-115.
2. Commenges, D., *Multi-state models in epidemiology.* Lifetime Data Anal, 1999. **5**(4): p. 315-27.
3. Hougaard, P., *Multi-state models: a review.* Lifetime Data Anal, 1999. **5**(3): p. 239-64.
4. Meira-Machado, L., et al., *Multi-state models for the analysis of time-to-event data.* Statistical Methods in Medical Research, 2009. **18**(2): p. 195-222.
5. Abner, E.L., et al., *The Statistical Modeling of Aging and Risk of Transition Project: Data Collection and Harmonization Across 11 Longitudinal Cohort Studies of Aging, Cognition, and Dementia.* Observational studies, 2015. **1**(2015): p. 56.
6. Putter, H., M. Fiocco, and R.B. Geskus, *Tutorial in biostatistics: Competing risks and multi-state models.* Statistics in Medicine, 2007. **26**(11): p. 2389-2430.
7. Andersen, P.K., et al., *Statistical models based on counting processes*. 2012: Springer Science & Business Media.
8. Siannis, F., V.T. Farewell, and J. Head, *A multi-state model for joint modelling of terminal and non-terminal events with application to Whitehall II.* Statistics in Medicine, 2007. **26**(2): p. 426-442.
9. Abner, E.L., et al., *Mild cognitive impairment: statistical models of transition using longitudinal clinical data.* International Journal of Alzheimer's Disease, 2012. **2012**.
10. Kryscio, R.J., et al., *Adjusting for Mortality when Identifying Risk Factors for Transitions to Mild Cognitive Impairment and Dementia.* Journal of Alzheimers Disease, 2013. **35**(4): p. 823-832.
11. Commenges, D., et al., *Incidence and mortality of Alzheimer's disease or dementia using an illness-death model.* Stat Med, 2004. **23**(2): p. 199-210.
12. Hsieh, H.J., T.H. Chen, and S.H. Chang, *Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan.* Stat Med, 2002. **21**(22): p. 3369-82.
13. Putter, H., et al., *Estimation and prediction in a multi-state model for breast cancer.* Biometrical Journal, 2006. **48**(3): p. 366-380.
14. Cadarso-Suarez, C., et al., *Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data.* Statistical Modelling, 2010. **10**(3): p. 291-314.
15. Andersen, P.K., S. Esbjerg, and T.I.A. Sorensen, *Multi-state models for bleeding episodes and mortality in liver cirrhosis.* Statistics in Medicine, 2000. **19**(4): p. 587-599.

16. Frydman, H., *A Nonparametric-Estimation Procedure for a Periodically Observed 3-State Markov Process, with Application to Aids.* Journal of the Royal Statistical Society Series B-Methodological, 1992. **54**(3): p. 853-866.

17. Joly, P. and D. Commenges, *A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS.* Biometrics, 1999. **55**(3): p. 887-890.

18. Keiding, N., J.P. Klein, and M.M. Horowitz, *Multi-state models and outcome prediction in bone marrow transplantation.* Statistics in Medicine, 2001. **20**(12): p. 1871-1885.

19. Klein, J.P. and Y.Y. Shu, *Multi-state models for bone marrow transplantation studies.* Statistical Methods in Medical Research, 2002. **11**(2): p. 117-139.

20. Commenges, D., *Inference for multi-state models from interval-censored data.* Stat Methods Med Res, 2002. **11**(2): p. 167-82.

21. Kalbfleisch, J.D. and J.F. Lawless, *The Analysis of Panel Data under a Markov Assumption.* Journal of the American Statistical Association, 1985. **80**(392): p. 863-871.

22. van den Hout, A. and F.E. Matthews, *Multi-state analysis of cognitive ability data: a piecewise-constant model and a Weibull model.* Stat Med, 2008. **27**(26): p. 5440-55.

23. Guo, W.S. and G. Marshall, *Ordmkv - a Computer-Program Fitting Proportional Odds Model for Multistate Markov Process.* Computer Methods and Programs in Biomedicine, 1995. **46**(3): p. 257-263.

24. Marshall, G., W.S. Guo, and R.H. Jones, *Markov - a Computer-Program for Multistate Markov-Models with Covariables.* Computer Methods and Programs in Biomedicine, 1995. **47**(2): p. 147-156.

25. Wu, H.M., M.F. Yen, and T.H.H. Chen, *SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression.* Computer Methods and Programs in Biomedicine, 2004. **75**(2): p. 95-105.

26. Machado, L.M., C. Cadarso-Suarez, and J. de Una-Alvarez, *tdc.msm: An R library for the analysis of multi-state survival data.* Computer Methods and Programs in Biomedicine, 2007. **86**(2): p. 131-140.

27. Jackson, C.H., *Multi-State Models for Panel Data: The msm Package for R.* Journal of Statistical Software, 2011. **38**(8): p. 1-28.

28. Lin, D.Y. and Z. Ying, *Cox Regression with Incomplete Covariate Measurements.* Journal of the American Statistical Association, 1993. **88**(424): p. 1341-1349.

29. Little, R.J., *Regression with missing X's: a review.* Journal of the American Statistical Association, 1992. **87**(420): p. 1227-1237.

30. Little, R.J. and D.B. Rubin, *Statistical analysis with missing data*. 2014: John Wiley & Sons.

31. Rubin, D.B., *Inference and Missing Data.* Biometrika, 1976. **63**(3): p. 581-590.

32. Robins, J.M., A. Rotnitzky, and L.P. Zhao, *Estimation of Regression-Coefficients When Some Regressors Are Not Always Observed.* Journal of the American Statistical Association, 1994. **89**(427): p. 846-866.

33. Vach, W. and M. Schumacher, *Logistic-Regression with Incompletely Observed Categorical Covariates - a Comparison of 3 Approaches.* Biometrika, 1993. **80**(2): p. 353-362.

34. Ibrahim, J.G., *Incomplete Data in Generalized Linear-Models.* Journal of the American Statistical Association, 1990. **85**(411): p. 765-769.

35. Horton, N.J. and N.M. Laird, *Maximum likelihood analysis of generalized linear models with missing covariates.* Statistical Methods in Medical Research, 1999. **8**(1): p. 37-50.

36. Ibrahim, J.G., et al., *Missing-data methods for generalized linear models: A comparative review.* Journal of the American Statistical Association, 2005. **100**(469): p. 332-346.

37. Prentice, R.L., *Covariate Measurement Errors and Parameter-Estimation in a Failure Time Regression-Model.* Biometrika, 1982. **69**(2): p. 331-342.

38. Paik, M.C., *Multiple imputation for the Cox proportional hazards model with missing covariates.* Lifetime Data Analysis, 1997. **3**(3): p. 289-298.

39. Buck, S.F., *A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic-Computer.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1960. **22**(2): p. 302-306.

40. Schemper, M. and T.L. Smith, *Efficient Evaluation of Treatment Effects in the Presence of Missing Covariate Values.* Statistics in Medicine, 1990. **9**(7): p. 777-784.

41. Rubin, D.B. *Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse*. in *Proceedings of the survey research methods section of the American Statistical Association*. 1978. American Statistical Association.

42. Rubin, D.B. *An overview of multiple imputation*. in *Proceedings of the survey research methods section of the American statistical association*. 1988.

43. Schafer, J.L., *Analysis of incomplete multivariate data*. 1997: CRC press.

44. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data Via Em Algorithm.* Journal of the Royal Statistical Society Series B-Methodological, 1977. **39**(1): p. 1-38.

45. Institute, S., *SAS/IML 9.3 User's Guide*. 2011: SAS Institute.

46. Katz, M.J., et al., *Age-specific and Sex-specific Prevalence and Incidence of Mild Cognitive Impairment, Dementia, and Alzheimer Dementia in Blacks and Whites A Report From the Einstein Aging Study.* Alzheimer Disease & Associated Disorders, 2012. **26**(4): p. 335-343.

47. Song, C., et al., *Multi‐stage transitional models with random effects and their application to the Einstein aging study.* Biometrical Journal, 2011. **53**(6): p. 938-955.

48. Berg, L., et al., *Mild Senile Dementia of Alzheimer Type - Research Diagnostic-Criteria, Recruitment, and Description of a Study Population.* Journal of Neurology Neurosurgery and Psychiatry, 1982. **45**(11): p. 962-968.

49. Hui-Min, W., Y. Ming-Fang, and T.H.-H. Chen, *SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression.* Computer methods and programs in biomedicine, 2004. **75**(2): p. 95-105.

50. Kaye, J., et al., *Exceptional Brain Aging in a Rural Population-Based Cohort.* Journal of Rural Health, 2009. **25**(3): p. 320-325.

51. Abner, E.L., et al., *Self-Reported Head Injury and Risk of Late-Life Impairment and AD Pathology in an AD Center Cohort.* Dementia and Geriatric Cognitive Disorders, 2014. **37**(5-6): p. 294-306.

52. Kryscio, R.J., et al., *Self-reported memory complaints Implications from a longitudinal cohort with autopsies.* Neurology, 2014. **83**(15): p. 1359-1365.

Vita

Wenjie Lou

Education

M.S. in Statistics, University of Kentucky, 2010-2012

B.S. in Statistics, Zhejiang University, 2006-2010

Employment

Teaching Assistant, Aug 2010-May 2011

Department of Statistics, University of Kentucky

Research Assistant, May 2011-Feb 2016

Sanders-Brown Center on Aging, University of Kentucky

Publications

Kryscio, R.J., Abner, E.L., Jicha, G.A., Nelson, P.T., Smith, C.D., Van Eldik, L.J., **Lou, W.,** Fardo, D.W., Cooper, G.E. and Schmitt, F.A., 2015. Self-reported memory complaints: A comparison of demented and unimpaired outcomes. Alzheimer's & Dementia, 11(7), pp.P383-P384.

Abner, E.L., Schmitt, F.A., Nelson, P.T., **Lou, W.,** Wan, L., Gauriglia, R., Dodge, H.H., Woltjer, R.L., Yu, L., Bennet, D.A. and Schneider, J.A., 2015. The Statistical Modeling of Aging and Risk of Transition Project: Data Collection and Harmonization Across 11 Longitudinal Cohort Studies of Aging, Cognition, and Dementia. Observational studies, 1(2015), p.56.