An Investigation of Methods for Missing Data in Hierarchical Models for Discrete Data

by

Muhammad Rashid Ahmed

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics – Biostatistics

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Hierarchical models are applicable to modeling data from complex surveys or longitudinal data when a clustered or multistage sample design is employed. The focus of this thesis is to investigate inference for discrete hierarchical models in the presence of missing data. This thesis is divided into two parts: in the first part, methods are developed to analyze the discrete and ordinal response data from hierarchical longitudinal studies. Several approximation methods have been developed to estimate the parameters for the fixed and random effects in the context of generalized linear models. The thesis focuses on two likelihood-based estimation procedures, the pseudo likelihood (PL) method and the adaptive Gaussian quadrature (AGQ) method.

The simulation results suggest that AGQ is preferable to PL when the goal is to estimate the variance of the random intercept in a complex hierarchical model. AGQ provides smaller biases for the estimate of the variance of the random intercept. Furthermore, it permits greater flexibility in accommodating user-defined likelihood functions.

In the second part, simulated data are used to develop a method for modeling longitudinal binary data when non-response depends on unobserved responses. This simulation study modeled three-level discrete hierarchical data with 30% and 40% missing data using a missing not at random (MNAR) missing-data mechanism. It focused on a monotone missing data-pattern. The imputation methods used in this thesis are: complete case analysis (CCA), last observation carried forward (LOCF), available case missing value (ACMVPM) restriction, complete case missing value (CCMVPM) restriction, neighboring case missing value (NCMVPM) restriction, selection model with predictive mean matching method (SMPM), and Bayesian pattern mixture model. All three restriction methods and the selection model used the predictive mean matching method

to impute missing data. Multiple imputation is used to impute the missing values. These m imputed values for each missing data produce m complete datasets. Each dataset is analyzed and the parameters are estimated. The results from the m analyses are then combined using the method of Rubin (1987), and inferences are made from these results. Our results suggest that restriction methods provide results that are superior to those of other methods. The selection model provides smaller biases than the LOCF methods but as the proportion of missing data increases the selection model is not better than LOCF. Among the three restriction methods the ACMVPM method performs best. The proposed method provides an alternative to standard selection and pattern-mixture modeling frameworks when data are not missing at random. This method is applied to data from the third Waterloo Smoking Project, a seven-year smoking prevention study having substantial non-response due to loss-to-follow-up.

Acknowledgements

I would like to acknowledge many people for helping me during my doctoral work. I would especially like to thank my advisor, Dr. Stephen Brown, for his support, encouragement, and confidence during my PhD. He provided great inspiration and gentle guidance, both of which encouraged me to pursue my interests throughout my research. I am also grateful to an extraordinary doctoral committee and wish to thank Drs. Mary Thompson and Richard Cook for their continual support and encouragement.

I gratefully acknowledge the Population Health Research Group staff, especially Dr. Paul McDonald and Pete Driezen, for their support and encouragement.

Finally, I would like to thank my family members, my Dad Bahri Hamid Hadi, my mother Asiya Khatoon, my lovely wife Tooba Ahmed, my sons Saad and Saif, my siblings Reshma, Wajid, Khalid, Zahid, and Shahid for their understanding, love, and patience.

I dedicate this thesis to my late father Bahri Hamid Hadi.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In many studies looking at the effectiveness of public health interventions, data are collected in a hierarchical manner (e.g., students are in classes that are in schools that are in communities) and information can also be collected over time on the same individual. In educational research, students within schools or students within classes share some common characteristics which need to be accounted for when performing statistical analysis. Traditional linear or generalized linear models account for only a single source of variation between observational units and ignore correlation structures where individuals belong to the same class or school. Similarly, in a repeated observation scenario the correlation within the same individual is ignored.

Any analysis which does not recognize the hierarchical structure of the data (i.e., a student-level analysis that does not take into account class- or school-level correlation) will encounter seri-

ous technical problems. For example, ignoring the hierarchical structure will generally cause the standard error of regression coefficients to be underestimated (Goldstein, 1986). Traditionally, clustering has been handled using design-based procedures which are ad hoc corrections to account for the sampling design (Skinner et al., 1989). In this technique, the survey design variables are regarded as nuisance variables which need to be taken into consideration to obtain robust standard errors.

Hierarchical modeling has a variety of names in the statistical literature including multilevel modeling (Goldstein, 1995; Mason et al., 1984), random effects modeling (Laird & Ware, 1982), general mixed linear modeling (Goldstein, 1986), variance component modeling (Longford, 1986), random coefficient modeling (de Leeuw & Kreft, 1986; Longford, 1993), and hierarchical linear modeling (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 1986, 2002).

A model which does not have a clear hierarchical structure (known as a cross-classified model) can also be handled using the hierarchical model or multilevel structure. Examples include data for (i) a large number of students from one community attending many schools and (ii) students from the same classes attending different courses. For example, in a study looking at students over time, a survey may be administered first in Grade 5 and then in each subsequent grade, through Grade 12. In such a study, students will typically move from one elementary school to different high schools over time. To model this change, the cross-classified structure of student movement needs to be incorporated into the model estimation since the correlation structure of the students has changed from elementary school to high school (over time). In other words, variation in different communities, schools, and students can be cross-classified and must be accounted for.

A similar type of model is the multiple membership model. In this model lower-level units are influenced by more than one higher-level unit from the same classification. The difference between the multiple membership model and the cross-classified model is that in the latter the data are not nested and in the former the observation does not belong to just one member of a classification. For example, a group of students from the same class (lower-level unit) can attend many courses offered in school (higher-level unit) and the student can be classified as a member of multiple courses. In addition, the marks and content of each course will affect the overall individual grade in each class to which they belong (Hill & Goldstein, 1998). Furthermore, in this hierarchical structure, students cross-classified by school and community are all regarded as important sources of variation which must be taken into account (Hill & Goldstein, 1998; Rasbash & Goldstein, 1994).

Hierarchical modeling not only accurately estimates the parameters by focusing on the hierarchical structure of the design, but also provides detailed information about the variance contribution at different stages of the design. For example, in a school smoking survey, the school board is first selected, followed by the school, and finally the students in eligible classes. In this setting level 1 relates to the student information, level 2 relates to the class, level 3 relates to the school, and level 4 relates to the school board. Furthermore, it is easy to account for covariates measured at any level, for example, school-level or class -level covariates which indicate whether the school or the class is in the treatment condition.

A transitional model can also be used in a hierarchical modeling structure. A transitional model is used to estimate the conditional mean of the current outcome given its past outcome. A transitional model combines the dependence of Y (outcome variable) on covariate X and allows for correlation within individuals over time. Furthermore, when the transitional Markov model is

correctly specified then the transitional events become conditionally independent ; hence transitional models can be used to make inferences in longitudinal data (Diggle et al., 1994; Zeger & Qaqish, 1988).

Another important aspect of hierarchical modeling is the structure of the outcome variables. If the outcome variable is continuous, the linear mixed model technique (Laird & Ware, 1982) can be used to handle these correlated data by extending the general linear model. Software such as the SAS procedure MIXED (Little & Wang, 1996; Molenberghs et al., 1997) can be used to analyze this kind of model. When the outcome variable is discrete (e.g., counts) or categorical (nominal or ordinal data), software such as the SAS procedure GLIMMIX based on the Wolfinger & O'Connell (1993) method can be used to fit the model.

For discrete hierarchical models a Bayesian approach can also be used through iterative simulations such as Gibbs sampling (Zeger & Karim, 1991). As a result of recent developments in computing power the Bayesian approach is increasingly being used for discrete hierarchical models. Software such as WinBUGS has been shown to solve a wide range of complicated problems (Gelman et al., 1995).

Missing data are common in longitudinal data; an individual may drop out for many reasons. Little & Rubin (1987) describe these dropouts using three missing-data mechanisms. The first mechanism is missing completely at random (MCAR); the missing data processes are independent of the observed and unobserved data. The second is missing at random (MAR); the missing data processes do not depend on the unobserved data. The third is missing not at random (MNAR); the missing data processes depend on the unobserved data.

There can be monotone patterns of missing data (responses are available for an individual until a

certain time and then missing for all subsequent times) and intermittent non-monotone patterns (responses are missing for a few time points).

If individuals are missing for self-selection reasons (MAR or MNAR) then there are at least two consequences: (i) the loss of power due to missing information and (ii) the possibility of a biased estimate. Some methods discard the incomplete data by default, resulting in decreased statistical power for detecting treatment effects. In addition to the lost power, the sample may no longer be representative of the population being studied, and analytic procedures under these conditions may lead to biased estimates and misleading conclusions.

Our focus in this thesis is on the likelihood-based methods (methods using the adaptive Gaussian quadrature (AGQ)) that are readily available in existing software such as SAS and STATA. Hierarchical models are increasingly used in educational and health policy research and users need information on which estimation procedure to use. This thesis provides the user with the pros and cons of using the approximate method and AGQ.

Most currently available software deals with missing data in hierarchical models based on the ignorable missing-data assumption. That is, it is assumed that the probability of dropout does not depend on the unobserved response (Little & Rubin, 1987; Rubin, 1976). In this thesis we develop methods for handling dropout in hierarchical models where the dropout depends on the unobserved response, i.e., the missing data are non-ignorable (Little & Rubin, 1987; Rubin, 1976).

Fitzmaurice & Laird (2000) used the pattern mixture model for handling non-ignorable dropout for a variety of discrete and continuous longitudinal outcomes. This thesis extends their methods by imputing the missing data under pattern mixture and selection models using the predictive

mean matching method. The other important feature of the proposed method is that it can be implemented in a variety of situations with only minor modifications, using existing statistical software for analyzing discrete hierarchical models, such as SAS and STATA.

1.2 Statement of the Problem

As stated earlier, missing data in a hierarchical structure or a multilevel setting are particularly troublesome because they result in a loss of information and reduce the power of statistical tests, especially when dealing with discrete longitudinal outcomes. This creates serious problems for researchers who use hierarchical models with existing software such as SAS, STATA, MLWIN, and HLM. The objectives of this study are:

- To develop methods for analyzing discrete and ordinal response data from hierarchical longitudinal studies with missing outcome values.
- To use the above estimation methods in complete case analysis and in missing data cases where data are missing based on the past, current, and future outcomes.
- To develop an imputation method for hierarchical models to impute missing values using the predictive mean matching method under two models for non-ignorable missing data: a pattern mixture model and a selection model; and to compare the results with two standard methods: complete case analysis and last observation carried forward (LOCF).
- To apply the above techniques to an existing dataset, the Waterloo Smoking Prevention Project 3 (WSPP3), and to compare the results with those from the simulated dataset.

- To analyze the WSPP3 datasets using a Bayesian pattern mixture model and to compare the results with those from the pattern mixture model and the selection model.

SAS macros and programs written in SAS have been developed to analyze the WSPP3 data. These programs can be generalized to perform similar analyses for different datasets.

1.3 Description of Dataset

This section describes a longitudinal dataset that includes missing observations. This dataset will be used throughout the thesis as a key example to illustrate the techniques developed and to compare them with other frequently used methods.

1.3.1 Waterloo Smoking Project (WSPP3)

The Waterloo Smoking Prevention Project 3 (WSPP3) was conducted by the Population Health Research Group (previously known as the Health Behavior Research Group) at the University of Waterloo (Brown & Cameron, 1997; Cameron et al., 1999). The purpose of WSPP3 was to evaluate a social-influence smoking prevention program at the elementary level (grades 6 through 8) followed by an activity-based program at the secondary level (grades 9 and 10). In addition, students were followed at grades 11 and 12 to assess the long-term impact of the intervention.

Seven school boards in Southwestern Ontario, Canada participated in this study, which included 100 eligible elementary schools (fifteen from each of six boards and ten from the seventh board). One hundred participating schools were randomly assigned in a four-to-one ratio to receive either

an intensive anti-smoking public health education program (treatment) or their standard school curriculum (control). A detailed description is given by Driezen (2001).

In the first phase of the study, elementary school data (grades 6 to 8) were collected over a period of three years from 1989 to 1992. The schools were randomly assigned to an experimental condition based on their school risk score (i.e., the smoking rate among the senior students in the school prior to intervention).

The second phase of the study started when the cohort was in grade 9. Six of the seven school boards participating in the elementary school study agreed to participate in the second phase. Secondary schools which were projected to receive at least 30 students from the elementary school cohort were eligible to participate. Of the 35 eligible schools, 30 agreed to participate. Students who attended other schools (e.g., with fewer than 30 cohort members) were also followed. The secondary schools were pair-matched within school board by size, number of cohort students planning to attend the school, and proportion of cohort students from the elementary schools. The pairs of schools were then randomly assigned to either an intervention or a control condition.

During the final phase of the study, cohort members were followed through Grades 11 and 12 and surveyed to assess the long-term impact of the treatment. Table 1 shows the number of students recruited and their distribution between the experimental and control conditions.

Students were classified into one of five smoking categories: never smoked, tried once, quitter, experimental smoker (smoked less than once a week), and regular smoker (smoked weekly). In this study, the analysis will be restricted to three categories: smoker (experimental smoker or regular smoker), nonsmoker (never smoked, tried once), and quitter.

Table 1.1: Number of students providing data for WSPP3*

Study	Grade	# providing data	# participating in intervention study	# in same treatment condition for entire study
Elementary School Trial	6 (1990)	4466	4466	3821
	7 (1991)	5455	5333	
	8 (1992)	5593	5305	
Secondary School Trial	9 (1993)	4703	2670	2439
	10 (1994)	4999	2643	
Follow-up	11 (1995)	4420	Not Applicable	Not Applicable
	12 (1996)	4204		

1.4 Organization of Thesis

The remainder of the thesis is organized as follows. Chapter 2 provides a description of the hierarchical model for analyzing longitudinal or clustered data and discusses various approaches to parameter estimation. The chapter begins with a formalization of the two-level hierarchical model and its assumptions for both continuous and discrete outcomes. Two prominent estimation approaches are discussed: adaptive Gaussian quadrature (AGQ) and pseudo likelihood (PL).

Chapter 3 begins with the specification of the simulation model and the parameter values. A series of simulations were conducted on a three-level hierarchical model to examine the performance of the parameter estimates and their standard errors obtained from the different estimation procedures. Based on the simulation results, the chapter discusses the pros and cons of two estimation procedures: PL and AGQ.

Chapter 4 describes missing-data mechanisms with an emphasis on non-ignorable missing data. Two models for non-ignorable missing data—selection and pattern mixture models—are de-

scribed, together with their advantages and limitations in the context of hierarchical models. For comparison purposes, a Bayesian pattern mixture model is also described.

Chapter 5 discusses simulation studies and the specification of the simulation model and parameter values. A series of simulation studies were conducted using the three-level discrete hierarchical model and three missing-data mechanisms were established (missing completely at random, missing at random, and missing not at random). Furthermore, the predictive mean matching method was used to impute the missing data under the pattern mixture and selection models. Three restriction methods were employed under the pattern mixture model: the complete case missing value (CCMV), available case missing value (ACMV), and neighboring case missing value (NCMV). Lastly, aggregate parameter estimates were obtained under the three restriction methods and the selection model using the multiple imputation method (Rubin, 1987).

In Chapter 6 an analysis of the Waterloo Smoking Prevention Project 3 is conducted using the proposed methodology for handling the informative missing data.

Lastly, Chapter 7 briefly summarizes the overall findings and outlines future work.

Chapter 2

Hierarchical Models

2.1 Basic Hierarchical Model

The *i.i.d.* (independently identical distribution) assumption of y_i given x_i is key for inferences in a simple regression model. When the sampling design is based on a hierarchical or multilevel structure, many assumptions of the simple regression model do not hold including the *i.i.d.* assumption. For example, it is expected that individual schools from the WSPP3 datasets would have distinct features in terms of school smoking policy, administration, and community. If a simple regression model is fitted to these datasets without taking into account the correlation between students within schools, it may result in inflated effect size estimates and spuriously small standard errors for the parameter estimates (Snijders & Bosker, 1999). Hierarchical models allow the appropriate modeling of the correlation and correct the estimates for the violation of *i.i.d.*.

Let Y_{ij} be the outcome variable and X_{ij} be a covariate for subject i (student in j^{th} school) in cluster j. A suitable model is:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + e_{ij} \tag{2.1}$$

The difference between this model and the simple regression model is the added u_{0j} term (random shift in intercept for different schools), which accounts for mean differences among schools. If we assume that the schools are randomly sampled, which in this case they are, then the school effect can be treated as a random variable and the above equation can be represented as a two-level model. The complete specification of the model is as follows:

$$Y_{ij} = \eta_{ij} + e_{ij} \tag{2.2}$$

where

$$\eta_{ij} = \beta_{0j} + \beta_1 X_{ij}; \qquad \beta_{0j} = \beta_0 + u_{0j}$$
$$u_{0j} \sim N(0, \sigma_u^2); \qquad e_{ij} \sim N(0, \sigma_e^2)$$
$$Cov(u_{0j}, e_{ij}) = 0$$

The parameter β_0 is defined as an average intercept for all students and u_{0j} is the random shift in intercept for different schools. So we can write a simplified model as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + e_{ij} \tag{2.3}$$

with

$$u_{0j} \sim N(0, \sigma_u^2); \quad e_{ij} \sim N(0, \sigma_e^2); \quad Cov(u_{0j}, e_{ij}) = 0$$

Model 2.3 is known as a random intercept model (Goldstein, 1995) and can be extended further by treating the slope of the line relating Y_{ij} and X_{ij} as a random coefficient (Goldstein, 1995). This allows us to estimate the variability in the regression coefficients (both intercepts and slopes) across the second level. That is, suppose

$$\beta_{0j} = \beta_0 + u_{0j} \qquad \beta_{1j} = \beta_1 + u_{1j}$$

where

$$u_{0j} \sim N(0, \sigma_u^2) \qquad u_{1j} \sim N(0, \sigma_{u1}^2)$$
$$e_{ij} \sim N(0, \sigma_e^2) \qquad Cov(u_{0j}, u_{1j}) = \sigma_{01}^2$$
$$E(u_{0j}, e_{ij}) = 0$$

Substituting the above model into Eq. (2.1)

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij}) + (u_{0j} + u_{1j} X_{ij} + e_{ij}) \tag{2.4}$$

The first part of Eq. (2.4) is considered fixed and the second part is considered random. Equation (2.3) can easily be extended by adding any second-level or school-level predictor such as the treatment conditions of the schools C_j as a $N \times 1$ matrix the entries of which have a value of one for treatment schools and a value of zero for control schools:

$$\beta_{0j} = \beta_0 + \beta_2 C_j + u_{0j}$$
$$\beta_{1j} = \beta_1 + \beta_3 C_j + u_{1j}$$

Substituting into Eq. (2.2), we obtain

$$Y_{ij} = (\beta_0 + \beta_1 X_{ij} + \beta_2 C_j + \beta_3 C_j X_{ij}) + (u_{0j} + u_{1j} X_{ij} + e_{ij}) \tag{2.5}$$

As usual the first part is considered fixed and the second part random. Equation (2.5) can be extended to level three or higher. As the levels increase the expression gets more complicated. This thesis only focuses on the random intercept model. In matrix notation, the random intercept model can be written as:

$$Y = X\beta + Zu + e \tag{2.6}$$

where

$$u \sim N(0, D)$$

$$e \sim N(0, \Sigma)$$

where u and e are independent. Model 2.6 is generally referred to in the literature as the Laird-Ware model or the linear mixed model (Laird & Ware, 1982). In model 2.6, Y is the $N \times 1$ response vector for subjects i in the j^{th} school where $1 \leq i \leq n_j$, n_j is the total number of students in the j^{th} school, and $N = \sum_j n_j$. X is the $N \times p$ design matrix for the fixed effects, including second-level fixed effects, β is the vector of fixed-effect coefficients, Z is the $N \times J$ design matrix for the random effects, u is the $J \times 1$ vector of random-effect coefficients, e is the $N \times 1$ vector of random errors for the subjects, D is the $J \times J$ covariance matrix for the random effects, and Σ is the $N \times N$ covariance matrix.

The first two moments in the Laird-Ware model are

$$E(Y) = X\beta \tag{2.7}$$

$$\begin{aligned} Cov(Y) &= ZDZ^T + \Sigma \\ &= V(\alpha) \end{aligned}$$

where α denotes the vector of all variance and covariance parameters.

2.2 Inference for Hierarchical Models

Inference for linear mixed models can be based on procedures such as maximum likelihood (ML) methods or empirical Bayes methodology which yields restricted maximum likelihood (REML) estimates (Harville, 1977; Jennrich & Schluchter, 1986; Laird & Ware, 1982). The difference between ML and REML approaches lies primarily in the treatment of the likelihood function. In the ML method the variance components are estimated by values that maximize the likelihood function over the parameter space. In contrast, REML partitions the likelihood into pieces and maximizes the portion which does not include the fixed effects. Browne & Draper (2000) concluded that REML is at least as good as ML and sometimes better, especially in estimating the variance components. ML methods lead to biased estimates of the variances in small samples, while REML estimates of the variances are less biased since they take into account the loss of degrees of freedom due to the estimation of the regression parameters.

2.2.1 Maximum Likelihood (ML) Methods

Let $\theta = (\alpha, \beta)$ be the vector of all the parameters in the Laird-Ware model. Estimates of the parameters in the Laird-Ware model can be obtained by maximizing the joint likelihood function with respect to θ=(α,β):

$$L(\alpha, \beta) = (2\pi)^{-N/2}|V(\alpha)|^{-1/2}exp\left(-\frac{1}{2}(Y - X\beta)^T V^{-1}(\alpha)(Y - X\beta)\right) \tag{2.8}$$

Assuming that the variance parameters α are known, the maximum likelihood estimate (MLE) of β can be obtained by maximizing Eq. (2.8) conditional on the variance parameters, α (Laird & Ware, 1982):

$$\hat{\beta}(\alpha) = (X^T V^{-1}(\alpha)X)^{-1}(X^T V^{-1}(\alpha)Y) \tag{2.9}$$

Harville (1977) shows that $\hat{\beta}$ is unbiased under the assumption that the mean structure is correctly specified and since $\hat{\beta}$ is linear in Y then the variance of the estimator is easily determined:

$$V(\hat{\beta}) = [(X^T V^{-1}(\alpha)X)^{-1}XV^{-1}(\alpha)]V(Y)[(X^T V^{-1}(\alpha)X)^{-1}XV^{-1}(\alpha)]^T$$

$$V(\hat{\beta}) = (X^T V^{-1}(\alpha)X)^{-1} \tag{2.10}$$

$$\hat{\beta} \sim N(\beta, (X^T V^{-1}(\alpha)X)^{-1}) \tag{2.11}$$

Liang & Zeger (1986) propose using a “sandwich estimator" for $V(\hat{\beta})$. It provides consistent estimates of the covariance for parameter estimates even when a parametric model fails to hold or the variance structure of the parameter estimate is mis-specified. In practice the components $V(\alpha)$ are not known and must be estimated from the datasets. Most often a nonlinear opti-

mization program is used to obtain a maximum likelihood estimate of the variance parameters. The program searches over the values of these parameters until it finds the values that minimize $-2ln(\theta)$. The MLE of α is obtained by maximizing (2.8) after β is replaced by $\hat{\beta}(\alpha)$:

$$l_{ML}(\alpha, \hat{\beta}) = -2ln(\alpha, \hat{\beta}) = Nln(2\pi) + ln|V(\alpha)| + (Y - X\hat{\beta})^T V^{-1}(\alpha)(Y - X\hat{\beta})$$

$$l_{ML}(\alpha, \hat{\beta}) = Nln(2\pi) + ln|V(\alpha)| + Y^T V^{-1}(\alpha)Y - Y^T V^{-1}(\alpha)X\hat{\beta} - \hat{\beta}X^T V^{-1}(\alpha)(Y - X\hat{\beta}) \tag{2.12}$$

The ML estimators are asymptotically unbiased and require a large sample size to have accurate variance estimates. Van der Leeden & Busing (1994) showed that when the sample size is small the ML procedure provided downward biased estimates of variance components and in some cases spurious results because there is no adjustment for the degrees of freedom lost by estimating the regression coefficients. Furthermore, they observed that the variance components at level two were often underestimated. This can lead to false significance of the covariates in the model due to the underestimation of the variance of β.

2.2.2 Restricted Maximum Likelihood (REML) Method

Restricted maximum likelihood (REML) estimation corrects the underestimation of the variance component by explicitly taking into account the loss of the degrees of freedom by maximizing the likelihood of a set of residual contrasts (Diggle et al., 1994). The SAS procedure MIXED uses REML as the default. The REML estimator $\hat{\alpha}$ is obtained by minimizing the following

$-2ln$ likelihood function:

$$l_{REML}(\hat{\beta}(\alpha), \alpha) = Nln(2\pi) + ln|X^T V^{-1}(\alpha) X| + ln|V(\alpha)| + (Y - X\hat{\beta})^T V^{-1}(\alpha)(Y - X\hat{\beta}) + N - p \tag{2.13}$$

where p is the number of regression coefficients.

Equation (2.13) represents the likelihood function of the error contrasts. When a REML estimate of α is available, the REML estimate of β is the same as Eq. (2.10) for maximum likelihood estimation.

There are no closed-form solutions for α, therefore iterative methods are needed to calculate the ML or REML estimates of β and $V(\alpha)$. Harville (1977) suggested using a Newton-Raphson scoring algorithm to estimate the parameters. The EM (expectation and maximization) algorithm also provides a convenient approach to computation for the random effects model in which the unobservable random subject effects and within-subject errors are treated as missing observations (Dempster et al., 1977; Laird et al., 1987; Laird & Ware, 1982). However, the Newton-Raphson algorithm is preferred over the EM algorithm because of its faster convergence (Dennis & Schnabel, 1983; Lindstrom & Bates, 1988), although there are example when the EM algorithm is more accurate. In this thesis, the Newton-Raphson algorithm is used.

2.3 Generalized Linear Mixed Models

Generalized linear models can be used for independent discrete and continuous outcomes. Non-Gaussian but exponential family response variables can be modeled using a linear model through a link function, and this methodology is referred to as the generalized linear model (GLM) for in-

dependent observations. Members of the exponential family have this general probability density or mass function (McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972):

$$f(Y|\theta, \varrho) = exp\left(\left[\frac{Y\theta - b(\theta)}{a(\varrho)}\right] + c(Y, \varrho)\right) \tag{2.14}$$

where θ represents the canonical parameter for an exponential family when the dispersion parameter ϱ is known. Functions a(.), b(.) and c(.) are specified according to different distributions (McCullagh & Nelder, 1989).

The generalized linear model has a link function that relates the linear combination of the covariates (η) to the expectation of Y:

$$g(\mu) = \eta = X\beta$$

where

$$\mu = E(Y) = b'(\theta)$$

$$Var(Y) = b''(\theta)a(\varrho)$$

In the generalized linear model Y is the response vector, X is the design matrix for the fixed effects, and β is the vector of fixed-effect parameters. The canonical link functions are log and logit for Poisson and binary data, respectively. For the log function $g(\mu) = log(\mu)$ and for the logit function $g(\mu) = log\left[\frac{\mu}{1-\mu}\right]$.

The log-likelihood function for the generalized linear model is

$$l(Y, \mu, \varrho) = \sum_i \log(f(Y_i, \mu_i(\beta), \varrho))$$

Maximum likelihood estimates for β can be obtained by weighted least squares by iteration solving: at the k^{th} step

$$X^T V_{k-1} X \beta_k = X^T V_{k-1} Y^*_{k-1}$$

where

$$\mu = E(Y); \Sigma = Var(Y); D = \frac{\partial \mu}{\partial \eta}; V_{k-1} = D_{k-1} \Sigma^{-1}_{k-1} D_{k-1};$$

$$and$$

$$Y^*_{k-1} = \eta_{k-1} + (Y - \mu_{k-1}) D^{-1}_{k-1}$$

GLM has been extended to non-exponential family distributions by the use of quasi-likelihood. The concept of quasi-likelihood was introduced by Wedderburn (1974) and discussed in detail by McCullagh & Nelder (1989). The use of maximum likelihood in parameter estimation requires exact specification of the distribution in order to construct the likelihood function. Quasi-likelihood provides an alternative for problems where the distribution of the response variable may not be known, but its variance function can be expressed as a function of the mean. The quasi-likelihood function for the i^th individual $Q(\mu_i, Y_i)$ is defined by the equation

$$U_i = \frac{\partial Q(\mu_i; Y_i)}{\partial \mu_i} = \frac{Y_i - \mu_i}{\sigma^2 V(\mu_i)}$$

where U_i has the following properties in common with a log likelihood derivative:

$$E(U_i) = 0$$
$$Var(U_i) = \frac{1}{\sigma^2 V(\mu_i)}$$

and

$$-E\left[\frac{\partial U_i}{\partial \mu_i}\right] = \frac{1}{\sigma^2 V(\mu_i)}$$

Since the likelihood function is based on these three properties, we can assume that quasi-likelihood behaves like a log-likelihood function for estimating β. So $Q(\mu_i; Y_i)$ can be defined as:

$$Q(\mu_i; Y_i) = \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\sigma^2 V(t)} dt$$

The quasi-likelihood estimates of β can be obtained by the Newton-Raphson algorithm with a quasi-score function and a quasi-Fisher information matrix (McCullagh & Nelder, 1989).

The joint quasi-likelihood of the independent observations is the sum of the individual contributions to the quasi-likelihood function:

$$Q(\mu; Y) = \sum_i Q(\mu_i; Y_i)$$

2.3.1 Marginal Methods

Generalized estimating equations (GEE) were first proposed by Liang & Zeger (1986) as an estimation technique to estimate the marginal model for the analysis of longitudinal data. GEE is based on a multivariate version of quasi-likelihood (McCullagh & Nelder, 1989; Wedderburn, 1974). This method requires the specification of only the first two moments of a distribution. The procedure involves fitting a generalized linear model to the marginal distribution of the repeated measures and adjusting for correlation between observations on the same subject. It estimates the

population average parameter through a known link function. In a marginal model we define only the specifications of the marginal mean, the variance of the response, and the correlation structure of the response. Liang & Zeger (1986) show that if the model for the mean is correctly specified then the regression estimates are consistent and efficient even if the working correlation structure is mis-specified. In this approach the emphasis is on estimating the regression parameters while treating the response correlation parameters as nuisance parameters.

Let Y_{it} denote the response for the i^{th} individual, measured at time t. The response variables for the i^{th} subject are $Y_i = (Y_{i1}, ..., Y_{iT})$ be the mean of Y_i for the i^{th} individual. Let $X_{it} = (X_{it1}, X_{it2}, ..., X_{itp})$ be a $p \times 1$ vector of covariates associated with Y_{it}, which may be time dependent or fixed covariates. Then

$$\mu_{it} = E(Y_{it}|X_{it})$$

$$g(\mu_{it}) = X_{it}\beta$$

$$Var(Y_{it}|X_{it}) = V(\mu_{it})\varpi$$

where β is the vector of unknown coefficients, ϖ is a scale parameter and $t = 1, 2, ..., T$. Let θ be the vector of all the parameters in this model. The estimate of β is obtained as the solution to the quasi-score equations:

$$\psi_\beta(\theta) = \sum_i X_i^T (\frac{\partial \mu_i(\beta)}{\partial \mu_i})^T A_i V_i^{-1} [Y_i - \mu_i(\beta)] = 0$$

In the marginal model $V_i = Cov(Y_i) = A_i^{1/2} R_i A_i^{1/2}/\varpi$, where $A_i = diag(V(\mu_{i1}), ..., V(\mu_{iT}))$ and R_i is an $T \times T$ working correlation matrix. The solution of the above equation is obtained using iteratively re-weighted least squares. Liang & Zeger (1986) showed that the estimates obtained by GEE are consistent and asymptotically normal given only the correct specification of the mean and certain regularity conditions.

The estimating functions $\psi_\beta(\theta)$ are unbiased if the data are complete or missing completely at random (MCAR) and in this case the GEE approach produces consistent estimates for the mean parameters (Laird, 1988). Liang & Zeger (1986) showed that if the working covariance assumptions are correct and the GEE estimator and the model-based covariance matrix are consistent under MAR, then GEE becomes ML estimates. However, when the data are MAR or MNAR and the covariance assumption is wrong then the estimating equations are not unbiased and hence fail to produce consistent estimates (Fitzmaurice et al., 1995) for the variance. Robins et al. (1995) proposed an inverse-probability weighted GEE approach that yields unbiased equations and consistent estimates for the mean parameters. Estimation of these weights is possible when the data are MAR (Cook et al., 2002; Robins et al., 1995), but sensitivity analyses must be conducted when the data are MNAR (Rotnitzky et al., 1998). The price to be paid for incorporating weights is that a model must be specified for the missing-data mechanism. Weighted GEE can handle the MAR and MNAR mechanisms.

Let W_i be a $T \times T$ matrix whose t^{th} diagonal element is an estimate of the reciprocal of the probability that the t^{th} element of Y_i is observed. The weighted version of the estimating equation is

$$\psi_\beta(\theta) = \textstyle\sum_i X_i^T W_i [Y_i - \mu_i(\beta)] = 0$$

In practice, the probability that the t^{th} element of the i^{th} subject is missing is unknown and can, at best, be estimated by a logistic regression. The weighted estimating equation has been criticized by Little & Rubin (2002), who argue that the greater efficiency of fitting a fully parametric model may outweigh the associated potential for bias. In this thesis we focused on the parametric model and decided not to use the weighted estimating equation. However, the GEE estimate was used to estimate the initial values for model fitting using the SAS procedure GENMOD.

2.3.2 Binary Outcomes

Statistical methodology for the hierarchical data analysis of non-Gaussian data is less well developed than that for Gaussian data. This is especially true for binary-outcome data that lead to generalized linear mixed models (GLMM) with a nonlinear link function such as the logistic link. Hierarchical models not only take into account the correlation structure but also provide estimates for the cluster-specific covariates. The most common choice is the logistic normal model, also known as the hierarchical logistic model.

From this point, the focus will be on the hierarchical logistic model. For simplicity consider the case where there are two levels and a single predictor variable. Let Y_{ij} be the response random variable for the i^{th} subject within the j^{th} group (cluster), $Y_j = (Y_{j1}, Y_{j2}, ..., Y_{jn_j})^T$, X_{ij} a p vector of covariates associated with the i^{th} subject within the j^{th} group, $\pi_{ij} = Pr(Y_{ij} = 1)$ the probability of observing a successful event, and β a p vector of the regression coefficients. Let consider the logistic model

$$logit(\pi_{ij}) = log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) \tag{2.15}$$

$$= X_{ij}\beta + u_{0j} \tag{2.16}$$

Then

$$\pi_{ij} = \frac{exp(X_{ij}\beta + u_{0j})}{1 + exp(X_{ij}\beta + u_{0j})} \tag{2.17}$$

where $Y_{ij} \sim B(1, \pi_{ij})$ with $var(Y_{ij}) = \pi_{ij}/(1-\pi_{ij})$. The usual assumption of normal random effects $u_{0j} \sim N(0, \sigma_u^2)$ is for convenience. It does not create difficulties in the estimation of the fixed and random effects.

The hierarchical logistic model in Eq. (2.15) presents more challenges than the standard logistic model. ML estimates are often used to estimate the parameters in the standard logistic model. However, in hierarchical logistic models this is more difficult. Let $f(Y_j|u_{0j}, \beta, \alpha)$ denote the conditional probability density function for the response variable Y_j given u_{0j}, and let $h(u_{0j})$ denote the probability density function for the random effect u_{0j}, which is assumed to be normal, then

$$f(Y_j|X_{j1}, X_{j2}, ..., X_{jp}, \beta, \alpha, u_{0j}) = \prod_{i=1}^{n_j} \pi_{ij}^{Y_{ij}}(1-\pi_{ij})^{1-Y_{ij}} \tag{2.18}$$

$$= \prod_{i=1}^{n_j} \frac{[exp(X_{ij}\beta + u_{0j})]^{Y_{ij}}}{1 + exp(X_{ij}\beta + u_{0j})}$$

and

$$f(Y_j|X_{j1}, X_{j2}, ..., X_{jp}, \beta, \alpha) = \int f(Y_j|u_{0j}, \beta, \alpha)h(u_{0j})du_{0j} \tag{2.19}$$

In general, integral (2.19) does not have a closed-form expression when the model is nonlinear. Several approximate methods have been proposed to estimate the fixed and random effects in the context of generalized linear models. Some of these methods consist of taking a first-order Taylor expansion (Sheiner & Beal, 1980; Vonesh & Carter, 1992; Wolfinger & O'Connell, 1993). Others use a Gaussian quadrature method (Davidian & Gallant, 1992; Pinheiro & Bates, 1995). Wolfinger & O'Connell (1993) used a procedure called pseudo-likelihood (PL) in which a Taylor series approximation method was used to approximate the link function and integrated likelihood for the marginal distribution. However, Breslow & Clayton (1993) showed that such an approximation could be quite inaccurate under certain conditions. Furthermore, Gibbons et al. (1993) and Pinheiro & Bates (1995) developed accurate approximations to the ML using Gauss-Hermite quadrature, and these are now implemented in the SAS procedure NLMIXED and STATA and the package MIXOR. It should also be mentioned that the SAS procedure NLMIXED uses Adaptive Gaussian Quadrature (AGQ) to compute the integral over the random-effect distribution in order to obtain the approximate likelihood. An alternative approximation uses a higher-order Laplace transformation (Raudenbush & Bryk, 2002) and is implemented in the HLM program. In this thesis the focus is on PL (Wolfinger & O'Connell, 1993) and AGQ (Pinheiro & Bates, 1995). Both methods use an iterative approach to estimate the parameters.

2.3.3 Pseudo-Likelihood Method

The Pseudo-Likelihood (PL) approach for nonlinear models uses an approximation based on a linear mixed model with current values of the covariance parameter estimate. The resulting linear mixed model is then fitted again which is itself an iterative process. On convergence the new parameter estimates are used to update the linearization which results in a new linear mixed model. The process stops when the change in the parameter estimates between the successive linear fittings is within a specified tolerance. Wolfinger & O'Connell (1993) present the above approximate method by fitting model (2.19) using same notation as in equation 2.6 in matrix form. The procedure begins with

$$\eta = g(\pi) = logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = X\beta + Zu \tag{2.20}$$

where $\eta = g(\pi)$ is referred to as the link function. It is assumed that u has a normal distribution with zero mean and covariance matrix D. Then

$$E(Y|u) = g^{-1}(X\beta + Zu) = g^{-1}(\eta) = \mu$$

where $u \sim N(0, D)$ and $V(Y|u) = \Sigma$. Following Wolfinger & O'Connell (1993) a first-order Taylor series of μ about $\tilde{\beta}$ and $\tilde{u}$ yields

$$g^{-1}(\eta) \cong g^{-1}(\tilde{\eta}) + \tilde{\triangle}X(\beta - \tilde{\beta}) + \tilde{\triangle}Z(u - \tilde{u})$$

$$\tilde{\triangle} = \left[\frac{\partial g^{-1}(\eta)}{\partial \eta}\right]_{\tilde{\beta},\tilde{u}}$$

where $\tilde{\triangle}$ is a diagonal matrix of derivatives of the conditional mean evaluated at the expansion locus. Then

$$\tilde{\triangle}^{-1}(\mu - g^{-1}(\tilde{\eta})) + X\tilde{\beta} + Z\tilde{u} \cong X\beta + Zu \cong P$$

and $V[P|u] = \tilde{\triangle}^{-1}\Sigma\tilde{\triangle}^{-1}$ which is a linear mixed model with pseudo-response P, fixed effect β, and random effect u. Define the marginal variance in the linear mixed pseudo-model to be $V[\alpha] = ZDZ^T + \tilde{\triangle}^{-1}\Sigma\tilde{\triangle}^{-1}$. The log pseudo-likelihood (MPL) and restricted log pseudo-likelihood (RPL) for P are then

$$l(\alpha, p) = -\frac{1}{2}log|V(\alpha)| - \frac{1}{2}\gamma^T V^{-1}(\alpha)\gamma - \frac{N}{2}log(2\pi);$$

and

$$l_R(\alpha, p) = -\frac{1}{2}log|V(\alpha)| - \frac{1}{2}\gamma^T V^{-1}(\alpha)\gamma - \frac{1}{2}log|X^T V^{-1}(\alpha)X| - \frac{N-q}{2}log(2\pi)$$

with $\gamma = P - X(X^T V^{-1}(\alpha)X)^{-1}X^T V^{-1}(\alpha)P$ and where q denotes the rank of X.

The fixed-effect parameters β and the random effects u can be estimated from these expressions:

$$\hat{\beta} = \left(X^T V^{-1}(\hat{\alpha})X\right)^{-1} X^T V^{-1}(\hat{\alpha})P$$

$$\hat{u} = \hat{D}Z^T V^{-1}(\hat{\alpha})\hat{\gamma}$$

where α is estimated by the dual quasi-Newton optimizing technique and the objective function for minimization is $-2l(\alpha, p)$ or $-2l_R(\alpha, p)$. Once the convergence is achieved using the

optimization techniques, the regression parameters are estimated using

$$\hat{\beta} = (X^T V^{-1}(\hat{\alpha}) X)^{-1} X^T V^{-1}(\hat{\alpha}) P$$

and the random effects are predicted as

$$\hat{u} = \hat{D} Z^T V^{-1}(\hat{\alpha}) \hat{\gamma}$$

This process continues until the relative change between the parameter estimates at two successive iterations is sufficiently small. The GLIMMIX procedure in SAS version 9.1.3 allows for the estimation of binary hierarchical models by expanding upon the properties of MIXED (which considers the linear model as the response variable).

The advantage of the linearization-based method is the relatively simple form of the linearized model that typically fits the model based on only the mean and the variance in the linearized form. Furthermore, it is computationally efficient. Therefore, most often this method is used to provide starting values for other procedures. The potential disadvantage of this approach is the absence of a true objective function for the overall optimization process which could potentially bias the estimates of the covariance parameters, especially in the binary response model, because of the double iterative process. Furthermore, this method deteriorates as the distribution of the response variable departs further from normality or if large variance components are present. The parameter estimates are then negatively biased (Breslow & Lin, 1995). Another disadvantage is that PL does not directly involve the likelihood. Thus, this method cannot use likelihood-based inference such as likelihood ratio tests and likelihood-based confidence intervals.

2.3.4 Adaptive Gaussian Approach

The Gaussian quadrature method approximates the integral in Eq. (2.19) by a weighted sum over predefined abscissas for the random effects. A good approximation can usually be obtained with an adequate number of quadrature points as well as appropriate centering and scaling of the abscissas. The weights and abscissas used in Gaussian quadrature rules for the most common distributions can be obtained from the tables of Abramowitz & Stegun (1964) or by using the algorithm proposed by Bjorck & Golub (1973). A problem related to multiple integrations can be transformed to successive applications of simple one-dimensional Gaussian quadrature rules.

Lindstrom & Bates (1990) show how adaptive Gaussian quadrature works in discrete hierarchical models. Let Y_{ij} be the response random variable for i^{th} subject with the j^th group (cluster). The probability of observing a successful event is defined as $\pi_{ij} = Pr(Y_{ij} = 1)$. In this case the cluster-specific parameter vector is modeled as a random intercept model

$$\eta_{ij} = logit(\pi_{ij}) = X_{ij}\beta + u_{0j}, \qquad u_{0j} \sim N(0, \sigma_u^2)$$

where β is a p-dimensional vector of fixed population parameters, u_{0j} is a random effect associated with the j^{th} cluster, and X_{ij} is the design vector for all the fixed effects covariates for individual i. For the above nonlinear mixed effect model AGQ centers the quadrature points around the empirical Bayes estimates $\hat{u_{0j}}$ of the random effects, where the empirical Bayes estimates are calculated from the function

$$-log\left\{\prod_{j}^{J}\left[\prod_{i=1}^{n_j} f(Y_{ij}|u_{0j},\beta,\alpha)h(u_{0j})\right]\right\}$$

and scaled using the final negative Hessian value D from the optimization of this function. Let k_q and w_q denote the standard Gauss-Hermite abscissas and weights (Golub & Welsch, 1969). The quadrature points k_q are then adjusted to be $a_q = \hat{u_{0j}} + \sqrt{2}Dk_q$ for q = 1,2,.....,Q. The standard weights w_q are scaled to be $w_q^* = w_q e^{k_q^2}$. Then

$$L(\beta, \alpha) = \int_u \prod_{j=1}^{J} \left[\prod_{i=1}^{n_j} f(Y_{ij}|u_{0j}, \beta, \alpha) h(u_{0j}) \right] du$$
$$\approx \sqrt{(2)} D \sum_{q=1}^{Q} \prod_{j=1}^{J} \left[\prod_{i=1}^{n_j} f(Y_{ij}|u_{0j}, \beta, \alpha) \right] \left[h(\Phi(\hat{u}_{0j} + \sqrt{2}Dk_q)) \right] \phi(\hat{u}_{0j} + \sqrt{2}Dk_q) w_q e^{k_q^2}$$

Here $\phi(.)$ is the standard normal probability density function and $\Phi(.)$ is the standard normal cumulative density function. AGQ can be generalized to approximate any nonlinear mixed effect model.

The SAS procedure NLMIXED is recommended for the analysis of binary data that require accurate covariance parameter estimates (Murray et al., 2004). Murray et al. (2004) further suggest that the numerical integration maximum likelihood estimation method employed by NLMIXED is superior for multilevel analysis involving small groups, such as family studies.

The SAS procedure NLMIXED uses AGQ to compute the integral over the random effects in order to obtain the approximate likelihood. The number of quadrature points is adaptively selected by evaluating the log likelihood function at the starting values of the parameters until a relatively small change arises between two successive evaluations. This method permits more flexibility in accommodating user-defined likelihood functions than the PL methods. Furthermore, for a small cluster size, AGQ methods perform better than PL methods, but they become more complicated if the number of random effects is greater than two. For three-level hierarchical models, such as

WSPP3, the AGQ methods can be used under the Markov transition model.

2.4 Transitional Model

When the transition pattern in repeated binary data is of interest, a more appropriate approach is to model the transition probabilities over time. Various authors have considered modeling heterogeneous transitional data without missing data (Albert & Waclawiw, 1998; Cook, 1999). If the Markov transitional model is correctly specified then the transitional events become conditionally independent and transitional models can be used to make inferences about parameters (Diggle et al., 1994; Zeger & Qaqish, 1988). In such models individual movement to a given state at time $t+1$ is dependent upon the state at time t. Markov transition models combine the dependence of Y on covariates X and correlation within individuals over time, by regressing the current value of Y on X and previous values.

McCullagh & Nelder (1989) suggested two models that can be used to model transition probabilities: generalized logit models and proportional odds models. Throughout this thesis, the generalized logit model is used. The correlation across time within subjects is accounted for using Markov transition models and the correlation between subjects within clusters is incorporated using random effects.

Let $Q_{ij,t} = k$, $t = 1, 2, ..., T$ denote the status of the i^{th} subject in cluster j at time t with k possible states where $k = 1, 2, ...K$. We assume that the evolution of the status satisfies a first-order Markov chain with transitional probability from state k to state l at time t defined as (t $\geq$

2),

$$p_{ijt}(l|k) = Pr(Q_{ij,t} = l|Q_{ij,t-1} = k, X_{1ij,t-1}, C_j, \theta_{kl})$$

where θ_{kl} denotes the collection of all parameters.

For the i^{th} subject with previous stage k, we use the generalized logit model

$$log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_{0j|kl} + \beta_{1|kl}X_{1ij,t-1} + \beta_{2|kl}t$$

where

$$\beta_{0j|kl} = \beta_{0|kl} + \beta_{3|kl}C_j + u_{0j|kl}.$$

Yosef (1997) conducted a simulation study of a two-level mixed-effects logit model with a single random effect comparing the AGQ and PL methods. For AGQ, he used the MIXOR program of Hedeker and Gibbons and found that AGQ generally gives less biased estimates than PL. However, as the number of random effects increases, AGQ becomes more computationally complex and inefficient. Yosef (1997) concluded that AGQ performs well if the number of random effects is small. Breslow & Lin (1995) also show large differences in the estimate of the variance of the random effects using PL and AGQ with a two-level hierarchical structure. Their results suggest that PL produces higher biases for the variance estimate of the random intercept. Furthermore, their simulation studies show that PL deteriorates as the data depart from normal (e.g., binary) and as the variance component increases. Breslow & Lin (1995) conclude that it is better to use exact methods such as AGQ rather than approximation methods such as PL if the purpose is to estimate the variance of the random effect. A more recent study has shown similar results with small cluster sizes. AGQ tends to provide better results than PL for the variance estimate of the

random intercept (Browne & Draper, 2006).

However, AGQ is computationally intensive and leads to a corresponding difficulty in successful estimation for higher levels of random effect. Recently, the GLLAMM procedure in STATA has used AGQ and can be extended to more than two levels of random effect. However, PL has an advantage over AGQ because of its computational efficiency. The procedure is fast compared to AGQ. For the simulation analysis in Chapter 5, along with the GEE approach (used to find initial values for the fixed-effect parameters), PL was used to find the starting values for the variance estimate of the random intercept to use in AGQ. PL has an advantage over AGQ if complex models (e.g., with a large number of random effects and/or multiple hierarchies) are required.

In this thesis, we consider hierarchical model with correlation within subjects across time, and between subjects within schools. A Markov transitional model assumption was used to account for the correlation within individuals over time, and correlation between subjects is incorporated using the normal random effects. Under this assumption, the analysis of a three-level hierarchical model requires specification of only a single level of random effect. Chapter 3 will use a simulation study to assess the relative merits of AGQ and PL for this clustered Markov transitional model. Later in the thesis, we develop an imputation method for hierarchical models to impute the missing values using the predictive mean matching method when the missing data are classified as MNAR. Based on the simulation study reported in Chapter 3, we will determine which estimation method (AGQ or PL) should be used to estimate model parameters when dealing with missing data.

2.5 Summary

This chapter described hierarchical models for analyzing longitudinal and/or clustered data and discussed two approaches to parameter estimation. This thesis focuses on likelihood-based and approximate likelihood methods that are easily available in standard software such as SAS and STATA. The first estimation method used the PL procedure in which a Taylor series approximation method is used to approximate the link function and integrated likelihood for the marginal distribution. This method has been implemented in the SAS procedure GLIMMIX that is used in the subsequent chapters. The second estimation method is the AGQ in which a quadrature rule is used to approximate the likelihood function. The SAS procedure NLMIXED uses AGQ to estimate the model parameters. Comparisons between likelihood and Bayesian methods are performed in Chapter 6 with WSPP3 datasets. In Chapter 3 simulation studies are performed to examine the performance of the parameter estimates and standard errors obtained from likelihood and approximate likelihood methods.

Chapter 3

Evaluating the Estimation Procedures Using Simulation

This chapter describes the design of the simulation study to examine the performance of the parameter estimates and estimated standard deviations obtained from two estimation procedures: AGQ and PL. A series of simulations were conducted on a discrete hierarchical model with explanatory variables. SAS programs are used to generate the data, implement the estimation techniques, fit the specified models, and compute the estimation accuracy indices for both approaches. These simulations focus only on binary response variables with a logit link function with a normal random-effect distribution.

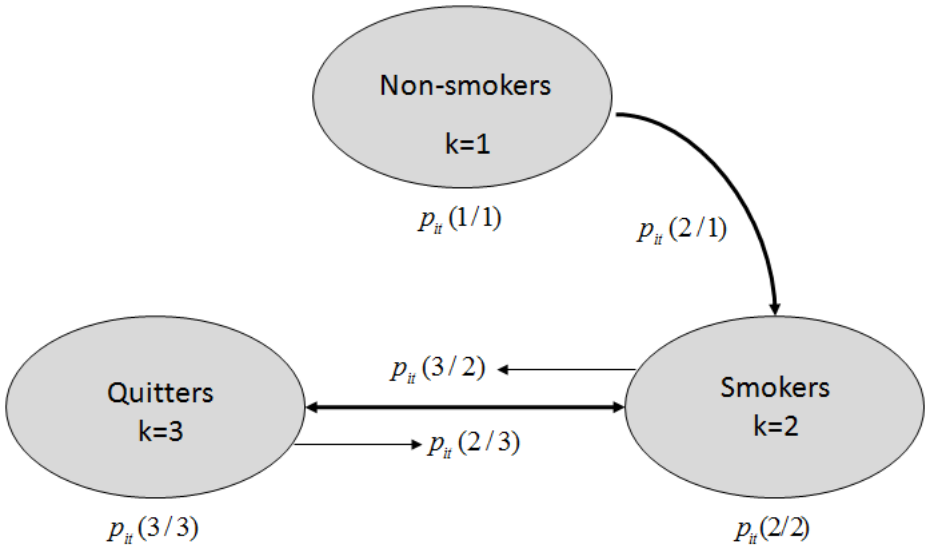
3.1 Simulation Model and Parameter Values

Data were simulated under a hypothetical experimental situation. The purpose of this simulation was to estimate the growth rate difference between two groups and the variability in intercept. As a starting point for the simulation, a hypothetical longitudinal study based on the WSPP3 study described earlier was created to examine smoking behavior among school-aged youth with smoking transitions as the primary outcomes. In this hypothetical study, schools were randomly assigned to either a control or treatment condition and students within those schools were followed across time.

This simulation uses a three-level hierarchical model with a baseline and six time points. At the baseline along with all the cluster information, the smoking status was used to construct individual transitions over time. The individual smoking status was assigned to one of three categories for any given individual. Individuals who never smoked a cigarette or smoked only once are considered nonsmokers; those who have smoked more than one cigarette and have smoked a cigarette in the last thirty days are considered smokers; and those who have smoked more than one cigarette but who did not smoke a cigarette in the last thirty days or those who quit smoking are considered quitters. The individual movement to a given state at time $t + 1$ is dependent upon the state at time t. We model the probability of individuals moving from one state to another between two given time points. For example, a nonsmoker at time t can become a smoker at time $t + 1$. In this simulation, time is considered discrete and the individual's state is determined based on the subject's assessment at a given time. Because of the discrete time-point assessment, it is possible that a nonsmoker at time t can move to the quitter state at time $t + 1$ by moving through two transitions (nonsmoker to smoker and smoker to quitter). For simplicity, we

assumed that the subject's transition occurred only at the discrete time points and that transition took place between the two assessment periods. Since we assumed that the individual can not have two or more transitions between time points, transition from nonsmoker to quitter and quitter to nonsmoker are considered invalid. Figure 3.1 shows the three smoking states and the allowable movements from one state to another.

Figure 3.1: Graph for possible transition states



If $Q_{ij,t} = k$ denotes the status of the i^{th} subject in cluster j at time t with k possible states; $i = 1, 2, ..., n_j$, $j = 1, 2, ..., J$, $t = 1, 2, ..., T$, and $k = 1, 2, ...K$, we define an indicator variable $Y_{ijt|kl}$ such that, for t$\geq$ 2,

$$Y_{ijt|kl} = \begin{cases} 1 & \text{if } Q_{ij,t} = l | Q_{ij,t-1} = k \\ 0 & \text{if } Q_{ij,t} = k | Q_{ij,t-1} = k \end{cases} \tag{3.1}$$

We assume that the evolution of the status satisfies a first-order Markov chain with transitional probability from state k to l defined as

$$\begin{aligned} p_{ijt}(l|k) &= Pr(Q_{ij,t} = l|Q_{ij,t-1} = k, X_{ijt}, \theta_{kl}) \\ &= Pr(Y_{ijt|kl} = 1|X_{ijt}, \theta_{kl}), \end{aligned}$$

where θ_{kl} denotes the collection of all the parameters and $l, k = 1, 2, 3$.

For the i^{th} subject with previous state k, we use the generalized logit model

$$log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_{0j|kl} + \beta_{1|kl}X_{1ij,t-1} + \beta_{2|kl}t$$

where $\beta_{0j|kl} = \beta_{0|kl} + \beta_{3|kl}C_j + u_{0j|kl}$

So in matrix notation,

$$logit(Pr(Y_{ijt|kl} = 1)) = X_{ijt|kl}\beta_{kl} + u_{0j|kl} \tag{3.2}$$

We model (k,l) pairs (1,2), (2,3), and (3,2),

where

$Y_{ijt|kl} \sim B(1, \pi_{ijt|kl})$;

$Y_{ijt|kl} = 0$ if there is no transition from state k to l;

$Y_{ijt|kl} = 1$ if there is a transition from state k to l;

t is a time measurement variable;

$X_{1ij,t-1}$ is a time-dependent covariate for the i^{th} subject in the j^{th} school at time $t-1$, where $X_{1ij,t-1} \sim N(5,1)$;

$\beta_{0j|kl}$ is the baseline score for the j^{th} cluster where $\beta_{0j|kl} = \beta_{0|kl} + \beta_{3|kl}C_j + u_{0j|kl}$;

$\beta_{1|kl}$ is the slope for time-dependent covariate ;

$\beta_{2|kl}$ is the effect of time;

C_j is a binary variable for school j coded as 0 for the control and 1 for the intervention group;

$\beta_{3|kl}$ is a log odds of the transition for the intervention group compared to the control group given the covariates and time ;

$\beta_{0|kl}$ is a log odds of the transition for the control group at t=0 and $X_{1ij,t-1} = 0$;

$u_{0j|kl}$ is a random effect for the intercept and assumed to be independent of the level-two predictors;

$$\pi_{ijt|kl} = Pr(Y_{ijt|kl} = 1|X_{ijt}, \theta_{kl})$$

.

Each individual in the sample could pass through several states during the period of observation. Each model for Y is state specific, that is, observations and parameters in the model depend on the state the subject is in at times $(t-1)$ and t.

The data are generated using the interactive matrix language in SAS. Model (3.2) can be written with a fixed and a random part. If

$$\eta_{ijt|kl} = logit(\pi_{ijt|kl},)$$

then

$$\eta_{ijt|kl} = \beta_{0|kl} + \beta_{1|kl}X_{1ij,t-1} + \beta_{2|kl}t + \beta_{3|kl}C_j + (u_{0j|kl}), \tag{3.3}$$

and

$$\pi_{ijt|kl} = \frac{exp(\eta_{ijt|kl})}{1 + exp(\eta_{ijt|kl})} \tag{3.4}$$

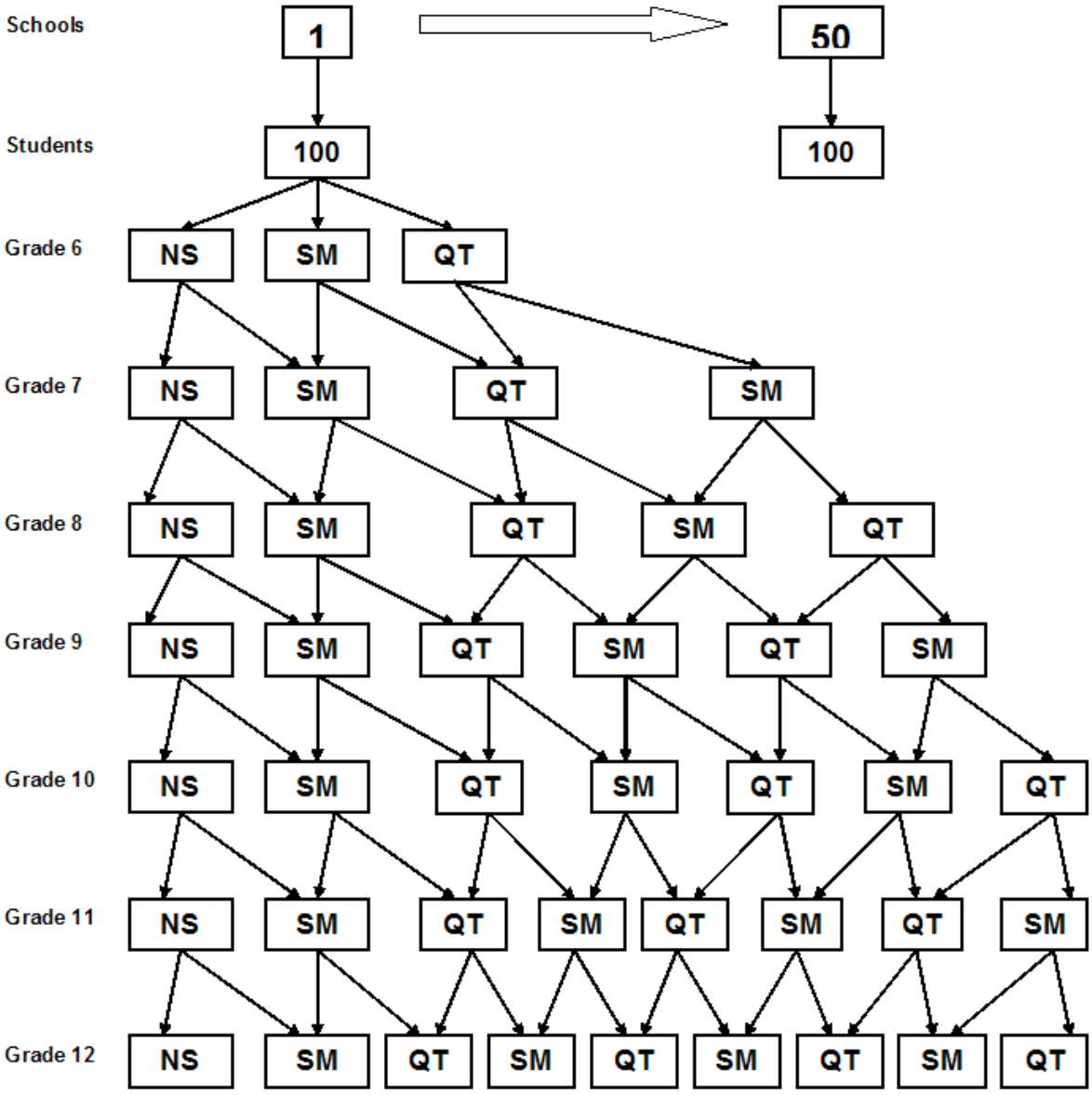
For the simulation study, data are generated to replicate the WSPP3 study, described previously, under ideal conditions. We performed 500 simulations. We generated 50 schools per simulation and in each school we generated 100 students and assigned a smoking status to each individual based on the baseline (grade 6) smoking-prevalence rates and the covariates from the WSPP3 study as shown in Table 6.1 and described in model (3.2). Each school was assigned to either the treatment or control group using a binomial random variable with $b \sim (n, 1, p = 0.55)$. The time variable (t) is an indicator from 1 to 7 as a proxy for time. Time dependent variable is created as a proxy for the number of smoking friend. Figure 3.2 shows how the simulated data were created. In each of the 50 schools, smoking transitions for 100 students were simulated over the seven discrete time points. Each point represents a school grade so that students were followed from grade 6 to grade 12. In each grade, students were classified into one of three possible states: nonsmoker, smoker, or quitter. Figure 3.2 also shows the possible transitions

students could make over the seven-year period in one school. For example, an individual who was a nonsmoker in grade 6 could remain a nonsmoker in grade 7 or could have started smoking by then. Similarly, a student who was a smoker in grade 6 could quit by grade 7 or remain a smoker.

Each time point represents a school year, starting from grade 6. At the baseline, within a cluster we assigned the smoking status by generating a trinomial distribution with probabilities 0.06, 0.08, and 0.86 respectively for being a quitter, smoker, or nonsmoker. These probabilities were based on the WSPP3 study and retained the hierarchy while creating the datasets. At each time point, we created data based on an individual's previous smoking status. At the baseline, we model the transition from nonsmoker to smoker using model (3.2) with the fixed-effects parameters set to $\beta_0 = -2.3, \beta_1 = 0.2, \beta_2 = 0.61$, and $\beta_3 = -4.1$ and the distribution of the random effect are assumed as $N(0, \sigma_u^2)$, where $\sigma_u^2 = 0.68$. We model the transition from smoker to quitter using model (3.2) with the fixed parameters set to $\beta_0 = 0.8, \beta_1 = -0.1, \beta_2 = -0.3$, and $\beta_3 = 0.2$ and the variance of the random effect set to $\sigma_u^2 = 0.68$. We model the transition from quitter to smoker using model (3.2) with the fixed parameters set to $\beta_0 = -1.7, \beta_1 = 0.3, \beta_2 = 0.1$, and $\beta_3 = -5.5$ and the variance of the random effect set to $\sigma_u^2 = 0.68$. Parameter values were derived using some of the information from the actual WSPP3 complete data analysis for grade 9-12 students. The transition probabilities for each state were determined using Eqs. (3.3) and (3.4). Finally, the outcome variable for each state was created using these transition probabilities (p) which were then converted to binary random variables. As an example, students who were nonsmokers at the baseline would be assigned 1 if they moved from nonsmoker to smoker, or 0 if they stayed in the nonsmoking state.

Figure 3.3 shows the individual transitions over time in the three states. Figure 3.3a shows the

Figure 3.2: Representation of the simulated data



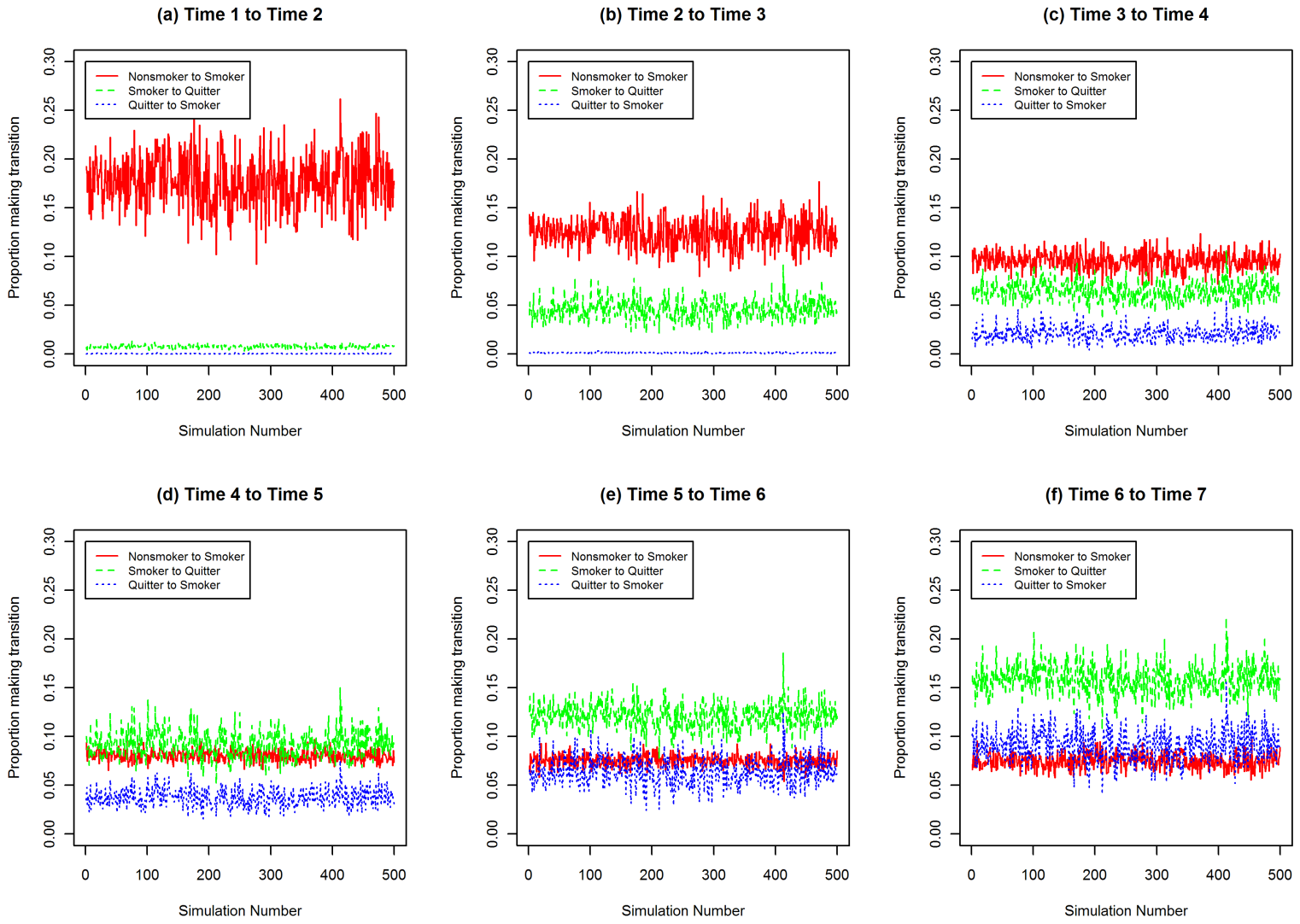
proportion of students who move from their baseline state to another state from time 1 (grade 6) to time 2 (grade 7). Figure 3.3b-f shows the other time points. Figure 3.3 clearly shows that initially most of the students make the nonsmoker to smoker transition and later they move from the smoker to the quitter state.

Figure 3.4 indicates how the school smoking rates change over the 7 time points for 10 randomly selected schools for 16 randomly selected simulations. Each graph shows a different simulation. The main purpose for these graphs is to show the variation in smoking rates between the schools.

Once the datasets were generated, PL and AGQ were used to estimate the parameters. If PL and AGQ did not converge, the datasets were replaced with new datasets to achieve 500 estimates for each parameter. PL is implemented in the SAS procedure GLIMMIX and AGQ is implemented in SAS procedure NlMIXED (SAS version 9.1.3). The two estimation methods were compared using the following criteria:

1. Numerical convergence
2. Average empirical bias of parameter estimate
3. Average standardized empirical bias of mean estimate
4. Root mean square error
5. Coverage Rates

Figure 3.3: Transition proportion for each smoking state



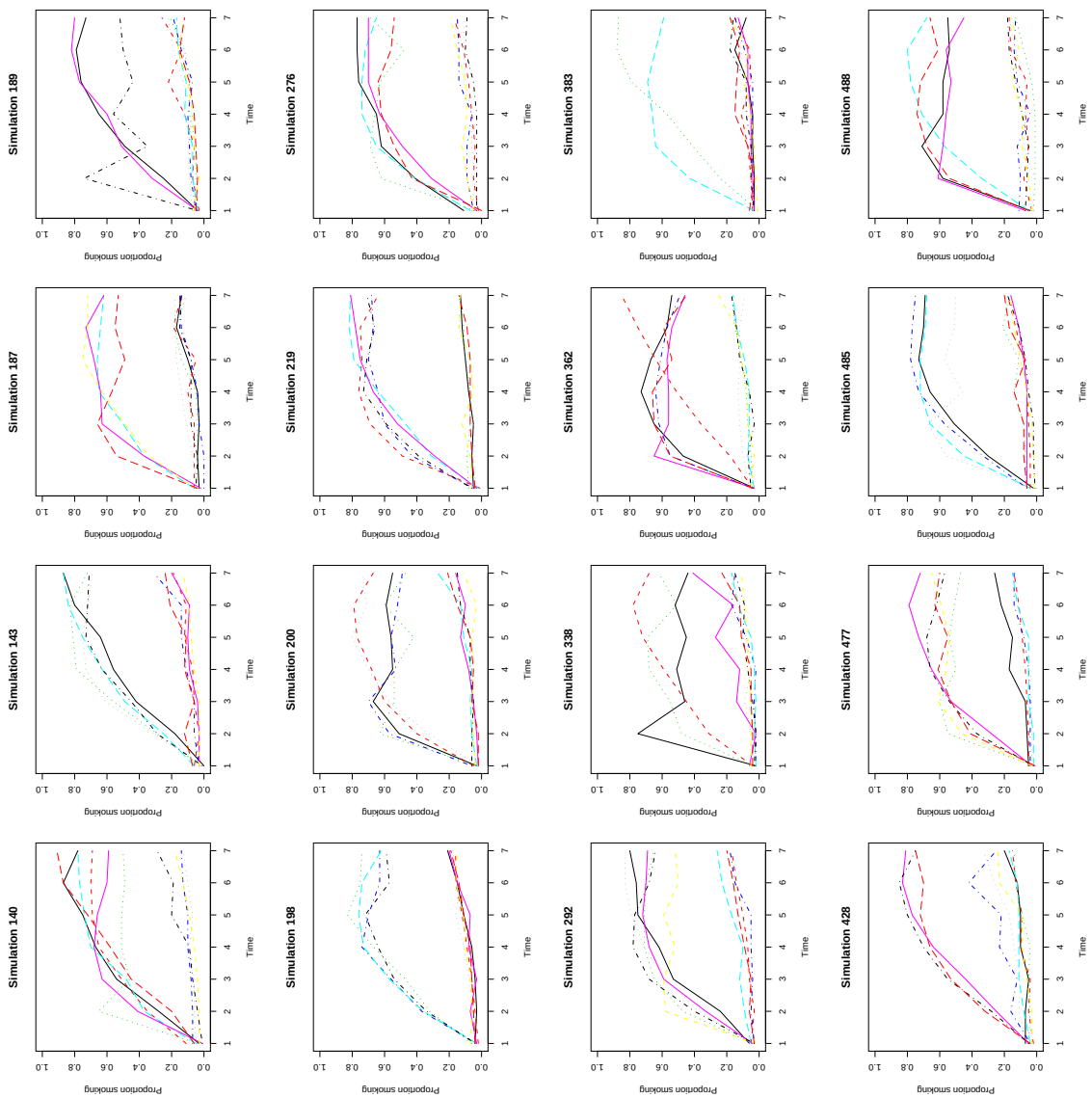


Figure 3.4: School smoking rates over time for sample simulations (10 schools)

3.2 Numerical Convergence

The numerical convergence is measured by the convergence rate. This rate was based on an indicator variable produced by PL and AGQ to show whether convergence was achieved. However, for counting purposes, the number of non-converging simulations was recorded and these simulations were replaced with the new simulated datasets to achieve 500 parameter estimates. In our SAS program, an indicator function was created to identify the convergence of the estimation procedures. If the procedure converges then it assigns the value 1, otherwise it assigns 0. Our results based on indicator variables show that PL did not converge 61 times and AGQ did not converge 19 times in the process of achieving 500 converging simulations which is a much smaller non-convergence rate than reported in other studies (e.g., Callens M (2005)). Non-convergence problems included: no estimated values; estimation only of fixed-effect parameters; and unreasonable estimated values. Upon closer inspection of the non-converging simulations, it was found that if only a few individuals move from one state to another, then the estimation procedures are not robust in calculating the variance estimate. Non-convergence was also more likely when the estimate for the variance of the random intercept was close to zero.

In AGQ the user is required to provide the initial values for the regression parameters as well as for the variance of the random intercept. Some of the non-convergence in AGQ was due to the initial values. If the provided initial values were far from the actual values, AGQ sometimes did not converge. To solve this problem the generalized estimating equation method (GEE, SAS procedure GENMOD) was used to provide the initial values for the fixed-effect parameters (Van Ness et al., 2007). The GEE approach is popular because the estimates of the mean parameters remain consistent even if the correlation or the covariance structure is mis-specified.

In contrast, PL uses a double iterative scheme in which the parameter starting values are generated from an iteratively derived approximating linear mixed model. These initial values are used to update the linearization, and are then applied to iteratively fit a final model.

3.3 Average Empirical Bias of Parameter Estimates

The average empirical bias of a parameter estimate is defined as the average difference between the parameter estimate, $\hat{\theta}_{mk}$, and the true parameter value, θ_k:

$$\text{Average Empirical Bias}_{\text{k}} = \frac{\sum_{m=1}^{500} (\hat{\theta}_{mk} - \theta_k)}{500}$$

For reporting purposes, the average biases and the empirical standard deviation for each parameter and each transition in all $m = 500$ simulations were recorded and are summarized. The results for the simulation study are reported in Table 3.1 in terms of biases.

The true values, average parameter estimates, and the corresponding empirical standard deviation for AGQ and PL are reported in Table 3.1 along with the average biases. The table includes all three transitions: nonsmoker to smoker, smoker to quitter, and quitter to smoker.

Table 3.1 shows that the fixed parameter estimates obtained using either method are close to the true values but the estimated empirical biases for the AGQ method are comparatively small.

Both methods show that the bias of the estimate of the model intercept (β_0) varies from one transition to another. For the two transitions (nonsmoker to smoker and smoker to quitter), the model intercept estimates are close to the true value. This is true for both AGQ and PL. However, for the second transition (smoker to quitter) the estimates based on the PL method are

Table 3.1: Average estimated empirical bias for all three transitions based on 500 simulations*

Parameter	TRUE	PL Estimate	PL ESD	PL EB	AGQ Estimate	AGQ ESD	AGQ EB
Non-smoker to smoker transition							
β_0	-2.300	-2.327	(0.202)	**-0.027**	-2.295	(0.218)	**0.005**
β_1	0.200	0.198	(0.025)	**-0.002**	0.201	(0.022)	**0.001**
β_2	0.610	0.606	(0.020)	**-0.004**	0.608	(0.019)	**-0.002**
β_3	-4.100	-4.064	(0.213)	**0.036**	-4.071	(0.220)	**0.029**
σ^2_μ	0.680	0.438	(0.092)	**-0.242**	0.671	(0.079)	**-0.009**
Smoker to quitter transition							
β_0	0.800	0.923	(0.203)	**0.123**	0.800	(0.233)	**0.000**
β_1	-0.100	-0.105	(0.024)	**-0.005**	-0.099	(0.028)	**0.001**
β_2	-0.300	-0.298	(0.018)	**0.002**	-0.304	(0.016)	**-0.004**
β_3	0.200	0.261	(0.208)	**0.061**	0.215	(0.147)	**0.015**
σ^2_μ	0.680	0.452	(0.111)	**-0.228**	0.675	(0.089)	**-0.005**
Quitter to smoker transition							
β_0	-1.700	-1.709	(0.266)	**-0.009**	-1.681	(0.326)	**0.019**
β_1	0.300	0.393	(0.038)	**0.093**	0.302	(0.064)	**0.002**
β_2	0.100	0.101	(0.028)	**0.001**	0.096	(0.028)	**-0.004**
β_3	-5.500	-4.211	(0.285)	**1.289**	-5.507	(0.324)	**-0.007**
σ^2_μ	0.680	0.430	(0.141)	**-0.250**	0.673	(0.133)	**-0.007**

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(l|l)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j|kl}$; EB= average empirical bias; ESD= empirical standard deviation

much higher than those from the AGQ. Similar results are shown for the estimated empirical biases. With few exceptions, the empirical biases for AGQ are smaller than the PL in all three transitions.

The results show that AGQ provides better estimates for the treatment condition parameter (β_3) in all three transitions (nonsmoker to smoker, smoker to quitter, and quitter to smoker) and PL overestimates the treatment condition parameter in the smoker-to-quitter transition. Estimates obtained from the PL method provide higher biases in the quitter to smoker transition than in the other two transitions. Our results are consistent with other studies which show that both AGQ and PL have similar results in terms of the fixed parameter estimates and their estimated biases.

The estimates for the time-dependent covariate parameter (β_1) and the variable time parameter (β_2)(a proxy for grade) are similar in both methods. Except in quitter to smoker transition, all the estimates are close to the true values in all transitions.

Lastly, both methods underestimate the variance of the random intercept in all three transitions; however, the AGQ estimates are close to the true values for all three transitions.

In conclusion, Table 3.1 shows that both AGQ and PL provide similar results with respect to the biases for the fixed-effect parameter estimates. For the variance estimate of the random intercept, AGQ provides parameter estimates close to the true values. In contrast PL methods provide estimates of the variance of parameter estimates that are much lower than the true values.

3.4 Average Standardized Empirical Bias

The average standardized empirical bias (ASEB) of the parameter estimate was also calculated. This is defined as follows:

$$\text{ASEB}_k = \frac{\sum_{m=1}^{500} \left(\frac{\hat{\theta}_{mk} - \theta_k}{SE(\hat{\theta}_{mk})} \right)}{500}$$

where $SE(\hat{\theta}_{mk})$ is the model-based standard error of the estimated parameter. The average standardized bias is useful for understanding the impact of the bias on interval estimates and statistical tests. The results from the simulation are reported in Table 3.2.

The results show that generally the PL method consistently has higher average standardized empirical biases as compared to the AGQ method, with average standardized empirical biases being higher for the variance of the random intercept. Furthermore, for the variance estimate of the random intercept, AGQ provides a lower standardized bias for all three transitions.

In conclusion, Table 3.2 shows that in general, AGQ method provides smaller standardized biases for all the parameters compared to the PL method.

3.5 Root Mean Square Error

The mean squared error (MSE) of an estimate is defined as the squared empirical bias plus its corresponding variance. The MSE is also called the squared error of the estimate and increases

Table 3.2: Average standardized empirical bias for all three transitions based on 500 simulations*

Parameter	TRUE	PL Estimate	PL ESD	PL ASEB	AGQ Estimate	AGQ ESD	AGQ ASEB
Non-smoker to smoker transition							
β_0	-2.300	-2.327	(0.202)	**-0.156**	-2.295	(0.218)	**0.027**
β_1	0.200	0.198	(0.025)	**-0.116**	0.201	(0.022)	**0.026**
β_2	0.610	0.606	(0.020)	**-0.253**	0.608	(0.019)	**-0.120**
β_3	-4.100	-4.064	(0.213)	**0.203**	-4.071	(0.220)	**0.140**
σ^2_μ	0.680	0.438	(0.092)	**-1.751**	0.671	(0.079)	**-0.116**
Smoker to quitter transition							
β_0	0.800	0.923	(0.203)	**0.788**	0.800	(0.233)	**-0.002**
β_1	-0.100	-0.105	(0.024)	**-0.250**	-0.099	(0.028)	**0.021**
β_2	-0.300	-0.298	(0.018)	**0.193**	-0.304	(0.016)	**-0.269**
β_3	0.200	0.261	(0.208)	**0.343**	0.215	(0.147)	**0.072**
σ^2_μ	0.680	0.452	(0.111)	**-1.749**	0.675	(0.089)	**-0.067**
Quitter to smoker transition							
β_0	-1.700	-1.709	(0.266)	**-0.080**	-1.681	(0.326)	**0.064**
β_1	0.300	0.393	(0.038)	**1.451**	0.302	(0.064)	**0.031**
β_2	0.100	0.101	(0.028)	**0.052**	0.096	(0.028)	**-0.147**
β_3	-5.500	-4.211	(0.285)	**1.490**	-5.507	(0.324)	**-0.026**
σ^2_μ	0.680	0.430	(0.141)	**-1.934**	0.673	(0.133)	**-0.066**

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j|kl}$; ESD= empirical standard deviation; ASEB=average standardized empirical bias

in value as the variance of an estimate increases. This can be a useful diagnostic component for selecting an estimator, since small MSE values indicate a small variance and bias. The square root of the mean squared error (RMSE) is defined as the positive square root of the mean squared error:

$$RMSE_k = \sqrt{\frac{\sum_{m=1}^{500}(\hat{\theta}_{mk} - \theta_k)^2}{500}}$$

The RMSE provides a more easily interpretable measure of the MSE by transforming the MSE to the same scale as the parameter. The RMSE results are shown in Table 3.3.

Table 3.3: Estimated RMSE for all three transitions based on 500 simulations*

Parameter	TRUE	PL Estimate	PL ESD	PL RMSE	AGQ Estimate	AGQ ESD	AGQ RMSE
Non-smoker to smoker transition							
β_0	-2.300	-2.327	(0.202)	**0.204**	-2.295	(0.218)	**0.218**
β_1	0.200	0.198	(0.025)	**0.025**	0.201	(0.022)	**0.022**
β_2	0.610	0.606	(0.020)	**0.020**	0.608	(0.019)	**0.020**
β_3	-4.100	-4.064	(0.213)	**0.216**	-4.071	(0.220)	**0.222**
σ^2_μ	0.680	0.438	(0.092)	**0.259**	0.671	(0.079)	**0.080**
Smoker to quitter transition							
β_0	0.800	0.923	(0.203)	**0.237**	0.800	(0.233)	**0.233**
β_1	-0.100	-0.105	(0.024)	**0.025**	-0.099	(0.028)	**0.028**
β_2	-0.300	-0.298	(0.018)	**0.018**	-0.304	(0.016)	**0.016**
β_3	0.200	0.261	(0.208)	**0.216**	0.215	(0.147)	**0.147**
σ^2_μ	0.680	0.452	(0.111)	**0.254**	0.675	(0.089)	**0.089**
Quitter to smoker transition							
β_0	-1.700	-1.709	(0.266)	**0.266**	-1.681	(0.326)	**0.327**
β_1	0.300	0.393	(0.038)	**0.101**	0.302	(0.064)	**0.064**
β_2	0.100	0.101	(0.028)	**0.028**	0.096	(0.028)	**0.028**
β_3	-5.500	-4.211	(0.285)	**1.320**	-5.507	(0.324)	**0.324**
σ^2_μ	0.680	0.430	(0.141)	**0.287**	0.673	(0.133)	**0.134**

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j|kl}$;
ESD=empirical standard deviation; RMSE=root mean square error

The results show that generally AGQ method produces lower RMSE estimates for the treatment condition parameter (β_3) in all three transitions. Both methods (PL and AGQ) provide similar and higher estimates of the RMSE for all the parameters in the quitter-to-smoker transition. For AGQ the estimated RMSEs are smaller for all three transitions.

The RMSE estimate for the time-dependent covariates parameter (β_1) and variable time parameter (β_2) (a proxy for grade) are almost identical in both methods for non-smoker to smoker transition. The major difference can be seen in the variance estimate of the random intercept, where PL consistently has higher RMSE estimates as shown in Fig. 3.5. Figure 3.5 shows the RMSE estimates for the variance of the random intercept for all three transitions.

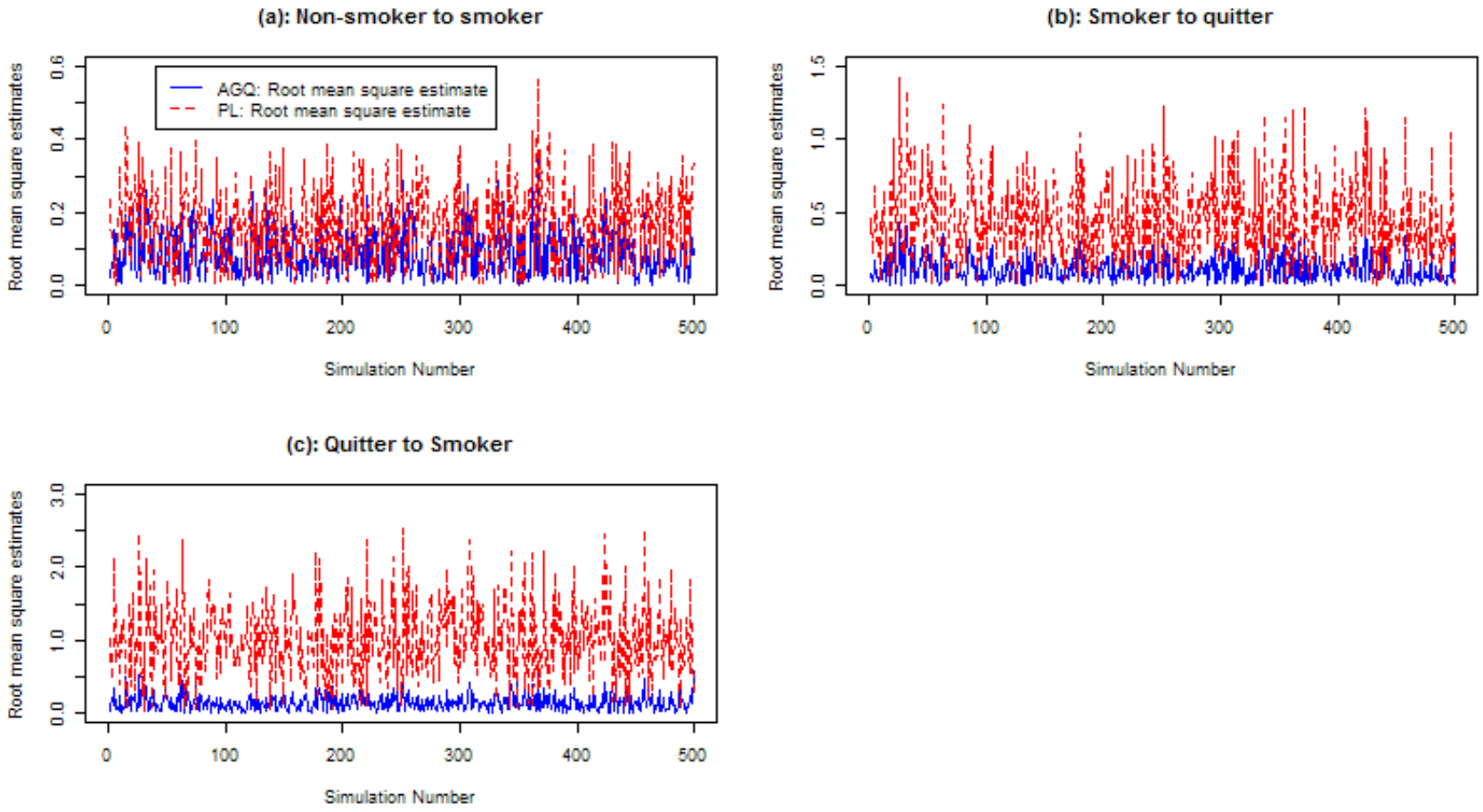
For both methods, the RMSE estimates for the variance of the random intercept are much higher in the quitter-to-smoker transition than in the other two transitions. As expected, AGQ provides a much smaller RMSE estimate.

In conclusion, Table 3.3 suggests that both methods provide similar RMSE results for the estimates of the regression parameters but for the variance of the random intercept, AGQ provides a much smaller RMSE.

3.6 Coverage Rates

Using the estimated parameters and their model based standard errors, 95% confidence intervals for 500 iterations were calculated for each parameter in each iteration. The coverage rate for a given method is defined as the ratio of the number of iterations in which the calculated confidence interval contains the true value of the parameter to the total number of iterations. For

Figure 3.5: RMSE for the variance of the random intercept



500 observations from a binomial distribution with probability of success equal to 0.95, a 95% probability interval would be (0.93,0.97). Any coverage rates that do not fall in this range should be considered as not agreeing with the nominal 95% rates. The nominal 95% coverage rates are reported in Table 3.4.

Based on these results, the coverage rates obtained from the AGQ method are much higher than the PL method. Except for the variance of the random effect in the quitter to smoker transition, all parameter estimates obtained from AGQ method have coverage rates more than 90%. In contrast the coverage rates for the variance of the random intercept obtained from PL method are close to 60% in all three cases.

Table 3.4: Nominal 95% coverage rates for parameters for all three transitions based on 500 simulations*.

		PL			AGQ		
Parameter	**TRUE**	**Estimate**	**ESD**	**cov**	**Estimate**	**ESD**	**cov**
Non-smoker to smoker transition							
Non-smoker to smoker transition							
β_0	-2.300	-2.327	(0.202)	**0.876**	-2.295	(0.218)	**0.930**
β_1	0.200	0.198	(0.025)	**0.828**	0.201	(0.022)	**0.970**
β_2	0.610	0.606	(0.020)	**0.828**	0.608	(0.019)	**0.962**
β_3	-4.100	-4.064	(0.213)	**0.882**	-4.071	(0.220)	**0.968**
σ^2_μ	0.680	0.438	(0.092)	**0.634**	0.671	(0.079)	**0.924**
Smoker to quitter transition							
β_0	0.800	0.923	(0.203)	**0.814**	0.800	(0.233)	**0.926**
β_1	-0.100	-0.105	(0.024)	**0.852**	-0.099	(0.028)	**0.920**
β_2	-0.300	-0.298	(0.018)	**0.826**	-0.304	(0.016)	**0.952**
β_3	0.200	0.261	(0.208)	**0.862**	0.215	(0.147)	**0.990**
σ^2_μ	0.680	0.452	(0.111)	**0.612**	0.675	(0.089)	**0.904**
Quitter to smoker transition							
β_0	-1.700	-1.709	(0.266)	**0.868**	-1.681	(0.326)	**0.932**
β_1	0.300	0.393	(0.038)	**0.806**	0.302	(0.064)	**0.964**
β_2	0.100	0.101	(0.028)	**0.840**	0.096	(0.028)	**0.964**
β_3	-5.500	-4.211	(0.285)	**0.862**	-5.507	(0.324)	**0.950**
σ^2_μ	0.680	0.430	(0.141)	**0.564**	0.673	(0.133)	**0.864**

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(l|l)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j|kl}$; ESD= empirical standard deviation; cov=coverage rates

3.7 Conclusion

In summary, this chapter compared two estimation methods for multilevel binary regression models in a longitudinal setting: PL and AGQ methods. These estimation methods are frequently used in the applied multilevel-modeling literature to estimate parameters in the multilevel logistic regression model. This simulation study shows that the AGQ method generally provides more accurate estimates than the PL method. It provides a smaller bias especially with the estimate of the variance of the random intercept. Our results indicate that with a three-level hierarchical model, the AGQ method can be used in the context of a transitional model, and the parameter estimates for the variance of the random intercept obtained are close to the true values. Coverage rates clearly show that parameter estimates obtained from the AGQ method are much superior than to the PL method for both fixed and random effect parameters.

In conclusion, while the AGQ method is slower than the PL (Breslow & Clayton, 1993) method, the parameter estimates from the PL method tend to be biased for binary dependent variables (e.g., Rodriguez & Goldman (1995, 2001)). Moreover, the PL method does not involve a likelihood which prohibits the use of likelihood-based inference such as likelihood ratio tests and likelihood-based confidence intervals. A further advantage of the AGQ method is that the precision can be increased by simply using more quadrature points (Raudenbush et al., 2000).

Finally, these simulation results confirmed that the AGQ method is preferred to the PL method when the goal is to estimate the variance of the random intercept in a complex hierarchical model. It can easily be used in standard software such as SAS and STATA. Based on this conclusion, Chapter 5 uses the AGQ method to estimate the parameters when dealing with missing data.

Chapter 4

Missing Values

In longitudinal studies individuals drop out for many reasons, creating added complications in a hierarchical model. These dropouts can have important consequences for the validity of the findings from any method of analysis. Standard statistical methods have been developed based on the assumption of complete datasets. When data are missing, the assumptions of these methods may be violated. For example, conclusions based on the assumption that the data are a random sample from a population may not be valid if the dropouts have certain characteristics. There is an added complication in a hierarchical setting because of the possibility of missing data at more than one level. For instance, at level one student information may be missing for several reasons: (i) a subject refuses to answer a question; (ii) a subject drops out of the study for whatever reason; (iii) the contact information was recorded incorrectly; or (iv) a subject misses the study visit. At level two, a school may drop out because of parental concerns, school work load, and sometimes time conflicts due to other surveys taking place in a given school. Missing data at level two are more problematic then missing data at level one. If a level-two unit has missing data, the

information from all individuals at that school is lost. Snijders & Bosker (1993) have discussed this in much more detail; the focus of this thesis is level-one missing data.

4.1 Terminology and Notation

Assume that in a longitudinal study there are N observations and the i^{th} subject is to be observed in the j^{th} cluster at times indexed by $t = 1, 2, ..., T$. Let the random variable Y be the $N \times 1$ response vector for all subjects. If n_{jt} is the the number of subjects in the j^{th} school at time t, then $N = \sum_{t=1}^{T} \sum_{j} n_{jt}$ is the total number of observations. Let X be a $N \times p$ design matrix for the fixed effects with parameter β, Z the $N \times J$ design matrix for the random effect, u the $J \times 1$ vector of random parameters, e the $N \times 1$ vector of errors for the subjects, D the $J \times J$ covariance matrix for the random effects, and Σ the $N \times N$ error covariance matrix. The model for a continuous random variable, Y, is

$$Y = X\beta + Zu + e \tag{4.1}$$

where

$$u \sim N(0, D)$$

$$e \sim N(0, \Sigma)$$

Furthermore, assume that $Y_i = (Y_i^o, Y_i^m)$ is a T-dimensional vector of all the scheduled measurements for subject i, where Y_i^o represents the observed part of Y_i and Y_i^m represents the missing part. Let $R_i = (r_{i1}, r_{i2}, ..., r_{iT})^T$ be the same length as Y_i and denote a vector of missing-data indicators for Y_i, where $R_{it} = 1$ if Y_{it} is missing and 0 otherwise. The full dataset (Y_i, R_i, X_i, Z_i)

consisting of the complete data together with the missing-data indicators has density function $f(Y_i, R_i|X_i, Z_i, \theta, \psi)$. Where $\theta = (\beta^T, \alpha^T)$ and ψ are vectors that describe the measurement and dropout processes. As usual β is the fixed-effect parameter vector and α is a variance components vector.

According to the Rubin (1976) taxonomy, the missing data or non-responses can be placed in three broad categories: (i) missing completely at random (MCAR); (ii) missing at random (MAR); or (iii) missing non-ignorable (MNAR). Each category is described below.

4.1.1 Missing Completely at Random (MCAR)

Under the MCAR mechanism, the probability of non-response does not depend on either observed or unobserved data. This implies that the missing-data indicator R is independent of both Y^o and Y^m. Mathematically, in the MCAR case

$$f(R|Y, X, Z) = f(R)$$

In an example related to the smoking survey, a subject may drop out of the study for reasons not related to his/her smoking behavior or other characteristics. Analyzing data from individuals with complete data may result in loss of power for our design, but the estimated parameters are not biased by the absence of data. The results obtained from this mechanism are always valid but in practice this assumption is too restrictive.

4.1.2 Missing at Random (MAR)

In the MAR case, the probability of non-response depends only on the observed part of the outcome or covariate vector, not on the unobserved components. Mathematically, in the MAR case

$$f(R|Y, X, Z) = f(R|Y^o, X, Z)$$

For example, perhaps males are more likely to drop out than females in a smoking study because they do not want to participate in a smoking cessation program. Under this mechanism a likelihood analysis using the correct model can provide valid inference because the joint likelihood can be factorized into two distinct parts: one for the complete data and one for the missing process (Laird, 1988).

4.1.3 Missing Not at Random (MNAR)

In the MNAR case, the probability of non-response depends on the unobserved observations. In practice, this is the most difficult process to handle. Mathematically, this process can be described as

$$f(R|Y, X, Z) = f(R|Y^o, Y^m, X, Z)$$

As an example, consider the following scenarios: 1. A non-smoker at $t-1$ may be more likely to have dropped out if he/she became a smoker between $t-1$ and t. 2. A smoker at $t-1$ who remains a smoker may be more likely to drop out than someone who quits between $t-1$ and t. In

these scenarios, individuals drop out from the study because of their future changes in behavior.

In this missing data mechanism a valid inference can be made only if the non-response process can be modeled explicitly. The analysis of such a process requires a joint model for the full data which contains responses for both observed and unobserved variables indicating the dropout process. To deal with the MNAR process Little & Rubin (1987) suggest using selection models or a pattern mixture model. These methods will be described later.

4.2 Dealing with Missing Data

Two common approaches to treating missing data in applied research are (i) to exclude the individual with missing observations from the statistical analysis or (ii) to estimate the missing values and use the estimated values in the analysis. Various methods to calculate an estimate of the missing values have been suggested and are described below.

4.2.1 Complete Case Analysis

Complete case analysis assumes that the data are MCAR and so those for whom data are missing are a random sample from the study population. In this procedure subjects who have missing data are excluded and only those subjects who have complete data over the course of the study are included. The procedure is simple to implement and may not cause problems in the statistical analysis if the data truly are MCAR. Under MCAR, complete case analysis yields unbiased parameter estimates and reasonably efficient results (Anderson et al., 1983; Roth & Switzer, 1995). However, Buck (1960) observed that ignoring the missing units and using only complete

cases for analysis results in a loss of potentially useful data. In addition, the reduced sample size can lead to a loss of error degrees of freedom which in turn yields a loss of statistical power and inefficient parameter estimation (Cohen & Cohen, 1983). Little & Rubin (1987) suggest that the bias in the estimates can be severe if the data are not MCAR.

4.2.2 Weighting Method

The weighting method adopts the idea of survey design weights, which are inversely proportional to the probability of selection of an individual. This method is commonly used in complex survey data with non-response (Yi & Cook, 2002). Once the weights are computed this method is simple to adopt and almost all standard software can incorporate these weights to estimate the model parameters. If the computed weight does not depend on the response variable then an unbiased estimate of the model parameter can be obtained. The computation of weights requires full information about the survey design and also the information related to the missing-value process. This method works well when the data are MAR, but is not needed with likelihood-based analysis. Most often the logistic model is used to obtain the weights by predicting the probability of non-response on response variables (Y_{ij}) given covariates (X_{ij}).

4.2.3 Mean Substitution

This method uses the sample mean of a variable to replace the missing value (Collins et al., 2001). It involves a single imputation not based on a predictive model. It treats the imputed values as observed values and does not consider the variation in the imputed values which results in an underestimate of variance. It yields an inconsistent estimate of some parameters even under

the MCAR assumption. In a binary case analysis, mean estimates can not replace the binary outcome. To estimate a valid response in the binary case, mean estimates are used to estimate the probability p and then a cutoff value p_0 is selected based either on the data or on experience with prior studies. If the probability p is greater than the cutoff point p_0 then the missing response variable takes the value 1, otherwise it is set to 0.

4.2.4 Last Observation Carried Forward (LOCF)

This method is commonly used in clinical trials (Molenberghs & Verbeke, 2005). Clinicians use this method based on the assumption that the subject profile is unchanged from the previous assessment. It also requires the strong assumption that the individual outcome profile remains at the level of the last observed measurement throughout the remainder of the follow-up. However, it is possible that the individual outcome changes as soon as the individual stops the treatment. This method can produce substantial biases in the estimator of treatment effects, inflated type I error rates of the associated tests, and coverage probabilities that are far from the nominal level (Molenberghs & Verbeke, 2005). In general this method requires serious sensitivity analysis and has a differential effect on the results, depending on the missing-data scenario and the method of analysis used (Little & Yau, 1996).

4.2.5 Maximum-Likelihood-Model-Based Procedures

The procedures outlined above are relatively simple to implement and may yield satisfactory results when the data are MCAR. However, their performance is unreliable and whether or not they have performed satisfactorily may not be easily determined (Little & Rubin, 1987). Maximum

likelihood procedures require the less restrictive MAR assumption. They provide unbiased parameter estimates under both MCAR and MAR. In addition, they provide more efficient estimates than the listwise and pairwise deletion under MCAR. However, these procedures are computationally intensive and until recently have been little used. As a result of vastly improved computing speed these procedures are gaining popularity among applied researchers. Rubin (1976) showed that if the separability condition (the model no longer depends on the unobserved data, at least in terms of the probability model, or inference can be based on only the marginal observed data density) is satisfied within the likelihood framework, ignorability is equivalent to the union of MAR and MCAR. Molenberghs et al. (2004) showed that linear mixed model estimates based on the likelihood approach are alternatives for complete case analysis and LOCF in MAR.

4.2.6 Multiple Imputation

Multiple imputation (MI), first proposed by Rubin in the early 1970s (Rubin, 1976) as a way to address survey non-response, addresses the issues associated with single imputation. It involves replacing missing values by M $(M \geq 2)$ imputed values to create M complete datasets. Furthermore, these multiple imputed datasets are analyzed using standard procedures assuming no missing data. The parameter estimates obtained from each dataset are aggregated using Rubin multiple imputation techniques (Rubin, 1976). The process of combining these results generated from multiple imputed datasets is independent of the analytic procedure used to estimate the parameters. The major advantages of multiple imputation, as indicated by Rubin (1987), are that complete data methods are used to analyze each complete dataset; moreover, the ability to use the data analyst's knowledge when handling the missing values is not only retained but actually

enhanced. In addition, multiple imputations allow users to reflect their uncertainty as to which values to impute. The disadvantages include: the time intensiveness of imputing five to ten data sets, the need to test models for each data set separately, and the need to recombine the model results into one summary. The procedures for MI are as follows:

- Missing data are filled in M times to generate M complete datasets;
- The M complete datasets are then analyzed by standard procedures;
- Finally, the results from the M complete data sets are combined for inference.

For example, if regression coefficients and their standard errors are estimated then an MI regression coefficient is computed by averaging across the M imputed datasets using the usual formula:

$$\hat{\gamma} = \frac{1}{M} \sum_{m=1}^{M} \hat{\gamma}'_m$$

where $\hat{\gamma}'_m$ is the regression estimate from the m^{th} imputed dataset. The estimated standard error for each parameter is comprised of the within-imputation variability, the between-imputation variability, and a correction factor to account for simulation error. The within-imputation variance is estimated as the mean of the estimated variances across the M imputation datasets:

$$\bar{U} = \sum_{m=1}^{M} \frac{\hat{\sigma^2_m}}{M}$$

The between-imputation variance is the sample variance of the estimates calculated by:

$$B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\gamma}'_m - \hat{\gamma} \right)^2$$

This value is multiplied by the correction factor $1 + \frac{1}{M}$ and added to the within-imputation variance. The overall estimated standard error associated with the regression coefficient is the positive square root of the total variance:

$$SE = \left[\bar{U} + (1 + \frac{1}{M})B\right]^{1/2}$$

Several techniques involved in MI are discussed by Rubin (1987), Little & Rubin (1987), and Schafer (1999).

4.2.7 Propensity Score Method

A propensity score is generally defined as the conditional probability of an individual having a missing-response measurement given a vector of observed covariates (W) (Rosenbaum & Rubin, 1983). In this procedure a propensity score is generated for each individual with missing values to indicate the probability of that observation being missing. The observations are then grouped based on these propensity scores, and Approximate Bayesian Bootstrap (ABB) imputation (Rubin, 1987) is applied to each group. This procedure does not have any distributional constraint on the missing variables. The steps are as follows:

- For each variable Y_{ijt}, $t = 2, ..., T$, create a corresponding indicator variable R_{ijt}

$$R_{ijt} = \begin{cases} 1 & \text{if } Y_{ijt} \text{ is missing} \\ 0 & \text{if } Y_{ijt} \text{ is observed} \end{cases}$$

- Fit a hierarchical model with a logit link function

$$logit(\pi_{ijt}) = W_{ijt}\psi + u_{0j}$$

where $\pi_{ijt} = Pr(R_{ijt} = 1|W_{ijt}, u_{0j})$ and $logit(\pi_{ijt}) = log\left(\frac{\pi_{ijt}}{1-\pi_{ijt}}\right)$.

- Use the conditional probability to create a propensity score for each observation to indicate the probability of it being missing.

- Sort the observations by propensity score and divide them into a fixed number of groups based on the propensity scores.

- Apply ABB to impute the missing response values in each group. In group k, let Y^o denote the n_1 observations with non-missing Y values and Y^m denote the n_0 observations with missing Y values. ABB first draws n_0 observations randomly with replacement from Y^o and uses these values as the n_o imputed values for the missing response vector, Y^m, which is combined with the non-missing data to create a new dataset Y^*.

- Repeat the process M times to create M new datasets.

- Analyze these M imputed datasets separately as complete datasets. Use a hierarchical logistic model to estimate the parameters. Combine the results from each analysis using techniques from Rubin (1987) as explained in Section 4.2.6. Finally, compare the parameter estimate with the complete dataset estimates and those from other imputed methods.

4.2.8 Predictive Mean Matching Methods

The predictive mean matching methods are similar to propensity score (PS) methods and are a combination of the hot-deck method (Little & Rubin, 2002) and the regression imputation method. The hot-deck method matches the cases based on the similarity between specific covariates X (Little & Rubin, 2002). The difference between the PS method and the predicted mean matching method is the relationship between the probability of a missing response variable. In the PS method, missingness as a function of the relevant covariates is modeled and the cases are divided into different groups based on the predicted probability of being missing. In the predicted mean matching method the relationship between the response variable per se and the relevant covariates is modeled for complete data, and the cases are divided into different groups based on the predicted value for the response variable.

Compared with the hot-deck method which matches the cases based on similarity between specific covariates X, the predicted mean matching method matches cases based on a linear combination of covariates. It also makes it possible to replace each missing value with several observed values, which in turn can account for the variation in the imputed values. The ABB is used in the predicted mean matching method to impute each missing value. The steps are as follows:

- Establish a predictive model for response variable Y based on X and Z for the complete cases:

$$logit(Pr(Y_{ijt} = 1|X, Z, u)) = X\beta + Zu.$$

- The predicted probability of Y being 1 is

$$logit(\hat{Pr}(Y_{ijt} = 1|X, Z, u)) = X\hat{\beta} + Z\hat{u}$$

  where $\hat{u}$ is the empirical Bayes estimate of u described in Chapter 2.

- Divide the cases into different groups based on the predicted probabilities (in this thesis, the predicted probabilities were divided based on quintiles).

- Apply ABB to impute the missing response values in each group. In group k, let Y^o denote the n_1 observations with non-missing Y values and Y^m denote the n_0 observations with missing Y values. ABB first draws n_0 observations randomly with replacement from Y^o and uses these values as the n_0 imputed values for the missing response vector, Y^m, which is combined with the non-missing data to create a new dataset Y^*. Repeat this process M times and create M imputed datasets.

- Analyze these M complete imputed datasets separately; combine the results using the Rubin procedure (Rubin, 1987) as explained in Section 4.2.6.

4.3 Methods for Nonignorable Missing Data

In practice the hypothesis of random dropout is essentially untestable; it cannot be verified or contradicted by examination of the observed data (Little & Rubin, 1987). If this assumption is doubtful alternative procedures should be developed, especially when the degree of departure

from MAR is thought to be severe. That is, one needs to model the joint distribution of longitudinal response and the dropout. From the likelihood point of view, two models based on different factorizations of the joint distribution are pattern mixture and selection models.

4.3.1 Pattern Mixture Models

Pattern mixture models were first defined by Little (1993), and first model the marginal distribution of the missing-data indicators, and then the conditional distribution of the observed data given the dropout pattern. The population of the observed data then becomes a mixture of distributions, weighted by the probabilities of the dropout patterns. Little (1995) defined two pattern mixture models for non-ignorable dropout: outcome-dependent dropout and random-effect-dependent dropout. In outcome-dependent models, subjects are grouped according to their dropout times and identifying restrictions are placed on the missing-value distributions for those groups (Little & Wang, 1996; Little, 1993; Molenberghs et al., 1998). In random-effect dropout models a random coefficient model is formulated with summaries of the dropout time included as a subject-level covariate (Fitzmaurice et al., 2001; Hedeker & Gibbons, 1997; Wu & Bailey, 1989). Little (1995) suggested that outcome-dependent models are appropriate when the reasons for dropout seem closely related to the response variable itself, whereas random-effect-dependent models ascribe dropout to an underlying process (such as the progression of a disease) which the outcome variable measures imperfectly.

Assume that in a longitudinal study there are N subjects, indexed by $i = 1, 2, ..., N$, and the i^{th} subject is to be observed T times, $t = 1, 2, ..., T$. Assume the conditional distribution of Y_{it} is Bernoulli given the covariate vectors X including both fixed and random effect. Then the model

can generally be expressed for Y_{it} as a conditional mixed model

$$f(Y_{it}|X_{it}, Z_{it}, u_i) = exp[Y_{it}\eta_{it} - log(1 + exp(\eta_{it}))];$$

where

$$\eta_{it} = X_{it}\beta + Z_{it}u_i$$

It is assumed that $Y_i = (Y_i^o, Y_i^m)$ is a T-dimensional vector of all the scheduled measurements for subject i, where Y_i^o represents the observed part of Y_i and Y_i^m represents the missing part. Let $R_i = (r_{i1}, r_{i2}, ..., r_{iT})^T$ be the same length as Y_i and denote a vector of missing-data indicators for Y_i, where $R_{it} = 1$ if Y_{it} is missing and 0 otherwise. We assume that the number of missing-data patterns will be equal to T, the number of time points, because of the monotone missing-data assumptions (as shown in Table 4.2). The full data (Y_i, R_i) consist of the complete data together with the missing data indicators, so $f(Y_i, R_i|X_i, W_i, \theta, \psi)$, the distribution of (Y_i, R_i) conditional on X_i and W_i can be written:

$$f(Y_i, R_i|X_i, W_i, \theta, \psi) = f(Y_i|R_i, X_i, \theta) \times f(R_i|W_i, \psi)$$

where $f(Y_i|R_i, X_i, \theta)$ models the conditional distribution of the data given the pattern of missing data and $f(R_i|W_i, \psi)$ models the distribution of the dropout pattern; W_i is a design matrix for the missing-data process; $\theta = (\beta^T, \alpha^T)^T$ is the vector of all the parameters described in Section 2.1 and ψ is a parameter vector that describes the dropout processes. As defined earlier in Section (2.1), β is a fixed-effect parameter vector and α denotes the vector of all variance and covariance parameters in $Eq.(2.7)$. The missing pattern for a subject is defined through the vector $M_i =$

$(m_{i1}, ..., m_{iP})^T$, where

$$m_{ip} = \begin{cases} 1 & \text{if individual } i \text{ belongs to pattern } p, \quad p = 1, ..., P \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

The subjects are divided into groups based on having a similar pattern M_{ip}. In pattern mixture models, the complete data set consists of P different missing-data patterns.

To explain further, consider a four-year study. If all subjects are included provided they have baseline measures, there are eight possible missing-data patterns. A typical data structure is illustrated in Table 4.1.

Table 4.1: Missing-data patterns

Pattern	Baseline	Year 1	Year 2	Year 3
1	O	O	O	O
2	O	O	O	$*$
3	O	O	$*$	O
4	O	$*$	O	O
5	O	O	$*$	$*$
6	O	$*$	O	$*$
7	O	$*$	$*$	O
8	O	$*$	$*$	$*$

$*$ missing observation: O non-missing observation

The fully specified PMM is always underidentified (Daniels & Hogan, 2000) because of the need to estimate many pattern-specific parameters, and constraints are needed to make the model identifiable. Either restriction methods must be set for some patterns, additional assumptions must be made, or information must be borrowed from the observed data. Little (1994) solves

this problem using identifying restrictions. This is done by setting inestimable parameters of the incomplete patterns to functions of the parameters describing the distribution of the individuals, known as completers, which have complete data at all the time points studied. In this aspect, the assumptions made in pattern mixture models are no less stringent than those made in selection models. However, pattern mixture models with identifying restrictions are much easier to work with than selection models (Section 4.3.2) and have the advantage that non-identifiable parameters are clearly specified.

Several restriction methods are used for pattern mixture models. These will be illustrated using the WSPP3 study and assuming that every individual falls in one of the seven missing-data patterns, conforming exactly to a monotone dropout pattern. The monotone missing-data patterns for the WSPP3 dataset are shown in Table 4.2. In all cases, predictive mean matching was used to impute the missing values. From this point on, the word $impute$ in this thesis will refer to imputation under the predictive mean matching method.

Table 4.2: A tabulation of possible Monotone missing-data patterns over seven years of assessments

Pattern	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
P1	O	O	O	O	O	O	O
P2	O	O	O	O	O	O	$*$
P3	O	O	O	O	O	$*$	$*$
P4	O	O	O	O	$*$	$*$	$*$
P5	O	O	O	$*$	$*$	$*$	$*$
P6	O	O	$*$	$*$	$*$	$*$	$*$
P7	O	$*$	$*$	$*$	$*$	$*$	$*$

$*$ missing observation; O non-missing observation

Complete Case Missing Value (CCMV) Restriction: Little (1993) proposed the complete case missing value restriction (CCMV) for the pattern mixture model. CCMV uses data from the subjects who are in pattern 1 to impute the missing observations for the remaining patterns. In this method, we discard cases with any missing values and unavailable information is always borrowed only from completers (Little, 1993). The advantage of this method is that it is simple to implement and a valid inference is obtained when dropout depends on the regressors (Little, 1993). A disadvantage is that it is an unnecessary waste of information to discard all the incomplete cases. This is especially true if the number of covariates is large, so eliminating individuals based on missing data can result in a considerable number of incomplete cases (Little, 1992). CCMV is considered to be a useful baseline method for comparison with other methods (Little, 1992) and leads to valid inference when the majority of subjects have complete data.

Steps for CCMV:

- P2: Impute the missing values at year 7 using all the observed cases in P1.
- P3: Impute the missing values at years 6 & 7 using all the observed cases in P1.
- P4: Impute the missing values at years 5, 6, & 7 using all the observed cases in P1.
- P5: Impute the missing values at years 4, 5, 6, & 7 using all the observed cases in P1.
- P6: Impute the missing values at years 3, 4, 5, 6, & 7 using all the observed cases in P1.
- P7: Impute the missing values at years 2, 3, 4, 5, 6, & 7 using all the observed cases in P1.

Available Case Missing Value (ACMV) Restrictions: ACMV methods use the largest sets of available cases for estimating individual parameters. An advantage of the ACMV methods is that they make use of the incomplete cases in a plausible way. A disadvantage is that the estimated covariance matrix of the covariates is not necessarily positive definite (Little, 1992). Little (1992) found ACMV estimates to be more accurate than CCMV estimates, with the exception that ACMV estimates are less successful for datasets that contain high multicollinearity among the independent variables.

Steps for ACMV:

- P2: Impute the missing values at year 7 using all the observed cases in P1.
- P3: Impute the missing values at year 6 using all the observed cases in P1 and P2. Impute the missing values at year 7 using the observed cases in P1 and the imputed values for year 6 in P3.
- P4: Impute the missing values at year 5 using all the observed cases in P1, P2, and P3. Impute the missing values at year 6 using all the observed cases in P1 and P2, and the imputed values for year 5 in P4. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P4.
- P5: Impute the missing values at year 4 using all the observed cases in P1, P2, P3, and P4. Impute the missing values at year 5 using all the observed cases in P1, P2, and P3, and the imputed values for year 4 in P5. Impute the missing values at year 6 using all the observed cases in P1 and P2, and the imputed values for year 5 in P5. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P5.

- P6: Impute the missing values at year 3 using all the observed cases in P1, P2, P3, P4, and P5. Impute the missing values at year 4 using all the observed cases in P1, P2, P3, and P4, and the imputed values for year 3 in P6. Impute the missing values at year 5 using all the observed cases in P1, P2, and P3, and the imputed values for year 4 in P6. Impute the missing values at year 6 using all the observed cases in P1 and P2 and the imputed values for year 5 in P6. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P6.

- P7: Impute the missing values at year 2 using all the observed cases in P1, P2, P3, P4, P5, and P6. Impute the missing values at year 3 using all the observed cases in P1, P2, P3, P4, and P5, and the imputed values for year 2 in P7. Impute the missing values at year 4 using all the observed cases in P1, P2, P3, and P4, and the imputed values for year 3 in P7. Impute the missing values at year 5 using all the observed cases in P1, P2, and P3, and the imputed values for year 4 in P7. Impute the missing values at year 6 using all the observed cases in P1 and P2, and the imputed values at year 5 in P7. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P7.

Neighboring Case Missing Value (NCMV) Restrictions This restriction makes use of the available data from subjects in the neighboring patterns. This restriction assumes that the subjects who drop out in a given pattern are similar to those in neighboring patterns. The NCMV methods borrow information from the closest available pattern (Kenward et al., 2003).

Steps for NCMV:

- P2: Impute the missing values at year 7 using all the observed cases in P1.

- P3: Impute the missing values at year 6 using all the observed cases in P2. Impute the missing values at year 7 using the observed cases in P1 and the imputed values in P2.

- P4: Impute the missing values at year 5 using all the observed cases in P3. Impute the missing values at year 6 using all the observed cases in P2, and the imputed values for year 5 in P4. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P4.

- P5: Impute the missing values at year 4 using all the observed cases in P4. Impute the missing values at year 5 using all the observed cases in P3 and the imputed values for year 4 in P5. Impute the missing values at year 6 using all the observed cases in P2 and the imputed values for year 5 in P5. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P6.

- P6: Impute the missing values at year 3 using all the observed cases in P5. Impute the missing values at year 4 using all the observed cases in P4 and the imputed values for year 3 in P6. Impute the missing values at year 5 using all the observed cases in P3 and the imputed values for year 4 in P6. Impute the missing values at year 6 using all the observed cases in P2 and the imputed values for year 5 in P6. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P6.

- P7: Impute the missing values at year 2 using all the observed cases in P6. Impute the missing values at year 3 using all the observed cases in P5 and the imputed values for year 2 in P7. Impute the missing values at year 4 using all the observed cases in P4 and the imputed values for year 3 in P7. Impute the missing values at year 5 using all the observed

cases in P3 and the imputed values for year 4 in P7. Impute the missing values at year 6 using all the observed cases in P2 and the imputed values for year 5 in P7. Impute the missing values at year 7 using all the observed cases in P1 and the imputed values for year 6 in P7.

The steps for pattern mixture models are:

- Select an identification method of choice (CCMV, ACMV, NCMV).
- Based on the identification method and the transition being considered, estimate the conditional distribution of the unobserved outcomes, given the observed outcomes.
- Draw multiple imputations for the unobserved components, based on the predictive mean matching method.
- Analyze the multiple imputed datasets using the method of choice.

4.3.2 Two-Stage Heckman Selection Model

Selection models have been used most often in the econometrics literature. Heckman (1979) described the selection model in an application to the estimation of the labor supply function, and used the model to correct the sample selection bias which can arise for many reasons. The selection model specifies a model for the missing-data mechanism by factoring the joint distribution as follows:

$$f(Y_i, R_i|X_i, Z_i, \theta, \psi) = f(Y_i|X_i, Z_i, \theta) \times f(R_i|Y_i, X_i, Z_i, \psi) \tag{4.3}$$

where $f(Y_i, R_i|X_i, Z_i, \theta, \psi)$ is the joint distribution of Y and R, $f(Y_i|X_i, Z_i, \theta)$ models the complete data, and $f(R_i|Y_i, X_i, Z_i, \psi)$ models the missing-data mechanism. The joint distribution of (Y^o, R) is obtained by integrating out Y^m from the right-hand side of the equation:

$$\begin{aligned} f(Y^o, R) &= \int f(Y^o, Y^m) \times f(R|Y^o, Y^m) dY^m \\ &= \int f(Y^m|Y^o) \times f(Y^o) \times f(R|Y^o, Y^m) dY^m \\ &= f(Y^o) \int f(Y^m|Y^o) \times f(R|Y^0, Y^m) dY^m \\ &= f(Y^o) E_{(Y^m|Y^o)} \left[f(R|Y^o, Y^m) \right] \end{aligned}$$

The above equation shows that the selection model explicitly specifies the dependency of the missing-data mechanism on its corresponding missing value which in fact is a limitation in the selection model. The primary advantage of using the selection model is that it directly models the marginal distribution of the response Y and the dropout process conditional on Y. The Heckman model was originally designed for normally distributed variables but it can easily be extended to non-normal data as well. Later we will describe how to use the Heckman procedure for hierarchical binary data.

Consider first the situation where we have a random sample of N observations, with response variables $Y = (Y_1, Y_2, ..., Y_N)^T$ and assume there exists a second variable $V = (V_1, V_2, ..., V_N)^T$ that contains information about the missingness in Y. Our objective is to create a predictive model for Y using information on V. To develop a valid predictive model for Y, we first need to construct a predictive model for V.

Assume that Y obeys the regression model

$$Y = X\beta + \varepsilon_1 \tag{4.4}$$

and V obeys the regression model

$$V = W\psi + \varepsilon_2, \tag{4.5}$$

where $Y = (y_i)$, $V = (v_i)$, $X = (x_{ip})$, $W = (w_{iq})$, $i = 1, 2, 3, ...N$, $p = 1, 2, ...P$, and $q = 1, 2, ...Q$; β and ψ are $P \times 1$ and $Q \times 1$ vector of parameters respectively; and ε_1 and ε_2 are random error terms, with

$$\begin{aligned} E(\varepsilon_{1i}) &= 0 \qquad & E(\varepsilon_{2i}) &= 0 \\ var(\varepsilon_{1i}) &= \sigma_{11} \qquad & var(\varepsilon_{2i}) &= \sigma_{22} \end{aligned}$$

$$cov(\varepsilon_{1i}, \varepsilon_{2i}) = \begin{cases} \sigma_{12} & i = i' \\ 0 & i \neq i' \qquad i, i' = 1, 2, ..., N \end{cases}$$

For the regression function of Y the expected value is $E(Y|X) = X\beta$. However, because of selection bias caused by missing data in Y the regression function for the response variable Y can be written as

$$E(Y|X, S) = X\beta + E(\varepsilon_1|S) \tag{4.6}$$

where S represents the selection criteria. If there are no missing values or if the missing data are MCAR then the conditional expectation $E(\varepsilon_1|S) = 0$ which means there is no selection bias.

Since it was assumed that $(\varepsilon_1, \varepsilon_2)$ has a normal distribution, using well-known results (Johnson & Wichern, 2001) for the conditional distribution for the bivariate normal distribution (see Appendix A), we have

$$E(\varepsilon_{1i}|\varepsilon_{2i} \geq -W_i^T\psi) = \frac{\sigma_{12}}{\sqrt{\sigma_{22}}}\lambda_i \tag{4.7}$$

$$E(\varepsilon_{2i}|\varepsilon_{2i} \geq -W_i^T\psi) = \frac{\sigma_{22}}{\sqrt{\sigma_{22}}}\lambda_i \tag{4.8}$$

where $\lambda_i = \dfrac{\phi(C_i)}{1 - \Phi(C_i)}$, ϕ is the standard normal probability density function, Φ is the standard normal cumulative density function, and $C = \dfrac{-W_i^T\psi}{\sqrt{\sigma_{22}}}$. Economists call the parameter λ the inverse Mill's ratio or the hazard rate. The final model, then, is

$$E(Y|X, S) = X\beta + \frac{\sigma_{12}}{\sqrt{\sigma_{22}}}\lambda \tag{4.9}$$

The general steps estimating the parameters in the selection model are:

- Estimate the parameter ψ in the selection model using Equation (4.5)
- Use the estimated $\hat{\psi}$ from Equation (4.5) to estimate λ_i
- Include $\hat{\lambda}_i$ in Equation (4.9) and estimate parameters β and the coefficient for $\hat{\lambda}$.

4.3.3 The Selection Model with Correlated Binary Response Data

Assume that $(Y_{ijt|kl}, R_{ijt|kl})$ is a pair of binary indicators. $Y_{ijt|kl}$ is an indicator of a transition from state k to state l for subject i in cluster j at time t and $R_{ijt|kl}$ is an indicator of whether that observation is missing. Thus, define an indicator variable $Y_{ijt|kl}$ such that, for $t \geq 2$

$$Y_{ijt|kl} = \begin{cases} 1 & \text{if } Q_{ij,t} = l | Q_{ij,t-1} = k, X_{ijt}, \theta_{kl} \\ 0 & \text{if } Q_{ij,t} = k | Q_{ij,t-1} = k, X_{ijt}, \theta_{kl} \end{cases} \tag{4.10}$$

where $Q_{ij,t} = k$ denotes the status of the i^{th} subject in cluster j at time t with k possible states; $i = 1, 2, ..., n_j$, $j = 1, 2, ..., J$, $t = 1, 2, ..., T$, and $k = 1, 2, ...K$. We assume that the evolution of the status satisfies a first-order Markov chain with transitional probability from state k to l defined as

$$\begin{aligned} p_{ijt}(l|k) &= Pr(Q_{ij,t} = l | Q_{ij,t-1} = k, X_{ijt}, \theta_{kl}) \\ &= Pr(Y_{ijt|kl} = 1 | X_{ijt}, \theta_{kl}), \end{aligned}$$

where θ_{kl} denotes the collection of all the parameters and $l, k = 1, 2, 3$.

For the i^{th} subject with previous state k, we use the generalized logit model to model the transition from state k to l.

$$logit(Pr(Y_{ijt|kl} = 1)) = X_{ijt}\beta_{kl} + u_{0j|kl} \tag{4.11}$$

where $u_{0j|kl} \sim N(0, \sigma^2_{u|kl})$

Each individual in the sample could pass through several states during the period of observation. Each model for Y and R is state specific, that is, observations and parameters in the model depend on the state the subject is in at times $(t-1)$ and t.

In what follows, we drop the subscripts relating to state for simplicity. The response variable of interest is observed only if $R_{ijt} = 0$. Because of the selection procedure, the sample information allows us to model $Pr\{Y_{ijt} = 1 | X_{ijt}, R_{ijt} = 0\}$ and $Pr\{R_{ijt} = 1 | X_{ijt}\}$ but not $Pr\{Y_{ijt} =$

$1|X_{ijt}\}$. It is assumed that the binary indicator Y_{ijt} is related to a continuous latent variable through the observation rule $Y_{ijt} = I(Y^*_{ijt} > 0)$ and that the latent random variables Y^*_{ijt} obey the regression model

$$Y^*_{ijt} = X_{ijt}\beta^* + u^*_{0j} + \varepsilon_{1ijt}, \tag{4.12}$$

where ε_{1ijt} are $i.i.d.$ errors distributed independently of the X_{ijt} and conditional on a random effect which is assumed to have zero mean and finite variance. Similarly, R_{ijt} is related to a continuous latent variable through the observation rule $R_{ijt} = I(R^*_{ijt} > 0)$ and it is assumed that the latent random variable R^*_{ijt} obeys the regression model

$$R^*_{ijt} = W_{ijt}\psi^* + u^*_{0j} + \varepsilon_{2ijt}, \tag{4.13}$$

where ε_{2ijt} are $i.i.d.$ errors distributed independently of the X_{ijt}, conditional on random effect which is assumed to have mean and finite variance. The errors, ε_{1ijt} and ε_{2ijt} are assumed to be correlated with correlation coefficient ρ, where

$$\begin{aligned} E(\varepsilon_{1ijt}) &= 0 \qquad & E(\varepsilon_{2ijt}) &= 0 \\ var(\varepsilon_{1ijt}) &= \sigma_{11} \qquad & var(\varepsilon_{2ijt}) &= \sigma_{22} \end{aligned}$$

$$cov(\varepsilon_{1ijt}, \varepsilon_{2ijt}) = \begin{cases} \sigma_{12} & i = i' \\ 0 & i \neq i' \qquad i, i' = 1, 2, ..., n_j \end{cases}$$

where

$$\begin{Bmatrix} \varepsilon_{1ijt} \\ \varepsilon_{2ijt} \end{Bmatrix} = N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{bmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}\sigma_{22}} \\ \rho\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} \end{bmatrix} \right]$$

For the regression function of Y^*_{ijt} the expected value is $E(Y^*_{ijt}|X_{ijt}, u^*_{0j}) = X_{ijt}\beta^* + u^*_{0j}$. Because of selection bias the regression function for the latent response variable Y^*_{ijt} can be written as

$$E(Y^*_{ijt}|X_{ijt}, u^*_{0j}, S) = X_{ijt}\beta^* + u^*_{0j} + E(\varepsilon_{1ijt}|S) \tag{4.14}$$

where S represents the selection criteria. If there are no missing values or if the missing data are MCAR then the conditional expectation $E(\varepsilon_{1ijt}|S) = 0$ which means there is no selection bias. Since it was assumed that $(\varepsilon_{1ijt}, \varepsilon_{2ijt})$ has a normal distribution, we have (see Appendix A)

$$E[\varepsilon_{1ijt}|\varepsilon_{2ijt} \geq -(X_{ijt}\beta^* + u^*_{0j})] = \frac{\sigma_{12}}{\sqrt{\sigma_{22}}}\lambda_{ijt} \tag{4.15}$$

$$E[\varepsilon_{2ijt}|\varepsilon_{2ijt} \geq -(W_{ijt}\psi^* + u^*_{0j})] = \frac{\sigma_{22}}{\sqrt{\sigma_{22}}}\lambda_{ijt} \tag{4.16}$$

where $\lambda_{ijt} = \dfrac{\phi(C_{ijt})}{1 - \Phi(C_{ijt})}$, ϕ is the standard normal probability density function, Φ is the standard cumulative density function and $C = \dfrac{-(W_{ijt}\psi^* + u^*_{0j})}{\sigma_{22}}$. The final model would include

$$E(Y^*_{ijt}|X_{ijt}, S) = X_{ijt}\beta + u^*_{0j} + \frac{\sigma_{12}}{\sqrt{\sigma_{22}}}\hat{\lambda_{ijt}} \tag{4.17}$$

Equation (4.17) is not identified because of σ_{22}. Equation (4.13) can be multiplied by any positive number without affecting any of the observed cases. The sign of $\frac{R^*_{ijt}}{\sigma_{22}}$ is the same as that of R^*_{ijt}. So the implied value of R_{ijt} is unaffected. Therefore, the variance, σ_{22}, is unidentified and can

be set to any arbitrary value. A convenient normalization, σ_{22} =1 (Dubin & Rivers, 1990), is used to estimate Equation (4.13) and is used in Equation (4.17).

In the two-stage Heckman model, we first estimate all the parameters related to R^*_{ijt} using Eq. (4.13) and then estimate λ_{ijt}, the effect of the selection bias from the non-random selection rule. A probit model can be used here because of the normality assumption for the continuous latent random variable. In the second stage we use the extra predictor $\hat{\lambda}_{ijt}$ in the model and estimate parameters using Equation (4.17). In the second stage, a logit model is used to compare our results with the other methods used in this thesis. The probit model and logit model tend to produce extremely similar results (Dubin & Rivers, 1990), except when there is a lot of data in the tails.

The general steps for the selection model are:

- First fit the probit model for R using AGQ to estimate the parameters in Equation (4.13).
- Estimate the selection bias $\lambda = \dfrac{\phi(C)}{1-\Phi(C)}$, where $C = \dfrac{-(W_{ijt}\psi^* + u^*_{0j})}{\sigma_{22}}$.
- Include $\hat{\lambda}$ in model (4.17) to estimate all the parameters included in the model by fitting a hierarchical logistic model.
- Using the parameter estimates from previous steps, use the predictive mean matching method to impute the non-response.
- After obtaining the m imputed datasets, analyze each dataset and then combine the resulting estimates using Rubin's formula described in Section 4.2.

4.3.4 Bayesian Hierarchical Model

In this section we introduce the Bayesian approach. The fundamental difference between Bayesian analysis and classical analysis is that in the latter the parameters are assumed to be fixed whereas in the former the parameters are assumed to be random variables. This allows us to define the distribution of parameters in the Bayesian setting. Bayesian inferences on parameters are based on the posterior distribution of the parameters. The posterior distribution is the conditional density of the parameters given the observed data. To derive a posterior distribution, prior distributions must initially be assumed for the parameters. The prior distribution is the parameters' marginal distribution, which is defined in terms of other parameters, known as hyper-parameters. Assume that the prior distribution for parameter θ is $\Pi(\theta)$ and $f(Y|\theta)$ is the distribution of the data, Y, given θ. If $\theta \in \Theta$, where Θ is the complete parameter space, then using Bayes' rule, the posterior distribution is defined as (e.g. (Ghosh et al., 2006)):

$$\Pi(\theta|Y) = \frac{\Pi(\theta)f(Y|\theta)}{\int_{\Theta}\Pi(\theta)f(Y|\theta)d\theta} \tag{4.18}$$

Using a Bayesian approach, missing data can be considered random variables and can therefore have posterior distributions. The unknown values of the missing data Y^m can be predicted using the observed data Y^o and inferences about θ can be made using the posterior predictive distribution (Gelman et al., 1995):

$$\Pi(\theta|Y^o) \propto \Pi(\theta)f(Y^o|\theta) \tag{4.19}$$

The observed-data posterior density can be obtained by

$$\Pi(\theta|Y^o) = \int \Pi(\theta|Y^o, Y^m)\Pi(Y^m|Y^o)dY^m \tag{4.20}$$

$$\propto \int f(Y|\theta)\Pi(\theta)\Pi(Y^m|Y^o)dY^m \tag{4.21}$$

where $f(Y|\theta)$ is the complete-data likelihood and $\Pi(\theta|Y^o, Y^m) \propto f(Y|\theta)\Pi(\theta)$ is the complete-data posterior density for θ. Since

$$\Pi(Y^m|Y^o) = \int_\Theta \Pi(Y^m|Y^o, \theta) \times \Pi(\theta|Y^o)d\theta \tag{4.22}$$

after substituting Eq. (4.22) into Eq. (4.21) the observed-data posterior density becomes

$$\Pi(\theta|Y^o) \propto \int f(Y|\theta)\Pi(\theta)[\int_\Theta \Pi(Y^m|Y^o, \theta)\Pi(\theta|Y^o)d\theta]dY^m \tag{4.23}$$

A major limitation of a Bayesian approach is that obtaining the posterior distribution requires the integration of high-dimensional functions (Evans & Swartz, 1995) and can be computationally difficult. To solve this problem, Markov Chain Monte Carlo (MCMC) techniques are used to approximate the posterior distribution. This method is based on iteratively drawing samples from a Markov chain. Under certain conditions, the distribution of the samples becomes closer to the posterior distribution $\Pi(\theta|Y)$ (Gelman et al., 1995). After a finite number of iterations the MCMC will produce random samples from a stationary distribution which approximates the posterior distribution. In some situations, a large number of iterations of the MCMC is required before convergence occurs. The Metropolis-Hastings algorithm and the Gibbs' sampler

are the two most common methods used in MCMC algorithms. A detailed description of these algorithms can be found in Gelman et al. (1995).

Advances in computing have made it possible to analyze complex Bayesian hierarchical models using numerical approximation and simulation techniques. This has increased the range of problems and the complexity of analysis that now can be performed using Bayesian techniques. The WinBUGS (windows-based Bayesian inference using Gibbs sampling) software package implements a computational technique that uses MCMC simulation for Bayesian hierarchical modeling (Lunn et al., 2000). WinBUGS was produced by the BUGS (Bayesian inference using Gibbs sampling) project, which is a joint program of the Medical Research Council's Biostatistics Unit in Cambridge and the Department of Epidemiology and Public Health of Imperial College at St. Mary's Hospital in London. The developers of the software warn users to exercise caution when using the software because it is not perfect and MCMC is inherently less robust than analytical statistical methods.

4.3.5 Bayesian Pattern Mixture Models

In Bayesian approaches, pattern mixture models factor the joint distribution of indicators for missing data patterns and the conditional distribution of the data given these patterns (Gatsonis et al., 2002). The parameters of the conditional data model are typically under-identified; assumptions regarding the missing data mechanism can lead to models that allow for the estimation of these parameters.

In this section, we define pattern mixture models under the Bayesian framework. Rubin (1976) introduced this idea by classifying subjects as respondents or nonrespondents and specified a

prior distribution for the nonrespondents' mean centered around the respondents' mean. In this way he was able to calculate a probability interval for the mean of the entire population, including both respondents and nonrespondents.

The Bayesian pattern mixture model requires prior information for each missing-data pattern. As the number of patterns increases, more assumptions are needed for both the identified and unidentified parameters. This can be problematic for researchers because they need to correctly specify prior information otherwise their results will be biased. Typically, researchers will use only two missing-data patterns, distinguishing subjects with complete data versus those with at least one missing data point (e.g. Kaciroti et al. (2006)).

In the analysis of WSPP3 data (Chapter 6), this thesis uses a similar approach by specifying two missing-data patterns. The first pattern includes all subjects with complete data for all time points. The second pattern includes those subjects with at least one missing data point. This method is similar to the CCMV method described in Section 4.3.1.

For comparison purposes, the seven patterns defined in Table 4.2 are also used along with the two-pattern model, for the simulated datasets only. For model fitting, the same hierarchical model was used for all patterns but the model parameters were allowed to differ. Results from the Bayesian analyses are compared with the methods that used all seven patterns.

For the two-pattern model, assume that $\theta^{(1)}$ is the identified parameter vector corresponding to pattern 1 that includes all subjects with complete data for all time points and $\theta^{(2)}$ is the unidentified parameter vector corresponding to pattern 2 that includes those subjects with at least one missing data point. In the two-pattern model, a hierarchical logit model is used (again dropping the subscript representing the states for simplicity)

$$logit(\pi_{ijt}^{(p)}) = logit(Pr(Y_{ijt}^{(p)} = 1|X_{ijt}, \theta^{(p)})) = \beta_0^{(p)} + \beta_1^{(p)} X_{1ij,t-1} + \beta_2^{(p)} t + \beta_3^{(p)} C_j + (u_{0j}), \tag{4.24}$$

$$\text{with } u_{0j} \sim N(0, \sigma_u^2), \text{ and}$$

where p represents the patterns such that $p = 1, 2$ correspond to pattern 1 and pattern 2, and all other symbols are as defined in previous chapters.

As mentioned earlier, because of the missing data, the pattern mixture model is underidentified. Thus some restriction, or prior information on the parameters, is needed in the model to identify the parameters. Little & Rubin (2002) defined the prior distribution $\Pi(\theta^{(2)}|\theta^{(1)})$, on the parameters of the missing pattern ($p = 2$) conditioned on the parameters of the complete observed pattern ($p = 1$). Our approach is somewhat different but similar to that of Kaciroti et al. (2006). Kaciroti et al. (2006) constructed the prior distribution $\Pi(\theta^{(2)}|\theta^{(1)})$ by relating the distribution of the missing data to the distribution of the observed data. Kaciroti et al. (2006) used this method for ordinal outcome data, where they related the cumulative odds of the missing data with the cumulative odds of the observed data using a Bayesian approach. In his model,

$$\pi_{ijt,l}^{(p)} = Pr(Y_{ijt}^{(p)} \leq l|X_{ijt}, \theta^{(p)}) \tag{4.25}$$

such that for $p = 1, 2$,

$$\frac{\pi_{ijt}^{(2)}}{1 - \pi_{ijt}^{(2)}} = \hat{h}^{ijt,(2)} \frac{\pi_{ijt}^{(1)}}{1 - \pi_{ijt}^{(1)}}, \tag{4.26}$$

where $\hat{\hbar}^{ijt,(2)}$ is the cumulative odds ratio statistic between the missing data pattern and complete-data pattern and can be considered as a measure of the departure from random dropout. In our case such as for the binary data, $\hat{\hbar}^{ijt,(2)}$ is the odds ratio statistic between the missing data pattern and the complete data pattern which measures the intensity of departure from missing at random (MAR).

In our approach, for the two-pattern model, a hierarchical logit model is used for pattern 1 that includes all subjects having complete data for all time points.

$$logit(Pr(Y_{ijt}^{(1)} = 1|X_{ijt}, \theta^{(1)})) = \beta_0^{(1)} + \beta_1^{(1)} X_{1ij,t-1} + \beta_2^{(1)} t + \beta_3^{(1)} C_j + (u_{0j}) \tag{4.27}$$

$$u_{0j} \sim N(0, \sigma_u^2),$$

where $\theta^{(1)}$ denotes the collection of all the regression parameters related to pattern 1. Flat priors are assumed for all the parameters, $\theta^{(1)}$, for pattern 1 and a gamma prior is assumed for the variance of the random effect. For the prior information for pattern 2, a model for the missing data indicator(R) is developed first, such that,

$$logit(Pr(R_{ijt} = 1|X_{ijt}, R_{ij,t-1} = 0, \psi)) = \beta_0^m + \beta_1^m X_{1ij,t-1} + \beta_2^m t + \beta_3^m C_j + (u_{0j}), \tag{4.28}$$

where ψ denotes the collection of all the regression parameters related to the missing data indicator. Flat priors are also assumed for all the parameters ψ related to the missing data indicator (R) and a gamma prior is assumed for the variance of the random effect.

Similar to the approach of Kaciroti et al. (2006), parameter estimates from the missing data indicator model (Equation 4.28) are used to provide prior information for the pattern 2 model, such that $\pi(\theta^{(2)}|\wp) = N(\hat{\wp}, 1)$. Several variance estimates were tried but the convergence of the model is relatively fast assuming a unit variance. For pattern 2, a hierarchical logit model is used,

$$logit(Pr(Y_{ijt}^{(2)} = 1|X_{ijt}, \theta^{(2)})) = \beta_0^{(2)} + \beta_1^{(2)} X_{1ij,t-1} + \beta_2^{(2)} t + \beta_3^{(2)} C_j + (u_{0j}) \tag{4.29}$$

$$u_{0j} \sim N(0, \sigma_u^2)$$

where $\theta^{(2)}$ denotes the collection of all the regression parameters related to pattern 2. The posterior estimates obtained from both patterns are combined using Rubin's formula described earlier.

Finally, for comparison purposes, all seven patterns were also used under the Bayesian pattern mixture model. The same methods are used with seven-pattern models. For example, for pattern 2 where data are missing for the last time point, parameter estimates from the missing data model are used to provide the prior information for the parameters in the model for pattern 2 data. For pattern 3 where data are missing for the last two time points, parameter estimates from the missing data model are used to provide the prior information for the parameters in the pattern 3 model. Similarly, estimates for all the other patterns are obtained. In all models, hierarchical designs are incorporated by introducing a normal random-effects distribution with the same gamma prior assumed for the variance of the random effect.

The posterior estimates obtained from all patterns are combined using Rubin's formula described earlier. Results from the Bayesian analysis for both two and seven patterns are compared with the ACMVPM method that uses all seven patterns in Section 5.9.

In Chapter 5, we compare the results from a Bayesian analysis with those from the pattern mixture model described in Section 4.3.1, using simulated data. In Chapter 6, we use WinBUGS to perform a Bayesian hierarchical analysis of the WSPP3 dataset under the non-ignorable missing-data mechanism.

4.4 Summary

This chapter described the handling of missing data for analyzing longitudinal clustered data as well as various imputation approaches such as predictive mean matching methods under the pattern mixture, selection, and Bayesian pattern mixture models. In Chapter 5 simulation studies are performed to examine the performance of the parameter estimates and their standard errors obtained from these imputation methods.

Chapter 5

Simulation

5.1 Simulation Model and Parameter Values

The purpose of the simulation was to extend the method for handling missing data in hierarchical models by exploring the performance of different techniques under various missing-data conditions. The focus was on the monotone missing-data pattern. Monotone missing data occur when responses are available for an individual until a certain time and then missing for all subsequent times. Data with missing observations were generated and various simulation settings were manipulated, including the missing-data rate and missing-data mechanism. This section compares the performance of standard methods (i.e., complete case analysis and last observation carried forward) under MCAR, MAR, and MNAR missing-data mechanisms against the pattern mixture model (PM) and the selection model (SM). The PM model is used under the three restriction methods (CCMV, ACMV, NCMV) together with the predictive mean matching method to im-

pute the missing response values. The SM model is also used with the predictive mean matching method to impute the missing response values. Note that the PM and SM models are used only with the MNAR missing-data mechanism.

In this chapter, we used longitudinal binary datasets and simulated the percentage of missing data, the pattern of missing data, and the degree of systematic non-response. Datasets, with 30% or 40% missing data, were created using the same design and parameters as described in Chapter 3. For each dataset, missing data indicators were created according to the models described below. As a starting point, we considered a hypothetical longitudinal study examining smoking behavior among school-aged youth with smoking behavior as the primary outcome. Schools were randomly assigned to either a control or treatment condition. The data were generated under the generalized logit model (5.1) below.

$$log\left(\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right) = \beta_{0|kl} + \beta_{1|kl}X_{1ij,t-1} + \beta_{2|kl}t + \beta_{3|kl}C_j + (u_{0j|kl}) \tag{5.1}$$

where

$$u_{0j|kl} \sim N(0, \sigma^2_{u|kl})$$

To generate the data for the simulation, a person-period dataset was first created so that unique subjects appeared in the dataset multiple times, once for each measurement point. Since there were seven measurement points, each subject had seven records (or seven rows of data) as well as an indicator for the measurement time point (t=1 to 7). Simulated data were then generated under the following assumptions. First, all subjects had to possess valid data for the first measurement time (t=1). Then, a missing data indicator was generated for all subsequent time points (t=2 to 7). Previous studies have used 10% to 40% missing data for simulation purposes (e.g., Gibson

& Olejnik (2003)). Based on published results and our own trials indicating that the effects of missing data were most pronounced when proportion of missing data are 20%, a decision was made to generate 30% and 40% missing data under a monotone missing-data pattern.

To generate the monotone missing-data pattern for the MCAR mechanism, a binomial random variable was created for time points 2 through 7 for each subject. This indicator variable was assigned a value of 1 (missing) using the probabilities of 0.097 and 0.135 for 30% and 40% missing data respectively, and 0 otherwise. In other words for the 30% missing data simulations, $Pr(R_{it} = 1) = 0.097$. When a missing data point first occurred, the data for all subsequent time points were set to missing, thereby creating the monotone missing-data pattern. For example, if $R_{i3} = 1$, then the values for time points 3 through 7 were also set to missing. Furthermore, if more than one time point was assigned a missing value, then the earliest time point was used to create the monotone missing-data pattern. That is, if $R_{i2} = 1$ and $R_{i4} = 1$, then the values for time points 2 through 7 were set to 1. In total, 500 iterations of the simulation were conducted; the creation of the simulated dataset was independent at each iteration.

For MAR and MNAR, the missing data were created using a state specific logit model as shown in Eq. 5.2.

$$\begin{aligned} logit\left(P(R_{ijt|kl} = 1|X_{ij,t-1}, Y_{ij,t-1|kl}, Y_{ijt|kl})\right) &= m_{0|kl} + m_{1|kl} \times X_{ij,t-1} + m_{2|kl} \times Y_{ij,t-1|kl} \\ &\quad + m_{3|kl} \times Y_{ijt|kl} + u_{0j|kl}; t \geq 2 \end{aligned} \tag{5.2}$$

In Eq. 5.2., $Y_{ij,1}$ represents the initial state for an individual. The parameters were chosen

to ensure that 30% and 40% of the data were assigned as missing (as described below). The parameters used for different missing-data rates under the MNAR model for the three different transition models are as follows:

A. For 30% missing data (MNAR simulations)

a. Nonsmoker to Smoker (k=1)

$m_{0,12} = -2.6$, $m_{1,12} = -0.1$, $m_{2,12} = 0.2$, $m_{3,12} = 0.3$

b. Smoker to Quitter (k=2)

$m_{0,23} = -2.74$, $m_{1,23} = -0.1$, $m_{2,23} = 0.2$, $m_{3,23} = 0.3$

c. Quitter to Smoker (k=3)

$m_{0,32} = -3.24$, $m_{1,32} = -0.1$, $m_{2,32} = 0.2$, $m_{3,32} = 0.3$

B. For 40% missing data (MNAR simulations)

a. Nonsmoker to Smoker

$m_{0,12} = -2.15$, $m_{1,12} = -0.1$, $m_{2,12} = 0.2$, $m_{3,12} = 0.3$

b. Smoker to Quitter

$m_{0,23} = -2.4$, $m_{1,23} = -0.1$, $m_{2,23} = 0.2$, $m_{3,23} = 0.3$

c. Quitter to Smoker

$m_{0,32} = -2.85$, $m_{1,32} = -0.1$, $m_{2,32} = 0.2$, $m_{3,32} = 0.3$

The coefficients m_{31}, m_{32}, and m_{33} differentiate the missing-data mechanisms between MNAR, MAR, and MCAR. To create missing data under MAR, where the missing data do not depend on Y_{ijt}, we assume that $m_{3,12} = 0$, $m_{3,23} = 0$, and $m_{3,32} = 0$. To achieve the desired proportion

of missing data (30% or 40%), other parameters need to be changed, and for simplicity we vary the intercept coefficients ($m_{0,12}$, $m_{0,23}$, and $m_{0,32}$). Note that this creates the same missing data rates for all three transitions. In practice, some transition may have a higher proportion missing, and others a lower proportion leading to an overall rate of $30\% \sim 40\%$.

To illustrate the results of the modeling, Tables 5.1 and 5.2 show the number and proportion of individuals having missing data at each time point for a particular simulated dataset for the MNAR missing data mechanism. For the 30%-missing-data simulation (Table 5.1), 48.98% of individuals (2449/5000) had at least one missing data point. For the 5000 individuals over 6 time points there are a total of 30,000 observations of which 9,186 were assigned as missing, leading to a 30.62% missing data rate (9,186/30,000 = 30.62). Similarly, for the 40%-missing-data simulation (Table 5.2), 63.04% of individuals (3152/5000) have at least one missing data point, leading to 40.65% missing data overall. Table 5.3 shows the percentage of missing data for all three missing data mechanisms.

Table 5.1: Sample simulation showing the number and percentage of missing data at each time point with 30% missing data under MNAR

	Individual missing data	**Overall missing observations**
	n(%)	n(%)
Time	**n=5000**	**n=30,000**
Time 2	560* (11.2)	560** (1.87)
Time 3	380 (7.60)	940 (3.13)
Time 4	416 (8.32)	1356 (4.52)
Time 5	377 (7.54)	1733 (5.78)
Time 6	415 (8.30)	2148 (7.16)
Time 7	301 (6.02)	2449 (8.16)
Overall	2449 (48.98)	9186 (30.62)

$*$ number of new individuals missing at a given time point; $**$ total number of missing observations at a given time point

Table 5.2: Sample data set showing the number and percentage of missing data at each time point with 40% missing data under MNAR

	Individual missing data	**Overall missing observations**
	n(%)	n(%)
Time	**n=5000**	**n=35,000**
Time 2	788* (15.76)	788** (2.63)
Time 3	506 (10.12)	1294 (4.31)
Time 4	540 (10.80)	1834 (6.11)
Time 5	480 (9.60)	2314 (7.71)
Time 6	498 (9.96)	2812 (9.37)
Time 7	340 (6.80)	3152 (10.51)
Overall	3152 (63.04)	12194 (40.65)

$*$ number of new individuals missing at a given time point; $**$ total number of missing observations at a given time point

Table 5.3: Sample simulation showing the percentage of missing data for all three missing data mechanisms

Missing Data Proportion for all three cases		**Non-smoker to Smoker**	**Smoker to Quitter**	**Quitter to Smoker**	**Overall missing observation**
30% missing data	MCAR	30.45	31.93	30.35	30.91
	MAR	31.37	31.17	29.33	30.62
	MNAR (m_{3j}=0.3)	31.81	30.17	29.89	30.62
30% missing data	MCAR	28.75	30.07	32.41	30.41
	MAR	30.04	29.74	29.66	29.81
	MNAR (m_{3j}=0.9)	31.32	30.15	30.45	30.64
40% missing data	MCAR	39.22	40.70	41.18	40.36
	MAR	40.54	40.07	40.54	40.38
	MNAR (m_{3j}=0.3)	40.94	41.21	40.63	40.92

Table 5.4 lists the proportion of students in the three smoking categories for each time point, for the same simulated dataset. When the cohort was at time 1 (grade 6), the majority of students had no smoking experience, furthermore, smoking prevalence increased over the time. At time 1

Table 5.4: Sample simulation showing the percentage of subjects in each state over time

Smoking Status		Time 1	2	3	4	5	6	7
Missing	Prevalence (%)	–	11.2	18.8	27.1	34.7	43.0	49.0
	(Frequency)	–	(560)	(940)	(1356)	(1733)	(2148)	(2449)
Nonsmoker	Prevalence (%)	86.4	61.8	47.8	37.2	29.2	21.9	15.4
	(Frequency)	(4319)	(3089)	(2388)	(1858)	(1458)	(1093)	(770)
Smoker	Prevalence (%)	7.6	19.8	20.2	21.7	22.1	22.8	24.0
	(Frequency)	(378)	(988)	(1009)	(1083)	(1103)	(1139)	(1199)
Quitter	Prevalence (%)	6.1	7.3	13.3	14.1	14.1	12.4	11.6
	(Frequency)	(303)	(363)	(663)	(703)	(706)	(620)	(582)

only 7.6% of 5000 students were classified as smokers; at time 7, 24% of students were classified as smokers.

For the same simulated dataset, Table 5.5 summarizes the transition of individuals from one smoking category to another. It shows that the among non-smokers in grade 6, 21.9% remained non-smokers over time. Most transitions involve the non-smoker to smoker transition and relatively few individuals experienced the quitter to smoker transition.

Table 5.5: Sample simulation showing the transitions between states over time for students starting in time 1 (grade 6) to time 7 (grade 12)

		Status at t									
		Missing		Nonsmoker		Smoker		Quitter		Overall at t-1	
Time	Status at t-1	%	(n)	%	(n)	%	(n)	%	(n)	%	(n)
2	Nonsmoker	10.7	(464)	71.5	(3089)	17.7	(766)	–	–	86.4	(4319)
	Smoker	15.3	(58)	–	–	44.2	(167)	40.5	(153)	7.6	(378)
	Quitter	12.5	(38)	–	–	18.2	(55)	69.3	(210)	6.1	(303)
3	Missing	100.0	(560)	–	–	–	–	–	–	11.2	(560)
	Nonsmoker	7.5	(232)	77.3	(2388)	15.2	(469)	–	–	61.8	(3089)
	Smoker	11.1	(110)	–	–	49.7	(491)	39.2	(387)	19.8	(988)
	Quitter	10.5	(38)	–	–	13.5	(49)	76.0	(276)	7.3	(363)
4	Missing	100.0	(940)	–	–	–	–	–	–	18.8	(940)
	Nonsmoker	9.4	(225)	77.8	(1858)	12.8	(305)	–	–	47.8	(2388)
	Smoker	11.6	(117)	–	–	56.2	(567)	32.2	(325)	20.2	(1009)
	Quitter	11.2	(74)	–	–	31.8	(211)	57.0	(378)	13.3	(663)
5	Missing	100.0	(1356)	–	–	–	–	–	–	27.1	(1356)
	Nonsmoker	9.3	(173)	78.5	(1458)	12.2	(227)	–	–	37.2	(1858)
	Smoker	11.2	(121)	–	–	61.7	(668)	27.1	(294)	21.7	(1083)
	Quitter	11.8	(83)	–	–	29.6	(208)	58.6	(412)	14.1	(703)
6	Missing	100.0	(1733)	–	–	–	–	–	–	34.7	(1733)
	Nonsmoker	9.9	(144)	75.0	(1093)	15.2	(221)	–	–	29.2	(1458)
	Smoker	12.0	(132)	–	–	66.1	(729)	21.9	(242)	22.1	(1103)
	Quitter	19.7	(139)	–	–	26.8	(189)	53.5	(378)	14.1	(706)
7	Missing	100.0	(2148)	–	–	–	–	–	–	43.0	(2148)
	Nonsmoker	8.6	(94)	70.4	(770)	21.0	(229)	–	–	21.9	(1093)
	Smoker	11.5	(131)	–	–	71.3	(812)	17.2	(196)	22.8	(1139)
	Quitter	12.3	(76)	–	–	25.5	(158)	62.3	(386)	12.4	(620)

Furthermore, for each simulation a second set of datasets was created by setting the m_{31}, m_{32}, and m_{33} parameters to more extreme values. We increase the values on average from about 0.30 to 0.90 to produce situations where the missing data are more dependent on the future observations. To achieve the desired proportion of missing data (30%), other parameters need to be changed, and for simplicity we vary the intercept coefficients (m_{01}, m_{02}, and m_{03}). These latest simulated data were only created for 30% missing datasets and allow us to show that our methods can handle this extreme situation without further assumptions. After generating the data with missing observations, we used four analytic techniques. The first was a complete case analysis and the remaining techniques employed the following imputation methods:

- LOCF
- Pattern mixture models
- Selection model

5.2 Complete Case Analysis (CCA)

In this procedure, subjects with missing data were excluded and the analysis was based only on those subjects with complete data for the entire study. The procedure is simple to implement and it is used in most standard statistical software packages. If the MCAR assumption is wrong then the CCA method may provide biased estimates.

5.3 Last Observation Carried Forward (LOCF)

This method is commonly used in clinical trials. It assumes that the subject profile is unchanged from the previous assessment. It requires the strong assumption that the individual outcome profile remains at the level of the last observed measurement throughout the remainder of the follow-up. However, it is possible that the outcome profile changes as soon as the individual stops the treatment. In this study, missing values for the time-dependent covariates are replaced by their last reported observation.

5.4 Predictive Mean Matching

In the predictive mean matching process, the relationship between the response variable and the relevant covariates is modeled for the complete data cases, and the incomplete data cases are divided into different groups based on the predicted values of the response variable. This method matches the cases based on a linear combination of covariates $X\beta + Zu$. It is not constrained by the continuous and normality assumptions and replaces each missing value with several observed values based on the matching, which in turn accounts for the variation in imputed values. The Approximate Bayesian Bootstrap (ABB) is used in the predictive mean matching method to impute each missing value multiple times. Because of the variation in the imputed values this method provides a more valid inference than would a single imputation. The SAS procedure NLMIXED uses maximum likelihood estimation to fit the model and provides empirical Bayes estimates of the random effects u. The steps are:

- Establish a predictive model for response variable Y based on X and C for the complete

cases:

$$logit(Pr(Y^o = 1|X)) = X\beta + u \tag{5.3}$$

- The predicted probability of Y being 1 is

$$\Pr(\hat{Y^o} = 1|X) = \frac{Exp\left(X\hat{\beta} + \hat{u}\right)}{1 + Exp\left(X\hat{\beta} + \hat{u}\right)} \tag{5.4}$$

- Divide the cases into different groups based on the predicted probabilities.

- Apply ABB to impute the missing response values in each group. In group k, let Y^o denote the n_1 observations with non-missing Y values and Y^m denote the n_0 observations with missing Y values. ABB first draws n_0 observations randomly with replacement from Y^o and uses these values as the n_o imputed values for the missing response vector, Y^m, which is combined with the non-missing data to create a new dataset Y^*. Repeat this process M times and create M imputed datasets.

- Analyze these M complete imputed datasets separately; combine the results using the Rubin procedure incorporated in the SAS procedure MIANALYZE (Rubin, 1987).

5.5 Pattern Mixture Model

For imputation and modeling, the pattern mixture model with the predictive mean matching method was restricted to a monotone missing-data pattern because the WSPP3 dataset which we

replicated has relatively few individuals in the intermittent missing-data categories. Table 5.6 shows that there are only seven possible monotone missing-data patterns. Individuals with complete data at all time points follow pattern 1. Similarly, individuals with data only at time 1 follow pattern 7. The goal of this study was to improve the existing restriction-case pattern mixture model using a predictive mean matching method within each pattern. To do this, three existing restriction techniques were used in the pattern mixture model: complete case missing value (CCMVPM), available case missing value (ACMVPM), and neighboring case missing value (NCMVPM).

Table 5.6: Monotone missing-data pattern

Pattern	t_1	t_2	t_3	t_4	t_5	t_6	t_7
P1	O	O	O	O	O	O	O
P2	O	O	O	O	O	O	$*$
P3	O	O	O	O	O	$*$	$*$
P4	O	O	O	O	$*$	$*$	$*$
P5	O	O	O	$*$	$*$	$*$	$*$
P6	O	O	$*$	$*$	$*$	$*$	$*$
P7	O	$*$	$*$	$*$	$*$	$*$	$*$

$*$ missing observation: O non-missing observation

5.5.1 Steps for Complete Case Missing Value with Predictive Mean Matching Approach (CCMVPM)

This restriction method was used on the complete data for pattern 1 to impute the means for the missing observations in the remaining patterns. It assumes that the missing value distributions are the same as the complete case distribution.

- For the missing data in pattern P2, we used the predictive mean matching method to impute the missing values at t_7 using all the observed cases from pattern P1.
- For the missing data in pattern P3, we used the predictive mean matching method to impute the missing values at t_6 and t_7 using all the observed cases from pattern P1.
- For the missing data in pattern P4, we used the predictive mean matching method to impute the missing values at t_5, t_6, and t_7 using all the observed cases from pattern P1.
- For the missing data in pattern P5, we used the predictive mean matching method to impute the missing values at t_4, t_5, t_6, and t_7 using all the observed cases in P1.
- For the missing data in pattern P6, we used the predictive mean matching method to impute the missing values at t_3, t_4, t_5, t_6, and t_7 using all the observed cases in P1.
- For the missing data in pattern P7, we used the predictive mean matching method to impute the missing values at t_2, t_3, t_4, t_5, t_6, and t_7 using all the observed cases in P1.
- After creating m imputed datasets for each pattern, we combined the resultant estimates using the techniques from Rubin (1987) incorporated in the SAS procedure MIANALYZE.
- All parameter estimates were then compared with those from the full datasets.

5.5.2 Steps for Available Case Missing Value with Predictive Mean Matching Approach (ACMVPM)

This restriction method uses the information contained in all the missing-data patterns for all available subjects to impute the means for the missing observations in the remaining patterns.

- For the missing data in pattern P2, we used the predictive mean matching method to impute the missing values at t_7 using all the observed cases in pattern P1.

- For the missing data in pattern P3, we used the predictive mean matching method to impute the missing values at t_6 using all the observed cases in patterns P1 and P2. To impute the missing values at t_7, we used the cases in P1 and the imputed values for t_6 in P3.

- For the missing data in pattern P4, we used the predictive mean matching method to impute the missing values at t_5 using all the observed cases in P1, P2, and P3. To impute the missing values at t_6 we used the cases in P1 and P2 and the imputed values for t_5 in P4. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P4.

- For the missing data in pattern P5, we used the predictive mean matching method to impute the missing values at t_4 using all the observed cases in P1, P2, P3, and P4. To impute the missing values at t_5 we used the cases in P1, P2, and P3 and the imputed values for t_4 in P5. To impute the missing values at t_6 we used the cases in P1 and P2 and the imputed values for t_5 in P5. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P5.

- For the missing data in pattern P6, we used the predictive mean matching method to impute the missing values at t_3 using all the observed cases in P1, P2, P3, P4, and P5. To impute the missing values at t_4 we used the cases in P1, P2, P3, and P4 and the imputed values for t_3 in P6. To impute the missing values at t_5 we used the cases in P1, P2, and P3 and the imputed values for t_4 in P6. To impute the missing values at t_6 we used the cases in P1

and P2 and the imputed values for t_5 in P6. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P6.

- For the missing data in pattern P7, we used the predictive mean matching method to impute the missing values at t_2 using all the observed cases in P1, P2, P3, P4, P5, and P6. To impute the missing values at t_3 we used the observed cases in P1, P2, P3, P4, P5, and the imputed values for t_2 in P7. To impute the missing values at t_4 we used the cases in P1, P2, P3, and P4 and the imputed values for t_3 in P7. To impute the missing values at t_5 we used the cases in P1, P2, and P3 and the imputed values for t_4 in P7. To impute the missing values at t_6 we used the cases in P1 and P2 and the imputed values for t_5 in P7. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P7.
- For each pattern the data were combined using Rubin's formulae described earlier.
- The parameter estimates were then compared with those from the full datasets.

5.5.3 Steps for Neighboring Case Missing Value (NCMVPM)

This restriction method uses the information from subjects in a neighboring pattern to impute the means for the missing observation in the remaining patterns. The algorithm follows:

- For the missing data in pattern P2, we used the predictive mean matching method to impute the missing values at t_7 using all the observed cases in P1.
- For the missing data in pattern P3, we used the predictive mean matching method to impute the missing values at t_6 using all the observed cases in P2. To impute the missing values at t_7, we used the cases in P1 and the imputed values for t_6 in P3.

- For the missing data in pattern P4, we used the predictive mean matching method to impute the missing values at t_5 using all the observed cases in P3. To impute the missing values at t_6 we used the cases in P2 and the imputed values for t_5 in P4. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P4.

- For the missing data in pattern P5, we used the predictive mean matching method to impute the missing values at t_4 using all the observed cases in P4. To impute the missing values at t_5 we used the cases in P3 and the imputed values for t_4 in P5. To impute the missing values at t_6 we used the cases in P2 and the imputed values for t_5 in P5. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P5.

- For the missing data in pattern P6, we used the predictive mean matching method to impute the missing values at t_3 using all the observed cases in P5. To impute the missing values at t_4 we used the cases in P4 and the imputed values for t_3 in P6. To impute the missing values at t_5 we used the cases in P3 and the imputed values for t_4 in P6. To impute the missing values at t_6 we used the cases in P2 and the imputed values for t_5 in P6. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P6.

- For the missing data in pattern P7, we used the predictive mean matching method to impute the missing values at t_2 using all the observed cases in P6. To impute the missing values at t_3 we used the cases in P5 and the imputed values for t_2 in P7. To impute the missing values at t_4 we used the cases P4 and the imputed values for t_3 in P7. To impute the missing values at t_5 we used the cases in P3 and the imputed values for t_4 in P7. To impute the missing values at t_6 we used the cases in P2 and the imputed values for t_5 in P7. To impute the missing values at t_7 we used the cases in P1 and the imputed values for t_6 in P7.

- For each pattern the data were combined using Rubin's formulae described earlier.
- The parameter estimates were then compared with those from the full datasets.

5.5.4 Example: Use of Pattern Mixture Model for Transition Data

In order to demonstrate a practical implementation of the pattern mixture model, consider four examples based on the WSPP3 data. A number of transitions might be possible for any one individual. However, due to the monotone missing data assumption, at most one transition will have a missing response. Thus, any individual could have more than one complete transition sequence, but could have at most one sequence that is incomplete. It is this incomplete sequence that is imputed. For simplicity all the examples considered here are for individuals who have missing data at time 4.

Examples of sample sequences include:

- non-smoker → non-smoker → smoker → smoker → quitter → quitter → smoker: (NNSSQQS)
- non-smoker → non-smoker → non-smoker → missing → missing → missing → missing (NNN.)
- non-smoker → non-smoker → smoker → missing → missing → missing → missing (NNS.)
- non-smoker → smoker → quitter → missing → missing → missing → missing (NSQ.)

In order to implement the pattern mixture model, the steps involved include:

- Determine the pattern for the individual based on Table 4.2, using the rule that the pattern for the whole data record is determined by the first missing response value.
- Break out the sequences for each individual based on when a transition to another state occurs where missing data first occurs.
- Determine the complete sequences for each individual based on the state to which they belong and identify them so they are not imputed.
- Determine the incomplete sequence for an individual(if any) and identify it for imputation.
- Apply all three restriction methods to impute the missing responses.

Case 1: (NNSSQQS)

In this case, the individual has complete data. Based on Table 4.2, this individual falls under pattern P1. Since all three transitions are represented in the sequence, all three transition models will be fit for this sequence. If a value of 96 represents a valid missing response for a particular sequence at time t, either because it is the first element of the transition sequence, because the sequence terminated at $t-1$ or earlier, or because the sequence had not begun at time t, then

- Non-smoker to smoker transition: Sequence NNS is represented by: $(96, 0, 1, 96, 96, 96, 96)$
- Smoker to quitter transition: Sequence SSQ is represeneted by: $(96, 96, 96, 0, 1, 96, 96)$
- Quitter to smoker transition: Sequence QQS is represented by: $(96, 96, 96, 96, 96, 0, 1)$

For the non-smoker to smoker transition, the model is fit for two time points $(96, 0, 1)$ and no imputation is performed because this individual has complete data for this transition.

For the smoker to quitter transition, the model is only fit for time points 4 and 5 $(96, 96, 96, 0, 1)$. Time points 1,2,3, 6, and 7 are considered valid skips and no values are imputed. For the quitter to smoker transition, the model is only fit for time points 6 and 7 $(96, 96, 96, 96, 96, 0, 1)$ and all other time points are considered valid skips.

Case 2: (N N N .) In this case, the individual has missing data at time 4; based on Table 4.2, this individual falls under pattern P5. As a result, we only fit the non-smoker to smoker transition model to impute the missing data. The procedure for all three restriction methods can be easily applied to impute the missing data. Once the individual is assigned to be a smoker at any time point, no further imputation is performed for that individual.

Case 3: (N N S .) In this case, the individual has missing data at time 4; based on Table 4.2, this individual falls under pattern P5. We fit two models: (1) one for the non-smoker to smoker transition and (2) One for the smoker to quitter transition. For the non-smoker to smoker transition no imputation is performed because the individual has moved to the next state before the occurrence of missing data. For the smoker to quitter transition, this individual has missing data. Here, the model is fit for the smoker to quitter transition with only one time point (0,.) and then imputation is performed using pattern P5 from Table 4.2. Other sequences that involve the smoker to quitter transition that are included in pattern P5 are $(QSS....), (QQS....), (SSS....), (NSS....)$, and $(SQS....)$. Based on the imputation, once the individual is assigned as quitter at any time point, no further imputation is performed for that individual. In this case, the procedure for all three restriction methods will be applied assuming the individual falls in pattern P5.

Case 4: (NSQ.) In this case, the individual has missing data at time 4; based on Table 4.2, this individual falls under pattern P5. In this case all three transition models are fit. First, we fit

the non-smoker to smoker transition model. No imputation is performed because the individual has moved to the smoker state. Next, we fit the smoker to quitter transition and no imputation is performed because the individual has moved to the quitter state. Finally, we fit the quitter to smoker transition model and imputation is performed for the quitter to smoker transition.

5.6 Selection Model

The steps used in the selection model are:

- The dropout model was fitted as in Eq. (4.5) with all the covariates. Hierarchical logistic regression models were used to estimate the predicted values and parameters using AGQ as described in Chapter 3.
- The inverse Mills ratio was computed as described in Chapter 4.
- The inverse Mills ratio was used as one of the predictors in the outcome model.
- Using parameter estimates from the outcome model, we used the predictive mean matching method to impute the missing response values.
- After obtaining M imputed datasets, we combined the resulting estimates using Rubin's formulae (see section 4.2.6).

5.7 Simulation Results

In this section, the results for each imputation method are compared and their implications are discussed. Comparisons are made for 30% and 40% missing-value rates and three transitions (nonsmoker to smoker, smoker to quitter, quitter to smoker). Starting with the complete datasets from Chapter 3, five hundred simulations were performed separately for both the 30% and 40% missing-value rates. For each simulation AGQ was used to estimate the model parameters. The details of the simulation are discussed in Chapter 3 and section 5.1.

5.7.1 Parameter Estimates

Tables 5.7–5.9 report the average parameter estimates over 500 simulations under MCAR, MAR, and MNAR, estimates from LOCF under MCAR, MAR, and MNAR, estimates from the three restriction methods for the pattern mixture model (ACMVPM, CCMVPM, NCMVPM) with the predictive mean matching method under MNAR, and estimates from the selection model with the predictive mean matching method under MNAR. The parameter estimates from each method were compared with the true values under the 30% and 40% missing-data rates. The first column in Tables 5.7–5.9 gives the true values used to generate the simulated data. The estimates from the full data analysis (FDA) (missing data not removed) are reported in the second column. The third, fourth, and fifth columns report the parameter estimates using CCA under the assumption that the missing data mechanism is MCAR, MAR, and MNAR respectively. The sixth, seventh, and eighth columns report the parameter estimates from LOCF under MCAR, MAR, and MNAR, respectively. The ninth, tenth, and eleventh columns report the parameter estimates from ACMVPM, CCMVPM, and NCMVPM under MNAR respectively. The last col-

umn reports the parameter estimates from the selection model with the predictive mean matching method (SMPM) under MNAR. All the tables report the parameter estimates for 500 convergent simulations with the empirical standard deviation in parentheses. The results are discussed below.

Table 5.7: Average estimates with empirical standard deviations for parameters for the nonsmoker-to-smoker transition based on 500 simulations*

Parameter		TRUE	Complete Case Analysis FDA	MCAR	MAR	MNAR	LOCF MCAR	MAR	MNAR	Pattern Mixture (MNAR) ACMVPM	CCMVPM	NCMVPM	SMPM
30% Missing Values													
β_1	Mean	-2.3	-2.295	-2.244	-2.120	**-2.245**	-2.190	-2.189	**-2.026**	**-2.290**	-2.311	-2.358	-2.306
	ESD		(0.218)	(0.251)	(0.453)	**(0.518)**	(0.221)	(0.230)	**(0.185)**	**(0.223)**	(0.249)	(0.272)	(0.229)
β_1	Mean	0.2	0.201	0.194	0.182	**0.184**	0.169	0.169	**0.168**	**0.203**	0.200	0.202	0.204
	ESD		(0.022)	(0.025)	(0.038)	**(0.051)**	(0.028)	(0.023)	**(0.028)**	**(0.029)**	(0.026)	(0.029)	(0.029)
β_2	Mean	0.61	0.608	0.610	0.589	**0.561**	0.568	0.571	**0.549**	**0.608**	0.610	0.618	0.609
	ESD		(0.019)	(0.045)	(0.088)	**(0.084)**	(0.071)	(0.063)	**(0.059)**	**(0.024)**	(0.026)	(0.044)	(0.022)
β_3	Mean	-4.1	-4.071	-4.100	-4.023	**-4.176**	-3.927	-4.453	**-4.443**	**-4.090**	-4.082	-4.229	-4.086
	ESD		(0.187)	(0.220)	(0.248)	**(0.241)**	(0.222)	(0.229)	**(0.241)**	**(0.203)**	(0.221)	(0.217)	(0.203)
σ^2_μ	Mean	0.68	0.671	0.659	0.710	**0.632**	0.712	0.651	**0.621**	**0.663**	0.697	0.670	0.677
	ESD		(0.079)	(0.078)	(0.079)	**(0.074)**	(0.095)	(0.089)	**(0.072)**	**(0.084)**	(0.082)	(0.087)	(0.080)
40% Missing Values													
β_0	Mean	-2.3	-2.295	-2.232	-1.944	**-2.277**	-2.133	-2.132	**-1.979**	**-2.286**	-2.303	-2.407	-2.298
	ESD		(0.218)	(0.293)	(0.396)	**(0.565)**	(0.210)	(0.223)	**(0.180)**	**(0.231)**	(0.252)	(0.311)	(0.234)
β_1	Mean	0.2	0.201	0.192	0.174	**0.202**	0.164	0.165	**0.162**	**0.201**	0.200	0.204	0.202
	ESD		(0.022)	(0.031)	(0.043)	**(0.062)**	(0.028)	(0.024)	**(0.028)**	**(0.031)**	(0.029)	(0.029)	(0.030)
β_2	Mean	0.61	0.608	0.604	0.515	**0.499**	0.506	0.511	**0.495**	**0.609**	0.611	0.634	0.608
	ESD		(0.019)	(0.054)	(0.049)	**(0.055)**	(0.039)	(0.035)	**(0.037)**	**(0.025)**	(0.026)	(0.057)	(0.024)
β_3	Mean	-4.1	-4.071	-4.070	-3.992	**-4.098**	-3.850	-4.362	**-4.367**	**-4.089**	-4.086	-4.289	-4.083
	ESD		(0.187)	(0.222)	(0.247)	**(0.249)**	(0.210)	(0.212)	**(0.232)**	**(0.204)**	(0.220)	(0.227)	(0.202)
σ^2_μ	Mean	0.68	0.671	0.678	0.735	**0.632**	0.698	0.637	**0.609**	**0.664**	0.734	0.669	0.675
	ESD		(0.079)	(0.068)	(0.072)	**(0.081)**	(0.091)	(0.088)	**(0.071)**	**(0.086)**	(0.067)	(0.087)	(0.079)

*Model used: $log\left\{\frac{p_{ijt}(2|1)}{p_{ijt}(1|1)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; Mean estimate; ESD= empirical standard deviation

Table 5.8: Average estimates with empirical standard deviation for parameters for parameters for the smoker-to-quitter transition based on 500 simulations*

Parameter		TRUE	Complete Case Analysis FDA	Complete Case Analysis MCAR	Complete Case Analysis MAR	Complete Case Analysis MNAR	LOCF MCAR	LOCF MAR	Pattern Mixture (MNAR) MNAR	Pattern Mixture (MNAR) ACMVPM	Pattern Mixture (MNAR) CCMVPM	Pattern Mixture (MNAR) NCMVPM	Pattern Mixture (MNAR) SMPM
30% Missing Values													
β_0	Mean	0.8	0.800	0.822	0.747	**0.696**	0.792	0.851	**0.687**	**0.825**	0.776	0.788	0.790
	ESD		(0.233)	(0.338)	(0.328)	**(0.852)**	(0.275)	(0.347)	**(0.260)**	**(0.282)**	(0.240)	(0.274)	(0.623)
β_1	Mean	-0.1	-0.099	-0.095	-0.093	**-0.088**	-0.090	-0.085	**-0.088**	**-0.099**	-0.096	-0.094	-0.098
	ESD		(0.028)	(0.048)	(0.055)	**(0.036)**	(0.033)	(0.034)	**(0.037)**	**(0.033)**	(0.038)	(0.034)	(0.040)
β_2	Mean	-0.3	-0.304	-0.290	-0.271	**-0.260**	-0.351	-0.350	**-0.356**	**-0.300**	-0.290	-0.310	-0.304
	ESD		(0.016)	(0.038)	(0.033)	**(0.029)**	(0.021)	(0.021)	**(0.025)**	**(0.022)**	(0.019)	(0.029)	(0.022)
β_3	Mean	0.2	0.215	0.220	0.210	**0.242**	0.213	0.208	**0.190**	**0.204**	0.217	0.207	0.221
	ESD		(0.147)	(0.159)	(0.168)	**(0.175)**	(0.173)	(0.155)	**(0.156)**	**(0.149)**	(0.167)	(0.182)	(0.166)
σ^2_μ	Mean	0.68	0.675	0.657	0.658	**0.762**	0.653	0.761	**0.846**	**0.666**	0.668	0.672	0.728
	ESD		(0.089)	(0.091)	(0.085)	**(0.102)**	(0.088)	(0.105)	**(0.090)**	**(0.101)**	(0.096)	(0.094)	(0.094)
40% Missing Values													
β_0	Mean	0.8	0.800	0.795	0.702	**0.510**	0.762	0.823	**0.729**	**0.826**	0.771	0.779	0.648
	ESD		(0.233)	(0.326)	(0.309)	**(0.864)**	(0.291)	(0.415)	**(0.270)**	**(0.288)**	(0.251)	(0.272)	(0.632)
β_1	Mean	-0.1	-0.099	-0.087	-0.090	**-0.087**	-0.087	-0.081	**-0.087**	**-0.097**	-0.094	-0.100	-0.098
	ESD		(0.028)	(0.054)	(0.061)	**(0.039)**	(0.035)	(0.033)	**(0.038)**	**(0.035)**	(0.039)	(0.038)	(0.042)
β_2	Mean	-0.3	-0.304	-0.283	-0.247	**-0.244**	-0.361	-0.361	**-0.369**	**-0.300**	-0.290	-0.317	-0.303
	ESD		(0.016)	(0.048)	(0.026)	**(0.025)**	(0.019)	(0.022)	**(0.023)**	**(0.023)**	(0.021)	(0.036)	(0.022)
β_3	Mean	0.2	0.215	0.216	0.205	**0.240**	0.210	0.198	**0.190**	**0.208**	0.220	0.201	0.216
	ESD		(0.147)	(0.181)	(0.160)	**(0.178)**	(0.173)	(0.156)	**(0.160)**	**(0.150)**	(0.166)	(0.174)	(0.165)
σ^2_μ	Mean	0.68	0.675	0.656	0.656	**0.787**	0.651	0.785	**0.881**	**0.668**	0.668	0.672	0.741
	ESD		(0.089)	(0.091)	(0.085)	**(0.102)**	(0.086)	(0.102)	**(0.091)**	**(0.104)**	(0.098)	(0.097)	(0.090)

*Model used: $log\left\{\frac{p_{ijt(3|2)}}{p_{ijt(2|2)}}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; Mean estimate; ESD=empirical standard deviation

Table 5.9: Average estimates with empirical standard deviations for parameters for the quitter-to-smoker transition based on 500 simulations*

Parameter		TRUE	Complete Case Analysis				LOCF			Pattern Mixture (MNAR)			
			FDA	MCAR	MAR	MNAR	MCAR	MAR	MNAR	ACMVPM	CCMVPM	NCMVPM	SMPM
30% Missing Values													
β_0	Mean	-1.7	-1.681	-1.643	-1.634	**-1.599**	-1.678	-1.650	**-1.469**	**-1.669**	-1.682	-1.626	-1.634
	ESD		(0.326)	(0.495)	(0.449)	**(0.470)**	(0.459)	(0.532)	**(0.392)**	**(0.394)**	(0.373)	(0.403)	(0.437)
β_1	Mean	0.3	0.302	0.296	0.282	**0.210**	0.287	0.207	**0.209**	**0.299**	0.280	0.292	0.303
	ESD		(0.053)	(0.082)	(0.066)	**(0.045)**	(0.042)	(0.040)	**(0.039)**	**(0.055)**	(0.068)	(0.053)	(0.072)
β_2	Mean	0.1	0.096	0.087	0.080	**0.073**	0.084	0.081	**0.073**	**0.097**	0.092	0.088	0.077
	ESD		(0.028)	(0.046)	(0.135)	**(0.028)**	(0.096)	(0.037)	**(0.032)**	**(0.037)**	(0.033)	(0.040)	(0.034)
β_3	Mean	-5.5	-5.507	-5.622	-5.381	**-5.885**	-5.602	-5.726	**-5.885**	**-5.451**	-5.459	-5.468	-5.637
	ESD		(0.324)	(0.413)	(0.392)	**(0.432)**	(0.541)	(0.455)	**(0.450)**	**(0.383)**	(0.496)	(0.394)	(0.443)
σ^2_μ	Mean	0.68	0.673	0.642	0.610	**0.591**	0.658	0.822	**0.920**	**0.656**	0.654	0.661	0.644
	ESD		(0.133)	(0.180)	(0.151)	**(0.118)**	(0.173)	(0.194)	**(0.137)**	**(0.157)**	(0.154)	(0.160)	(0.140)
40% Missing Values													
β_0	Mean	-1.7	-1.681	-1.611	-1.619	**-1.635**	-1.633	-1.674	**-1.470**	**-1.671**	-1.674	-1.640	-1.643
	ESD		(0.326)	(0.542)	(0.506)	**(0.515)**	(0.461)	(0.537)	**(0.466)**	**(0.450)**	(0.377)	(0.433)	(0.506)
β_1	Mean	0.3	0.301	0.297	0.282	**0.200**	0.287	0.201	**0.202**	**0.303**	0.287	0.290	0.299
	ESD		(0.064)	(0.079)	(0.072)	**(0.047)**	(0.042)	(0.045)	**(0.048)**	**(0.059)**	(0.076)	(0.055)	(0.084)
β_2	Mean	0.1	0.096	0.110	0.136	**0.073**	0.127	0.083	**0.075**	**0.099**	0.092	0.088	0.079
	ESD		(0.028)	(0.048)	(0.120)	**(0.029)**	(0.086)	(0.041)	**(0.038)**	**(0.040)**	(0.033)	(0.040)	(0.028)
β_3	Mean	-5.5	-5.507	-5.593	-5.354	**-5.960**	-5.562	-5.791	**-5.974**	**-5.440**	-5.436	-5.470	-5.714
	ESD		(0.324)	(0.446)	(0.405)	**(0.468)**	(0.547)	(0.420)	**(0.447)**	**(0.408)**	(0.502)	(0.403)	(0.470)
σ^2_μ	Mean	0.68	0.673	0.624	0.596	**0.621**	0.642	0.905	**0.965**	**0.657**	0.655	0.663	0.666
	ESD		(0.133)	(0.182)	(0.148)	**(0.144)**	(0.171)	(0.226)	**(0.151)**	**(0.167)**	(0.159)	(0.166)	(0.142)

*Model used: $log\left\{\frac{p_{ijt}(2|3)}{p_{ijt}(3|3)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; Mean estimate; ESD=empirical standard deviation

Model intercept: Estimates under MCAR and MAR are similar to those of the full data analysis (FDA). Estimates under MNAR show the biggest difference when compared to those of the full data analysis. Differences in the model intercept estimates are seen for LOCF under all three missing-data mechanisms. The most noticeable differences are observed in the results for the MNAR missing data mechanism. All three restriction methods with the predictive mean matching approach provide similar estimates for the model intercept; generally, the restriction methods are less biased than the other methods. Among the restriction methods, ACMVPM estimates are less biased than the other two methods. The selection model with the predictive mean matching method is less biased than LOCF for MNAR. This is true for all three transitions and both missing-data proportions. Comparison shows little differences between restriction models and SMPM.

Treatment condition: All three restriction methods provide similar estimates for the treatment condition parameter (β_3); their estimates are much closer to the FDA estimates than are those of the other methods. The largest differences were observed when the parameters were estimated under MNAR for both complete case analysis and LOCF. Increasing the missing-data proportion from 30% to 40% widened the differences in the parameter estimates for all methods. Minor changes were observed among the restrictions methods but major changes were observed for the methods under the LOCF and complete case analysis. Compared to the LOCF and SMPM method, the restriction methods performed better in all three transitions.

Time (a proxy for grade): The estimates from the restriction methods are close to the FDA estimates. Among the restriction methods, the estimates from NCMVPM are much closer to the estimates from complete case analysis. This is true for all three transitions.

The parameter estimates for the time-dependent covariates (β_1) are similar for all three restriction methods and the selection model in all three transitions and the estimates are close to the FDA estimate.

Variance of random intercept: Both LOCF and complete case analysis under the MNAR missing mechanism underestimate the variance of the random intercept in the non-smoker to smoker transition. The estimates obtained from the restriction methods show much closer values to the FDA estimate for all three transitions. The estimates from the selection model are similar to the restriction methods. For the two transitions, the trend remains the same. Noticeable differences were observed when the missing-data proportion was increased from 30% to 40%. Estimates are more biased for both the LOCF and complete case analysis. Estimates from the restriction methods show little variation and remain close to the FDA analysis.

In conclusion, the results suggest that restriction methods with the predictive mean matching method perform better in general than the other methods studied with regards to parameter estimation. Furthermore, among the pattern mixture models, ACMVPM performs better than the other two. Comparison between restriction methods and selection models show few differences but generally restriction methods perform better than the SMPM.

5.7.2 Average Empirical Bias Estimates

The average empirical bias of a parameter estimate is defined as the average difference between the parameter estimate $\hat{\theta}_{mk}$ and the true parameter value θ_k for simulation m:

$$\text{Average Empirical Bias}_{\text{k}} = \frac{\sum_{m=1}^{500} (\hat{\theta}_{mk} - \theta_k)}{500}$$

Tables 5.10–5.12 report the average empirical bias estimates for each missing-data mechanism (MCAR, MAR, and MNAR) along with the imputation methods. The results were also compared for the 30% and 40% missing-data rates.

Model intercept: The average empirical bias estimates under MCAR and MAR are similar for all three transitions but, generally, smaller than the MNAR biases. The biases are also much higher under all the LOCF methods especially under the MNAR missing data mechanism. All three restriction methods using the predictive mean matching method provide similar estimates which are better than those of other methods under MNAR. The selection model with the predictive mean matching method performed better when the missing data proportion was increased to 40%, but only for the nonsmoker-to-smoker transition. Biases are to be higher for the smoker to quitter transition. The main reason could be that fewer individuals are moving into and out of this transition as shown in Table 5.5. Among the three restriction methods, ACMVPM provides less average empirical bias than the CCMVPM and NCMVPM methods.

Treatment condition: All three restriction methods and the selection model with the predictive mean matching methods provide similar estimates for the treatment condition parameter (β_3) and their biases are small compared to the LOCF methods. When the missing-data proportion is 40%, the selection model leads to less bias in the nonsmoker-to-smoker and quitter-to-smoker transitions. LOCF as usual provides much higher biases than the other methods. This is true for all three transitions and both missing-data proportions.

Table 5.10: Average empirical biases (EB) with ESD for parameters for the nonsmoker-to-smoker transition based on 500 simulations*

		Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter		**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values						
β_0	EB	0.056	0.180	**0.055**	0.110	0.111	**0.274**	**0.010**	-0.011	-0.058	-0.006
	ESD	(0.251)	(0.453)	**(0.518)**	(0.221)	(0.230)	**(0.185)**	**(0.223)**	(0.249)	(0.272)	(0.229)
β_1	Bias	-0.006	-0.018	**-0.016**	-0.031	-0.031	**-0.032**	**0.003**	0.000	0.002	0.004
	ESD	(0.025)	(0.038)	**(0.051)**	(0.028)	(0.023)	**(0.028)**	**(0.029)**	(0.026)	(0.029)	(0.029)
β_2	EB	0.000	-0.021	**-0.049**	-0.042	-0.039	**-0.061**	**-0.002**	0.000	0.008	-0.001
	ESD	(0.045)	(0.088)	**(0.084)**	(0.071)	(0.063)	**(0.059)**	**(0.024)**	(0.026)	(0.044)	(0.022)
β_3	EB	0.000	0.077	**-0.076**	0.173	-0.353	**-0.343**	**0.010**	0.018	-0.129	0.014
	ESD	(0.220)	(0.248)	**(0.241)**	(0.222)	(0.229)	**(0.241)**	**(0.203)**	(0.221)	(0.217)	(0.203)
σ^2_μ	EB	-0.021	0.030	**-0.048**	0.032	-0.029	**-0.059**	**-0.017**	0.017	-0.010	-0.003
	ESD	(0.078)	(0.079)	**(0.074)**	(0.095)	(0.089)	**(0.072)**	**(0.084)**	(0.082)	(0.087)	(0.080)
					40% Missing Values						
β_0	EB	0.068	0.356	**0.023**	0.167	0.168	**0.321**	**0.014**	-0.003	-0.107	0.002
	ESD	(0.293)	(0.396)	**(0.565)**	(0.210)	(0.223)	**(0.180)**	**(0.231)**	(0.252)	(0.311)	(0.234)
β_1	EB	-0.008	-0.026	**0.002**	-0.036	-0.035	**-0.038**	**0.001**	0.000	0.004	0.002
	ESD	(0.031)	(0.043)	**(0.062)**	(0.028)	(0.024)	**(0.028)**	**(0.031)**	(0.029)	(0.029)	(0.030)
β_2	EB	-0.006	-0.095	**-0.111**	-0.104	-0.099	**-0.115**	**-0.001**	0.001	0.024	-0.002
	ESD	(0.054)	(0.049)	**(0.055)**	(0.039)	(0.035)	**(0.037)**	**(0.025)**	(0.026)	(0.057)	(0.024)
β_3	EB	0.030	0.108	**0.002**	0.250	-0.262	**-0.267**	**0.011**	0.014	-0.189	0.017
	ESD	(0.222)	(0.247)	**(0.249)**	(0.210)	(0.212)	**(0.232)**	**(0.204)**	(0.220)	(0.227)	(0.202)
σ^2_μ	EB	-0.002	0.055	**-0.048**	0.018	-0.043	**-0.071**	**-0.016**	0.054	-0.011	-0.005
	ESD	(0.068)	(0.072)	**(0.081)**	(0.091)	(0.088)	**(0.071)**	**(0.086)**	(0.067)	(0.087)	(0.079)

*Model used: $log\left\{\frac{p_{ijt}(2|1)}{p_{ijt}(1|1)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; EB=$\hat{\theta}_{mk} - \theta_k$; ESD=empirical standard deviation

Table 5.11: Average empirical biases (EB) with ESD for parameters for the smoker-to-quitter transition based on 500 simulations*

		Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter		**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values						
β_0	EB	0.022	-0.053	**-0.104**	-0.008	0.051	**-0.113**	**0.025**	-0.024	-0.012	-0.010
	ESD	(0.338)	(0.328)	**(0.852)**	(0.275)	(0.347)	**(0.260)**	**(0.282)**	(0.240)	(0.274)	(0.623)
β_1	EB	0.005	0.007	**0.012**	0.010	0.015	**0.012**	**0.001**	0.004	0.006	0.002
	ESD	(0.048)	(0.055)	**(0.036)**	(0.033)	(0.034)	**(0.037)**	**(0.033)**	(0.038)	(0.034)	(0.040)
β_2	EB	0.010	0.029	**0.040**	-0.051	-0.050	**-0.056**	**0.000**	0.010	-0.010	-0.004
	ESD	(0.038)	(0.033)	**(0.029)**	(0.021)	(0.021)	**(0.025)**	**(0.022)**	(0.019)	(0.029)	(0.022)
β_3	EB	0.020	0.010	**0.042**	0.013	0.008	**-0.010**	**0.004**	0.017	0.007	0.021
	ESD.	(0.159)	(0.168)	**(0.175)**	(0.173)	(0.155)	**(0.156)**	**(0.149)**	(0.167)	(0.182)	(0.166)
σ^2_μ	EB	-0.023	-0.022	**0.082**	-0.027	0.081	**0.166**	**-0.014**	-0.012	-0.008	0.048
	ESD	(0.091)	(0.085)	**(0.102)**	(0.088)	(0.105)	**(0.090)**	**(0.101)**	(0.096)	(0.094)	(0.094)
					40% Missing Values						
β_0	EB	-0.005	-0.098	**-0.290**	-0.038	0.023	**-0.071**	**0.026**	-0.029	-0.021	-0.152
	ESD	(0.326)	(0.309)	**(0.864)**	(0.291)	(0.415)	**(0.270)**	**(0.288)**	(0.251)	(0.272)	(0.632)
β_1	EB	0.013	0.010	**0.013**	0.013	0.019	**0.013**	**0.003**	0.006	0.000	0.002
	ESD	(0.054)	(0.061)	**(0.039)**	(0.035)	(0.033)	**(0.038)**	**(0.035)**	(0.039)	(0.038)	(0.042)
β_2	EB	0.017	0.053	**0.056**	-0.061	-0.061	**-0.069**	**0.000**	0.010	-0.017	-0.003
	ESD	(0.048)	(0.026)	**(0.025)**	(0.019)	(0.022)	**(0.023)**	**(0.023)**	(0.021)	(0.036)	(0.022)
β_3	EB	0.016	0.005	**0.040**	0.010	-0.002	**-0.010**	**0.008**	0.020	0.001	0.016
	ESD	(0.181)	(0.160)	**(0.178)**	(0.173)	(0.156)	**(0.160)**	**(0.150)**	(0.166)	(0.174)	(0.165)
σ^2_μ	EB	-0.024	-0.024	**0.107**	-0.029	0.105	**0.201**	**-0.012**	-0.012	-0.008	0.061
	ESD	(0.091)	(0.085)	**(0.102)**	(0.086)	(0.102)	**(0.091)**	**(0.104)**	(0.098)	(0.097)	(0.090)

*Model used: $log\left\{\frac{p_{ijt}(3|2)}{p_{ijt}(2|2)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; EB=$\hat{\theta}_{mk} - \theta_k$; ESD=empirical standard deviation

Table 5.12: Average empirical biases (EB) with ESD for parameters for the quitter-to-smoker transition based on 500 simulations*

		Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter		**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values						
β_0	EB	0.057	0.066	**0.101**	0.022	0.050	**0.231**	**0.031**	0.018	0.074	0.066
	ESD	(0.495)	(0.449)	**(0.470)**	(0.459)	(0.532)	**(0.392)**	**(0.394)**	(0.373)	(0.403)	(0.437)
β_1	EB	-0.004	-0.018	**-0.090**	-0.013	-0.093	**-0.091**	**-0.001**	-0.020	-0.008	0.003
	ESD	(0.082)	(0.066)	**(0.045)**	(0.042)	(0.040)	**(0.039)**	**(0.055)**	(0.068)	(0.053)	(0.072)
β_2	EB	-0.013	-0.020	**-0.027**	-0.016	-0.019	**-0.027**	**-0.003**	-0.008	-0.012	-0.023
	ESD	(0.046)	(0.135)	**(0.028)**	(0.096)	(0.037)	**(0.032)**	**(0.037)**	(0.033)	(0.040)	(0.034)
β_3	EB	-0.122	0.119	**-0.385**	-0.102	-0.226	**-0.385**	**0.049**	0.041	0.032	-0.137
	ESD	(0.413)	(0.392)	**(0.432)**	(0.541)	(0.455)	**(0.450)**	**(0.383)**	(0.496)	(0.394)	(0.443)
σ^2_μ	EB	-0.038	-0.070	**-0.089**	-0.022	0.142	**0.240**	**-0.024**	-0.026	-0.019	-0.036
	ESD	(0.180)	(0.151)	**(0.118)**	(0.173)	(0.194)	**(0.137)**	**(0.157)**	(0.154)	(0.160)	(0.140)
					40% Missing Values						
β_0	EB	0.089	0.081	**0.065**	0.067	0.026	**0.230**	**0.029**	0.026	0.060	0.057
	ESD	(0.542)	(0.506)	**(0.515)**	(0.461)	(0.537)	**(0.466)**	**(0.450)**	(0.377)	(0.433)	(0.506)
β_1	EB	-0.003	-0.018	**-0.100**	-0.013	-0.099	**-0.098**	**0.003**	-0.013	-0.010	-0.001
	ESD	(0.079)	(0.072)	**(0.047)**	(0.042)	(0.045)	**(0.048)**	**(0.059)**	(0.076)	(0.055)	(0.084)
β_2	EB	0.010	0.036	**-0.027**	0.027	-0.017	**-0.025**	**-0.001**	-0.008	-0.012	-0.021
	ESD	(0.048)	(0.120)	**(0.029)**	(0.086)	(0.041)	**(0.038)**	**(0.040)**	(0.033)	(0.040)	(0.028)
β_3	EB	-0.093	0.146	**-0.460**	-0.062	-0.291	**-0.474**	**0.060**	0.064	0.030	-0.214
	ESD	(0.446)	(0.405)	**(0.468)**	(0.547)	(0.420)	**(0.447)**	**(0.408)**	(0.502)	(0.403)	(0.470)
σ^2_μ	EB	-0.056	-0.084	**-0.059**	-0.038	0.225	**0.285**	**-0.023**	-0.025	-0.017	-0.014
	ESD	(0.182)	(0.148)	**(0.144)**	(0.171)	(0.226)	**(0.151)**	**(0.167)**	(0.159)	(0.166)	(0.142)

*Model used: $log\left\{\frac{p_{ijt}(2|3)}{p_{ijt}(3|3)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; EB=$\hat{\theta}_{mk} - \theta_k$; ESD=empirical standard deviation

Time (a proxy for grade): Empirical bias estimates from the restriction methods are smaller than those from other methods. For the three restriction methods, the estimates from CCMVPM are much smaller than those from the other two methods, but only for the nonsmoker-to-smoker transition. ACMVPM performs better in quitter-to-smoker and smoker-to-quitter transition. The parameter estimates for the time-dependent covariates (β_1) are similar for all methods and all transitions.

Variance of random intercept: LOCF consistently has higher biases than the other methods for the variance estimate of the random intercept especially under the MNAR missing data mechanism. Similarly under the MNAR missing data mechanism, all three restriction methods produce smaller biases than the standard methods. Comparing the three restriction methods, the results are mixed; for example, in the non-smoker to smoker transition NCMVPM provides smaller biases than the ACMVPM method but these differences are very small. The biases from the selection model are similar to or greater than the biases from the pattern mixture methods.

In conclusion, these results suggest that the biases are low if we use restriction methods with predictive mean matching method. Furthermore, among the pattern mixture models, ACMVPM performed slightly better than the other two methods. The selection model produced slightly greater empirical bias than the pattern mixture methods.

5.7.3 Average Standardized Empirical Bias Estimates

In addition to measuring the raw bias, we calculated the average standardized empirical bias (ASEB) of the parameter estimate. The standardized empirical bias includes the model-based estimate of the standard error of the parameters, and is useful for understanding the impact of

bias on the interval estimates and statistical tests.

$$\text{ASEB}_{\text{k}} = \frac{\sum_{m=1}^{500} \left(\frac{\hat{\theta}_{mk} - \theta_k}{SE(\hat{\theta}_{mk})} \right)}{500}$$

The simulation results with 30% and 40% missing data are reported in Tables 5.13–5.15. The results are similar to those for the raw empirical bias estimates. In conclusion, The results suggest that the restriction methods consistently provide lower standardized empirical biases than the other methods. Of the restriction methods, ACMVPM performed the best. LOCF produced larger standardized biases than all other methods. Similar trends were observed for all transitions and both missing-data proportions.

Table 5.13: Average standardized empirical biases for the nonsmoker-to-smoker transition based on 500 simulations*

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values					
β_0	0.259	0.652	**0.185**	0.558	0.561	**1.404**	**0.053**	-0.046	-0.202	-0.023
β_1	-0.266	-0.588	**-0.695**	-1.148	-1.146	**-1.184**	**0.113**	0.010	0.063	0.125
β_2	-0.004	-0.436	**-1.155**	-0.812	-1.011	**-1.275**	**-0.067**	-0.001	0.239	-0.044
β_3	-0.003	0.327	**-0.389**	0.944	-1.910	**-1.870**	**0.042**	0.074	-0.452	0.075
σ^2_μ	-0.242	0.348	**-0.663**	0.484	-0.440	**-0.890**	**-0.167**	0.220	-0.128	-0.043
					40% Missing Values					
β_0	0.306	1.151	**0.077**	0.843	0.853	**1.636**	**0.068**	-0.010	-0.363	0.013
β_1	-0.337	-1.254	**0.082**	-1.338	-1.276	**-1.385**	**0.043**	-0.007	0.129	0.084
β_2	-0.152	-1.853	**-1.882**	-1.961	-1.860	**-1.848**	**-0.048**	0.027	0.663	-0.090
β_3	0.130	0.500	**0.005**	1.385	-1.441	**-1.476**	**0.052**	0.057	-1.251	0.093
σ^2_μ	-0.023	0.680	**-0.653**	0.273	-0.647	**-1.078**	**-0.162**	0.744	-0.145	-0.069

*Model used: $log\left\{\frac{p_{ijt}(2|1)}{p_{ijt}(1|1)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.14: Average standardized empirical biases for the smoker-to-quitter transition based on 500 simulations*

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values					
β_0	0.083	-0.203	**-0.279**	-0.040	0.197	**-0.621**	**0.089**	-0.100	-0.041	-0.013
β_1	0.161	0.162	**0.415**	0.318	0.488	**0.398**	**0.026**	0.123	0.165	0.079
β_2	0.319	0.927	**1.513**	-1.264	-1.359	**-1.527**	**-0.008**	0.576	-0.357	-0.175
β_3	0.106	0.060	**0.270**	0.081	0.043	**-0.080**	**0.029**	0.105	0.036	0.122
σ^2_μ	-0.284	-0.280	**0.866**	-0.334	0.942	**1.480**	**-0.139**	-0.131	-0.087	0.529
					40% Missing Values					
β_0	-0.016	-0.372	**-0.724**	-0.166	0.083	**-0.355**	**0.090**	-0.119	-0.084	-0.299
β_1	0.385	0.213	**0.471**	0.392	0.570	**0.400**	**0.076**	0.186	0.011	0.061
β_2	0.536	1.657	**2.077**	-1.433	-1.587	**-1.793**	**-0.001**	0.533	-0.539	-0.130
β_3	0.096	0.029	**0.253**	0.073	-0.017	**-0.086**	**0.058**	0.143	0.007	0.107
σ^2_μ	-0.335	-0.353	**1.127**	-0.378	1.181	**1.675**	**-0.123**	-0.143	-0.096	0.727

*Model used: $log\left\{\frac{p_{ijt}(3|2)}{p_{ijt}(2|2)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.15: Average standardized empirical biases for the quitter-to-smoker transition based on 500 simulations*

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values					
β_0	0.142	0.215	**0.347**	0.061	0.134	**0.895**	**0.083**	0.058	0.203	0.165
β_1	-0.070	-0.365	**-1.567**	-0.296	-1.938	**-1.880**	**-0.024**	-0.275	-0.134	0.030
β_2	-0.369	-0.218	**-1.020**	-0.211	-0.559	**-0.966**	**-0.073**	-0.299	-0.298	-0.661
β_3	-0.302	0.322	**-0.853**	-0.236	-0.691	**-0.986**	**0.133**	0.111	0.082	-0.329
σ^2_μ	-0.251	-0.564	**-0.835**	-0.165	0.971	**1.436**	**-0.159**	-0.167	-0.120	-0.271
					40% Missing Values					
β_0	0.216	0.259	**0.214**	0.202	0.063	**0.827**	**0.072**	0.074	0.159	0.142
β_1	-0.044	-0.354	**-1.679**	-0.296	-1.998	**-1.967**	**0.039**	-0.176	-0.163	0.061
β_2	0.285	0.406	**-1.158**	0.401	-0.487	**-0.963**	**-0.039**	-0.274	-0.305	-0.608
β_3	-0.227	0.383	**-0.979**	-0.144	-0.886	**-1.202**	**0.161**	0.162	0.074	-0.499
σ^2_μ	-0.376	-0.677	**-0.532**	-0.295	1.444	**1.590**	**-0.158**	-0.168	-0.112	-0.104

*Model used: $log\left\{\frac{p_{ijt}(2|3)}{p_{ijt}(3|3)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

5.7.4 Root Mean Square Error

The mean squared error (MSE) of an estimate is defined as the squared empirical bias plus the corresponding variance. This is a useful diagnostic component for selecting an estimator, since small MSE values indicate small variance as well as bias. The square root of the mean squared error (RMSE) is defined as the positive square root of the mean squared error.

Tables 5.16–5.18 show the RMSE for the simulated parameter estimates with 30% and 40% missing data. The results are similar to those for the empirical bias estimates. As usual, higher RMSE are observed for the treatment conditions in all methods. The RMSE for the restriction methods are lower than those for either the complete case or LOCF methods and the selection method. In conclusion, the results suggest that the restriction methods and the selection method consistently provide lower RMSEs than the other methods. Of the restriction methods, ACMVPM performed the best. LOCF produced larger RMSEs than any other method. Similar trends were observed for all transitions and both missing-data proportions.

Table 5.16: Root mean square error for the nonsmoker-to-smoker transition based on 500 simulations*

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values					
β_0	0.258	0.488	**0.520**	0.247	0.255	**0.331**	**0.223**	0.250	0.278	0.229
β_1	0.026	0.042	**0.053**	0.042	0.038	**0.042**	**0.029**	0.026	0.029	0.029
β_2	0.045	0.090	**0.097**	0.083	0.074	**0.085**	**0.024**	0.026	0.045	0.022
β_3	0.220	0.260	**0.253**	0.282	0.421	**0.420**	**0.203**	0.221	0.253	0.203
σ^2_μ	0.081	0.084	**0.088**	0.101	0.094	**0.093**	**0.086**	0.084	0.088	0.080
					40% Missing Values					
β_0	0.300	0.532	**0.565**	0.268	0.280	**0.368**	**0.232**	0.252	0.329	0.234
β_1	0.032	0.051	**0.062**	0.046	0.042	**0.047**	**0.031**	0.029	0.029	0.030
β_2	0.055	0.107	**0.124**	0.111	0.105	**0.121**	**0.025**	0.026	0.062	0.024
β_3	0.224	0.269	**0.249**	0.326	0.337	**0.354**	**0.204**	0.220	0.296	0.202
σ^2_μ	0.068	0.090	**0.094**	0.093	0.098	**0.100**	**0.087**	0.086	0.088	0.079

*Model used: $log\left\{\frac{p_{ijt}(2|1)}{p_{ijt}(1|1)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; RMSE=$\sqrt{\frac{\sum_{m=1}^{500}(\hat{\theta}_{mk}-\theta_k)^2}{500}}$

Table 5.17: Root mean square error for the smoker-to-quitter transition based on 500 simulations*

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values					
β_0	0.338	0.333	**0.858**	0.275	0.351	**0.283**	**0.284**	0.241	0.274	0.623
β_1	0.048	0.055	**0.038**	0.035	0.037	**0.039**	**0.033**	0.039	0.034	0.040
β_2	0.039	0.044	**0.049**	0.055	0.054	**0.062**	**0.022**	0.021	0.031	0.023
β_3	0.160	0.168	**0.180**	0.174	0.156	**0.157**	**0.149**	0.168	0.182	0.168
σ^2_μ	0.094	0.088	**0.131**	0.092	0.133	**0.189**	**0.102**	0.096	0.095	0.105
					40% Missing Values					
β_0	0.326	0.325	**0.911**	0.294	0.415	**0.279**	**0.289**	0.252	0.273	0.650
β_1	0.056	0.061	**0.042**	0.038	0.038	**0.040**	**0.035**	0.039	0.038	0.042
β_2	0.050	0.059	**0.061**	0.063	0.065	**0.073**	**0.023**	0.023	0.039	0.022
β_3	0.181	0.160	**0.183**	0.174	0.156	**0.160**	**0.150**	0.167	0.174	0.166
σ^2_μ	0.094	0.088	**0.148**	0.091	0.146	**0.220**	**0.105**	0.099	0.098	0.109

*Model used: $log\left\{\frac{p_{ijt}(3|2)}{p_{ijt}(2|2)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; RMSE=$\sqrt{\frac{\sum_{m=1}^{500}(\hat{\theta}_{mk}-\theta_k)^2}{500}}$

Table 5.18: Root mean square error for the quitter-to-smoker transition based on 500 simulations*

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values					
β_0	0.057	0.066	**0.101**	0.022	0.050	**0.231**	**0.031**	0.018	0.074	0.066
β_1	0.004	0.018	**0.090**	0.013	0.093	**0.091**	**0.001**	0.020	0.008	0.003
β_2	0.013	0.020	**0.027**	0.016	0.019	**0.027**	**0.003**	0.008	0.012	0.023
β_3	0.122	0.119	**0.385**	0.102	0.226	**0.385**	**0.049**	0.041	0.032	0.137
σ^2_μ	0.038	0.070	**0.089**	0.022	0.142	**0.240**	**0.024**	0.026	0.019	0.036
					40% Missing Values					
β_0	0.089	0.081	**0.065**	0.067	0.026	**0.230**	**0.029**	0.026	0.060	0.057
β_1	0.003	0.018	**0.100**	0.013	0.099	**0.098**	**0.003**	0.013	0.010	0.001
β_2	0.010	0.036	**0.027**	0.027	0.017	**0.025**	**0.001**	0.008	0.012	0.021
β_3	0.093	0.146	**0.460**	0.062	0.291	**0.474**	**0.060**	0.064	0.030	0.214
σ^2_μ	0.056	0.084	**0.059**	0.038	0.225	**0.285**	**0.023**	0.025	0.017	0.014

*Model used: $log\left\{\frac{p_{ijt}(2|3)}{p_{ijt}(3|3)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; RMSE=$\sqrt{\frac{\sum_{m=1}^{500}(\hat{\theta}_{mk}-\theta_k)^2}{500}}$

5.7.5 Coverage Rates

Using the estimated parameters and their model based standard errors, 95% confidence intervals were calculated for each parameter in each iteration. The coverage rate for a given method is defined as the ratio of the number of iterations in which the calculated confidence interval contains the true value of the parameter to the total number of iterations. The nominal 95% coverage rates are reported in Tables 5.24–5.26 in the next section.

As expected, LOCF under the MNAR missing data mechanism consistently provides the lowest coverage rates for almost all the parameters in all three transitions. The results from all three restriction methods show that the estimated coverage rates are much higher than the standard methods. The results from selection models are similar to those from the restriction methods. As the missing data proportion increase from 30% to 40%, the coverage rates for restriction methods and selection models continue to be better than those for the standard methods. In some cases the standard methods provide coverage rates less than 0.5.

5.8 Sensitivity Analysis

The purpose of this section is to investigate the situation where the probability of the missing data depends more on the unobserved data (i.e., to make the data more MNAR). In this section missing data were created with the values $m_{31} = 0.9$, $m_{32} = 0.9$, and $m_{33} = 0.9$ as defined earlier. To retain the 30% missing-data proportion, the intercepts were varied while keeping all other parameter values constant. Similarly to Section 5.1, this section compares the performance of standard methods (i.e., complete case analysis and LOCF) under MCAR, MAR, and MNAR

against the pattern mixture model and the selection model with the predictive mean matching method. The results are shown in Tables 5.19–5.22. Bayesian analysis was also performed on five simulated datasets using the WinBUGS software. Table 5.28 compares the Bayesian analysis with the pattern mixture and selection models.

Table 5.19 reports the average parameter estimates over 500 simulations under MCAR, MAR, and MNAR, estimates from LOCF under MCAR, MAR, and MNAR, estimates from the three restriction methods for the pattern mixture model (ACMVPM, CCMVPM, NCMVPM) with the predictive mean matching method under MNAR, and estimates from the selection model with the predictive mean matching method under MNAR.

Table 5.19: Average estimates for parameters for all three transitions based on 500 simulations* where (m_{3j}=0.9).

Parameter		TRUE	FDA	Complete Case Analysis MCAR	MAR	MNAR	LOCF MCAR	MAR	Pattern Mixture (MNAR) MNAR	ACMVPM	CCMVPM	NCMVPM	SMPM
Non-smoker to smoker transition at 30% missing data													
β_0	Mean	-2.3	-2.295	-2.243	-2.263	**-2.398**	-2.237	-2.233	**-2.068**	**-2.290**	-2.311	-2.299	-2.302
	ESD		(0.218)	(0.261)	(0.378)	**(0.397)**	(0.214)	(0.222)	**(0.190)**	**(0.221)**	(0.242)	(0.246)	(0.225)
β_1	Mean	0.2	0.201	0.189	0.185	**0.186**	0.173	0.173	**0.171**	**0.203**	0.200	0.194	0.202
	ESD		(0.022)	(0.028)	(0.038)	**(0.040)**	(0.028)	(0.025)	**(0.026)**	**(0.027)**	(0.026)	(0.026)	(0.029)
β_2	Mean	0.61	0.608	0.623	0.660	**0.635**	0.627	0.629	**0.602**	**0.609**	0.609	0.604	0.610
	ESD		(0.019)	(0.053)	(0.053)	**(0.058)**	(0.034)	(0.037)	**(0.033)**	**(0.024)**	(0.025)	(0.037)	(0.020)
β_3	Mean	-4.1	-4.071	-4.136	-4.060	**-4.246**	-3.996	-4.545	**-4.520**	**-4.088**	-4.082	-4.146	-4.082
	ESD		(0.187)	(0.205)	(0.221)	**(0.251)**	(0.218)	(0.213)	**(0.237)**	**(0.199)**	(0.214)	(0.204)	(0.195)
σ^2_μ	Mean	0.68	0.671	0.629	0.685	**0.616**	0.724	0.663	**0.631**	**0.664**	0.660	0.670	0.676
	ESD		(0.079)	(0.080)	(0.073)	**(0.080)**	(0.095)	(0.090)	**(0.073)**	**(0.085)**	(0.066)	(0.087)	(0.082)
Smoker to quitter transition at 30% missing data													
β_0	Mean	0.8	0.800	0.882	0.802	**0.835**	0.816	0.867	**0.703**	**0.836**	0.772	0.800	0.769
	ESD		(0.233)	(0.331)	(0.309)	**(0.803)**	(0.267)	(0.358)	**(0.264)**	**(0.276)**	(0.231)	(0.261)	(0.632)
β_1	Mean	-0.1	-0.099	-0.083	-0.093	**-0.089**	-0.090	-0.088	**-0.091**	**-0.098**	-0.097	-0.100	-0.100
	ESD		(0.028)	(0.048)	(0.051)	**(0.032)**	(0.033)	(0.033)	**(0.035)**	**(0.032)**	(0.037)	(0.033)	(0.037)
β_2	Mean	-0.3	-0.304	-0.290	-0.287	**-0.274**	-0.344	-0.341	**-0.345**	**-0.301**	-0.290	-0.306	-0.303
	ESD		(0.016)	(0.035)	(0.025)	**(0.024)**	(0.019)	(0.019)	**(0.018)**	**(0.020)**	(0.018)	(0.030)	(0.021)
β_3	Mean	0.2	0.215	0.215	0.196	**0.226**	0.215	0.216	**0.191**	**0.200**	0.215	0.209	0.222
	ESD		(0.147)	(0.168)	(0.162)	**(0.175)**	(0.167)	(0.154)	**(0.150)**	**(0.149)**	(0.161)	(0.164)	(0.164)
σ^2_μ	Mean	0.68	0.675	0.663	0.662	**0.752**	0.660	0.745	**0.809**	**0.670**	0.670	0.675	0.723
	ESD		(0.089)	(0.092)	(0.085)	**(0.100)**	(0.087)	(0.112)	**(0.094)**	**(0.103)**	(0.093)	(0.092)	(0.095)
Quitter to smoker transition at 30% missing data													
β_0	Mean	-1.7	-1.681	-0.870	-1.649	**-1.587**	-1.696	-1.692	**-1.427**	**-1.624**	-1.690	-1.612	-1.603
	ESD		(0.326)	(0.540)	(0.393)	**(0.507)**	(0.417)	(0.508)	**(0.360)**	**(0.390)**	(0.335)	(0.356)	(0.448)
β_1	Mean	0.3	0.296	0.299	0.270	**0.217**	0.287	0.214	**0.214**	**0.294**	0.279	0.293	0.296
	ESD		(0.054)	(0.077)	(0.063)	**(0.042)**	(0.042)	(0.034)	**(0.045)**	**(0.047)**	(0.067)	(0.049)	(0.066)
β_2	Mean	0.1	0.096	0.218	0.031	**0.069**	0.051	0.084	**0.070**	**0.093**	0.092	0.090	0.083
	ESD		(0.028)	(0.045)	(0.115)	**(0.031)**	(0.076)	(0.038)	**(0.029)**	**(0.035)**	(0.029)	(0.036)	(0.037)
β_3	Mean	-5.5	-5.507	-5.630	-5.406	**-5.819**	-5.591	-5.660	**-5.783**	**-5.432**	-5.436	-5.464	-5.583
	ESD		(0.324)	(0.410)	(0.378)	**(0.437)**	(0.503)	(0.467)	**(0.417)**	**(0.381)**	(0.465)	(0.379)	(0.412)
σ^2_μ	Mean	0.68	0.673	0.677	0.619	**0.552**	0.678	0.822	**0.889**	**0.651**	0.658	0.666	0.626
	ESD		(0.133)	(0.167)	(0.151)	**(0.104)**	(0.170)	(0.163)	**(0.140)**	**(0.151)**	(0.147)	(0.153)	(0.118)

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; Mean estimate; ESD=empirical standard deviation

Table 5.20: Average empirical biases for parameters for all three transitions based on 500 simulations* where (m_{3j}=0.9).

Parameter		Complete Case Analysis MCAR	MAR	MNAR	LOCF MCAR	MAR	MNAR	Pattern Mixture (MNAR) ACMVPM	CCMVPM	NCMVPM	SMPM
Non-smoker to smoker transition at 30% missing data											
β_0	EB	0.057	0.037	**-0.098**	0.063	0.067	**0.232**	**0.010**	-0.011	0.001	-0.002
	ESD	(0.261)	(0.378)	**(0.397)**	(0.214)	(0.222)	**(0.190)**	**(0.221)**	(0.242)	(0.246)	(0.225)
β_1	EB	-0.011	-0.015	**-0.014**	-0.027	-0.027	**-0.029**	**0.003**	0.000	-0.006	0.002
	ESD	(0.028)	(0.038)	**(0.040)**	(0.028)	(0.025)	**(0.026)**	**(0.027)**	(0.026)	(0.026)	(0.029)
β_2	EB	0.013	0.050	**0.025**	0.017	0.019	**-0.008**	**-0.001**	-0.001	-0.006	0.000
	ESD	(0.053)	(0.053)	**(0.058)**	(0.034)	(0.037)	**(0.033)**	**(0.024)**	(0.025)	(0.037)	(0.020)
β_3	EB	-0.036	0.040	**-0.146**	0.104	-0.445	**-0.420**	**0.012**	0.018	-0.046	0.018
	ESD	(0.205)	(0.221)	**(0.251)**	(0.218)	(0.213)	**(0.237)**	**(0.199)**	(0.214)	(0.204)	(0.195)
σ^2_μ	EB	-0.051	0.005	**-0.064**	0.044	-0.017	**-0.049**	**-0.016**	-0.020	-0.010	-0.004
	ESD	(0.080)	(0.073)	**(0.080)**	(0.095)	(0.090)	**(0.073)**	**(0.085)**	(0.066)	(0.087)	(0.082)
Smoker to quitter transition at 30% missing data											
β_0	EB	0.082	0.002	**0.035**	0.016	0.067	**-0.097**	**0.036**	-0.028	0.000	-0.031
	ESD	(0.331)	(0.309)	**(0.803)**	(0.267)	(0.358)	**(0.264)**	**(0.276)**	(0.231)	(0.261)	(0.632)
β_1	EB	0.017	0.007	**0.011**	0.010	0.012	**0.009**	**0.002**	0.003	0.000	0.000
	ESD	(0.048)	(0.051)	**(0.032)**	(0.033)	(0.033)	**(0.035)**	**(0.032)**	(0.037)	(0.033)	(0.037)
β_2	EB	0.010	0.013	**0.026**	-0.044	-0.041	**-0.045**	**-0.001**	0.010	-0.006	-0.003
	ESD	(0.035)	(0.025)	**(0.024)**	(0.019)	(0.019)	**(0.018)**	**(0.020)**	(0.018)	(0.030)	(0.021)
β_3	EB	0.015	-0.004	**0.026**	0.015	0.016	**-0.009**	**0.000**	0.015	0.009	0.022
	ESD	(0.168)	(0.162)	**(0.175)**	(0.167)	(0.154)	**(0.150)**	**(0.149)**	(0.161)	(0.164)	(0.164)
σ^2_μ	EB	-0.017	-0.018	**0.072**	-0.020	0.065	**0.129**	**-0.010**	-0.010	-0.005	0.043
	ESD	(0.092)	(0.085)	**(0.100)**	(0.087)	(0.112)	**(0.094)**	**(0.103)**	(0.093)	(0.092)	(0.095)
Quitter to smoker transition at 30% missing data											
β_0	EB	0.830	0.051	**0.113**	0.004	0.008	**0.273**	**0.076**	0.010	0.088	0.097
	ESD	(0.540)	(0.393)	**(0.507)**	(0.417)	(0.508)	**(0.360)**	**(0.390)**	(0.335)	(0.356)	(0.448)
β_1	EB	-0.001	-0.030	**-0.083**	-0.013	-0.086	**-0.086**	**-0.006**	-0.021	-0.007	-0.004
	ESD	(0.077)	(0.063)	**(0.042)**	(0.042)	(0.034)	**(0.045)**	**(0.047)**	(0.067)	(0.049)	(0.066)
β_2	EB	0.118	-0.069	**-0.031**	-0.049	-0.016	**-0.030**	**-0.007**	-0.008	-0.010	-0.017
	ESD	(0.045)	(0.115)	**(0.031)**	(0.076)	(0.038)	**(0.029)**	**(0.035)**	(0.029)	(0.036)	(0.037)
β_3	EB	-0.130	0.094	**-0.319**	-0.091	-0.160	**-0.283**	**0.068**	0.064	0.036	-0.083
	ESD	(0.410)	(0.378)	**(0.437)**	(0.503)	(0.467)	**(0.417)**	**(0.381)**	(0.465)	(0.379)	(0.412)
σ^2_μ	EB	-0.003	-0.061	**-0.128**	-0.002	0.142	**0.209**	**-0.029**	-0.022	-0.014	-0.054
	ESD	(0.167)	(0.151)	**(0.104)**	(0.170)	(0.163)	**(0.140)**	**(0.151)**	(0.147)	(0.153)	(0.118)

*Model used: $log\left\{\frac{p_{ijt}(k|l)}{p_{ijt}(3|3)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$; EB=$\hat{\theta}_{mk} - \theta_k$; ESD=empirical standard deviation

Table 5.21: Average standardized empirical biases for parameters for all three transitions based on 500 simulations* where (m_{3j}=0.9).

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
Non-smoker to smoker transition at 30% missing data										
β_0	0.269	0.150	**-0.375**	0.338	0.357	**1.255**	**0.055**	-0.050	0.013	-0.007
β_1	-0.459	-0.494	**-0.612**	-1.133	-1.273	**-1.544**	**0.128**	0.000	-0.237	0.081
β_2	0.399	1.212	**0.784**	0.792	0.807	**-0.468**	**-0.031**	-0.029	-0.205	-0.021
β_3	-0.162	0.182	**-0.757**	0.559	-2.377	**-2.260**	**0.060**	0.077	-0.211	0.110
σ^2_μ	-0.613	0.068	**-0.901**	0.596	-0.244	**-0.817**	**-0.185**	-0.298	-0.121	-0.052
Smoker to quitter transition at 30% missing data										
β_0	0.332	0.015	**0.096**	0.064	0.289	**-0.560**	**0.140**	-0.122	0.007	-0.054
β_1	0.484	0.192	**0.585**	0.374	0.482	**0.377**	**0.068**	0.097	0.011	-0.002
β_2	0.486	0.819	**1.504**	-1.855	-1.889	**-1.910**	**-0.075**	0.575	-0.212	-0.170
β_3	0.100	-0.033	**0.214**	0.128	0.127	**-0.094**	**0.009**	0.111	0.060	0.194
σ^2_μ	-0.232	-0.274	**0.852**	-0.299	0.861	**1.347**	**-0.110**	-0.123	-0.063	0.534
Quitter to smoker transition at 30% missing data										
β_0	1.170	0.192	**0.453**	0.006	0.020	**1.377**	**0.209**	0.046	0.275	0.245
β_1	-0.025	-0.617	**-1.491**	-0.328	-2.045	**-2.000**	**-0.120**	-0.387	-0.144	-0.092
β_2	1.413	-0.771	**-1.150**	-0.723	-0.594	**-1.228**	**-0.223**	-0.339	-0.325	-0.497
β_3	-0.367	0.296	**-0.820**	-0.219	-0.510	**-1.009**	**0.191**	0.175	0.105	-0.200
σ^2_μ	-0.019	-0.521	**-1.407**	-0.002	0.975	**1.347**	**-0.222**	-0.156	-0.105	-0.471

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.22: Root mean square error for parameters for all three transitions based on 500 simulations* where (m_{3j}=0.9).

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
Non-smoker to smoker transition at 30% missing data										
β_0	0.267	0.379	**0.409**	0.223	0.231	**0.300**	**0.221**	0.243	0.246	0.225
β_1	0.030	0.041	**0.042**	0.039	0.036	**0.039**	**0.027**	0.026	0.026	0.029
β_2	0.055	0.073	**0.063**	0.038	0.042	**0.034**	**0.024**	0.025	0.037	0.020
β_3	0.208	0.225	**0.291**	0.241	0.493	**0.482**	**0.200**	0.215	0.209	0.196
σ^2_μ	0.095	0.073	**0.102**	0.105	0.091	**0.087**	**0.086**	0.069	0.088	0.082
Smoker to quitter transition at 30% missing data										
β_0	0.341	0.309	**0.804**	0.267	0.365	**0.281**	**0.278**	0.232	0.261	0.633
β_1	0.051	0.051	**0.034**	0.035	0.035	**0.036**	**0.032**	0.037	0.033	0.037
β_2	0.036	0.028	**0.035**	0.048	0.045	**0.049**	**0.020**	0.021	0.030	0.021
β_3	0.169	0.162	**0.176**	0.168	0.155	**0.151**	**0.149**	0.162	0.164	0.166
σ^2_μ	0.094	0.087	**0.123**	0.089	0.129	**0.159**	**0.103**	0.093	0.092	0.105
Quitter to smoker transition at 30% missing data										
β_0	0.990	0.396	**0.519**	0.417	0.508	**0.452**	**0.397**	0.335	0.366	0.458
β_1	0.077	0.069	**0.093**	0.044	0.092	**0.097**	**0.048**	0.070	0.049	0.066
β_2	0.127	0.134	**0.044**	0.090	0.041	**0.042**	**0.035**	0.030	0.038	0.041
β_3	0.430	0.390	**0.541**	0.511	0.493	**0.504**	**0.387**	0.469	0.380	0.420
σ^2_μ	0.167	0.163	**0.165**	0.170	0.217	**0.251**	**0.154**	0.149	0.154	0.130

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.23: Comparison table: Average standardized empirical biases for parameters for all three cases and for all three transitions based on 500 simulations for different value of M_{3j}*

	30% Missing Values					
	$M_{3j} = 0.3$			$M_{3j} = 0.9$		
Parameter	**CCA**	**LOCF**	**ACMVPM**	**CCA**	**LOCF**	**ACMVPM**
Nonsmoker-to-smoker transition at 30% missing data						
β_0	0.185	1.404	0.053	-0.375	1.255	0.055
β_1	-0.695	-1.184	0.113	-0.612	-1.544	0.128
β_2	-1.155	-1.275	-0.067	0.784	-0.468	-0.031
β_3	-0.389	-1.870	0.042	-0.757	-2.260	0.060
σ^2_μ	-0.663	-0.890	-0.167	-0.901	-0.817	-0.185
Smoker-to-quitter transition at 30% missing data						
β_0	-0.279	-0.621	0.089	0.096	-0.560	0.140
β_1	0.415	0.398	0.026	0.585	0.377	0.068
β_2	1.513	-1.527	-0.008	1.504	-1.910	-0.075
β_3	0.270	-0.080	0.029	0.214	-0.094	0.009
σ^2_μ	0.866	1.480	-0.139	0.852	1.347	-0.110
Quitter-to-smoker transition at 30% missing data						
β_0	0.347	0.895	0.083	0.453	1.377	0.209
β_1	-1.567	-1.880	-0.024	-1.491	-2.000	-0.120
β_2	-1.020	-0.966	-0.073	-1.150	-1.228	-0.223
β_3	-0.853	-0.986	0.133	-0.820	-1.009	0.191
σ^2_μ	-0.835	1.436	-0.159	-1.407	1.347	-0.222

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

The results from a sensitivity analysis suggest the same trends as in the previous simulations where the missing data were less dependent on the future values. For the nonsmoker-to-smoker transition, comparing the mild and extreme values of the missing-data mechanism, in general the average empirical bias estimates (Table 5.20) for all parameters show an increase in bias. The pattern mixture model with the restriction methods shows a smaller bias increase than the other methods, especially in the estimate of variance of the random intercept. Similar trends are observed for the other two transitions.

Table 5.23 shows the comparative results for the average standardized empirical bias (ASEB) of an estimate. The results from Table 5.23 suggest the same trend as in previous simulations where the missing data were less dependent on the future values ($M_{3j} = 0.3$). Overall, in most cases, the ACMVPM methods show a smaller ASEB than the other methods, especially in the estimate of the variance of the random intercept.

The coverage-rate results (Tables 5.24–5.26) show that for the ACMVPM, all the estimated coverage rates are within the (0.93,0.97) interval mentioned earlier. The results from the selection model are similar to the restriction methods and have a much higher coverage rate than those of the other methods. These results demonstrate that the restriction methods produced better overall results relative to standard methods and that with data more dependent on the missing response, the restriction methods perform well.

Table 5.24: Nominal 95% coverage rates for parameters for all three cases based on 500 simulations: Nonsmoker-to-smoker transition*.

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values (with M_{3j}=0.3)					
β_0	0.904	0.778	**0.722**	0.886	0.858	**0.738**	**0.948**	0.920	0.922	0.936
β_1	0.944	0.946	**0.634**	0.772	0.836	**0.776**	**0.950**	0.970	0.940	0.954
β_2	0.920	0.732	**0.626**	0.762	0.686	**0.696**	**0.966**	0.952	0.854	0.966
β_3	0.972	0.918	**0.876**	0.788	0.512	**0.524**	**0.944**	0.940	0.968	0.952
σ^2_μ	0.954	0.942	**0.894**	0.822	0.826	**0.824**	**0.978**	0.926	0.912	0.934
					30% Missing Values (with M_{3j}=0.9)					
β_0	0.888	0.848	**0.802**	0.902	0.880	**0.778**	**0.948**	0.922	0.954	0.936
β_1	0.894	0.886	**0.730**	0.750	0.708	**0.616**	**0.940**	0.908	0.932	0.914
β_2	0.752	0.710	**0.684**	0.716	0.728	**0.698**	**0.952**	0.950	0.908	0.940
β_3	0.962	0.932	**0.802**	0.862	0.366	**0.426**	**0.934**	0.940	0.962	0.926
σ^2_μ	0.928	0.910	**0.802**	0.840	0.848	**0.796**	**0.942**	0.938	0.912	0.926
					40% Missing Values					
β_0	0.854	0.730	**0.682**	0.860	0.808	**0.666**	**0.940**	0.924	0.902	0.938
β_1	0.874	0.932	**0.552**	0.736	0.790	**0.722**	**0.944**	0.960	0.942	0.946
β_2	0.850	0.614	**0.556**	0.510	0.562	**0.586**	**0.956**	0.950	0.740	0.948
β_3	0.958	0.882	**0.824**	0.680	0.686	**0.622**	**0.936**	0.932	0.672	0.944
σ^2_μ	0.976	0.910	**0.850**	0.850	0.816	**0.786**	**0.972**	0.900	0.914	0.928

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.25: Nominal 95% coverage rates for parameters for all three cases based on 500 simulations: Smoker-to-quitter transition*.

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values (with M_{3j}=0.3)					
β_0	0.894	0.872	**0.642**	0.900	0.836	**0.792**	**0.956**	0.928	0.892	0.870
β_1	0.830	0.848	**0.842**	0.924	0.894	**0.886**	**0.944**	0.916	0.942	0.900
β_2	0.870	0.844	**0.678**	0.904	0.858	**0.734**	**0.946**	0.876	0.924	0.934
β_3	0.972	0.944	**0.894**	0.926	0.960	**0.898**	**0.978**	0.946	0.974	0.964
σ^2_μ	0.912	0.910	**0.834**	0.920	0.822	**0.716**	**0.934**	0.902	0.912	0.910
					30% Missing Values (with M_{3j}=0.9)					
β_0	0.854	0.852	**0.602**	0.910	0.762	**0.784**	**0.924**	0.934	0.896	0.880
β_1	0.818	0.824	**0.740**	0.866	0.854	**0.812**	**0.932**	0.896	0.928	0.894
β_2	0.720	0.686	**0.642**	0.556	0.544	**0.536**	**0.932**	0.872	0.910	0.922
β_3	0.902	0.846	**0.800**	0.834	0.864	**0.810**	**0.958**	0.916	0.942	0.860
σ^2_μ	0.880	0.854	**0.824**	0.866	0.768	**0.734**	**0.900**	0.892	0.888	0.870
					40% Missing Values					
β_0	0.916	0.886	**0.622**	0.876	0.768	**0.846**	**0.960**	0.924	0.880	0.876
β_1	0.774	0.828	**0.820**	0.902	0.896	**0.890**	**0.946**	0.918	0.922	0.894
β_2	0.778	0.670	**0.440**	0.862	0.728	**0.624**	**0.946**	0.870	0.898	0.950
β_3	0.920	0.966	**0.890**	0.886	0.912	**0.878**	**0.972**	0.910	0.952	0.946
σ^2_μ	0.876	0.864	**0.784**	0.906	0.758	**0.658**	**0.924**	0.882	0.898	0.878

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.26: Nominal 95% coverage rates for parameters for all three cases based on 500 simulations: Quitter-to-smoker transition*.

	Complete Case Analysis			LOCF			Pattern Mixture (MNAR)			
Parameter	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**ACMVPM**	**CCMVPM**	**NCMVPM**	**SMPM**
					30% Missing Values (with M_{3j}=0.3)					
β_0	0.886	0.824	**0.760**	0.816	0.794	**0.784**	**0.932**	0.956	0.922	0.920
β_1	0.866	0.854	**0.702**	0.928	0.514	**0.546**	**0.958**	0.938	0.966	0.950
β_2	0.864	0.794	**0.818**	0.858	0.886	**0.794**	**0.958**	0.906	0.942	0.902
β_3	0.926	0.918	**0.872**	0.866	0.814	**0.798**	**0.952**	0.874	0.950	0.928
σ^2_μ	0.900	0.858	**0.830**	0.842	0.776	**0.754**	**0.924**	0.916	0.914	0.924
					30% Missing Values (with M_{3j}=0.9)					
β_0	0.840	0.824	**0.660**	0.802	0.790	**0.648**	**0.918**	0.900	0.920	0.898
β_1	0.882	0.836	**0.738**	0.902	0.464	**0.496**	**0.956**	0.864	0.954	0.936
β_2	0.848	0.818	**0.766**	0.840	0.790	**0.750**	**0.916**	0.882	0.902	0.888
β_3	0.888	0.894	**0.836**	0.880	0.810	**0.702**	**0.938**	0.882	0.930	0.932
σ^2_μ	0.914	0.848	**0.696**	0.846	0.834	**0.778**	**0.898**	0.902	0.886	0.916
					40% Missing Values					
β_0	0.852	0.772	**0.744**	0.818	0.810	**0.748**	**0.920**	0.968	0.914	0.880
β_1	0.888	0.830	**0.656**	0.928	0.478	**0.498**	**0.960**	0.932	0.972	0.940
β_2	0.852	0.822	**0.748**	0.864	0.880	**0.744**	**0.948**	0.902	0.946	0.946
β_3	0.914	0.912	**0.832**	0.870	0.810	**0.730**	**0.946**	0.874	0.950	0.904
σ^2_μ	0.886	0.860	**0.822**	0.828	0.674	**0.724**	**0.902**	0.886	0.890	0.932

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

Table 5.27: Comparison table: Nominal 95% coverage rates for parameters for all three cases and for all three transitions based on 500 simulations*

	30% Missing Values						40% Missing Values		
	$M_{3j} = 0.3$			$M_{3j} = 0.9$			$M_{3j} = 0.3$		
Parameter	**CCA**	**LOCF**	**ACMVPM**	**CCA**	**LOCF**	**ACMVPM**	**CCA**	**LOCF**	**ACMVPM**
Nonsmoker-to-smoker transition									
β_0	0.722	0.738	0.948	0.802	0.778	0.948	0.682	0.666	0.940
β_1	0.634	0.776	0.950	0.730	0.616	0.940	0.552	0.722	0.944
β_2	0.626	0.696	0.966	0.684	0.698	0.952	0.556	0.586	0.956
β_3	0.876	0.524	0.944	0.802	0.426	0.934	0.824	0.622	0.936
σ^2_μ	0.894	0.824	0.978	0.802	0.796	0.942	0.850	0.786	0.972
Smoker-to-quitter transition									
β_0	0.642	0.792	0.956	0.602	0.784	0.924	0.622	0.846	0.960
β_1	0.842	0.886	0.944	0.740	0.812	0.932	0.820	0.890	0.946
β_2	0.678	0.734	0.946	0.642	0.536	0.932	0.440	0.624	0.946
β_3	0.894	0.898	0.978	0.800	0.810	0.958	0.890	0.878	0.972
σ^2_μ	0.834	0.716	0.934	0.824	0.734	0.900	0.784	0.658	0.924
Quitter-to-smoker transition									
β_0	0.760	0.784	0.932	0.660	0.648	0.918	0.744	0.748	0.920
β_1	0.702	0.546	0.958	0.738	0.496	0.956	0.656	0.498	0.960
β_2	0.818	0.794	0.958	0.766	0.750	0.916	0.748	0.744	0.948
β_3	0.872	0.798	0.952	0.836	0.702	0.938	0.832	0.730	0.946
σ^2_μ	0.830	0.754	0.924	0.696	0.778	0.898	0.822	0.724	0.902

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 X_{1ij,t-1} + \beta_2 t + \beta_3 C_j + u_{0j}$

At last, Table 5.27 shows the comparative results for the nominal 95% coverage rates for parameters, for all three cases, and for all three transitions. The three cases include the comparison between the 30% and 40% missing data rates and within the 30% missing data rates, two comparison were made for different values of m_{3j} (where the missing data is dependent on the future values). Overall, the ACMVPM methods shows a larger nominal coverage rates than the other methods, especially in the estimate of the variance of the random intercept.

5.9 Bayesian Analysis

In this section, pattern mixture models are used under the Bayesian framework and estimates are derived via the MCMC algorithm. The outcome variables and other model assumptions are as described in Section 5.1. The analysis was performed on five simulated datasets derived from the MNAR mechanism, created in Section 5.8.

As described in Section 4.3.4, two-pattern and seven-pattern models were fitted to data with 30% missing values. However, this thesis focuses on the two-pattern model because of the complexity of defining prior distributions for all the unidentified parameters in the seven-pattern model. In the two-pattern model, first an analysis was carried out for those individuals having complete data at all time points (pattern 1) and later an analysis was carried out on those individuals having at least one missing data point (pattern 2). Flat priors were assumed for all identified parameters except for the variance of the random effect. A dropout model (selection model) provided the prior information for unidentified parameters as described in Section 4.3.2. The analysis was conducted using WinBUGS 1.4.3.

For the seven-pattern model, the same hierarchical model was used for all seven patterns but the $\theta^{(p)}$ parameters were allowed to differ for each pattern p. As above, the parameter estimates from the selection model were used to provide the parameters for the prior distribution of unidentified parameters $\theta^{(p)}$ (p $>$ 1), while flat priors were assumed for the identified parameters $\theta^{(1)}$. Hierarchical designs were incorporated by introducing the normal random-effects distribution with a gamma prior assumed for the variance of the random effect.

For comparison purposes, five simulated data sets were used to estimate the parameters from the pattern mixture model and the Bayesian pattern mixture model. Within the pattern mixture model, the ACMVPM restriction with the predictive mean matching method was used to compare the results with the Bayesian pattern mixture model. The parameter estimates from each pattern were aggregated using Rubin's technique. Our results show that the proposed pattern mixture model produce estimates that, on average, were closer to the true values, than the Bayesian pattern mixture model.

Table 5.28 shows the parameter estimates from the pattern mixture model and the Bayesian pattern mixture model. The second column shows the true parameter values. The third and fourth columns show the parameter estimates based on the pattern mixture model with the ACMVPM restriction and the Bayesian pattern mixture model. The fifth column shows the parameter estimates from the Bayesian pattern mixture model using all seven patterns.

In the nonsmoker-to-smoker transition, the estimates from the pattern mixture model suggest that it performed better (i.e. less biased) than the Bayesian model, with the exception of the parameter estimate for time. The pattern mixture model with the ACMVPM restriction provides a better estimate for the variance of the random effect than the Bayesian pattern mixture

model. In conclusion, for the nonsmoker-to-smoker transition, the pattern mixture model with the ACMVPM restriction seems to perform better than the Bayesian pattern mixture model for this limited comparison.

For the smoker-to-quitter transition, the parameter estimate for the treatment condition has smaller bias for the pattern mixture model than for the two-pattern Bayesian pattern mixture model. Parameter estimates for the time dependent covariate and the variable time show similar results for both two-pattern Bayesian and pattern mixture models. As usual, when comparing the estimate of the variance for the random effect, the pattern mixture model provides a less biased estimate than the Bayesian pattern mixture model.

For the quitter-to-smoker transition, the parameter estimates for the time-dependent covariate and variable time (a proxy for grade) have smaller biases for the pattern mixture model than for the two-pattern Bayesian pattern mixture model. Similarly, when comparing the estimate of the variance for the random effect, the two-pattern Bayesian pattern mixture model provides a less biased estimate than the pattern mixture model and seven-pattern Bayesian pattern mixture models.

The results from all three transitions show that the overall parameter estimates for the intercept and the variance of the random effect have greater biases than the other parameters. These biases are greater in the quitter-to-smoker transition because only a few individuals return from the quitter to the smoker state. The other parameter estimates show similar results in terms of biases.

In general, the parameter estimates from the Bayesian pattern mixture models with the two-pattern model perform better than the seven-pattern model. The noticeable difference can be seen for the variance of the random effect estimates. The results from the seven-pattern model

shows more variability for variance of the random effect estimates than the two-pattern model.

Estimates obtained from the Bayesian analysis depend on a joint property of the data and the assumption of the prior distribution. Better and efficient inferences can be derived when the model incorporates more realistic prior assumptions (Kass & Raftery, 1995). For example, Kass & Raftery (1995) point out that choosing "non-informative" priors can force the hypothesis to favor the null hypothesis. Choosing an appropriate prior has always been an issue for researchers. As a result, in practice it is important to implement a sensitivity analysis to determine the influence of chosen priors. This becomes more complicated when there is an underlying assumption that missing data are missing not at random (MNAR), where inferences are sensitive to the proportion of missing data and the degree of dependence on the future outcome. In this thesis, other distributions were also assumed for the prior distributions for the regression parameters, including the log-normal distribution, similar to Kaciroti et al. (2006). During this process, we noticed that parameter estimates were sensitive to the choice of priors. Often convergence was not achieved, opposite signs of parameter estimates were obtained, and large biases were noticed for the parameter estimates. After several trials, a final solution was achieved using normal priors for the unidentified parameters and flat priors for the identified parameters. Convergence was always obtained when assuming a gamma prior for the variance of the random effect, so we did not check any other distributional form for the prior for this parameter.

Another limitation with the Bayesian analysis relates to the computational problems associated with estimating the posterior distributions. At present, there are only a few statistical software packages that offer Bayesian inference (e.g. WinBUGS and some of the SAS procedures) and these require the user to program complicated models. The method studied in this thesis does not require any advanced programming skills and does not incorporate any subjective assump-

tions related to prior distributions. It is straightforward to implement using a standard statistical software package (e.g. SAS).

Table 5.28: Average parameter estimates from the ACMVPM pattern mixture model and Bayesian pattern mixture models for five simulated datasets

	Actual Parameter	Pattern Mixture Model*	Bayesian Analysis (two pattern)	Bayesian Analysis (seven pattern)
Non-smoker to smoker transition				
Intercept	-2.30	-2.460 (0.187)	-2.341 (0.391)	-2.714 (0.335)
X1	0.20	0.206 (0.029)	0.182 (0.071)	0.276 (0.299)
Time	0.61	0.612 (0.025)	0.625 (0.340)	0.538 (0.298)
Condition	-4.10	-4.145 (0.159)	-4.171 (0.274)	-4.683 (0.325)
σ_u^2	0.68	0.679 (0.114)	0.721 (0.082)	0.615 (0.032)
Smoker to quitter transition				
Intercept	0.80	0.609 (0.237)	1.088 (0.145)	1.164 (0.141)
X1	-0.10	-0.096 (0.034)	-0.129 (0.024)	-0.092 (0.023)
Time	-0.30	-0.296 (0.023)	-0.311 (0.015)	-0.247 (0.015)
Condition	0.20	0.163 (0.116)	0.200 (0.068)	0.213 (0.068)
σ_u^2	0.68	0.685 (0.137)	0.738 (0.055)	0.811 (0.062)
Quitter to smoker transition				
Intercept	-1.7	-1.970 (0.331)	-1.493 (0.274)	-1.157 (0.250)
X1	0.3	0.303 (0.056)	0.294 (0.044)	0.233 (0.039)
Time	0.1	0.104 (0.041)	0.117 (0.032)	0.176 (0.028)
Condition	-5.5	-5.555 (0.300)	-5.628 (0.284)	-5.752 (0.292)
σ_u^2	0.68	0.685 (0.157)	0.809 (0.185)	0.878 (0.174)

* Empirical standard deviation shown in parentheses.

5.10 Simulation Conclusions

The purpose of this chapter was to extend the methods for handling missing data in binary hierarchical models by exploring the performance of different techniques under various missing-data conditions. The focus was on the monotone missing-data pattern. The methods investigated were:

1. Complete case analysis when data were generated under MCAR, MAR, or MNAR;
2. LOCF analysis when data were created for all three missing-data mechanisms;
3. Restriction methods when data were generated under MNAR, including:
   (a) CCMV with predictive mean (PM) method,
   (b) ACMV with PM method,
   (c) NCMV restriction with PM method;
4. Selection model with PM method when data were generated under MNAR;
5. Bayesian pattern mixture model when data were generated under MNAR (for five simulated datasets).

The results suggest that parameter estimates obtained by the different methods are dependent on the missing-data mechanism. In many cases, the standard method shows an increasing trend in bias occurring as the proportion of missing data increases from 30% to 40%. In some cases, LOCF method under the MNAR provides smaller biases as we increased the missing data proportion. This could be explained by fewer numbers of transitions in that state. Overall, the results

suggest that the restriction methods provide better results than the other methods studied. The selection model provides much smaller biases than LOCF and CCA. As the proportion of missing data increases, estimates from the selection model show more variation in their parameter estimates.

The biases vary with the individual transition. ACMVPM and NCMVPM are closer to each other when the transition is from nonsmoker to smoker. Of the restriction methods ACMVPM with the predictive mean matching approach produced the lowest average bias estimate because of its inherent ability to incorporate all the available data.

The results for the selection model were expected; that is, we expected that the selection model would perform better than standard methods such as CCA and LOCF. This is because the selection model is designed to adjust for the selection bias when non-ignorable missing data exist. Standard methods assume that the missing data are MAR or MCAR. In general, the results from the selection model were not as good as those from the pattern mixture model. The reason could be that the selection model was originally designed for non-ignorable missing data in a normally distributed situation. Heckman (1979) showed that this method can be used for non-normal distributions such as the logit and probit models studied in this thesis. For the discrete outcome variables, normal approximation methods (Lee, 1983) are used to calculate the inverse Mill's ratio from the first stage of the Heckman model. This approximation may be one reason that the Heckman two-stage selection model did not perform as well as the pattern mixture model in our simulation study for non-ignorable missing data.

Bayesian methods were applied to a few simulated datasets. The WinBUGS program was used to estimate the parameters and summarize the results. The parameter estimates are shown in

Table 5.28. All the parameter estimates computed by restriction methods are closer to the true values. In the seven pattern model, the parameter estimates from the Bayesian analysis are not as close to the restriction methods, when comparing with the true values. In non-smoker to smoker transition, Bayesian analysis provides lower estimated value for the variance of the random effect as compared to the restrictions methods. In the other two transitions, Bayesian analysis provides higher estimated values for the variance of the random effect.

A sensitivity analysis was performed by changing the missing-data parameter so the probability that the data were missing depending more on the unobserved response variable. This led to slightly more variation in the parameter estimates. Sensitivity analysis shows that the restriction method with predictive mean matching performs better than the other methods. Restriction methods are straightforward to use and our simulation shows some promising results with regard to reduced bias and RMSE. One advantage of restriction methods is that standard statistical methods can easily be applied, once the data are imputed and considered to be complete. Further research could be done to generalize these restriction methods for non-monotone missing data in the context of a hierarchical linear model.

It has been argued in the literature that fitting both a selection model and pattern mixture model can be a valuable sensitivity analysis tool (Michiels et al., 1999). Our results show some differences in parameter estimates between the selection model and the pattern mixture model. However, the conclusions from both models seem to be the same, increasing our confidence in our techniques and results.

The method proposed in this thesis seems to be superior to the standard methods and the Bayesian approach when dealing with non-ignorable missing data. The advantage of the proposed restric-

tion methods is that they allow an easy implementation of the discrete hierarchical model using a standard statistical software package (e.g. SAS). Furthermore, the proposed method can easily be extended to more complicated models such as cross-classified models and multiple membership models.

The implementation of the Bayesian approach has several drawbacks.

1. It is computationally expensive to run such a complex model.
2. The Bayesian analysis requires advanced programming skills to code both the model and the estimation procedures.
3. Convergence of the Bayesian analysis is an issue (for a more detailed review see: Cowles and Carlin, 1996)

Chapter 6

Application to WSPP3 Data

In this chapter, we consider the application of the proposed methodology to the third Waterloo Smoking Prevention Project (WSPP3). WSPP3 was conducted by the Population Health Research Group (previously known as the Health Behavior Research Group) at the University of Waterloo (Brown & Cameron, 1997; Cameron et al., 1999). The purpose of WSPP3 was to evaluate a social-influence smoking prevention program at the elementary level (grades 6 through 8) and an activity-based tobacco control program at the secondary level (grades 9 and 10). In addition, a longitudinal cohort of students was followed at grades 11 and 12 to assess the long-term impact of the intervention.

One hundred elementary schools from seven school boards in Southwestern Ontario, Canada participated in this study (fifteen schools from each of six boards and ten from the seventh board). These schools were randomly assigned in a four-to-one ratio to receive either an intensive anti-smoking public health education program or their standard school curriculum. A

detailed description is given by Driezen (2001).

Upon completion of the baseline data collection, students were classified into one of five smoking categories: never smoked, tried once, quitter, experimental smoker (smoked less than once a week), and regular smoker (smoked weekly). For the current analysis, only three categories were used: smoker (experimental smoker or regular smoker), nonsmoker (never smoked, tried once), and quitter. The data for this study are restricted to those students who participated in the WSPP3 study in grade 6. Individuals who were enrolled in later grades are excluded from this analysis.

The data for WSPP3 were collected based on a hierarchical design where the school was the unit of randomization, and variability in the smoking rates between schools existed. Any analysis of this dataset should take into account the variation between schools. In this analysis we capture such variation by introducing the random-effect distribution at the school level to capture the between-school variation.

Table 6.1 lists the proportion of students in the three smoking categories for each year of WSPP3. When the cohort was in grade 6, the majority of students had no smoking experience. Table 6.1 shows that smoking prevalence increased by grade. In grade 6, only 5.2% of 4456 students were classified as a smokers; in grade 12, 34% of students were classified as smokers.

Table 6.2 summarizes the transition of individuals from one smoking category to another. Among nonsmokers in grade 6, 24.5% remained nonsmokers over time. Students who reported regular smoking at grade 6 rarely moved out of this category. By the time the cohort reached high school, the probability that a regular smoker remained a regular smoker from one time point to the next was greater than 0.75.

Table 6.1: Smoking prevalence by grade

	Grade						
Smoking Status	*6*	*7*	*8*	*9*	*10*	*11*	*12*
% Missing	–	7.7	11.2	24.3	20.2	29	31.9
(n)	–	(342)	(500)	(1084)	(900)	(1292)	(1421)
% Nonsmoker	88.7	71.8	57.6	39.8	32.7	24.5	20.5
(n)	(3952)	(3200	(2567)	(1773)	(1458)	(1092)	(915)
% Quitter	6.1	11.1	13.5	13.2	13.1	13.4	14.0
(n)	(274)	(493)	(602)	(589)	(584)	(597)	(622)
% Smoker	5.2	9.4	17.7	22.7	34.0	33.1	33.6
(n)	(230)	(421)	(787)	(1010)	(1514)	(1475)	(1498)

Table 6.2 raises doubts about MAR by illustrating that a significantly greater percentage of smokers dropped out of the study compared to nonsmokers at each time point. Lichtenstein (1992) cited empirical evidence that in smoking cessation studies participants who were followed up after dropout were more likely to have experienced relapse. This suggest that MAR is not likely to occur. This is consistent with Hedeker & Gibbons (2006), who made a similar observation in their research, where they assumed that if a subject is missing at a particular time point it is because they are smoking, hence the missing data mechanism can be assumed to be MNAR rather than MAR. Use of MNAR models is typically done in situations where there is strong suspicion that the data violate MAR.

There are a variety of reasons why MNAR may exist among the non-smokers in this analyses. However, based on the existing literature (Lichtenstein, 1992) the most likely explanation for this finding is that non-smokers at $t-1$ who have become smokers at t may not respond at time t because, in general, missing data is more likely to be missing in smokers than the non-smokers (Table 6.2, (Hedeker & Gibbons, 2006)). In both cases, this implies that an individual dropout decision may be related to the future outcome and the assumption regarding to the MAR is in

doubt.

Table 6.3 shows the proportion of missing data for all seven time points. The first column shows the number of individuals with missing data at the given time point. The second column shows the total missing data at that time point. Overall, 23% of the observations are missing and about 44% of the subjects have at least one missing data point.

Table 6.2: Smoking transitions over time for students starting in grade 6 (time=1, n=4456)

		Status at t								*Overall at Time t-1*	
		Missing		*Nonsmoker*		*Quitter*		*Smoker*			
Time	*Status at t-1*	*%*	*(n)*	*%*	*(n)*	*%*	*(n)*	*%*	*(n)*	*%*	*(n)*
2	Nonsmoker	6.5	(256)	81.0	(3200)	6.1	(242)	6.4	(254)	88.7	(3952)
	Quitter	14.2	(39)	–	–	59.5	(163)	26.3	(72)	6.1	(274)
	Smoker	20.4	(47)	–	–	38.3	(88)	41.3	(95)	5.2	(230)
3	Missing	65.8	(225)	16.7	(57)	4.7	(16)	12.9	(44)	7.7	(342)
	Nonsmoker	5.7	(181)	78.3	(2505)	6.5	(209)	9.5	(305)	71.8	(3200)
	Quitter	8.9	(44)	0.4	(2)	54.6	(269)	36.1	(178)	11.1	(493)
	Smoker	11.9	(50)	0.7	(3)	25.7	(108)	61.8	(260)	9.4	(421)
4	Missing	76.6	(383)	9.8	(49)	4.8	(24)	8.8	(44)	11.2	(500)
	Nonsmoker	15.2	(390)	67.2	(1724)	5.9	(151)	11.8	(302)	57.6	(2567)
	Quitter	19.9	(120)	–	–	43.4	(261)	36.7	(221)	13.5	(602)
	Smoker	24.3	(191)	–	–	19.4	(153)	56.3	(443)	17.7	(787)
5	Missing	68.2	(739)	10.5	(114)	5.2	(56)	16.1	(175)	24.3	(1084)
	Nonsmoker	3.3	(58)	75.8	(1344)	5.1	(90)	15.8	(281)	39.8	(1773)
	Quitter	4.4	(26)	–	–	50.4	(297)	45.2	(266)	13.2	(589)
	Smoker	7.6	(77)	–	–	14.0	(141)	78.4	(792)	22.7	(1010)
6	Missing	86.9	(782)	3.7	(33)	2.2	(20)	7.2	(65)	20.2	(900)
	Nonsmoker	12.3	(180)	72.6	(1059)	4.3	(63)	10.7	(156)	32.7	(1458)
	Quitter	16.4	(96)	–	–	53.3	(311)	30.3	(177)	13.1	(584)
	Smoker	15.5	(234)	–	–	13.4	(203)	71.1	(1077)	34.0	(1514)
7	Missing	92.1	(1190)	1.2	(15)	1.1	(14)	5.7	(73)	29.0	(1292)
	Nonsmoker	4.5	(49)	82.4	(900)	4.2	(46)	8.9	(97)	24.5	(1092)
	Quitter	6.4	(38)	–	–	61.1	(365)	32.5	(194)	13.4	(597)
	Smoker	9.8	(144)	–	–	13.4	(197)	76.9	(1134)	33.1	(1475)

Table 6.3: Number and proportion of missing data at each time point

	Individual missing data	**Overall missing observations**
	n(%)	n(%)
Time	**n=4456**	**n=31192**
Time 2	342* (7.67)	342** (1.10)
Time 3	275 (13.84)	617 (1.20)
Time 4	660 (28.65)	1277 (4.10)
Time 5	134 (31.66)	1411 (4.52)
Time 6	392 (40.46)	1803 (5.78)
Time 7	159 (44.03)	1962 (6.29)
Overall	1962 (44.03)	7412 (23.76)

$*$ number of new individuals missing at a given time point; $**$ total number of missing observations at a given time point

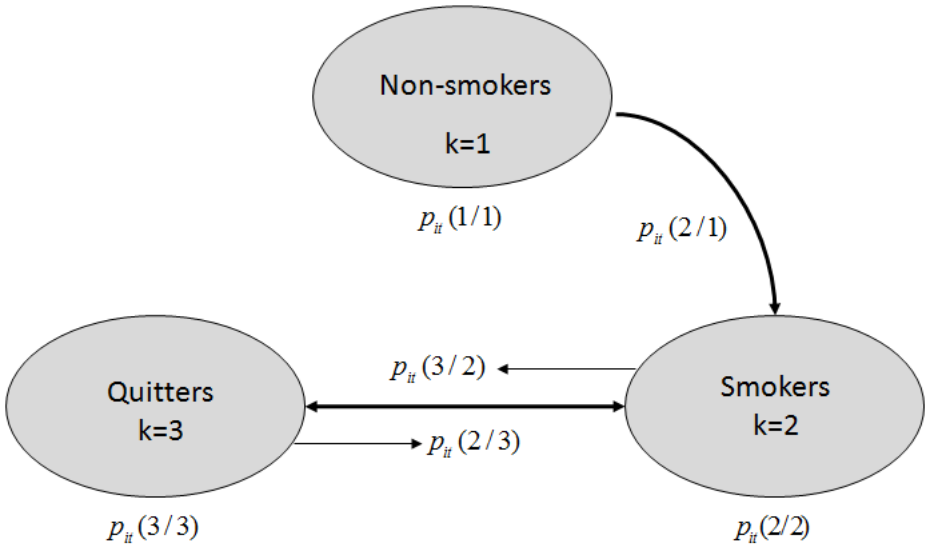
The present analyses are limited to a small subset of variables including treatment condition (level 2), sex, and smoking among five closest friends (time-dependent level 1). The treatment or intervention condition was defined as a binary variable where 0 represents the control group and 1 represents the intervention group. Sex was defined as a binary variable where 0 represents male and 1 represents female. The number of close friends who smoke cigarettes (X_1) was treated as continuous and ranged from 0 to 5, where 0 indicates that none of the participant's five closest friends smokes and 5 indicates that all of the participant's five closest friends smoke.

The purpose of this study was to evaluate the effect of intervention on individual transitions from one state to another. Because of the large amount of missing data within the cohort the objectives were: (1) to deal with the missing data and (2) to fit an appropriate model to assess the effectiveness of treatment on individual transitions from one state to another.

The analysis used a three-level model with grade 6 as the baseline. Six additional time points were used, one for each year of the study. The change in smoking state from one time point to the

next was used to model the probability of individual transitions over time. Individual movement to a given state at time $t+1$ is dependent upon the state at time t. Because of the discrete time-point assessment, it is possible that a nonsmoker at time t can move to the quitter state at time $t+1$ by moving through two transitions (nonsmoker to smoker and smoker to quitter, or smoker to quitter and quitter to smoker). For simplicity, we assumed that the transition occurred at the discrete time points and no transitions took place between the assessment periods. Furthermore, we assumed that the transition from nonsmoker to quitter and quitter to nonsmoker are invalid.

Figure 6.1: Graph for possible transition states



Let $Q_{ij,t} = k$ denote the status of the i^{th} subject in cluster j at time t with K possible states; $i = 1, 2, ..., n_j$, $j = 1, 2, ..., J$, $t = 1, 2, ..., T$, and $k = 1, 2, ...K$. We assume that the evolution of the status satisfies a first-order Markov chain with transitional probability from state k to l defined as

$$p_{ijt}(l|k) = Pr(Q_{ij,t} = l|Q_{ij,t-1} = k, X_{ijt}, \theta); t \geq 2,$$

where θ denotes the collection of all the parameters.

For the i^{th} subject with previous state k, we used the generalized logit model to model the transition from state k to l.

$$log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_{0j|kl} + \beta_{1|kl}t + \beta_{2|kl}X_{1ij,t-1} + \beta_{4|kl}G_{ij},$$

with $\beta_{0j|kl} = \beta_{0|kl} + \beta_{3|kl}C_j + u_{0j|kl}$ where $\beta's$ are different for different transitions.

For transition from state k to l, we define an indicator variable $Y_{ijt|kl}$ such that for $t \geq 2$,

$$Y_{ijt|kl} = \begin{cases} 1 & \text{if } Q_{ij,t} = l|Q_{ij,t-1} = k, X_{ijt}, \theta \\ 0 & \text{if } Q_{ij,t} = k|Q_{ij,t-1} = k, X_{ijt}, \theta \end{cases} \tag{6.1}$$

then the above logit model for state k to l transitions becomes:

$$logit(Pr(Y_{ijt|kl} = 1)) = \beta_{0j|kl} + \beta_{1|kl}t + \beta_{2|kl}X_{1ij,t-1} + \beta_{4|kl}G_{ij} \tag{6.2}$$

where

$$\beta_{0j|kl} = \beta_{0|kl} + \beta_{3|kl}C_j + u_{0j|kl}, \text{ with } \quad u_{0j|kl} \sim N(0, \sigma^2_{u|kl})$$

and where

$Y_{ijt|kl} \sim B(1, \pi_{ijt|kl})$;

$Y_{ijt|kl} = 0$ if there is no transition from state k to l;

$Y_{ijt|kl} = 1$ if there is a transition from state k to l;

t: an indicator for time (ranges from 1 to 7);

$X_{1ij,t-1}$: Time-varying covariate (smoking among five closest friends: range 0 to 5);

G_{ij}: Sex 1= Male and 0 = Female

$\beta_{4|kl}$ is the effect of sex;

$\beta_{1|kl}$ is the effect of time;

$\beta_{2|kl}$ is the slope for time-dependent covariate ;

C_j is a binary variable for school j coded as 0 for the control and 1 for the intervention group;

$\beta_{3|kl}$ is a log odds of the transition for the intervention group compared to the control group given the covariates and time ;

$\beta_{0|kl}$ is a log odds of the transition for the control group at t=0 and $X_{1ij,t-1} = 0$;

$u_{0j|kl}$ is a random effect for the intercept and assumed to be independent of the level-two predictors.

To achieve our objective, a pattern mixture model, a selection model, and a Bayesian pattern mixture model were used under the MNAR assumption and compared to the complete case (CCA) and last observation carried forward (LOCF) analysis. Of interest was estimating the treatment effect and the variation between the schools. Three restriction methods were used to fit the pattern

mixture model with the predictive mean matching method. The selection model was also used with the predictive mean matching method. These methods were compared with standard techniques such as CCA and LOCF. To maintain the monotone missing data pattern, any individual who has a non-monotone missing data pattern was removed from the analysis. As an example, any individual who follows this pattern $[O, O, O, M, M, O, O]$ where O means observed and M means missing data point, was removed from the analysis. The number of subjects with each missing-data pattern is shown in Table 6.4.

Table 6.4: Pattern of missing data

Pattern	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12	n
1	O	O	O	O	O	O	O	2635
2	O	O	O	O	O	O	M	175
3	O	O	O	O	O	M	M	439
4	O	O	O	O	M	M	M	458
5	O	O	O	M	M	M	M	145
6	O	O	M	M	M	M	M	268
7	O	M	M	M	M	M	M	336

M missing observation; O non-missing observation

6.1 Nonsmoker to Smoker

Table 6.5 reports the parameter estimates from CCA, LOCF, the three restriction methods for the pattern mixture model (CCMVPM, ACMVPM, NCMVPM) with the predictive mean matching method, the selection model with the predictive mean matching method, and the Bayesian pattern mixture method. The parameter estimates for all three transitions are reported with their standard

Table 6.5: Parameter estimates and empirical standard deviations for the WSPP3 dataset*

	CCA		LOCF		ACMVPM		CCMVPM		NCMVPM		SMPM		Bayesian	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Nonsmoker-to-smoker transition														
β_0	-3.894	(0.113)	-3.744	(0.117)	-3.771	(0.139)	-3.099	(0.087)	-3.726	(0.140)	-3.125	(0.128)	-2.775	(0.338)
β_1	0.136	(0.017)	0.053	(0.016)	0.152	(0.015)	0.106	(0.014)	0.132	(0.016)	0.155	(0.021)	0.089	(0.037)
β_2	0.786	(0.019)	0.811	(0.018)	0.732	(0.017)	0.694	(0.016)	0.733	(0.018)	0.645	(0.017)	0.725	(0.035)
β_3	0.082	(0.096)	0.054	(0.103)	0.243	(0.132)	0.070	(0.074)	0.164	(0.133)	0.112	(0.152)	0.039	(0.335)
β_4	-0.157	(0.061)	-0.143	(0.060)	-0.203	(0.056)	-0.078	(0.050)	-0.098	(0.057)	-0.054	(0.058)	-0.063	(0.119)
σ_u^2	0.199	(0.049)	0.241	(0.048)	0.398	(0.046)	0.131	(0.040)	0.400	(0.047)	0.324	(0.051)	0.541	(0.079)
Smoker-to-quitter transition														
β_0	-1.380	(0.147)	-1.196	(0.139)	-1.419	(0.169)	-1.410	(0.183)	-1.330	(0.163)	-1.338	(0.174)	-1.526	(0.325)
β_1	0.130	(0.023)	0.063	(0.022)	0.141	(0.023)	0.138	(0.024)	0.135	(0.023)	0.114	(0.027)	0.298	(0.302)
β_2	-0.502	(0.023)	-0.487	(0.022)	-0.535	(0.023)	-0.514	(0.035)	-0.519	(0.023)	-0.496	(0.054)	-0.214	(0.326)
β_3	-0.076	(0.093)	-0.131	(0.085)	-0.054	(0.129)	-0.108	(0.150)	-0.130	(0.122)	-0.132	(0.142)	-0.024	(0.321)
β_4	-0.042	(0.073)	-0.068	(0.072)	0.073	(0.075)	0.028	(0.079)	0.007	(0.076)	0.009	(0.089)	-0.019	(0.330)
σ_u^2	0.092	(0.091)	0.082	(0.068)	0.306	(0.058)	0.356	(0.064)	0.276	(0.057)	0.325	(0.085)	0.601	(0.064)
Quitter-to-smoker transition														
β_0	-2.060	(0.178)	-1.931	(0.143)	-2.447	(0.178)	-2.625	(0.204)	-2.564	(0.191)	-2.546	(0.187)	-2.037	(0.113)
β_1	0.025	(0.027)	-0.032	(0.021)	0.052	(0.026)	0.058	(0.028)	0.055	(0.026)	0.047	(0.027)	0.244	(0.019)
β_2	0.448	(0.118)	0.465	(0.021)	0.448	(0.028)	0.462	(0.042)	0.453	(0.029)	0.464	(0.031)	0.263	(0.019)
β_3	0.004	(0.027)	-0.051	(0.096)	-0.119	(0.117)	0.038	(0.128)	0.014	(0.133)	0.075	(0.123)	-0.031	(0.099)
β_4	0.076	(0.086)	0.020	(0.072)	0.118	(0.097)	0.123	(0.112)	0.112	(0.097)	0.111	(0.104)	0.091	(0.063)
σ_u^2	0.073	(0.076)	0.134	(0.063)	0.088	(0.155)	0.116	(0.147)	0.205	(0.071)	0.170	(0.092)	0.529	(0.018)

*Model used: $log\left\{\frac{p_{ijt}(l|k)}{p_{ijt}(k|k)}\right\} = \beta_0 + \beta_1 t + \beta_2 X_{1ij,t-1} + \beta_3 C_j + \beta_4 G_{ij} + u_{0j|kl}$; SE=standard error

errors in parentheses.

Table 6.5 shows that the estimates for the model intercept are similar for all the methods. The initial treatment condition has no significant effect on the transition from nonsmoker to smoker. The results for the variable time (a proxy for grade) and the time-dependent covariate (smoking among friends (X_1)), and sex are similar across the methods and are significant except for sex in SMPM and Bayesian analysis. The analysis shows that, after adjustment for other variables in the model, the probability of moving from non-smoker to a smoker increases with time and that males are more likely to make this transition. The estimate from the time-dependent covariate (X_1) shows a highly positive significant effect and shows that an individual with a higher number of smoking friends is more likely to move from the nonsmoker to the smoker state. The results are consistent for all methods except the Bayesian method. In general, the estimates from the Bayesian analysis do not agree with those of the other methods, perhaps due to the smaller number of transitions. The most noticeable differences were seen in the variance estimate for the random intercept. All methods show that the variance of the random effect is highly significant. All three restriction methods and the selection method have larger variance estimates than those of LOCF and CCA. The estimates from the restriction methods and the selection model are similar, supporting the robustness and the sensitivity of the parameter estimates (Michiels et al., 1999; Molenberghs & Verbeke, 2005).

6.2 Smoker to Quitter

For the smoker-to-quitter transition, the estimates for the model intercept are similar for all methods (Table 6.5) except in the Bayesian analysis method. The estimates for initial treatment condi-

tion parameters are similar across all the methods and have no significant effect on the individual transition from smoker to quitter. However, the estimates show that individuals in the treatment condition are less likely to move from smoker to quitter. Our results show that the Bayesian estimates do not perform as well as those of the restriction and selection models. Some of the reasons could be explained from the simulation studies in Chapter 5, where we simulated data under MNAR and compared the parameter estimates of the restriction methods and the Bayesian analysis. Our simulation results indicate that the restriction methods with predictive mean matching perform better than the Bayesian analysis.

The results for the variable time (a proxy for grade) are similar across the methods and are highly significant except in Bayesian analysis indicating an increasing probability of moving from smoking to quitting state with time, adjusting for the other variables. Sex differences were not found across the methods. However, the estimates for the time-dependent covariate (smoking among friends (X_1)) while different across the method are significant in all methods. This suggest that after the adjustment, the more smoking friends student has, the less likely the student is to make the transition from smoking to quitter state. Differences were also noted in the variance estimate of the random intercept. For all methods, the variance estimate of the random intercept was significant except in LOCF and CCA. Variance estimate of the random intercepts are higher among selection model and restriction methods than the other methods. The estimates from the restriction methods and the selection model are similar. The findings support our hypothesis that if the missing data are MNAR, then the estimates from the standard methods are not reliable.

6.3 Quitter to Smoker

Table 6.5 shows similar results for the model intercept for all estimation methods. Again, the results from the Bayesian analysis do not agree with those of other methods. The Bayesian analysis shows that the initial treatment condition has no significant effect on the probability of transition from the quitter to smoker state. The estimates from all methods other than in LOCF, ACMVPM, and Bayesian analysis show that individuals in the treatment condition are more likely to move from the quitter to smoker state. From the simulation studies in Chapter 5 the restriction methods perform better than the standard methods in terms of biases. Based on the simulation study, we conclude that the best estimate for the effect of the treatment condition is given by the ACMVPM method.

The estimates for the variable time (a proxy for grade) are similar in the restriction methods and Bayesian analysis and generally reflect a greater probability of moving from quitting to smoking with time. The estimates for the time-dependent covariate (smoking among friends (X_1)) are similar across methods and are significant indicating that students with more smoking friends are more likely to move from the quitting to the smoking state. Sex differences are similar in all methods and are not significant. The variance estimate of the random intercept was not significant for any other method except the NCMVPM method.

6.4 Conclusion

The WSPP3 data were analyzed using standard methods (CCA and LOCF), pattern mixture models including identifying restrictions, the selection model, and the Bayesian pattern mixture

model. The missing-data mechanism was assumed to be MNAR. The focus was on the monotone missing-data pattern. Imputation techniques such as the predictive mean matching method were used with the three restriction methods for the pattern mixture model and the selection model.

Tables 6.1 and 6.2 show that the proportion of students who never smoked dropped dramatically over time. The sharpest increase in the transition from nonsmoker to smoker occurred between grades 7 and 8. Our results indicate that treatment does not affect the individual transition from nonsmoker to smoker. The time-dependent covariate, smoking friends, positively affected the transition from nonsmoker to smoker (i.e., having more friends who smoke increases the probability of this transition). This variable negatively affected the transition from smoker to quitter (i.e., having more friends who smoke decreases the probability of this transition). The simulation results from Chapter 5 under the predictive mean matching method suggest that under the pattern mixture model ACMVPM performs better than any other method. The WSPP3 study was designed to test the hypothesis that students from the treatment schools are more likely to remain nonsmokers and less likely to become smokers. The results based on the restriction methods and the WSPP3 analysis in Chapter 6 are not consistent with this hypothesis and insignificant but show that the students from treatment schools are more likely to move from the nonsmoker to the smoker state.

Based on our simulation results and the WSPP3 application, we suggest that restriction methods with the predictive mean matching method should be used when the assumption is that the missing data are MNAR. The Bayesian analysis and the selection model have better results than CCA and LOCF but do not perform as well as the pattern mixture model. However the difference between the selection model and the restriction model seems to be small. Little (1995) argued that restriction methods for pattern mixture models do not require any explicit model assumption

for the dropout process; in contrast, the selection model uses all available information about the dropout process. In this study, the question of interest was to explore differences among the treatment and control conditions under a multilevel model with the MNAR assumption. Our results are similar to those of Molenberghs & Verbeke (2005): the pattern mixture model with identifiable restrictions (CCMVPM, ACMVPM, NCMPMV) is more appropriate than the selection model, CCF, or LOCF when the goal is to calculate the variance estimate for the random effect. Finally, the limited simulation in Chapter 5 and the WSPP3 application suggest that the Bayesian analysis also performs well in discrete hierarchical models but not as well as the restriction methods with predictive mean matching.

As indicated in earlier chapters, the study design of WSPP3 was based on repeated observations on individual students. These observations were correlated cross-sectionally, with clusters defined by schools; 100 schools were randomly assigned to either a control or treatment condition. The students within each school were followed across time which also complicated the dataset as students were able to move through the different smoking state categories over time (e.g., non-smoker to smoker state, smoker to quitter state, and quitter to smoker state). A third complication arose due to missing data at the student-level, a common occurrence among longitudinal data sets such as WSPP3. Researchers need simple yet comprehensive approaches for guiding their statistical modeling when working with such complex data. The results presented in this thesis may suggest that restriction methods appear more appropriate than standard methods and Bayesian approaches. Furthermore, the proposed methods are straightforward to implement using a standard statistical package (e.g. SAS) and can easily be extended to different settings such as cross-classified and multiple membership models.

Chapter 7

Conclusion and Future Research

This thesis set out to investigate inference for discrete hierarchical models in the presence of missing data. It is suggested that statistical methodology for hierarchical data analysis of non-Gaussian data is less well developed than that for Gaussian data, especially with binary outcome data which lead to generalized linear models with a nonlinear link function such as the logistic link. Furthermore, in a discrete hierarchical setting, the goal is to obtain good estimates of the marginal distribution of the data, which takes the form of an intractable integral and requires approximation techniques. As discussed, several approximation methods have been proposed to estimate the fixed and random effects in the context of generalized linear models. This thesis focused on two likelihood-based estimation procedures, the pseudo likelihood (Wolfinger & O'Connell, 1993) method and the adaptive Gaussian quadrature (Pinheiro & Bates, 1995) method, which are frequently used in applied hierarchical modeling. Furthermore, this thesis modeled the probability of individuals moving from one state to another at any given time point.

The simulation results suggest that AGQ is superior to PL for the estimation of random-effect parameters in that AGQ provides smaller biases for the estimation of random-effect parameters. Furthermore, AGQ permits greater flexibility in accommodating user-defined likelihood functions. In contrast, PL produces higher biases for the variance estimate of the random intercept; this result is consistent with other studies (Breslow & Lin, 1995). However, the computational efficiency (the time to complete one analysis) of PL is better.

Missing data is common in any longitudinal study regardless of the effort and pre-planning. In smoking-related studies, missing-data mechanisms are often either MAR or MNAR. Furthermore, data under the MNAR assumption are the most difficult to analyze for two reasons: a large number of potential models exists for these data and the hypothesis of random dropout can be neither confirmed nor repudiated.

In the second part of this study, simulations using the data from Chapter 3 were used to extend the method for handling missing data in binary hierarchical models by exploring the performance of different imputation techniques under various missing-data conditions. The simulation study used three-level discrete hierarchical data with 30% and 40% missing data under MNAR, and focused on the monotone missing-data pattern. The methods investigated were: Complete Case Analysis (CCA), Last Observation Carried Forward (LOCF), Complete Case Missing Value (CCMVPM) restriction, Available Case Missing Value (ACMVPM) restriction, Neighboring Case Missing Value (NCMVPM) restriction, and the selection model (SMPM). All three restriction methods and the selection model used the predictive mean matching method to impute the missing data. Once the data were imputed multiple imputation (Rubin, 1987) techniques were used to estimate the aggregated parameter estimates.

The results suggest that the parameter estimates obtained by CCA, LOCF, the restriction methods, and the selection model are highly dependent on the missing-data mechanism. Furthermore, it was shown that there is a consistent increasing trend in biases across all the methods as the proportion of missing data increases from 30% to 40%. The results further suggest that for MNAR, all three restriction methods (CCMVPM, ACMVPM, and CCMVPM) generally provide superior results compared to other methods including the Bayesian analysis. Furthermore, the selection model with the predictive mean matching approach provided similar results to the three restriction methods and was found to be superior to LOCF and CCA. However, if the proportion of missing data increases then the selection model is no longer a method of choice. In addition, among the three restriction methods, the parameter estimates from ACMVPM are closer to the true values.

One advantage of restriction methods is that standard statistical methods can easily be applied using available statistical software once the data have been imputed and are considered complete. However, further research could be done to generalize these restriction methods for non-monotone missing data in a context of a discrete hierarchical model.

The proposed methodology is not appropriate for instances where the sample size is too small in any given state and in those instances where an individual may move between more than one state in a given time point.

Unique challenges arise when using the above two estimation techniques (PL and AGQ) in a model which does not have a clear hierarchical structure (i.e., a cross-classified model). Examples include data on (i) a large number of students where individual students attend more than one school over time and (ii) students from the same classes who attend different courses. For

example, in a study of students over time, a survey may first be administered in Grade 5 and then in each subsequent grade through Grade 12. In such a study, students will typically move from one elementary school to different high schools over time. To model this change, the cross-classified structure of student movement needs to be incorporated in the model estimation since the correlation structure of the students has changed. The procedure becomes more complicated when there are missing data.

In the future, it would be useful to compare PL and AGQ for cross-classified data and multiple membership data. The imputation methods developed in this thesis could be used when there is a suspicion of MNAR. It would also be of interest to use AGQ for higher levels of random effects. However, the computational burden of AGQ increases rapidly with higher-dimensional models. STATA has incorporated AGQ for higher levels of random effects but issues related to the convergence and computational time remain. We managed to use a three-level model with a single random effect by incorporating the transitional model, but AGQ can be extended to higher-dimensional models (more than two levels of random effect). Bayesian estimation is another technique that could be used to deal with a higher number of random effects. Recent advances in Bayesian estimation avoid the need for numerical integration by repeatedly sampling from the posterior distribution of the parameters of interest and studies show that Bayesian models are increasingly being used for hierarchical models. More research is needed to implement the Bayesian analysis when dealing with missing data especially when the missing data are MNAR and there are issues related to the prior information.

Appendix A

Derivation of λ

Assume a random sample of N observations, with response variables $Y = (Y_1, ...Y_n)^T$ and $V = (V_1, V_2, ..., V_N)^T$. It is assumed that V contains information about the missingness in Y. Assume that Y obeys the regression model $\mathrm{Y} = \mathrm{X}\beta + \epsilon_1$ and V obeys the regression model $\mathrm{V} = \mathrm{W}\psi + \epsilon_2$, where $\mathrm{Y} = (y_i)$; $\mathrm{V} = v_i$; $\mathrm{X} = (x_{ip})$; $\mathrm{W} = w_{iq}$, $i = 1, 2, \ldots, \mathrm{N}$; $p = 1, 2, \ldots, \mathrm{P}$ and $q = 1, 2, \ldots, Q$.

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon_1 \tag{A.1}$$

$$\mathrm{V} = \mathrm{W}\psi + \epsilon_2 \tag{A.2}$$

where β and ψ are P$\times$1 and $Q\times$1 vectors of parameters, respectively, and ϵ_1 and ϵ_2 are the random error terms.

Furthermore, assume that ϵ_{1i} and ϵ_{2i} follow a bivariate normal distribution, that is

$$f(\epsilon_{1i}, \epsilon_{2i}) \sim N(0, \Sigma)$$

and

$$f(\mathrm{Y}_i, \mathrm{V}_i) \sim N(\mu_i, \Sigma)$$

where

$$\mu_i = \begin{pmatrix} \mathrm{X}_i^{\mathrm{T}}\beta \\ \mathrm{W}_i^{\mathrm{T}}\psi \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

For the regression function of Y the expected value is $E(Y|X) = X\beta$. Because of selection bias the regression function for the response variable Y can be written as

$$E(Y|X, S) = X\beta + E(\varepsilon_1|S),$$

where

$$E(\epsilon_{1i}|S) = \mathrm{E}(\epsilon_{1i}|\epsilon_{2i} \geq -W_i^T\psi),$$

where S represents the selection criteria. If there are no missing values or if the missing data is MCAR then the conditional expectation $E(\varepsilon_1|S) = 0$ which means there is no selection bias. Based on the joint bivariate normal density,

$$f(\epsilon_{1i}, \epsilon_{2i}) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}} \times \exp\left\{ -\frac{\sigma_{11}\sigma_{22}}{2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)}\left[\frac{\epsilon_{1i}^2}{\sigma_{11}} - 2\frac{\sigma_{12}\epsilon_{1i}\epsilon_{2i}}{\sigma_{11}\sigma_{22}} + \frac{\epsilon_{2i}^2}{\sigma_{22}}\right]\right\}$$

$$E(\epsilon_{1i}|\epsilon_{2i} \geq -W_i^T\psi) = \int_{-\infty}^{\infty} \epsilon_{1i} f(\epsilon_{1i}|\epsilon_{2i} \geq -W_i^T\psi) d\epsilon_{1i}$$

$$= \int_{-\infty}^{\infty} \epsilon_{1i} \frac{f(\epsilon_{1i}, \epsilon_{2i} \geq -W_i^T\psi)}{f(\epsilon_{2i} \geq -W_i^T\psi)} d\epsilon_{1i}$$

$$E(\epsilon_{1i}|\epsilon_{2i} \geq -W_i^T\psi) = \frac{\int_{-\infty}^{\infty} \epsilon_{1i} \int_{-W_i^T\psi}^{\infty} f(\epsilon_{1i}, \epsilon_{2i}) d\epsilon_{1i} d\epsilon_{2i}}{\int_{-W_i^T\psi}^{\infty} f(\epsilon_{2i}) d\epsilon_{2i}}, \tag{A.3}$$

where

$$\int_{-\mathrm{W}_i^{\mathrm{T}}\psi}^{\infty} f(\epsilon_{2i}) d\epsilon_{2i} = 1 - \Phi\left(\frac{-\mathrm{W}_i^{\mathrm{T}}\psi}{\sqrt{\sigma_{22}}}\right), \tag{A.4}$$

where Φ is the standard normal cumulative distribution function.

$$\int_{-\mathrm{W}_i^{\mathrm{T}}\psi}^{\infty} \int_{-\infty}^{\infty} \epsilon_{1i} f(\epsilon_{1i}, \epsilon_{2i}) d\epsilon_{1i}\epsilon_{2i} = \int_{-\mathrm{W}_i^{\mathrm{T}}\psi}^{\infty} \int_{-\infty}^{\infty} \epsilon_{1i} f(\epsilon_{1i}|\epsilon_{2i}) \times f(\epsilon_{2i}) d\epsilon_{1i} d\epsilon_{2i}$$
$$= \int_{-\mathrm{W}_i^{\mathrm{T}}\psi}^{\infty} \mathrm{E}(\epsilon_{1i}|\epsilon_{2i}) f(\epsilon_{2i}) d\epsilon_{2i}$$

Based on the Bivariate normal distribution (e.g. Wichern & Johnson (2001)), if

$$(\epsilon_{1i}, \epsilon_{2i}) \sim N\left[\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right]$$

then

$$(\epsilon_{1i}|\epsilon_{2i} = a) \sim N\left(\eta_1 + \frac{\sigma_{12}}{\sigma_{22}}a, \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{12}}\right)$$

In our case, $\eta_1 = \eta_2 = 0$ and the conditional expectation is

$$\mathrm{E}(\epsilon_{1i}|\epsilon_{2i}) = \frac{\sigma_{12}}{\sigma_{22}}\epsilon_{2i}$$

Therefore,

$$\begin{aligned} \int_{-\mathrm{W}_i^{\mathrm{T}}\Psi}^{\infty} \mathrm{E}(\epsilon_{1i}|\epsilon_{2i}) f(\epsilon_{2i}) d\epsilon_{2i} &= \int_{-\mathrm{W}_i^{\mathrm{T}}\Psi}^{\infty} \frac{\sigma_{12}}{\sigma_{22}}\epsilon_{2i} f(\epsilon_{2i}) d\epsilon_{2i} \\ &= \frac{\sigma_{12}}{\sigma_{22}} \int_{-\mathrm{W}_i^{\mathrm{T}}\psi}^{\infty} \frac{\epsilon_{2i}}{\sqrt{2\pi\sigma_{22}}} \times \exp - \frac{\epsilon_{21}^2}{2\sigma_{22}^2} d\epsilon_{2i} \end{aligned}$$

Let

$$\mathrm{V} = \frac{\epsilon_{2i}}{\sqrt{\sigma_{22}}}$$

then

$$\mathrm{V} \sim N(0, 1)$$

and

$$\int_{-\mathrm{W}_i^{\mathrm{T}}\Psi}^{\infty} \mathrm{E}(\epsilon_{1i}|\epsilon_{2i}) f(\epsilon_{2i}) d\epsilon_{2i} = \frac{\sigma_{12}}{\sigma_{22}} \int_{\frac{-\mathrm{W}_i^{\mathrm{T}}\Psi}{\sqrt{\sigma_{22}}}}^{\infty} \sqrt{\sigma_{22}} \frac{\mathrm{V}}{\sqrt{2\pi}} \exp -\frac{\mathrm{V}^2}{2} d\mathrm{V}$$

After integrating and substituting the limits, we end up with a normal pdf.

$$\int_{-\mathrm{W}_i^{\mathrm{T}}\Psi}^{\infty} \mathrm{E}(\epsilon_{1i}|\epsilon_{2i}) f(\epsilon_{2i}) d\epsilon_{2i} = \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} \phi\left(\frac{-\mathrm{W}_i^{\mathrm{T}}\Psi}{\sqrt{\sigma_{22}}}\right), \tag{A.5}$$

where ϕ is the standard normal probability density function. Substituting Eq. (4) and Eq. (5) in Eq. (3), therefore,

$$\begin{aligned} E(\epsilon_{1i}|\epsilon_{2i} \geq -\mathrm{W}_i^{\mathrm{T}}\Psi) &= \frac{\frac{\sigma_{12}}{\sqrt{\sigma_{22}}} \phi\left(\frac{-\mathrm{W}_i^{\mathrm{T}}\Psi}{\sqrt{\sigma_{22}}}\right)}{1 - \Phi\left(\frac{-\mathrm{W}_i^{\mathrm{T}}\Psi}{\sqrt{\sigma_{22}}}\right)} \\ &= \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} \lambda_i \end{aligned}$$

where

$$\begin{aligned} \lambda_i &= \frac{\phi(\mathrm{B}_i)}{1 - \Phi(\mathrm{B}_i)} = \frac{\phi(\mathrm{B}_i)}{\Phi(-\mathrm{B}_i)} \\ \mathrm{B}_i &= -\frac{\mathrm{W}_i^{\mathrm{T}}\psi}{\sqrt{\sigma_{22}}} \end{aligned}$$

λ_i is called the selection bias, or inverse of Mill's ratio, or the hazard rate. Similarly,

$$E(\epsilon_{2i}|\epsilon_{1i} \geq -\mathrm{W}_i^{\mathrm{T}}\psi) = \frac{\sigma_{22}}{\sqrt{\sigma_{22}}}\lambda_i$$

Appendix B

SAS Program Code

Listing B.1: WSPP3_pattern_ACMVPM_NS_SK_Analysis.sas

```
/******** WSPP3: Data Analysis: Pattern Mixture Model: ACMVPM: NS to SMK ***/

libname ras "C:\Users\r4ahmed\Desktop\final_thesis\data";
data data;
    set ras.long; /*** Specify your own dataset here ***/
    id=pid;          /*** Individual ID ***/
    sid=initschool;  /*** School identifier ***/
    timeclss=time;   /*** Proxy for time: time =1 ,2 ,3,4 ,5,6,7 ***/
    const=1;  /*** To model the intercept ****/
run;

/**** Arrange the data by individual time points ***/
proc transpose data=data(keep=pid my1 my2 my3 time) out=_a;
    by pid;
    id time;
    var my1 my2 my3;
run;

/*** Create the seven missing data patterns;***/

data _b;
    set _a;
    array a[*] _1--_7;
    pattern=.;
    do i=1 to dim(a);
        if pattern=. then do;
            if a[i]=. then pattern=9-i;
        end;
```

```
    end;
    if i>7 & pattern=. then pattern=1;
run;

/*** We created three patterns becuase we are interested in three transitions:
     (1) NS to SM, (2) SK to QT, and (3) QT to SM. The number of patterns can
     easily be extended ****/
data transi;
    merge data(in=a)
        _b(in=b keep=pid _name_ pattern rename=(pattern=p2) where=(_name_="MY1"))
        _b(in=c keep=pid _name_ pattern rename=(pattern=p3) where=(_name_="MY2"))
        _b(in=d keep=pid _name_ pattern rename=(pattern=p4) where=(_name_="MY3"));
    by pid;
    drop _name_;
    /*** Never Smoker to Smoker ***/
    my1=96;
    if time=2 and prevsmoke in (1,2) and smoke in (4,5) then my1=1;if time=2 and
        prevsmoke in (1,2) and smoke in (.) then my1=.;
    if time=3 and prevsmoke in (1,2) and smoke in (4,5) then my1=1;if time=3 and
        prevsmoke in (1,2) and smoke in (.) then my1=.;
    if time=4 and prevsmoke in (1,2) and smoke in (4,5) then my1=1;if time=4 and
        prevsmoke in (1,2) and smoke in (.) then my1=.;
    if time=5 and prevsmoke in (1,2) and smoke in (4,5) then my1=1;if time=5 and
        prevsmoke in (1,2) and smoke in (.) then my1=.;
    if time=6 and prevsmoke in (1,2) and smoke in (4,5) then my1=1;if time=6 and prevsmoke
        in (1,2) and smoke in (.) then my1=.;
    if time=7 and prevsmoke in (1,2) and smoke in (4,5) then my1=1;if time=7 and prevsmoke
        in (1,2) and smoke in (.) then my1=.;

    if time=2 and prevsmoke in (1,2) and smoke in (1,2) then my1=0; if time=2 and lag(my1)=.
        then my1=.;
    if time=3 and prevsmoke in (1,2) and smoke in (1,2) then my1=0; if time=3 and lag(my1)=.
        then my1=.;
    if time=4 and prevsmoke in (1,2) and smoke in (1,2) then my1=0; if time=4 and lag(my1)=.
        then my1=.;
    if time=5 and prevsmoke in (1,2) and smoke in (1,2) then my1=0; if time=5 and lag(my1)=.
        then my1=.;
    if time=6 and prevsmoke in (1,2) and smoke in (1,2) then my1=0; if time=6 and lag(my1)=.
        then my1=.;
    if time=7 and prevsmoke in (1,2) and smoke in (1,2) then my1=0; if time=7 and lag(my1)=.
        then my1=.;

    /*** Smoker to Quitter ***/
    my2=96;
    if time=2 and prevsmoke in (4,5) and smoke=3 then my2=1;if time=2 and prevsmoke in (4,5)
        and smoke=. then my2=.;
    if time=3 and prevsmoke in (4,5) and smoke=3 then my2=1;if time=3 and prevsmoke in (4,5)
        and smoke=. then my2=.;
    if time=4 and prevsmoke in (4,5) and smoke=3 then my2=1;if time=4 and prevsmoke in (4,5)
        and smoke=. then my2=.;
    if time=5 and prevsmoke in (4,5) and smoke=3 then my2=1;if time=5 and prevsmoke in (4,5)
        and smoke=. then my2=.;
    if time=6 and prevsmoke in (4,5) and smoke=3 then my2=1;if time=6 and prevsmoke in (4,5)
        and smoke=. then my2=.;
    if time=7 and prevsmoke in (4,5) and smoke=3 then my2=1;if time=7 and prevsmoke in (4,5)
        and smoke=. then my2=.;
    if time=2 and prevsmoke in (4,5) and smoke in (4,5) then my2=0; if time=2 and lag(my2)=.
```

```
        then my2=.;
    if time=3 and prevsmoke in (4,5) and smoke in (4,5) then my2=0; if time=3 and lag(my2)=.
        then my2=.;
    if time=4 and prevsmoke in (4,5) and smoke in (4,5) then my2=0; if time=4 and lag(my2)=.
        then my2=.;
    if time=5 and prevsmoke in (4,5) and smoke in (4,5) then my2=0; if time=5 and lag(my2)=.
        then my2=.;
    if time=6 and prevsmoke in (4,5) and smoke in (4,5) then my2=0; if time=6 and lag(my2)=.
        then my2=.;
    if time=7 and prevsmoke in (4,5) and smoke in (4,5) then my2=0; if time=7 and lag(my2)=.
        then my2=.;
    /*** Quitter to Smoker ***/
    my3=96;
    if time=2 and prevsmoke in (3) and smoke in (4,5) then my3=1;if time=2 and prevsmoke in (3)
        and smoke in (.) then my3=.;
    if time=3 and prevsmoke in (3) and smoke in (4,5) then my3=1;if time=3 and prevsmoke in (3)
        and smoke in (.) then my3=.;
    if time=4 and prevsmoke in (3) and smoke in (4,5) then my3=1;if time=4 and prevsmoke in (3)
        and smoke in (.) then my3=.;
    if time=5 and prevsmoke in (3) and smoke in (4,5) then my3=1;if time=5 and prevsmoke in (3)
        and smoke in (.) then my3=.;
    if time=6 and prevsmoke in (3) and smoke in (4,5) then my3=1;if time=6 and prevsmoke in (3)
        and smoke in (.) then my3=.;
    if time=7 and prevsmoke in (3) and smoke in (4,5) then my3=1;if time=7 and prevsmoke in (3)
        and smoke in (.) then my3=.;

    if time=2 and prevsmoke in (3) and smoke in (3) then my3=0; if time=2 and lag(my3)=. then my3=.;
    if time=3 and prevsmoke in (3) and smoke in (3) then my3=0; if time=3 and lag(my3)=. then my3=.;
    if time=4 and prevsmoke in (3) and smoke in (3) then my3=0; if time=4 and lag(my3)=. then my3=.;
    if time=5 and prevsmoke in (3) and smoke in (3) then my3=0; if time=5 and lag(my3)=. then my3=.;
    if time=6 and prevsmoke in (3) and smoke in (3) then my3=0; if time=6 and lag(my3)=. then my3=.;
    if time=7 and prevsmoke in (3) and smoke in (3) then my3=0; if time=7 and lag(my3)=. then my3=.;
    ny2=my1;   /*** Binary indicator showing individual transition from non-smoker to smoker ***/
    ny3=my2;   /*** Binary indicator showing individual transition from smoker to quitter ***/
    ny4=my3;   /*** Binary indicator showing individual transition from quitter to smoker ***/
    x1=locfx1; /*** Time dependent covariate: If individual is missing then take the LOCF ***/
    if ny2 not in (96);/*** Only for non-smoker to smokers transition ****/
run;

/*
proc print data=transi;
    where pid in (24234,20033,20027,22241);
    *where pid in (20001,20003,20004,20007,20222,21673,21802,20230,22022,32120);
    var id sid timeclss prevsmoke smoke my1 ny2 my2 ny3 my3 ny4 p2 ;
run;
*/

/**** Example for data structure *****//*
```

Obs	id	sid	timeclss	prevsmoke	smoke	MY1	ny2	MY2	ny3	MY3	ny4	p2
134	20027	1	1	.	2	96	96	96	96	96	96	5
135	20027	1	2	2	2	0	0	96	96	96	96	5
136	20027	1	3	2	4	1	1	96	96	96	96	5
137	20027	1	4	4	.	96	96	.	.	96	96	5
138	20027	1	5	.	.	96	96	.	.	96	96	5
139	20027	1	6	.	.	96	96	.	.	96	96	5
140	20027	1	7	.	.	96	96	.	.	96	96	5

```
 162     20033   1      1          .          1        96      96      96      96      96      96      5
 163     20033   1      2          1          1         0       0      96      96      96      96      5
 164     20033   1      3          1          1         0       0      96      96      96      96      5
 165     20033   1      4          1          .         .       .      96      96      96      96      5
 166     20033   1      5          .          .         .       .      96      96      96      96      5
 167     20033   1      6          .          .         .       .      96      96      96      96      5
 168     20033   1      7          .          .         .       .      96      96      96      96      5
5006     22241  12      1          .          1        96      96      96      96      96      96      5
5007     22241  12      2          1          4         1       1      96      96      96      96      5
5008     22241  12      3          4          3        96      96       1       1      96      96      5
5009     22241  12      4          3          .        96      96      96      96       .       .      5
5010     22241  12      5          .          .        96      96      96      96       .       .      5
5011     22241  12      6          .          .        96      96      96      96       .       .      5
5012     22241  12      7          .          .        96      96      96      96       .       .      5
8373     24234  28      1          .          1        96      96      96      96      96      96      1
8374     24234  28      2          1          2         0       0      96      96      96      96      1
8375     24234  28      3          2          4         1       1      96      96      96      96      1
8376     24234  28      4          4          5        96      96       0       0      96      96      1
8377     24234  28      5          5          3        96      96       1       1      96      96      1
8378     24234  28      6          3          3        96      96      96      96       0       0      1
8379     24234  28      7          3          4        96      96      96      96       1       1      1


*/

/*** Fitting only Non-Smnokers to Smoker Transition ****/

/*

 Prob of moving to NS to SM


 Obs    id       sid   timeclss     prevsmoke     smoke     MY1     ny2     MY2     ny3     MY3     ny4     p2

 66     20027    1       2            2           2         0       0      96      96      96      96      5
 67     20027    1       3            2           4         1       1      96      96      96      96      5
 78     20033    1       2            1           1         0       0      96      96      96      96      5
 79     20033    1       3            1           1         0       0      96      96      96      96      5
 80     20033    1       4            1           .         .       .      96      96      96      96      5
 81     20033    1       5            .           .         .       .      96      96      96      96      5
 82     20033    1       6            .           .         .       .      96      96      96      96      5
 83     20033    1       7            .           .         .       .      96      96      96      96      5
2595    22241   12       2            1           4         1       1      96      96      96      96      5
4529    24234   28       2            1           2         0       0      96      96      96      96      1
4530    24234   28       3            2           4         1       1      96      96      96      96      1


* [ ---- 2 Macro definitions ------------------------------------------------ ];

/*** %geeGlimmix: Provide initial values for NLMIXED ***/

/*--- %geeGlimmix ------------------------------------------------------------\
  |                                        |                                 |
  |                                                                          |
  | Arguments:                                                               |
  |      outcome = specifies the outcome varible for analysis                |
  |      pattern = specifies the pattern being modeled. Pattern gets         |
```

```
|              used in a where statement, so enclose it in a       |
|              %str() to make sure it is interpreted correctly     |
|     data    = specifies the dataset to use                       |
\------------------------------------------------------------------*/
%macro geeGlimmix(outcome=,pattern=,data=);
    %put NOTE: %nrstr(%geeGlimmix is running --- estimating initial values for NLMIXED.);
    /* These macro variables get used in the other macros, so need to
       declare their scope as global... */
    %global c0 c1 c2 c3 c4 sd;

    /* Need starting values for proc nlmixed, run a GEE model to estimate
       initial values for parameters of interest. In this case cond, time,
       x1 (time dependent covariate) and sex */
    proc genmod data=&data descending;
        where &pattern;
        class sid id timeclss;
        model &outcome = cond time x1 sex / dist=bin link=logit;
        repeated subject=sid(id) / withinsubject=timeclss;
        ods output GEEEmpPEst = parm;
    run;

    /* Store the parameter estimates in the global macro variables defined above */
    data _null_;
        set parm;
        estimate=round(estimate,0.001);
        parm=lowcase(parm);
        if Parm="intercept" then call symput("c0",Estimate);
        else if Parm="cond" then call symput("c1",Estimate);
        else if Parm="time" then call symput("c2",Estimate);
        else if Parm="x1" then call symput("c3",Estimate);
        else if Parm="sex" then call symput("c4",Estimate);
    run;

    /* Also need a starting value for the variance estimate of the random effect.
       Fit using proc glimmix, using same outcome and covariates */
    proc glimmix data=&data;
        where &pattern;
        class sid;
        model &outcome = cond time x1 sex / solution link=logit dist=bin;
        random intercept / subject = sid;
        ods output covparms=varest;
    run;

    /* store */
    data _null_;
        set varest;
        /* if glimmmix fails and you do not get an estiamte, provide a default value
                for estimate */
        if stderr=. then estimate=0.5;
        call symput("sd",Estimate);
    run;

    /* print the initial values to the log file, as a check to make sure everything
               is working  correctly */
     %put NOTE: %nrstr(%geeGlimmix --- ) starting value for c0 is &c0..;
     %put NOTE: %nrstr(%geeGlimmix --- ) starting value for c1 is &c1..;
     %put NOTE: %nrstr(%geeGlimmix --- ) starting value for c2 is &c2..;
```

```
     %put NOTE: %nrstr(%geeGlimmix ---- ) starting value for c3 is &c3..;
     %put NOTE: %nrstr(%geeGlimmix ---- ) starting value for c4 is &c4..;
     %put NOTE: %nrstr(%geeGlimmix ---- ) starting value for Variance is &sd..;
%mend geeGlimmix;

/*** %nlmixedPredict: Fitting a transitional model and computing the predicted probabilities ***/

/*--- %nlmixedPredict ------------------------------------------------------\
 |                                    |                                    |
 |                                                                         |
 | Arguments:                                                              |
 |     outcome = specifies the outcome varible for analysis                |
 |     pattern = specifies the pattern being modeled. Pattern gets         |
 |               used in a where statement, so enclose it in a             |
 |              %str() to make sure it is interpreted correctly            |
 |     data    = specifies the dataset to use                              |
 |     outpred = specifies an output dataset to store predicted            |
 |               values. This dataset is used by the macro                 |
 |              %predictiveMean                                            |
 |     title   = an optional title statement, so you know what             |
 |               transition is being modeled                               |
 \-------------------------------------------------------------------------*/
%macro nlmixedPredict(outcome=,pattern=,data=,outpred=,title=);
    %put NOTE: %nrstr(%nlmixedPredict is running.) \
         Fitting a model for the &outcome. transition.;
    %put NOTE: Using the following starting values: c0 = &c0, c1 = &c1, c2 = &c2.;
    %put NOTE: c3 = &c3, c4 = &c4, sd = &sd;

    /* Fit the transition model via NLMIXED */
    %if %length(&title) > 0 %then title "&title"; ;
    proc nlmixed data=&data(where=(&pattern));
        parms c0=%sysevalf(&c0) c1=%sysevalf(&c1) c2=%sysevalf(&c2) c3=%sysevalf(&c3)
              c4=%sysevalf(&c4) SD=%sysevalf(&sd);
        Z = c0 + c1*cond + c2*time + c3*x1 + c4*sex + U;
        bounds SD >= 0;
        p = exp(z)/(1+exp(z));
        model &outcome ~ binary(p);
        random U ~ normal(0,sd*sd) subject=sid;
        ods output ParameterEstimates=nlparm(keep=Parameter Estimate);
        predict p out=predP;
        predict u out=predU;
    run;
    title;

    /* Compute the predicted probabilities */
    data predU1;
        set predU;
        u_head = pred;
        keep sid u_head;
    run;

    proc sort data=predU1;by sid;run;
    proc sort data=&data;by sid;run;

    data _null_;
        set nlparm;
        parameter=lowcase(parameter);
```

```
        if parameter="c0" then call symput("cc0",estimate);
        else if parameter="c1" then call symput("cc1",estimate);
        else if parameter="c2" then call symput("cc2",estimate);
        else if parameter="c3" then call symput("cc3",estimate);
        else if parameter="c4" then call symput("cc4",estimate);
    run;

    data &outpred;
        merge &data predU1;
        by sid;
        k=%sysevalf(&cc0) + %sysevalf(&cc1)*cond + %sysevalf(&cc2)*time + %sysevalf(&cc3)*x1
                + %sysevalf(&cc4)*sex + u_head;
        p = exp(k)/(1+exp(k));
        keep sid id time p;
    run;
%mend nlmixedPredict;

/*** %predictiveMean: Steps are clearly written in Thesis ***/

/*--- %predictiveMean ------------------------------------------------\
 |                                    |                               |
 |                                                                    |
 | Arguments:                                                         |
 |        outcome = specifies the outcome varible for analysis        |
 |     newOutcome = specifies the name of a new outcome variable to   |
 |                  be used in subsequent calls to %geeGlimmix,       |
 |                  %nlmixedPredict and %predictiveMean               |
 |  imputePattern = specifies the new pattern to be modeled.          |
 |                  As before, enclose it in a %str()                 |
 |           data = specifies the dataset to use                      |
 |         predIn = specifies an input dataset storing predicted      |
 |                  mean values needed for imputation                 |
 |            out = name of an output dataset to be used as input in  |
 |                  subsequent calls to %geeGlimmix, %nlmixedPredict  |
 |                  and %predictiveMean                               |
 \--------------------------------------------------------------------*/
%macro predictiveMean(origOutcome=,outcome=,newOutcome=,imputePattern=,data=,predIn=,out=);
    proc means data=&predIn noprint;
        class sid;
        var p;
        output out=quant(where=(_type_=1)) p25=q1 p50=q2 p75=q3;
    run;

    proc sort data=&predIn;by sid id time;run;
    proc sort data=&data;by sid id time;run;
    proc sort data=quant;by sid;run;

    data _tmp;
        merge &data quant;
        by sid;
    run;

    data grps;
        merge _tmp &predIn;
        by sid id time;

                if p <= q1 then pc=1;
```

```
        else if q1 < p <= q2 then pc=2;
        else if q2 < p <= q3 then pc=3;
        else if p > q3 then pc=4;
    run;

    /* find number of missing data points in each group */
    proc means data=grps noprint;
        class sid pc;
        var &origOutcome;
        output out=MissN(where=(_type_=3)) nmiss=_nsize_;
    run;
    /* make sure there are no zeros in _nsize_ */
    data MissN;
        set MissN;
        if _nsize_=0 then _nsize_=1;
    run;

    proc sort data=grps;by sid pc;run;

    /* using the predictive mean, select a sample from each group */
    proc surveyselect data=grps(where=(&origOutcome^=.)) noprint out=selgrps method=urs n=MissN;
        strata sid pc;
    run;

    data selgrps;
        set selgrps;
        tmpOutcome = &origOutcome;
        keep sid pc tmpOutcome;
    run;

    proc sort data=selgrps;by sid pc;run;
    proc sort data=grps;by sid pc;run;

    /* impute missing values using predictive mean matching methods */
    data &out;
       merge grps selgrps;
       by sid pc;
       &newOutcome = &outcome;
       if &imputePattern then do;
           if &origOutcome = . then &newOutcome=tmpOutcome;
       end;
       run;
%mend predictiveMean;

/*** %ppmMethod: Predictive mean matching method***/

/*--- %ppmMethod ------------------------------------------------------\
 |                                                                     |
 |                                                                     |
 | Arguments:                                                          |
 |    origOutcome = the outcome variable that originally had missing  |
 |                  values                                             |
 |       outcome = the outcome with imputed values (?)                |
 |    newOutcome = outcome with imputed values following predictive    |
 |                  mean matching                                      |
 |       title   = an optional title statement, so you know what       |
 |                  transition is being modeled                        |
```

```
|           data = initial input dataset to analyze                  |
|            out = output dataset from predictive mean matching      |
|        outparm = parameter estimates from NLMIXED usint &out       |
|                  dataset for using in multiple imputation anal-    |
|                  ysis                                              |
|      impEstOut = output dataset containing estimates from proc     |
|                  MIANALYZE                                         |
|        outfile = name and location (filename with complete file    |
|                  path) containing the results from                 |
|                  proc MIANALYZE                                    |
\-------------------------------------------------------------------*/
%macro pmmMethod(origOutcome=,outcome=,newOutcome=,title=,data=,out=,outparm=,impEstOut=,outfile=);
    /* initial values for parameter estimates */
    proc genmod data=&data descending;
        class sid id timeclss;
        model &outcome = cond time x1 sex / dist=bin link=logit;
        repeated subject=sid(id) / withinsubject=timeclss;
        ods output GEEEmpPEst = parm;
    run;

    data _null_;
        set parm;
        estimate=round(estimate,0.001);
        parm=lowcase(parm);
        if parm="intercept" then call symput("pmmc0",estimate);
        else if parm="cond" then call symput("pmmc1",estimate);
        else if parm="time" then call symput("pmmc2",estimate);
        else if parm="x1" then call symput("pmmc3",estimate);
        else if parm="sex" then call symput("pmmc4",estimate);
    run;

    /* initial value for variance of the random effect */
    proc glimmix data=&data;
        class sid;
        model &outcome = cond time x1 sex / solution link=logit dist=bin;
        random intercept / subject=sid;
        ods output covparms = varest;
    run;

    data _null_;
        set varest;
        if stderr=. then estimate=0.5;
        else if 0<stderr<0.05 then estimate=0.05;
        else if stderr>10 then estimate=1;
        call symput("pmmsd",estimate);
    run;

    %if %length(&title)>0 %then title "&title"; ;
    proc nlmixed data=&data;
        parms c0=%sysevalf(&pmmc0) c1=%sysevalf(&pmmc1) c2=%sysevalf(&pmmc2) c3=%sysevalf(&pmmc3)
              c4=%sysevalf(&pmmc4) sd=%sysevalf(&pmmsd);
        Z = c0 + c1*cond + c2*time + c3*x1 + c4*sex + U;
        bounds sd > 0;
        p=exp(z)/(1+exp(z));
        model &outcome ~ binary(p);
        random U ~ normal(0,sd*sd) subject=sid;
        ods output ParameterEstimates=nlparm(keep=Parameter Estimate);
```

```
    predict p out=predP;
    predict u out=predU;
run;
title;

data predU1;
    set predU;
    u_head=pred;
    keep sid u_head;
run;

proc sort data=predU1;by sid;run;
proc sort data=&data;by sid;run;

data _null_;
    set nlparm;
    if parameter="c0" then call symput("cc0",estimate);
    else if parameter="c1" then call symput("cc1",estimate);
    else if parameter="c2" then call symput("cc2",estimate);
    else if parameter="c3" then call symput("cc3",estimate);
    else if parameter="c4" then call symput("cc4",estimate);
run;

data pmm;
    merge &data predU1;
    by sid;
    k=%sysevalf(&cc0) + %sysevalf(&cc1)*cond + %sysevalf(&cc2)*time + %sysevalf(&cc3)*x1
            + %sysevalf(&c4)*sex + u_head;
    p=exp(k)/(1+exp(k));
    keep sid id time p;
run;

/* predictive mean matching method */
proc means data=pmm noprint;
    class sid;
    var p;
    output out=quant(where=(_type_=1)) p25=q1 p50=q2 p75=q3;
run;

proc sort data=pmm;by sid id time;run;
proc sort data=&data;by sid id time;run;
proc sort data=quant;by sid;run;

data tmp;
    merge &data quant;
    by sid;
run;

data grps;
    merge tmp pmm;
    by sid id time;
    if p <= q1 then pc=1;
    else if q1 < p <= q2 then pc=2;
    else if q2 < p <= q3 then pc=3;
    else if p > q3 then pc=4;
run;
```

```
proc means data=grps noprint;
    class sid pc;
    var &origOutcome;
    output out=MissN(where=(_type_=3)) nmiss=_nsize_;
run;

data MissN;
    set MissN;
    if _nsize_=0 then _nsize=_1;
run;

proc sort data=grps;by sid pc;run;

/* First Predictive Mean */
proc surveyselect data=grps(where=(&origOutcome^=.)) noprint out=selgrps1 method=urs n=MissN;
    strata sid pc;
run;

data selgrps1;
    set selgrps1;
    &outcome = &origOutcome;
    keep sid pc &outcome;
run;

proc sort data=selgrps1;by sid pc;run;
proc sort data=grps;by sid pc;run;

data &data.1;
    merge grps selgrps1;
    by sid pc;
    &newOutcome = &origOutcome;
    if &origOutcome=. then &newOutcome=&outcome;
    predictedMean = 1;
run;

/* Second predictive mean */
proc surveyselect data=grps(where=(&origOutcome^=.)) noprint out=selgrps2 method=urs n=MissN;
    strata sid pc;
run;

data selgrps2;
    set selgrps2;
    &outcome = &origOutcome;
    keep sid pc &outcome;
run;
proc sort data=selgrps2;by sid pc;run;
proc sort data=grps;by sid pc;run;

data &data.2;
    merge grps selgrps2;
    by sid pc;
    &newOutcome = &origOutcome;
    if &origOutcome=. then &newOutcome = &outcome;
    predictedMean = 2;
run;

/* Third predictive mean */
```

```
proc surveyselect data=grps(where=(&origOutcome^=.)) noprint out=selgrps3 method=urs n=MissN;
    strata sid pc;
run;

data selgrps3;
    set selgrps3;
    &outcome = &origOutcome;
    keep sid pc &outcome;
run;
proc sort data=selgrps3;by sid pc;run;
proc sort data=grps;by sid pc;run;

data &data.3;
    merge grps selgrps3;
    by sid pc;
    &newOutcome = &origOutcome;
    if &origOutcome=. then &newOutcome = &outcome;
    predictedMean = 3;
run;

/* Fourth predictive mean */
proc surveyselect data=grps(where=(&origOutcome^=.)) noprint out=selgrps4 method=urs n=MissN;
    strata sid pc;
run;

data selgrps4;
    set selgrps4;
    &outcome = &origOutcome;
    keep sid pc &outcome;
run;
proc sort data=selgrps4;by sid pc;run;
proc sort data=grps;by sid pc;run;

data &data.4;
    merge grps selgrps4;
    by sid pc;
    &newOutcome = &origOutcome;
    if &origOutcome=. then &newOutcome = &outcome;
    predictedMean = 4;
run;

/* Fifth predictive mean */
proc surveyselect data=grps(where=(&origOutcome^=.)) noprint out=selgrps5 method=urs n=MissN;
    strata sid pc;
run;

data selgrps5;
    set selgrps5;
    &outcome = &origOutcome;
    keep sid pc &outcome;
run;
proc sort data=selgrps5;by sid pc;run;
proc sort data=grps;by sid pc;run;

data &data.5;
    merge grps selgrps5;
    by sid pc;
```

```
        &newOutcome = &origOutcome;
        if &origOutcome=. then &newOutcome = &outcome;
        predictedMean = 5;
    run;

    data &out;
        set &data.1 &data.2 &data.3 &data.4 &data.5;
        keep sid id time timeclss x1 x2 cond &newOutcome &origOutcome y2 sex predictedMean;
    run;

/***  Data after imputation look like this:
Obs        id      sid     timeclss     ny2     nnny2
33       20033      1          2          0        0
34       20033      1          3          0        0
35       20033      1          4          .        0
36       20033      1          5          .        0
37       20033      1          7          .        0
143      20033      1          6          .        0
14177    20033      1          2          0        0
14178    20033      1          3          0        0
14179    20033      1          4          .        0
14180    20033      1          5          .        0
14181    20033      1          7          .        0
14287    20033      1          6          .        0
28321    20033      1          2          0        0
28322    20033      1          3          0        0
28323    20033      1          4          .        0
28324    20033      1          5          .        0
28325    20033      1          7          .        0
28431    20033      1          6          .        0
42465    20033      1          2          0        0
42466    20033      1          3          0        0
42467    20033      1          4          .        0
42468    20033      1          5          .        0
42469    20033      1          7          .        0
42575    20033      1          6          .        0
56609    20033      1          2          0        0
56610    20033      1          3          0        0
56611    20033      1          4          .        0
56612    20033      1          5          .        0
56613    20033      1          7          .        0
56719    20033      1          6          .        0



/* get starting values for NLMIXED based on the first predictive mean only:*/
    proc genmod data=&out descending;
        where predictedMean=1;
        class sid id timeclss;
        model &newOutcome = cond time x1 sex / dist=bin link=logit;
        repeated subject = sid(id) / withinsubject=timeclss;
        ods output GEEEmpPEst = parm;
    run;

    data _null_;
        set parm;
        parm=lowcase(parm);
```

```
        estimate=round(estimate,0.001);
        if parm="intercept" then call symput("fc0",estimate);
        else if parm="cond" then call symput("fc1",estimate);
        else if parm="time" then call symput("fc2",estimate);
        else if parm="x1" then call symput("fc3",estimate);
        else if parm="sex" then call symput("fc4",estimate);
    run;

    proc glimmix data=&out;
        where predictedMean=1;
        class sid;
        model &newOutcome = cond time x1 sex / solution link=logit dist=bin;
        random intercept / subject=sid;
        ods output covparms = varest;
    run;

    data _null_;
        set varest;
        if stderr=. then estimate=0.5;
        if 0<stderr<0.05 then estimate=0.05;
        if stderr>10 then estimate=1;
        call symput("fsd",Estimate);
    run;

    /* used all 5 datasets to estimate the parameters */
    proc sort data=&out;by predictedMean;run;
    proc nlmixed data=&out;
        parms c0=%sysevalf(&fc0) c1=%sysevalf(&fc1) c2=%sysevalf(&fc2) c3=%sysevalf(&fc3)
              c4=%sysevalf(&fc4) sd=%sysevalf(&fsd);
        Z = c0 + c1*cond + c2*time +c3*x1 +c4*sex+U;
        bounds sd >= 0;
        p=exp(z)/(1+exp(z));
        MODEL &newOutcome ~ binary(p);
        RANDOM U ~ NORMAL(0,sd*sd) SUBJECT=sID;
        by predictedMean;
        ods output ParameterEstimates=nlparm;
    run;

    data &outparm;
        set nlparm;
        _Imputation_ = predictedMean;
        stderr = StandardError;
    run;

    ods rtf file=&outfile style=sasweb;
    proc mianalyze parms=&outparm;
        modeleffects c0 c1 c2 c3 c4 sd;
        ods output ParameterEstimates=&impEstOut;
    run;
    ods rtf close;
%mend pmmMethod;

* [ ---- 3 Modeling ------------------------------------------------------- ];
%geeGlimmix(outcome=ny2,pattern=%str(p2=1),data=transi);
%nlmixedPredict(outcome=ny2,pattern=%str(p2=1),data=transi,outpred=trPred,
 title=Prob of moving from NS to SM);
%predictiveMean(origOutcome=ny2,outcome=ny2,newOutcome=ny2b,imputePattern=%str(p2 in (1,2)),
```

```
data=transi , predIn=trPred , out=transi2 );

%geeGlimmix(outcome=ny2b , pattern=%str(p2 in (1,2)) , data=transi2 );
%nlmixedPredict(outcome=ny2b , pattern=%str(p2 in (1,2)) , data=transi2 , outpred=trPred2 ,
title=Prob of moving from NS to SM);
%predictiveMean(origOutcome=ny2 , outcome=ny2b , newOutcome=ny2c , imputePattern=%str(p2=3),
data=transi2 , predIn=trPred2 , out=transi3 );

%geeGlimmix(outcome=ny2c , pattern=%str(p2 in (1,2,3)) , data=transi3 );
%nlmixedPredict(outcome=ny2c , pattern=%str(p2 in (1,2,3)) , data=transi3 , outpred=trPred3 ,
title=Prob of moving from NS to SM);
%predictiveMean(origOutcome=ny2 , outcome=ny2c , newOutcome=ny2d , imputePattern=%str(p2=4),
data=transi3 , predIn=trPred3 , out=transi4 );

%geeGlimmix(outcome=ny2d , pattern=%str(p2 in (1,2,3,4)) , data=transi4 );
%nlmixedPredict(outcome=ny2d , pattern=%str(p2 in (1,2,3,4)) , data=transi4 , outpred=trPred4 ,
title=Prob of moving from NS to SM);
%predictiveMean(origOutcome=ny2 , outcome=ny2d , newOutcome=ny2e , imputePattern=%str(p2=5),
data=transi4 , predIn=trPred4 , out=transi5 );

%geeGlimmix(outcome=ny2e , pattern=%str(p2 in (1,2,3,4,5)) , data=transi5 );
%nlmixedPredict(outcome=ny2e , pattern=%str(p2 in (1,2,3,4,5)) , data=transi5 , outpred=trPred5 ,
title=Prob of moving from NS to SM);
%predictiveMean(origOutcome=ny2 , outcome=ny2e , newOutcome=ny2f , imputePattern=%str(p2=6),
data=transi5 , predIn=trPred5 , out=transi6 );

%geeGlimmix(outcome=ny2f , pattern=%str(p2 in (1,2,3,4,5,6)) , data=transi6 );
%nlmixedPredict(outcome=ny2f , pattern=%str(p2 in (1,2,3,4,5,6)) , data=transi6 , outpred=trPred6 ,
title=Prob of moving from NS to SM);
%predictiveMean(origOutcome=ny2 , outcome=ny2f , newOutcome=ny2g , imputePattern=%str(p2=7),
data=transi6 , predIn=trPred6 , out=transi7 );

/* At this point, all the missing values have been imputed using the ACMV restriction methods.
   The modeling below uses the predictive mean matching method */
data ny2_acmv;
    set transi7(keep=sid id time timeclss x1 x2 cond ny2g ny2 /*y*/ sex);
    set transi7;
    nny2=ny2g;
    drop ny2g;
run;
%pmmMethod(origOutcome=ny2 , outcome=nny2 , newOutcome=nnny2 , data=ny2_acmv , out=ny2_acmv_imp ,
outparm=impParm , impEstOut=impEst );
```

Listing B.2: data_creation.sas

```
/*************************************************************************\
 *** data_creation.sas: Generates the datasets for the simulation       ***
\*************************************************************************/
options nocenter ps=5000;* nonotes formdlim=" " ;
proc datasets library=work memtype=data kill;
quit;
data _a;
 retain seed %sysevalf(&simseed);
    /* create 50 schools */
    do sid=1 to 50;
        call ranbin(seed,1,0.55,cond); * Assign schools to treatment or control conditions;
        call rannor(seed,U0);
```

```
    do id=1 to 100;                 * need 100 students per school;
        *call rannor(seed,U1);       * level 2 data;
    /* level 2 data = 1000 students randomly assigned to treatment (cond=1) or
           control (cond=0). Treatment condition: generate standard normal variates
           with correlation of rho for bivariate normal distribution
         */

        /* generate level 1 data: 7 time points for each individual */
    do time=1 to 7;
        /* first level normal variate */
            call rannor(seed,z1);
            call rannor(seed,z2);

            /* Model parameters */

            b01=-2.3; b11= -4.1;  b21=  0.61;  b31=  0.20; u12=0.68;
            b02= 0.8; b12=  0.2;  b22= -0.3;   b32= -0.1; u23=0.68;
            b03=-1.7; b13= -5.5;  b23=  0.1;   b33=  0.3;  u32=0.68;


        /* Missing value parameter */
    /* 30% missing *
            m01= -2.6; m11= -0.1; m21=0.2; m31=0.3;
            m02= -2.74; m12= -0.1; m22=0.2; m32=0.3;
            m03= -3.24; m13= -0.1; m23=0.2; m33=0.3;
            m01_1=-2.35 ; m02_1=-2.12; m03_1=-2.2;/*** Intercept are only for MAR missign data ***
            k1=0.097;k2=0.116; k3=0.127;

        /* 30% missing *
            m01= -3.1; m11= -0.1; m21=0.2; m31=0.9;
            m02= -3.94; m12= -0.1; m22=0.2; m32=0.9;
            m03= -5.58; m13= -0.1; m23=0.2; m33=0.9;
            m01_1=-2.35 ; m02_1=-2.13; m03_1=-2.19;/*** Intercept are only for MAR missign data ***
            k1=0.097;k2=0.115; k3=0.13;
    /*** Only for MAR to keep 30% missing data ***/
          /* 40% missing */
            m01= -2.15;  m11=-0.1;  m21 =0.2; m31=0.3;
            m02= -2.4;  m12=-0.1;  m22 =0.2; m32=0.3;
            m03= -2.85;  m13=-0.1;  m23 =0.2; m33=0.3;
/*** Only for MAR to keep 40% missing data ****/
            m01_1=-1.92 ; m02_1=-1.75; m03_1=-1.8 ;
         k1=0.135; k2=0.165; k3=0.181;

    /* Random Intercept Model:*/
            if time=1 then do;
                 x1=(z1+5);x2=abs(z2);
        /* multinomial probabilities of smoking at baseline:
           1  = nonsmoker = 86%
           2  = quitter = 6%
           3  = smoker = 8% */

                mu_1=0;mu_2=0;mu_3=0;
                y = rantbl(seed, 0.86, 0.08, 0.06);

                                  xlag=.; ylag=.; x2lag=.;
                call symput("ylag",trim(left(put(y,8.))));
                call symput("xlag",(put(x1,8.2)));
```

```
        call symput("x2lag",(put(x2,8.2)));
        r11=0;r21=0; r31=0; r41=0; r12=0; r22=0; r32=0;
        r42=0; r13=0; r23=0; r33=0; r43=0;
        output;
    end;
    if time=2 then do;
        mu_1=.;mu_2=.;mu_3=.;y2=.;y3=.;y4=.;
        ylag=input(symget("ylag"),best.); * creates lag variable for Y;
        x1=(z1+5);
        x2=time*abs(z2)+ylag;
        xlag=input(symget("xlag"),best.); * creates lag variable for X;
        x2lag=input(symget("x2lag"),best.);
/* Probability of moving from non-smokers to smokers
          U12*u0 just multiplying the standard normal to the given
          variance: account for school level effect
   U1: account for individual correlation over time */
        if ylag=1 then do;
            eta12 = b01 + (b11*cond) + (b21*time) + (b31*xlag) + U12*u0 ;
            /* to convert to binary outcomes */
            mu_1 = (exp(eta12))/(1+exp(eta12));call ranbin(seed,1,mu_1,y2);
/* creates missing values */
            call ranbin(seed,1,k1,r11);    * MCAR;
            r31=(exp(m01_1+m11*x2lag+m21*ylag))/(1+exp(m01_1+m11*x2lag+m21*ylag));
            call ranbin(seed,1,r31,r31); * MAR;
            r41=(exp(m01+m11*x2lag+m21*ylag+m31*y))/(1+exp(m01+m11*x2lag+m21*ylag+m31*y));
            call ranbin(seed,1,r41,r41); * MNAR;
        end;

        /* Probability of moving from smokers to quitters */
        if ylag=2 then do;
            eta23 = b02 + (b12*cond) + (b22*time) + (b32*xlag)  + U23*u0;
            mu_2 = (exp(eta23))/(1+exp(eta23));
            call ranbin(seed,1,mu_2,y3);
            /* create missing value */
call ranbin(seed,1,k2,r12);    * MCAR;
            r32=(exp(m02_1+m12*x2lag+m22*ylag))/(1+exp(m02_1+m12*x2lag+m22*ylag));
            call ranbin(seed,1,r32,r32); * MAR;
            r42=(exp(m02+m12*x2lag+m22*ylag+m32*y))/(1+exp(m02+m12*x2lag+m22*ylag+m32*y));
            call ranbin(seed,1,r42,r42); * MNAR;
        end;

/* Probability of moving from Quitter to smokers */
if ylag=3 then do;
            eta32 = b03 + (b13*cond) + (b23*time) + (b33*xlag) +  U32*u0;
            mu_3 = (exp(eta32))/(1+exp(eta32));
            call ranbin(seed,1,mu_3,y4);
            /* create missing value */
    call ranbin(seed,1,k3,r13);    * MCAR;
            r33=(exp(m03_1+m13*x2lag+m23*ylag))/(1+exp(m03_1+m13*x2lag+m23*ylag));
            call ranbin(seed,1,r33,r33); * MAR;
            r43=(exp(m03+m13*x2lag+m23*ylag+m33*y))/(1+exp(m03+m13*x2lag+m23*ylag+m33*y));
            call ranbin(seed,1,r43,r43); * MNAR;
        end;

        /* store all three transitions in a single variable */
        if y2=1 then y=2;/*** started smoking ****/
        if y2=0 then y=1;/*** remained a non-smoker **/
```

```
        if y3=1 then y=3;/*** quit smoking ***/
        if y3=0 then y=2;/*** remained a smoker ***/
        if y4=1 then y=2;/*** relapse: returned to smoking ***/
        if y4=0 then y=3;/*** remained a quitter ***/

        t11=0;t21=0; t31=0; t41=0; t12=0; t22=0; t32=0;
        t42=0; t13=0; t23=0; t33=0; t43=0;
        call symput("ylag",trim(left(put(y,8.))));
        call symput("xlag",(put(x1,8.2)));
        call symput("x2lag",(put(x2,8.2)));
        output;
    end;
    if time=3 then do;
        mu_1=.;mu_2=.;mu_3=.;y2=.;y3=.;y4=.;
        ylag=input(symget("ylag"),best.);
        x1=(z1+5);
        x2=abs(z2)+ylag;

        xlag=input(symget("xlag"),best.);
        x2lag=input(symget("x2lag"),best.);
        if ylag=3 then y2_d1=1;
        if ylag=1 then y2_d2=1;
/* non-smoker to smoker transition */
        if ylag=1 then do;
            eta12 = b01 + (b11*cond) + (b21*time) + (b31*xlag) + U12*u0;
            mu_1 = (exp(eta12))/(1+exp(eta12));
            call ranbin(seed,1,mu_1,y2);
            /* create missing values */
            /* MCAR */
            call ranbin(seed,1,k1,r11);
            /* MAR */
            r31=(exp(m01_1+m11*x2lag+m21*ylag))/(1+exp(m01_1+m11*x2lag+m21*ylag));
            call ranbin(seed,1,r31,r31);
/* MNAR */
            r41=(exp(m01+m11*x2lag+m21*ylag))/(1+exp(m01+m11*x2lag+m21*ylag+m31*y));
            call ranbin(seed,1,r41,r41);
        end;
        /* smoker to quitter transition */
        if ylag=2 then do;
            eta23 = b02 + (b12*cond) + (b22*time) + (b32*xlag) + U23*u0;
            mu_2 = (exp(eta23))/(1+exp(eta23));
            call ranbin(seed,1,mu_2,y3);
/* create missing values */
            /* MCAR */
            call ranbin(seed,1,k2,r12);
/* MAR */
r32=(exp(m02_1+m12*x2lag+m22*ylag))/(1+exp(m02_1+m12*x2lag+m22*ylag));
            call ranbin(seed,1,r32,r32);
/* MNAR */
r42=(exp(m02+m12*x2lag+m22*ylag+m32*y))/(1+exp(m02+m12*x2lag+m22*ylag+m32*y));
            call ranbin(seed,1,r42,r42);
        end;
        /* quitter to smoker transition */
if ylag=3 then do;
            eta32 = b03 + (b13*cond) + (b23*time) + (b33*xlag) +  U32*u0;
            mu_3 = (exp(eta32))/(1+exp(eta32));
            call ranbin(seed,1,mu_3,y4);
```

```
            /* create missing values */
/* MCAR */
call ranbin(seed,1,k3,r13);
/* MAR */
            r33=(exp(m03_1+m13*x2lag+m23*ylag))/(1+exp(m03_1+m13*x2lag+m23*ylag));
            call ranbin(seed,1,r33,r33);
/* MNAR */
            r43=(exp(m03+m13*x2lag+m23*ylag+m33*y))/(1+exp(m03+m13*x2lag+m23*ylag+m33*y));
            call ranbin(seed,1,r43,r43);
        end;

        if y2=1 then y=2; if y2=0 then y=1;
        if y3=1 then y=3; if y3=0 then y=2;
        if y4=1 then y=2; if y4=0 then y=3;

        call symput("ylag",trim(left(put(y,8.))));
        call symput("xlag",(put(x1,8.2)));
        call symput("x2lag",(put(x2,8.2)));
        output;
    end;
    if time=4 then do;
        mu_1=.;mu_2=.;mu_3=.;y2=.;y3=.;y4=.;
        ylag=input(symget("ylag"),best.);
        x1=(z1+5);
        x2=abs(z2)+ylag;
        xlag=input(symget("xlag"),best.);
        x2lag=input(symget("x2lag"),best.);
        /* non-smoker to smoker */
        if ylag=1 then do;
            eta12 = b01 + (b11*cond) + (b21*time) + (b31*xlag) + U12*u0;
            mu_1 = (exp(eta12))/(1+exp(eta12));
            call ranbin(seed,1,mu_1,y2);
            /* missing values */
            /* MCAR */
            call ranbin(seed,1,k1,r11);
            /* MAR */
            r31=(exp(m01_1+m11*x2lag+m21*ylag))/(1+exp(m01_1+m11*x2lag+m21*ylag));
            call ranbin(seed,1,r31,r31);
            /* MNAR */
            r41=(exp(m01+m11*x2lag+m21*ylag+m31*y))/(1+exp(m01+m11*x2lag+m21*ylag+m31*y));
            call ranbin(seed,1,r41,r41);
        end;
        /* smoker to quitter */
        if ylag=2 then do;
            eta23 = b02 + (b12*cond) + (b22*time) + (b32*xlag) + U23*u0;
            mu_2 = (exp(eta23))/(1+exp(eta23));
            call ranbin(seed,1,mu_2,y3);
            /* missing values */
            /* MCAR */
            call ranbin(seed,1,k2,r12);
            /* MAR */
            r32=(exp(m02_1+m12*x2lag+m22*ylag))/(1+exp(m02_1+m12*x2lag+m22*ylag));
            call ranbin(seed,1,r32,r32);
            /* MNAR */
            r42=(exp(m02+m12*x2lag+m22*ylag+m32*y))/(1+exp(m02+m12*x2lag+m22*ylag+m32*y));
            call ranbin(seed,1,r42,r42);
        end;
```

```
        /* quitter to smoker */
if ylag=3 then do;
            eta32 = b03 + (b13*cond) + (b23*time) + (b33*xlag) +  U32*u0;
            mu_3 = (exp(eta32))/(1+exp(eta32));
            /* missing values */
    call ranbin(seed,1,mu_3,y4);
            /* MCAR */
call ranbin(seed,1,k3,r13);
/* MAR */
            r33=(exp(m03_1+m13*x2lag+m23*ylag))/(1+exp(m03_1+m13*x2lag+m23*ylag));
            call ranbin(seed,1,r33,r33);
            /* MNAR */
            r43=(exp(m03+m13*x2lag+m23*ylag+m33*y))/(1+exp(m03+m13*x2lag+m23*ylag+m33*y));
            call ranbin(seed,1,r43,r43);
        end;
        if y2=1 then y=2; if y2=0 then y=1;
        if y3=1 then y=3; if y3=0 then y=2;
        if y4=1 then y=2; if y4=0 then y=3;
        call symput("ylag",trim(left(put(y,8.))));
        call symput("xlag",(put(x1,8.2)));
        call symput("x2lag",(put(x2,8.2)));
        output;
    end;
    if time=5 then do;
        mu_1=.;mu_2=.;mu_3=.;y2=.;y3=.;y4=.;
        ylag=input(symget("ylag"),best.);
        x1=(z1+5);
        x2=abs(z2)+ylag;
        xlag=input(symget("xlag"),best.);
        x2lag=input(symget("x2lag"),best.);
        /* non-smoker to smoker */
        if ylag=1 then do;
            eta12 = b01 + (b11*cond) + (b21*time) + (b31*xlag) + U12*u0;
            mu_1 = (exp(eta12))/(1+exp(eta12));call ranbin(seed,1,mu_1,y2);
            /* missing values */
            /* MCAR */
            call ranbin(seed,1,k1,r11);
            /* MAR */
            r31=(exp(m01_1+m11*x2lag+m21*ylag))/(1+exp(m01_1+m11*x2lag+m21*ylag));
            call ranbin(seed,1,r31,r31);
            /* MNAR */
            r41=(exp(m01+m11*x2lag+m21*ylag+m31*y))/(1+exp(m01+m11*x2lag+m21*ylag+m31*y));
            call ranbin(seed,1,r41,r41);
        end;
        /* smoker to quitter */
        if ylag=2 then do;
            eta23 = b02 + (b12*cond) + (b22*time) + (b32*xlag) + U23*u0;
            mu_2 = (exp(eta23))/(1+exp(eta23));
            call ranbin(seed,1,mu_2,y3);
            /* missing values */
            /* MCAR */
            call ranbin(seed,1,k2,r12);
            /* MAR */
            r32=(exp(m02_1+m12*x2lag+m22*ylag))/(1+exp(m02_1+m12*x2lag+m22*ylag));
            call ranbin(seed,1,r32,r32);
            /* MNAR */
            r42=(exp(m02+m12*x2lag+m22*ylag+m32*y))/(1+exp(m02+m12*x2lag+m22*ylag+m32*y));
```

```
        call ranbin(seed,1,r42,r42);
    end;
    /* quitter to smoker */
    if ylag=3 then do;
        eta32 = b03 + (b13*cond) + (b23*time) + (b33*xlag) +  U32*u0 ;
        mu_3 = (exp(eta32))/(1+exp(eta32));
        /* missing values */
        call ranbin(seed,1,mu_3,y4);
        /* MCAR */
        call ranbin(seed,1,k3,r13);
        /* MAR */
        r33=(exp(m03_1+m13*x2lag+m23*ylag))/(1+exp(m03_1+m13*x2lag+m23*ylag));
        call ranbin(seed,1,r33,r33);
        /* MNAR */
        r43=(exp(m03+m13*x2lag+m23*ylag+m33*y))/(1+exp(m03+m13*x2lag+m23*ylag+m33*y));
        call ranbin(seed,1,r43,r43);
    end;
    if y2=1 then y=2; if y2=0 then y=1;
    if y3=1 then y=3; if y3=0 then y=2;
    if y4=1 then y=2; if y4=0 then y=3;
    call symput("ylag",trim(left(put(y,8.))));
    call symput("xlag",(put(x1,8.2)));
    call symput("x2lag",(put(x2,8.2)));
    output;
end;
if time=6 then do;
    mu_1=.;mu_2=.;mu_3=.;y2=.;y3=.;y4=.;
    ylag=input(symget("ylag"),best.);
    x1=(z1+5);
    x2=abs(z2)+ylag;
    xlag=input(symget("xlag"),best.);
    x2lag=input(symget("x2lag"),best.);
    /* non-smoker to smoker */
    if ylag=1 then do;
    eta12 = b01 + (b11*cond) + (b21*time) + (b31*xlag) + U12*u0 ;
        mu_1 = (exp(eta12))/(1+exp(eta12));
        call ranbin(seed,1,mu_1,y2);
        /* missing values */
        /* MCAR */
        call ranbin(seed,1,k1,r11);
        /* MAR: CDD */
        r21=(exp(m01_1+m11*x2lag))/(1+exp(m01_1+m11*x2lag));
        call ranbin(seed,1,r21,r21);
        /* MAR */
        r31=(exp(m01_1+m11*x2lag+m21*ylag))/(1+exp(m01_1+m11*x2lag+m21*ylag));
        call ranbin(seed,1,r31,r31);
        /* MNAR */
        r41=(exp(m01+m11*x2lag+m21*ylag+m31*y))/(1+exp(m01+m11*x2lag+m21*ylag+m31*y));
        call ranbin(seed,1,r41,r41);
    end;
    /* smoker to quitter */
    if ylag=2 then do;
        eta23 = b02 + (b12*cond) + (b22*time) + (b32*xlag)  + U23*u0 ;
        mu_2 = (exp(eta23))/(1+exp(eta23));call ranbin(seed,1,mu_2,y3);
        /* missing values */
        /* MCAR */
        call ranbin(seed,1,k2,r12);
```

```
            /* MAR */
            r32=(exp(m02_1+m12*x2lag+m22*ylag))/(1+exp(m02_1+m12*x2lag+m22*ylag));
            call ranbin(seed,1,r32,r32);
            /* MNAR */
            r42=(exp(m02+m12*x2lag+m22*ylag+m32*y))/(1+exp(m02+m12*x2lag+m22*ylag+m32*y));
            call ranbin(seed,1,r42,r42);
        end;
        /* quitter to smoker */
        if ylag=3 then do;
            eta32 = b03 + (b13*cond) + (b23*time) + (b33*xlag) +  U32*u0 ;
            mu_3 = (exp(eta32))/(1+exp(eta32));
            /* missing values */
call ranbin(seed,1,mu_3,y4);
            /* MCAR */
            call ranbin(seed,1,k3,r13);
            /* MAR */
            r33=(exp(m01_1+m13*x2lag+m23*ylag))/(1+exp(m01_1+m13*x2lag+m23*ylag));
            call ranbin(seed,1,r33,r33);
            /* MNAR */
            r43=(exp(m01+m13*x2lag+m23*ylag+m33*y))/(1+exp(m01+m13*x2lag+m23*ylag+m33*y));
            call ranbin(seed,1,r43,r43);
        end;
        if y2=1 then y=2; if y2=0 then y=1;
        if y3=1 then y=3; if y3=0 then y=2;
        if y4=1 then y=2; if y4=0 then y=3;
        call symput("ylag",trim(left(put(y,8.))));
        call symput("xlag",(put(x1,8.2)));
        call symput("x2lag",(put(x2,8.2)));
        output;
    end;
    if time=7 then do;
        mu_1=.;mu_2=.;mu_3=.;y2=.;y3=.;y4=.;
        ylag=input(symget("ylag"),best.);
        x1=(z1+5);
        x2=abs(z2)+ylag;
        xlag=input(symget("xlag"),best.);
        x2lag=input(symget("x2lag"),best.);
        /* non-smoker to smoker */
        if ylag=1 then do;
            eta12 = b01 + (b11*cond) + (b21*time) + (b31*xlag) + U12*u0 ;
            mu_1 = (exp(eta12))/(1+exp(eta12));
            /* missing values */
            call ranbin(seed,1,mu_1,y2);
            /* MCAR */
            call ranbin(seed,1,k1,r11);
            /* MAR */
            r31=(exp(m01_1+m11*x2lag+m21*ylag))/(1+exp(m01_1+m11*x2lag+m21*ylag));
            call ranbin(seed,1,r31,r31);
            /* MNAR */
            r41=(exp(m01+m11*x2lag+m21*ylag+m31*y))/(1+exp(m01+m11*x2lag+m21*ylag+m31*y));
            call ranbin(seed,1,r41,r41);
        end;
        /* smoker to quitter */
        if ylag=2 then do;
            eta23 = b02 + (b12*cond) + (b22*time) + (b32*xlag)  + U23*u0 ;
            mu_2 = (exp(eta23))/(1+exp(eta23));
            /* missing values */
```

```
                        call ranbin(seed,1,mu_2,y3);
                        /* MCAR */
                        call ranbin(seed,1,k2,r12);
                        /* MAR */
                        r32=(exp(m02_1+m12*x2lag+m22*ylag))/(1+exp(m02_1+m12*x2lag+m22*ylag));
                        call ranbin(seed,1,r32,r32);
                        /* MNAR */
                        r42=(exp(m02+m12*x2lag+m22*ylag+m32*y))/(1+exp(m02+m12*x2lag+m22*ylag+m32*y));
                        call ranbin(seed,1,r42,r42);
                    end;
                    /* quitter to smoker */
                    if ylag=3 then do;
                        eta32 = b03 + (b13*cond) + (b23*time) + (b33*xlag) +  U32*u0 ;
                        mu_3 = (exp(eta32))/(1+exp(eta32));
                        /* missing values */
            call ranbin(seed,1,mu_3,y4);
                        /* MCAR */
                        call ranbin(seed,1,k3,r13);
            /* MAR */
                        r33=(exp(m03_1+m13*x2lag+m23*ylag))/(1+exp(m03_1+m13*x2lag+m23*ylag));
                        call ranbin(seed,1,r33,r33);
                        /* MNAR */
                        r43=(exp(m03+m13*x2lag+m23*ylag+m33*y))/(1+exp(m03+m13*x2lag+m23*ylag+m33*y));
                        call ranbin(seed,1,r43,r43);
                    end;
                    if y2=1 then y=2; if y2=0 then y=1;
                    if y3=1 then y=3; if y3=0 then y=2;
                    if y4=1 then y=2; if y4=0 then y=3;
                    call symput("ylag",trim(left(put(y,8.))));
                    call symput("xlag",(put(x1,8.2)));
                    call symput("x2lag",(put(x2,8.2)));
                    output;
                end;
            end;
        end;
    end;

run;
data _b;
    set _a;

    /* need to retain a var here
    retain cmiss;*/

    y2=96; y3=96; y4=96;

        if time=1 then do;
            if y=1  then y2=0; * NS ;
            if y=2  then y3=0; * SMK ;
            if y=3  then y4=0; * QT;
        end;
        if time=2 then do;
            if y=1 and ylag=1 then y2=0;        /*** remained non-smokers***/
            if y=2 and ylag=1 then y2=1;/*** became smokers ***/

            if y=2 and ylag=1 then y3=0; /*** smoker state ***/
            if y=2 and ylag=2 then y3=0; /*** remained smokers ***/
```

```
    if y=3 and ylag=2 then y3=1; /*** quit smoking ***/

    if y=3 and ylag=2 then y4=0; /*** quitter state ***/
    if y=3 and ylag=3 then y4=0; /*** remained quit ****/
    if y=2 and ylag=3 then y4=1;/*** relapse: started smoking again ***/
end;
if time=3 then do;
    if y=1 and ylag=1 then y2=0;        /*** remained non-smokers***/
    if y=2 and ylag=1 then y2=1;/*** became smokers ***/

    if y=2 and ylag=1 then y3=0; /*** smoker state ***/
    if y=2 and ylag=2 then y3=0; /*** remained smokers ***/
    if y=3 and ylag=2 then y3=1; /*** quit smoking ***/

    if y=3 and ylag=2 then y4=0; /*** quitter state ***/
    if y=3 and ylag=3 then y4=0; /*** remained a quitter ****/
    if y=2 and ylag=3 then y4=1;/*** relapse: started smoking again ***/
end;
if time=4 then do;
    if y=1 and ylag=1 then y2=0;/*** remained non-smokers***/
    if y=2 and ylag=1 then y2=1;/*** became smokers ***/

    if y=2 and ylag=1 then y3=0;/*** smoker state ***/
    if y=2 and ylag=2 then y3=0;/*** remained smokers ***/
    if y=3 and ylag=2 then y3=1;/*** quit smoking ***/

    if y=3 and ylag=2 then y4=0;/*** quitter state ***/
    if y=3 and ylag=3 then y4=0;/*** remained quit ****/
    if y=2 and ylag=3 then y4=1;/*** relapse: started smoking again ***/
end;
if time=5 then do;
    if y=1 and ylag=1 then y2=0;/*** remained non-smokers***/
    if y=2 and ylag=1 then y2=1;/*** became smokers ***/

    if y=2 and ylag=1 then y3=0;/*** smoker state ***/
    if y=2 and ylag=2 then y3=0;/*** remained smokers ***/
    if y=3 and ylag=2 then y3=1;/*** quit smoking ***/

    if y=3 and ylag=2 then y4=0;/*** quitter state ***/
    if y=3 and ylag=3 then y4=0;/*** remained quit ****/
    if y=2 and ylag=3 then y4=1;/*** relapse: started smoking again ***/
end;
if time=6 then do;
    if y=1 and ylag=1 then y2=0;/*** remained non-smokers***/
    if y=2 and ylag=1 then y2=1;/*** became smokers ***/

    if y=2 and ylag=1 then y3=0;/*** smoker state ***/
    if y=2 and ylag=2 then y3=0;/*** remained smokers ***/
    if y=3 and ylag=2 then y3=1;/*** quit smoking ***/

    if y=3 and ylag=2 then y4=0;/*** quitter state ***/
    if y=3 and ylag=3 then y4=0;/*** remained quit ****/
    if y=2 and ylag=3 then y4=1;/*** relapse: started smoking again ***/
end;
if time=7 then do;
    if y=1 and ylag=1 then y2=0;/*** remained non-smokers***/
    if y=2 and ylag=1 then y2=1;/*** became smokers ***/
```

```
        if y=2 and ylag=1 then y3=0;/*** smoker state ***/
        if y=2 and ylag=2 then y3=0;/*** remained smokers ***/
        if y=3 and ylag=2 then y3=1;/*** quit smoking ***/

        if y=3 and ylag=2 then y4=0;/*** quitter state ***/
        if y=3 and ylag=3 then y4=0;/*** remained quit ****/
        if y=2 and ylag=3 then y4=1;/*** relapse: started smoking again ***/
    end;


Label y2 = " Moving from NS to SM";
Label y3 = " Moving from SM to QT";
Label y4 = " Moving from QT to SM";

timeclss=time;
const=1;

id=(sid*1000)+id;

/* assign missing values to the data */
if r11=1 then r11=.; if r11=0 then r11=1;
if r12=1 then r12=.; if r12=0 then r12=1;
if r13=1 then r13=.; if r13=0 then r13=1;

if r31=1 then r31=.; if r31=0 then r31=1;
if r32=1 then r32=.; if r32=0 then r32=1;
if r33=1 then r33=.; if r33=0 then r33=1;

if r41=1 then r41=.;if r41=0 then r41=1;
if r42=1 then r42=.;if r42=0 then r42=1;
if r43=1 then r43=.;if r43=0 then r43=1;

/* assign missing data to response variable */
cy2=y2*r11; cy3=y3*r12; cy4=y4*r13; * NS -> SM;
ay2=y2*r31; ay3=y3*r32; ay4=y4*r33; * SM -> QT;
ny2=y2*r41; ny3=y3*r42; ny4=y4*r43; * QT -> SM;

/* set all missing data to binary (0,1) values */
if r11=1 then r11=0;if r11=. then r11=1;
if r12=1 then r12=0;if r12=. then r12=1;
if r13=1 then r13=0;if r13=. then r13=1;

if r31=1 then r31=0;if r31=. then r31=1;
if r32=1 then r32=0;if r32=. then r32=1;
if r33=1 then r33=0;if r33=. then r33=1;

if r41=1 then r41=0;if r41=. then r41=1;
if r42=1 then r42=0;if r42=. then r42=1;
if r43=1 then r43=0;if r43=. then r43=1;


label
    cy2 = " Y2: MCAR" cy3= "Y3: MCAR"  cy4 = " Y4: MCAR"
    ay2 = " Y2: MAR"  ay3= "Y3: MAR"   ay4 = " Y4: MAR"
    ny2 = " Y2: MNAR" ny3= "Y3: MNAR"  ny4 = " Y4: MNAR"
    y2 = " Moving from NS-SM"
```

```
        y3 = " Moving from SM to QT"
        y4 = " Moving from QT to SM"
        ;
run;



/* create missing data patterns: MNAR */
proc transpose data=_b(keep=id time ny2 ny3 ny4) out=_nc;
    by id;
    id time;
    var ny2 ny3 ny4;
run;

data _nc;
    set _nc;
    array a[*] _1—_7;
    pattern=.;
    do i=1 to dim(a);
        if pattern=. then do;
            if a[i]=. then pattern=i;
        end;
    end;
    if i > 7 & pattern=. then pattern=0;
run;

/* MCAR */
proc transpose data=_b(keep=id time cy2 cy3 cy4) out=_cc;
    by id;
    id time;
    var cy2 cy3 cy4;
run;

data _cc;
    set _cc;
    array a[*] _1—_7;
    pattern=.;
    do i=1 to dim(a);
        if pattern=. then do;
            if a[i]=. then pattern=i;
        end;
    end;
    if i > 7 & pattern=. then pattern=0;
run;

/* MAR */
proc transpose data=_b(keep=id time ay2 ay3 ay4) out=_ac;
    by id;
    id time;
    var ay2 ay3 ay4;
run;

data _ac;
    set _ac;
    array a[*] _1—_7;
    pattern=.;
    do i=1 to dim(a);
```

```
        if pattern=. then do;
            if a[i]=. then pattern=i;
        end;
    end;
    if i > 7 & pattern=. then pattern=0;
run;

/* combine all missing data */
data sim;
    merge _b(in=a)
        _cc(in=bb keep=id _name_ pattern rename=(pattern=cp2) where=(_name_="cy2"))
        _cc(in=cc keep=id _name_ pattern rename=(pattern=cp3) where=(_name_="cy3"))
        _cc(in=dd keep=id _name_ pattern rename=(pattern=cp4) where=(_name_="cy4"))
        _ac(in=ab keep=id _name_ pattern rename=(pattern=ap2) where=(_name_="ay2"))
        _ac(in=ac keep=id _name_ pattern rename=(pattern=ap3) where=(_name_="ay3"))
        _ac(in=ad keep=id _name_ pattern rename=(pattern=ap4) where=(_name_="ay4"))
        _nc(in=b keep=id _name_ pattern rename=(pattern=p2) where=(_name_="ny2"))
        _nc(in=c keep=id _name_ pattern rename=(pattern=p3) where=(_name_="ny3"))
        _nc(in=d keep=id _name_ pattern rename=(pattern=p4) where=(_name_="ny4"));
    by id;
    drop _name_;

    array m[9,3]
        cp2 r11 cy2 cp3 r12 cy3 cp4 r13 cy4
        ap2 r31 ay2 ap3 r32 ay3 ap4 r33 ay4
        p2  r41 ny2 p3  r42 ny3 p4  r43 ny4;
    do i=1 to dim1(m);
        if m[i,1]^=0 then do;
            if time > m[i,1] then do;
                m[i,2]=1;
                m[i,3]=.;
            end;
        end;
    end;

    if cy2=. then cy3=96; if cy2=. then cy4=96;
    if cy3=. then cy4=96; if cy4=. then cy3=96.;

    if ay2=. then ay3=96; if ay2=. then ay4=96;
    if ay3=. then ay4=96; if ay4=. then ay3=96.;

    if ny2=. then ny3=96; if ny2=. then ny4=96;
    if ny3=. then ny4=96; if ny4=. then ny3=96.;


    *if ny2=96 then p2=96;
    *if ny3=96 then p3=96;
    *if ny4=96 then p2=96;
*if ny2=. then r41=1;

old_x1=x1; /*** to keep the original x1 ***/
x1=xlag;/*** to use lagx or X_{t-1} in subsequent analysis ***/
old_x2=x2;
x2=x2lag;

if x2=. then x2=old_x2;
if x1=. then x1=old_x1;
```

```
run;

data sim;
set sim;

*if ny2=96 then p2=96;
*if ny3=96 then p3=96;
*if ny4=96 then p4=96;

if time>=2 and (ny2=. and ny3=. and ny4=.) then do;
if ylag=1 and y=1 then ny3=96;if ylag=1 and y=1 then ny4=96;
if ylag=1 and y=2 then ny4=96;

if ylag=2 and y=2 then ny2=96;if ylag=2 and y=2 then ny4=96;
if ylag=2 and y=3 then ny2=96;

if ylag=3 and y=3 then ny2=96;if ylag=3 and y=3 then ny3=96;
if ylag=3 and y=2 then ny2=96;
end;

if time>=2 and (ay2=. and ay3=. and ay4=.) then do;
if ylag=1 and y=1 then ay3=96;if ylag=1 and y=1 then ay4=96;
if ylag=1 and y=2 then ay4=96;

if ylag=2 and y=2 then ay2=96;if ylag=2 and y=2 then ay4=96;
if ylag=2 and y=3 then ay2=96;

if ylag=3 and y=3 then ay2=96;if ylag=3 and y=3 then ay3=96;
if ylag=3 and y=2 then ay2=96;
end;

if time>=2 and (cy2=. and cy3=. and cy4=.) then do;
if ylag=1 and y=1 then cy3=96;if ylag=1 and y=1 then cy4=96;
if ylag=1 and y=2 then cy4=96;

if ylag=2 and y=2 then cy2=96;if ylag=2 and y=2 then cy4=96;
if ylag=2 and y=3 then cy2=96;

if ylag=3 and y=3 then cy2=96;if ylag=3 and y=3 then cy3=96;
if ylag=3 and y=2 then cy2=96;
end;
run;


proc datasets library=work nodetails nolist;
    delete _:;
quit;

/*** Data structure for LOCF method ****/
data locf;
set sim;
retain lcy2 lcy3 lcy4 lay2 lay3 lay4 lny2 lny3 lny4;
array vals[9,2]
    cy2 lcy2
    cy3 lcy3
    cy4 lcy4
    ay2 lay2
```

```
    ay3 lay3
    ay4 lay4
    ny2 lny2
    ny3 lny3
    ny4 lny4;
do i=1 to dim1(vals);
    if vals[i,1]^=. then vals[i,2]=vals[i,1];
    else vals[i,1]=vals[i,1];
end;drop i;

run;
data locf;
set locf;

if ny2=1 and ny3=0 then lny3=96;
if ny3=1 and ny4=0 then lny4=96;
if ny4=1 and ny3=0 then lny3=96;

if cy2=1 and cy3=0 then lcy3=96;
if cy3=1 and cy4=0 then lcy4=96;
if cy4=1 and cy3=0 then lcy3=96;

if ay2=1 and ay3=0 then lay3=96;
if ay3=1 and ay4=0 then lay4=96;
if ay4=1 and ay3=0 then lay3=96;

if y2=96 then lcy2=96;if y2=96 then lay2=96;if y2=96 then lny2=96;
if y3=96 then lcy3=96;if y3=96 then lay3=96;if y3=96 then lny3=96;
if y4=96 then lcy4=96;if y4=96 then lay4=96;if y4=96 then lny4=96;


if time=1 then do;
y2=96;y3=96;y4=96;
lny2=96;lny3=96;lny4=96;
lcy2=96;lcy3=96;lcy4=96;
lay2=96;lay3=96;lay4=96;
end;

if y2=1 and y3=0 then y3=96;
if y3=1 and y4=0 then y4=96;
if y4=1 and y3=0 then y3=96;

run;
data sim;
set sim;
if time=1 then do;
y2=96;y3=96;y4=96;
ny2=96;ny3=96;ny4=96;
cy2=96;cy3=96;cy4=96;
ay2=96;ay3=96;ay4=96;
end;

if y2=1 and y3=0 then y3=96;
if y3=1 and y4=0 then y4=96;
if y4=1 and y3=0 then y3=96;
```

```
if ny2=1 and ny3=0 then ny3=96;
if ny3=1 and ny4=0 then ny4=96;
if ny4=1 and ny3=0 then ny3=96;

if cy2=1 and cy3=0 then cy3=96;
if cy3=1 and cy4=0 then cy4=96;
if cy4=1 and cy3=0 then cy3=96;

if ay2=1 and ay3=0 then ay3=96;
if ay3=1 and ay4=0 then ay4=96;
if ay4=1 and ay3=0 then ay3=96;
run;
```

Bibliography

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of Mathematical Functions. National Bureau of Standards Applied Mathematics Series, Number 55*. U.S. Government Printing Office.

Albert, P., & Waclawiw, M. (1998). A two-state Markov chain for heterogeneous transitional data: A quasi-likelihood approach. *Statistics in Medicine*, *17*, 1481Ű1493.

Anderson, A., Basilevsky, A., & Hum, D. (1983). *Handbook of Survey Research*, chap. Measurement: Theory and Techniques. Academic Press.

Bjorck, A., & Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Math Comp*, *27*, 579–594.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, *82*, 81–91.

Brown, K. S., & Cameron, R. (1997). Long term evaluation of an elementary and secondary

school smoking intervention [final report]. Tech. rep., Ottawa, Canada: National Health Research and Development Program, Health Canada.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multi-level models. *Computational Statistics*, *15*, 391–420.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1(3)*, 473–514.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Buck, S. F. (1960). 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 (Methodological)*, *22*(2), 302–306.

Callens M, C. C. (2005). Performance of likelihood-based estimation methods for multilevel binary regression models. *Journal of Statistical Computation and Simulation*, *75*, 1003–1017.

Cameron, R., Brown, K. S., Best, J. A., Pelkman, C. L., Madill, C. L., Manske, S. R., & Payne, M. E. (1999). Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. *The American Journal of Public Health*, *89*(12), 1827–1831.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences*. Erlbaum, 2nd ed.

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

Cook, R. (1999). A mixed model for two-state Markov processes under panel observations. *Biometrics*, *55*, 915–920.

Cook, R., Kalbfleisch, J., & Yi, G. (2002). A generalised mover-stayer model for panel data. *Biostatistics*, *3*, 407–420.

Daniels, M. J., & Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, *56(4)*, 1241–8.

Davidian, M., & Gallant, A. R. (1992). Smooth nonparametric maximum likelihood for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Biopharmaceutics*, *20*(5), 529–556.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, *11(1)*, 57–85.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Dennis, J. E., & Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall.

Diggle, P. J., Liang, K., & Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Claredon Press.

Driezen, P. (2001). *The Development of Youth Smoking*. Master's thesis, University of Western Ontario.

Dubin, J. A., & Rivers, D. (1990). Selection bias in linear regression logit and probit models. *Sociological Methods and Research*, *18, Nos. 2 & 3*, 360–390.

Evans, M., & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, *10*, 254–272.

Fitzmaurice, G., & Laird, N. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics*, *1*, 141–156.

Fitzmaurice, G. M., Laird, N. M., & Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in Medicine*, *20*, 1009–1021.

Fitzmaurice, M. G., Molenberghs, G., & Lipsitz, S. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistal Society B*, *57(4)*, 691–704.

Gatsonis, C., Kass, R. E., & Carlin, B. (2002). *Case Studies in Bayesian Statistics, Volume 5*. Springer Science.

Gelman, A., Carlin, J., Stern, S., & Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Ghosh, J., Delampady, M., & Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer Science.

Gibbons, R. D., Hedeker, D., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics*, *18*, 271–279.

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement*, *63*, 204–238.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, *73*, 43–56.

Goldstein, H. (1995). *Multilevel Statistical Models*. Halstead Press, 2nd ed.

Golub, G. H., & Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Math. Comput.*, *23*, 221–30.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–340.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Hedeker, D., & Gibbons, R. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*(1), 64–78.

Hedeker, D., & Gibbons, R. (2006). *Longitudinal Data Analyses*. Wiley.

Hill, P., & Goldstein, H. (1998). Multilevel modeling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural Statistics*, *23*, 117–128.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, *42*, 805–20.

Johnson, R., & Wichern, D. (2001). *Applied Multivariate Statistical Analysis*. Prentice-Hall.

Kaciroti, N. A. T., Raghunathan, E. M., Schork, A., & Clark, N. M. (2006). A Bayesian approach for clustered longitudinal ordinal outcome with nonignorable missing data: Evaluation of an asthma education program. *Journal of the American Statistical Association*, *101*, 435–446.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the A*, *90*, 773–795.

Kenward, M. G., Molenberghs, G., & Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika*, *90*, 53–71.

Laird, N. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, *7*, 305–15.

Laird, N., Lange, N., & Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, *82 (397)*, 97–105.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lee, L. F. (1983). Generalized econometric models with selectivity. *Econometrica*, *51(2)*, 507–512.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73(1)*, 13–22.

Lichtenstein, G. R. E., E. (1992). Smoking cessation: What have we learned over the past decade. *Journal of Consulting and Clinical Psychology*, *60 (4)*, 518–527.

Lindstrom, M., & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association*, *83*, 1014–1022.

Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, *46*(3), 673–687.

Little, R., & Wang, Y. X. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, *52*, 98–111.

Little, R., & Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, *52(4)*, 1324–1333.

Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, *87*, 1227–1237.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125–134.

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

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measure studies. *Journal of the American Statistical Association*, *90*, 1112–1121.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, 2nd ed.

Longford, N. T. (1986). Variance components as a method for routine regression analysis of survey data. *Compstat, Physica-Verlag, Heidelberg*, (pp. 69–74).

Longford, N. T. (1993). *Random Coefficient Models*. Oxford University Press.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Mason, W. M., Wong, G. Y., & Entwistle, B. (1984). *Contextual Analysis Through the Multi-Level Linear Model. Sociological Methodology*. Jossey-Bass.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models. 2nd Ed.*. Chapman and Hall.

Michiels, B., Molenberghs, G., & & Lipsitz, S. R. (1999). Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, *55*, 978–983.

Molenberghs, G., Kenward, M. G., & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, *84*(1), 33–44.

Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, *52*(2), 153–161.

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Science and Business Media, Inc.

Molenberghs, T. H. J., Beunkens, I., Kenward, C., Mallinkrodt, M. G., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, *5*, 445–464.

Murray, D. M., Catellier, D. J., Hannan, P. J., Treuth, M. S., Stevens, J., Schmitz, K. H., Rice, J. C., & Conway, T. L. (2004). School-level intraclass correlation for physical activity in adolescent girls. *Med. Sci. Sports Exerc.*, *36*, 876–882.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, *185*, 370–84.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*(1), 12–35.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and crossclassified random structures using a multilevel model. *Journal of Educational and Behavioural Statistics*, *19*, 337–50.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models*. Thousand Oaks, 2nd ed.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM 5: Hierarchical Linear and Nonlinear Modeling*. Scientific Software International.

Robins, J. M., Rotnitzky, A., & Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106–121.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A*, *158*, 73–89.

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case study. *Journal of the Royal Statistical Society Series A*, *164*, 339–355.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Roth, P. L., & Switzer, F. S. (1995). A Monte Carlo analysis of missing data techniques in an HRM setting. *Journal of Management*, *21*, 1003–1023.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, *93*, 1321Ű39.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Inc.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15.

Sheiner, L. B., & Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: Routine clinical pharmacokinetic data. *J. Pharmacokin. Biopharmac.*, *8*, 553–71.

Skinner, C., Holt, D., & Smith, T. (1989). *Analysis of Complex Surveys*. John Wiley and Sons.

Snijders, T., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259.

Snijders, T. A. B., & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage.

Van der Leeden, R., & Busing, F. (1994). First iteration versus IGLS/RIGLS estimates in two level models: A Monte Carlo study with ML3. Unpublished manuscript, Department of Psychometrics and Research Methodology, Leiden University.

Van Ness, P. H., O'Leary, J., Byers, A., Fried, T., & Dubin, J. (2007). Fitting longitudinal mixed effect logistic regression models with the NLMIXED procedure. *SUGI 29: Statistics and Data Analysis*.

Vonesh, E. F., & Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, *48*, 1–17.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo likelihood approach. *Journal of Statistical Computation and Simulation*, *48*, 233–243.

Wu, M. C., & Bailey, K. R. (1989). Estimation and comparisons of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*, *45*(3), 939–955.

Yi, G. Y., & Cook, R. J. (2002). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association,*, *97*, 1071Ű1080.

Yosef, M. (1997). Two-level hierarchical mixed-effects logistic regression analysis: A comparison of maximum likelihood and penalized quasi-likelihood estimates. Unpublished apprenticeship paper, College of Education, Michigan State University.

Zeger, S., & Karim, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79–86.

Zeger, S. L., & Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, *44*, 1019–31.