Statistical Methods for Multi-State Analysis of Incomplete Longitudinal Data

by

Baojiang Chen

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

Waterloo, Ontario, Canada, 2008

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

Analyses of longitudinal categorical data are typically based on semiparametric models in which covariate effects are expressed on marginal probabilities and estimation is carried out based on generalized estimating equations (GEE). Methods based on GEE are motivated in part by the lack of tractable models for clustered categorical data. However such marginal methods may not yield fully efficient estimates, nor consistent estimates when missing data are present. In the first part of the thesis I develop a Markov model for the analysis of longitudinal categorical data which facilitates modeling marginal and conditional structures. A likelihood formulation is employed for inference, so the resulting estimators enjoy properties such as optimal efficiency and consistency, and remain consistent when data are missing at random. Simulation studies demonstrate that the proposed method performs well under a variety of situations. Application to data from a smoking prevention study illustrates the utility of the model and interpretation of covariate effects.

Incomplete data often arise in many areas of research in practice. This phenomenon is common in longitudinal data on disease history of subjects. Progressive models provide a convenient framework for characterizing disease processes which arise, for example, when the state represents the degree of the irreversible damage incurred by the subject. Problems arise if the mechanism leading to the missing data is related to the response process. A naive analysis might lead to biased results and invalid inferences. The second part of this thesis begins with an investigation of progressive multi-state models for longitudinal studies with incomplete observations. Maximum likelihood estimation is carried out based on an EM algorithm, and variance estimation is provided using Louis method. In general, the maximum likelihood estimates are valid when the missing data mechanism is missing completely at random or missing at random. Here we provide likelihood based method in that the parameters are identifiable no matter what the missing

data mechanism. Simulation studies demonstrate that the proposed method works well under a variety of situations.

In practice, we often face data with missing values in both the response and the covariates, and sometimes there is some association between the missingness of the response and the covariate. The proper analysis of this type of data requires taking this correlation into consideration. The impact of attrition in longitudinal studies depends on the correlation between the missing response and missing covariate. Ignoring such correlation can bias the statistical inference. We have studied the proper method that incorporates the association between the missingness of the response and missing covariate through the use of inverse probability weighted generalized estimating equations. The simulation illustrates that the proposed method yields a consistent estimator, while the method that ignores the association yields an inconsistent estimator.

Many analyses for longitudinal incomplete data focus on studying the impact of covariates on the mean responses. However, little attention has been directed to address the impact of missing covariates on the association parameters in clustered longitudinal studies. The last part of this thesis mainly addresses this problem. Weighted first and second order estimating equations are constructed to obtain consistent estimates of mean and association parameters.

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisors, Professors Richard J. Cook and Grace Y. Yi. Their wide knowledge and logical way of thinking have been of great value for me. Their understanding, encouraging and personal guidance have provided a good basis for the present thesis.

My special thanks go to Professors Mary E. Thompson, Jerry F. Lawless, John Petkau and Janice Husted for their valuable advice and for serving as thesis committee members.

I wish to thank all faculty members, administrative staff and my fellow graduate students for their help rendered to me during my studies. Especially, I owe my thanks to Ker-Ai Lee for help with the statistical computing.

Finally, I am forever indebted to my parents for their understanding, endless patience and encouragement when it was most required.

Contents

List of Tables **x**

List of Figures **xiv**

1 Introduction **1**

1.1 Overview 1

1.2 Mechanisms and Methods for Incomplete Longitudinal Data 7

1.2.1 Likelihood-Based Methods 8

1.2.2 Marginal Methods 11

1.2.3 Modeling the Missing Data Process 14

1.3 Outline of Thesis 16

2 Likelihood Analysis of Joint Marginal and Conditional Models for Longitudinal Categorical Data **18**

2.1 Introduction 18

2.2 Model Formulation 20

2.2.1 Marginal and Conditional Models 20

2.2.2 Estimation and Inference 23

2.3 Numerical Studies 27
2.3.1 Performance of the Proposed Method 27
2.3.2 Comparison of the Proposed Method and GEE 30
2.4 Inference with Missing Data 33
2.4.1 A Scoring Method 36
2.4.2 An EM Algorithm 38
2.5 Application to Waterloo Smoking Prevention Project 41
2.6 Derivatives of the Log-Likelihood 49
2.7 Log-Linear Parametrization of Categorical Model 54

3 Progressive Multi-State Models for Incomplete Longitudinal and Life History Data 56

3.1 Overview 56
3.2 Modeling Transition Probabilities 58
3.2.1 Notation and Model Formulation 59
3.2.2 Model Identifiability 62
3.2.3 EM Algorithm 62
3.2.4 Simulation Studies 64
3.3 Modeling Transition Intensities 67
3.3.1 Continuous Time Progressive Multi-State Models 76
3.3.2 Asymptotic Bias Under Dependent Inspection 80
3.3.3 Maximum Likelihood Estimate and EM Algorithm 82
3.3.4 Identifiability of the Model 86

3.3.5 Simulation Studies 87
3.4 Applications 94
3.4.1 Application to a Smoking Prevention Project 94
3.4.2 Application to Psoriatic Arthritis Data 97
3.5 Proof of the Identifiability of the Model 105

4 Marginal Methods for Longitudinal Data Analysis with Missing Response and Missing Covariates 113

4.1 Introduction 113
4.2 Notation and Model Formulation 115
4.3 Estimation and Inference 118
4.3.1 Estimating Equations for Response Parameters 118
4.3.2 Estimation of Parameters for the Missing Data Processes . . 121
4.3.3 Estimation and Inference 122
4.4 More Efficient Estimation via Augmented IPWGEE 124
4.5 Empirical Studies and Applications 126
4.5.1 Simulation Studies for Comparison of Procedures 126
4.5.2 Study of Asymptotic Bias under Misspecification of Association Structure for Missing Data Procedures 133
4.5.3 Application to a Smoking Prevention Project 134
4.5.4 Application to a Study of Patients with Skeletal Metastases 141
4.6 Extension to Accommodate Multiple Missing Covariates 148
4.7 Estimation of the Asymptotic Covariance Matrix 152
4.8 Some Proof for the Efficient Estimate via Augmented IPWGEE . . 153

5 Association Studies for Longitudinal Data Arising in Clusters with Missing Covariates **155**

5.1 Introduction 155

5.2 Cross-Sectional Studies 158

5.2.1 Notation and Model Assumptions 158

5.2.2 Estimation Procedures 161

5.2.3 Estimation and Inference 163

5.2.4 Simulation Studies 166

5.2.5 Asymptotic Studies 167

5.3 Clustered Longitudinal Data 173

5.3.1 Notation and Model Assumptions 173

5.3.2 Methods of Estimation 177

5.3.3 Estimation and Inference 179

5.3.4 Simulation Studies 182

5.3.5 Intermittently Missing Data 183

6 Discussion and Future Research **190**

6.1 Likelihood Analysis of Joint Marginal and Conditional Models for Longitudinal Categorical Data 190

6.2 Progressive Multi-State Models for Incomplete Longitudinal and Life History Data 192

6.3 Longitudinal Data Analysis with Incomplete Response and Covariates 194

Bibliography **196**

List of Tables

2.1 Simulation results under the three scenarios using Newton-Raphson and Fisher-scoring method 31

2.2 Comparison of the frequency properties of estimators of regression coefficients by the proposed method and GEE method with correctly specified models 34

2.3 Comparison of the frequency properties of estimators of regression coefficients by the proposed method and GEE method with misspecified conditional model: weak dependence 35

2.4 Comparison of the frequency properties of estimators of regression coefficients by the proposed method and GEE method with misspecified conditional model: strong dependence 35

2.5 Sample data from two schools participating in the Waterloo Smoking Prevention Project 42

2.6 Complete case analysis of the Waterloo Smoking Prevention Project data 46

2.7 Available data analysis of the Waterloo Smoking Prevention Project data 48

3.1 Simulation results under MAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with no temporal effect (i.e. $u_\beta = 1.0$) 68

3.2 Simulation results under MAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with temporal effect (i.e. $u_\beta = 1.2$) 69

3.3 Simulation results under MAR: about 30% missingness (i.e. $u_\alpha = 2$) with no temporal effect (i.e. $u_\beta = 1.0$) 70

3.4 Simulation results under MAR: about 30% missingness (i.e. $u_\alpha = 2.0$) with temporal effect (i.e. $u_\beta = 1.2$) 71

3.5 Simulation results under MNAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with no temporal effect (i.e. $u_\beta = 1.0$) 72

3.6 Simulation results under MNAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with temporal effect (i.e. $u_\beta = 1.2$) 73

3.7 Simulation results under MNAR: about 30% missingness (i.e. $u_\alpha = 2.0$) with no temporal effect (i.e. $u_\beta = 1.0$) 74

3.8 Simulation results under MNAR: about 30% missingness (i.e. $u_\alpha = 2.0$) with temporal effect (i.e. $u_\beta = 1.2$) 75

3.9 Empirical performance of regression estimators by various methods for the case without covariates: $J = 3$ 90

3.10 Empirical performance of regression estimators by various methods for the case with covariates: $J = 3$ 91

3.11 Empirical performance of regression estimators by various methods for the case without covariates: $J = 5$ 92

3.12 Empirical performance of regression estimators by various methods for the case with covariates: $J = 5$ 93

3.13 Analysis of the Waterloo Smoking Prevention Project data 98

3.14 Sample data of the psoriatic arthritis study 100

3.15 Analysis of the psoriatic arthritis data 102

4.1 Empirical bias, standard errors and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.6$) 130

4.2 Empirical bias, standard errors and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.3$) 131

4.3 Empirical bias, standard errors and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.0$) 132

4.4 Sample data from the Waterloo Smoking Prevention Project 136

4.5 Results of estimation based on unweighted and weighted GEE when analyzing data from the Waterloo Smoking Prevention Project: response models 139

4.6 Results of estimation based on unweighted and weighted GEE when analyzing data from the Waterloo Smoking Prevention Project: missing data models 140

4.7 Sample data from a bone metastases study 143

4.8 Results of estimation based on unweighted and weighted GEE when analyzing data from a bone metastases study: response models . . . 146

4.9 Results of estimation based on unweighted and weighted GEE when analyzing data from a bone metastases study: missing data models 147

5.1 Simulation results for the association study with missing covariates 168

5.2 Simulation results for the association study with missing covariates: about 20% missing (i.e. $\exp(\alpha_2) = 2.0, \exp(\alpha_3) = 1.5$) 184

5.3 Simulation results for the association study with missing covariates: about 25% missing (i.e. $\exp(\alpha_2) = 2.0, \exp(\alpha_3) = 1.0$) 185

5.4 Simulation results for the association study with missing covariates: about 30% missing (i.e. $\exp(\alpha_2) = 0.5, \exp(\alpha_3) = 1.5$) 186

5.5 Simulation results for the association study with missing covariates: about 35% missing (i.e. $\exp(\alpha_2) = 0.5, \exp(\alpha_3) = 1.0$) 187

List of Figures

2.1 Three states diagram of the transitions 27

2.2 Three-state diagram for the analysis of the Waterloo Smoking Prevention Project Data 44

3.1 K-state diagram for progressive process 59

3.2 A diagram of K-state progressive process 76

3.3 Asymptotic bias for missing not at random without covariates with 3 states. The first and the third horizonal rows are plots for the complete case analysis and the second and the fourth horizonal rows are plots for the available data analysis, with 3 and 5 observations, respectively. . . . 83

3.4 Asymptotic bias for missing not at random with one covariate with 3 states 3 observations. 84

3.5 Asymptotic bias for missing not at random with one covariate with 3 states 5 observations. 85

3.6 Survival functions for missing not at random without covariates with 3 states. The left figure is for 3 observations and the right is for 5 observations. 94

3.7 Three-state progressive diagram for the analysis of the Waterloo Smoking Prevention Project Data 95

3.8 Four-state progression diagram for psoriatic arthritis data 99

3.9 Transition probabilities for the analyses of the psoriatic arthritis data . 104

4.1 Asymptotic relative bias of regression coefficients under a misspecified models of the association structures for the missing covariate and response processes 135

5.1 Asymptotic relative bias of association parameter $\psi_{ijj'}$ in independence weights analysis with $\psi_{ijj'} = 4$ 170

5.2 Asymptotic relative bias of association parameter $\psi_{ijj'}$ in independence weights analysis with $\psi_{ijj'} = 2$ 171

5.3 Asymptotic relative bias of association parameter $\psi_{ijj'}$ in independence weights analysis with $\psi_{ijj'} = 1$ 172

Chapter 1

Introduction

1.1 Overview

Longitudinal studies are increasingly common in many areas of research including medicine, public health, and the social sciences. The defining characteristic of longitudinal studies is the repeated measurements on the same subject over time. The primary goal of a longitudinal study is often to characterize the change in responses over time as well as factors that influence this change.

During the past a few decades, statistical methods for the analysis of longitudinal data have been developed tremendously (e.g. Liang and Zeger, 1986; Prentice, 1988; Zhao and Prentice, 1990; Laird and Ware, 1982; Breslow and Clayton, 1993; Albert and Waclawiw, 1998; Albert, 2000). Despite this progress, there remains a need for further methodological research to develop analysis techniques suitable for different data and different analysis objectives. There are three broad classes of methods for the analysis of longitudinal data, namely, mixed effects models, marginal (typically semiparametric) methods, and transition models.

Mixed effects models are readily adapted if interest lies in cluster-specific or subject-specific inferences regarding covariate effects. Harville (1977) introduced a general class of linear mixed effects models for repeated measures and growth curves and Laird and Ware (1982) proposed to fit linear mixed effects models with the EM algorithm (Dempster et al., 1977). Cnaan et al. (1997) provided a detailed review of linear mixed effects models with an application to a schizophrenia clinical trial. Generally, the distribution of mixed effects is usually assumed to be normal. This assumption brings mathematical simplicity and convenience to estimation and inference for regression coefficients and also to the prediction of subject-specific random effects. The most common strategy used to deal with the mixed effects is to obtain the marginal likelihood by integrating the random effects out from the joint likelihood of the observable responses and random effects. Stiratelli et al. (1984) discussed an EM approach for the analysis of binary response data with Gaussian random effects and Longford (1993) discussed an approach based on direct maximization of the likelihood. A Gibbs sampling approach had been proposed by Zeger and Karim (1991) for the generalized linear mixed effects model. Anderson and Aitkin (1985) proposed to use adaptive Gaussian quadrature for the evaluation of integrals over the random effects, but, in practice, the calculation of the marginal likelihood can involve very intensive computation. As an alternative, Breslow and Clayton (1993) proposed to use Laplace approximations in the likelihood evaluation.

Marginal methods are commonly used to describe the dependence of the marginal, or "population averaged", features of a joint distribution on the explanatory variables through a specified link function. Estimations of parameters can be carried out without full distributional assumptions, but rather only require specification of a regression model for the mean response; estimation is based on generalized

estimating equations (GEE). The theoretical foundation for GEE can be found in Godambe (1960). Liang and Zeger (1986) and Zeger et al. (1988) proposed a class of GEEs for longitudinal data, now known as first order GEE. These methods do not require specification of the full joint distribution of the longitudinal response, but only specifications regarding of the marginal mean and variance of the response, and some "working" assumptions about the correlation of responses over time. Provided that the model for the mean is correctly specified, Liang and Zeger (1986) showed that this approach yields consistent estimates for regression parameters. They further showed that the estimates are robust to misspecification of the working correlation structure for the responses within subjects (Crowder, 2001).

Prentice (1988) and Zhao and Prentice (1990) proposed extensions of GEE to incorporate assumptions about higher-order moments for binary data. These methods are called GEE2 methods. The central idea is to model the marginal mean of each binary response and the association between pairs of response separately, and then construct a set of second-order joint estimating equations. Liang et al. (1992) discussed this class of estimating equations and extended it to consider the multivariate regression analysis for categorical data.

Transitional models examine the effect of covariates on the transition patterns across a binary or categorical response over time. With this approach, one models the probability distribution of the response at a particular time as a function of the covariates and the individual's past responses. Markov models are among the most convenient transition models, where one assumes that given the history, the conditional distribution of responses depends on only m prior observations, where the integer m is referred to as the order of the model. These models are particularly

attractive for categorical data that exhibit serial dependence since the coefficients of the past responses indicate how strongly the past outcomes are associated with the current response.

There are situations, however, where we do not want to condition on past outcomes to make inferences regarding a covariate effect (Diggle et al., 2002). For example, most clinical trials study the impact of treatment on the response at a fixed, final follow up time or on the entire response profile over time. In this case, we would not want to condition on past outcomes when making inferences regarding the effect of treatment since the earlier outcomes are internal potentially responsive "covariates". The attractive characterization of serial dependence that a transition model provides can be combined with a marginal regression structure by adopting the framework of marginalized transition models (MTM) (Azzalini, 1994; Heagerty, 2002). Azzalini (1994) introduced a binary Markov chain model to accommodate serial dependence arising in longitudinal studies. Heagerty and Zeger (2000) viewed the approach of Azzalini (1994) as combining a marginal mean model that captures systematic variation in the response as a function of covariates, with a conditional mean model that describes serial dependence and identifies the joint distribution of the current response. Inferences regarding the regression parameters are based on the likelihood method.

Incomplete Data

Longitudinal studies often feature incomplete data because of a missed study assessment or withdrawal. Problems arise if the mechanism leading to the missing data is related to the response process. Little and Rubin (1987) gave a general treatment of statistical analysis of missing data mechanisms, which includes a useful hierarchy of missing-value models. A missing-data mechanism is called Missing

Completely at Random (MCAR) if the missing data process is independent of any data, and Missing at Random (MAR) if the missing data process does not depend on the unobserved data. In contrast, data are Missing Not at Random (MNAR) if the missing data process depends on unobserved data. These notions will be discussed more completely in Section 1.2.

Likelihood methods based on the fully specified models and marginal methods based on GEE are two powerful statistical techniques that have been developed to accommodate missing data for longitudinal data analysis. Under the MCAR mechanism, the observed data are just a random sample of all the data, so a valid analysis can be obtained through a likelihood-based approach that omits data from individuals with incomplete data; this is known as a complete case (CC) analysis and, it is the technique that is most commonly used in most software packages. A CC analysis may lose efficiency because the smaller sample size will inflate the standard errors and reduce the power of tests, but no bias is introduced when the data are MCAR. MAR is a more realistic assumption than MCAR, and in most MAR scenarios, a CC analysis will be both inefficient and biased. When data are MAR or MCAR, and the parameters of the missing data mechanism are distinct from those of the sampling model, the data are said to be ignorably missing (Little and Rubin, 1987). In these cases, the missing data mechanism can be ignored in making likelihood-based inferences about the parameters in the sampling model. Under the MNAR mechanism, likelihood based methods are generally biased. Valid inferences generally require specifying the correct model for the missing data mechanism and identifiability of the parameters.

Difficulties with likelihood-based methods are that they require specification of the joint distributions of longitudinal responses, and sometimes, need specification

of the missing data process. However, in practice there is not a rich class of models for the joint distribution of longitudinal data, especially for discrete data, and it is not easy to specify the missing data model. In addition, for a MNAR mechanism, the likelihood based inferences are invalid and methods which attempt to correct for bias must rely on sensitivity analysis because parameters are not identifiable in general (Rotnitzky et al., 1998).

Marginal methods based on GEE are another alternative approach to accommodate problems with missing data. Under the MCAR mechanism, the GEE approach yields consistent estimates for the regression parameters. When the data are MAR or MNAR, an analysis based on GEE gives inconsistent estimates of parameters for the regression model. Robins and Rotnitzky (1995), and Robins et al. (1994, 1995) developed a class of estimators based on an Inverse Probability Weighted Generalized Estimating Equations (IPWGEE) in a regression setting when data are MAR. Rotnitzky and Robins (1995) extended this methodology to account for nonignorable nonresponse in the covariates or the outcomes. This approach involves modeling the missing data process and weighting the estimating equations by the inverse of a probability that is calculated based on the models for the missing data process. If the models for both the marginal mean of the response and the missing data process are correctly formulated, the IPWGEE approach corrects the bias and gives consistent estimates under the MAR mechanism.

1.2 Mechanisms and Methods for Incomplete Longitudinal Data

Let $Y_i = (Y_{i1}, \ldots, Y_{iJ})' = (Y_i^{(o)}, Y_i^{(m)})'$ be the vector of J measurements for subject i, $i = 1, \ldots, n$, where $Y_i^{(o)}$ represents the observed data part and $Y_i^{(m)}$ denotes the missing data part. Let $R_i = (R_{i1}, \ldots, R_{iJ})'$ be the corresponding missing data indicator vector, where $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ if Y_{ij} is missing. Let X_{ij} be the corresponding vector of covariates for subject i at time point j. Let $X_i = (X_{i1}', \ldots, X_{iJ}')'$. Rubin (1976) and Little and Rubin (1987) made the three classifications of missing data mechanisms, assuming X_i is always observed:

1. Missing Completely at Random (MCAR): Data are said to be MCAR if the probability of failure to observe a value does not depend on any observed or unobserved measurements, i.e.

$$P(R_i|Y_i, X_i) = P(R_i).$$

2. Missing at Random (MAR): Data are said to be MAR if, conditional on the observed data, the probability of failure to observe a value does not depend on the data that are unobserved. That is,

$$P(R_i|Y_i, X_i) = P(R_i|Y_i^{(o)}, X_i).$$

3. Missing Not at Random (MNAR): The missing data mechanism is said to be MNAR if the probability of failure to observe a value depends on the unobserved data, i.e.

$$P(R_i|Y_i, X_i) = P(R_i|Y_i^{(o)}, Y_i^{(m)}, X_i).$$

1.2.1 Likelihood-Based Methods

The likelihood for incomplete longitudinal data is developed by specifying the joint distribution of response variable Y_i and the missing data indicators R_i, given the covariates X_i. Two classes of likelihood-based models have been proposed based on alternative factorizations of the joint distribution. One is based on *selection models* (Little and Rubin, 1987), in which the joint distribution of Y_i and R_i is factorized as

$$f(R_i, Y_i|X_i; \beta, \alpha) = f(R_i|Y_i, X_i; \alpha)f(Y_i|X_i; \beta),$$

where the distribution of R_i, $f(R_i|Y_i, X_i; \alpha)$, is indexed by a vector of parameters α and the distribution of Y_i, $f(Y_i|X_i; \beta)$, is indexed by a vector of β. The other is called *pattern-mixture models* (Little, 1993; Glynn et al., 1986), in which the factorization of the joint distribution is

$$f(R_i, Y_i|X_i; \theta, \gamma) = f(Y_i|R_i, X_i; \gamma)f(R_i|X_i; \theta),$$

where $f(Y_i|R_i, X_i; \gamma)$, the distribution of Y_i, is defined separately for each missing data configuration and indexed by parameters γ, and the distribution of R_i, $f(R_i|X_i; \theta)$, is known up to parameters θ.

When we are concerned with the parameters of the marginal distribution of Y_i, averaged over the missing data patterns, it is more natural to use selection models, because people do not want to make inference conditional on the missing data indicators. In the followings, we focus on selection models.

There are two main methods for the likelihood-based methods; one is the observed likelihood method and the other is the joint modeling method. To outline this, we derive the joint density of the observed data $(Y_i^{(o)}, R_i)$ by integrating out

the missing data $Y_i^{(m)}$ in the joint distribution as

$$f(R_i, Y_i^{(o)}|X_i;\alpha,\beta) = \int f(R_i|Y_i^{(o)}, Y_i^{(m)}, X_i;\alpha) f(Y_i^{(o)}, Y_i^{(m)}|X_i;\beta) dY_i^{(m)}.$$

Then the joint likelihood for (α, β) is

$$L(\alpha,\beta;Y^{(o)},R) = \prod_{i=1}^{n} \int f(R_i|Y_i^{(o)}, Y_i^{(m)}, X_i;\alpha) f(Y_i^{(o)}, Y_i^{(m)}|X_i;\beta) dY_i^{(m)}. \quad (1.1)$$

When the missing data mechanism is MCAR or MAR, this likelihood becomes

$$\begin{aligned} L(\alpha,\beta;Y^{(o)},R) &= \prod_{i=1}^{n} \left\{ f(R_i|Y_i^{(o)}, X_i;\alpha) \int f(Y_i^{(o)}, Y_i^{(m)}|X_i;\beta) dY_i^{(m)} \right\} \\ &= \prod_{i=1}^{n} \left\{ f(R_i|Y_i^{(o)}, X_i;\alpha) f(Y_i^{(o)}|X_i;\beta) \right\}. \end{aligned}$$

Assuming the parameters α and β are functionally independent, then likelihood inference for β from the likelihood $L(\alpha, \beta; Y^{(o)}, R)$ is the same as a likelihood inference for β from the observed likelihood

$$L(\beta;Y^{(o)}) = \prod_{i=1}^{n} f(Y_i^{(o)}|X_i;\beta). \quad (1.2)$$

To get the maximum likelihood estimator, we aim to maximize the log likelihood

$$\ell(\beta;Y^{(o)}) = \sum_{i=1}^{n} \log f(Y_i^{(o)}|X_i;\beta)$$

using a Newton-Raphson algorithm

$$\beta^{(h+1)} = \beta^{(h)} + [I(\beta^{(h)})]^{-1} S(\beta^{(h)}), \quad h = 0, 1, 2, \ldots$$

or a Fisher-scoring algorithm

$$\beta^{(h+1)} = \beta^{(h)} + [J(\beta^{(h)})]^{-1} S(\beta^{(h)}), \quad h = 0, 1, 2, \ldots$$

until $\beta^{(h+1)}$ converges, where

$$I(\beta^{(h)}) = [-\partial^2 \ell(\beta;Y^{(o)})/\partial\beta\partial\beta']_{\beta=\beta^{(h)}},$$

$$S(\beta^{(h)}) = [\partial \ell(\beta; Y^{(o)})/\partial \beta]_{\beta=\beta^{(h)}}$$

and

$$J(\beta^{(h)}) = E[I(\beta^{(h)})].$$

One problem with the Newton-Raphson algorithm and the Fisher-scoring algorithm is that they require calculation of the second derivatives of the log likelihood and this can be complicated. The Expectation Maximization (EM) algorithm offers an alternative strategy to optimize the observed likelihood. Specifically, the EM algorithm iterates between the following two steps:

1. E-step: Find the expectation of the complete data log likelihood over the conditional distribution of the missing data, given the observed data and the current estimate $\beta^{(h)}$,

$$\begin{aligned} Q(\beta; \beta^{(h)}) &= E_{(Y^{(m)}|Y^{(o)};\beta^{(h)})}[\ell(\beta; Y)] \\ &= \int \ell(\beta; Y) f(Y^{(m)}|Y^{(o)}; \beta^{(h)}) dY^{(m)}. \end{aligned}$$

2. M-step: maximize $Q(\beta; \beta^{(h)})$ with respect to β to obtain the estimate $\beta^{(h+1)}$.

The EM algorithm is remarkably simple, both conceptually and computationally. Standard errors may be obtained by bootstrapping, or using Louis formula (Louis, 1982). However, the major drawbacks of the EM algorithm are that it can be very slow to converge when the missing data proportion is large, and the M step may be difficult (McLachlan and Krishnan, 1996).

When data are MNAR, the missing data model must be specified to make valid inference because the likelihood can not be simplified. The joint likelihood of the observed response $Y^{(o)}$ and the missing process R, (1.1), must be employed to make

inference and Newton-Raphson, Fisher-scoring and EM algorithms (Ibrahim et al., 2001) can also be employed in this setting.

1.2.2 Marginal Methods

Marginal models characterize how moments of the marginal response depend on explanatory variables. People often construct the generalized linear model as

$$g(\mu_{ij}) = X'_{ij}\beta,$$

where $\mu_{ij} = E(Y_{ij}|X_i)$ is the marginal mean and $g(\cdot)$ is a known link function. If the distribution of Y_i is fully specified, likelihood-based method is a good choice for the estimation of the parameters. However, in practice, there is not a rich class of models for the joint distribution of longitudinal data. From this point of view, GEE are appealing since they only require assumptions about the regression model for the marginal mean and the variance function. Let $\mu_i(\beta) = (\mu_{i1}, \ldots, \mu_{iJ})'$, then the GEE for β is given by

$$\sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} D'_i \cdot V_i^{-1} \cdot (Y_i - \mu_i(\beta)) = 0, \tag{1.3}$$

where $D_i = \partial\mu_i(\beta)/\partial\beta'$ and V_i is the covariance matrix of Y_i. Liang and Zeger (1986) suggest using a "working" covariance matrix to replace V_i, which can be modeled as

$$V_i = a(\phi)A_i^{1/2}G_i(\rho)A_i^{1/2},$$

where $a(\cdot)$ is a known function, ϕ is a scale parameter, A_i is a $J \times J$ diagonal matrix with elements $v_{ij} = \mathrm{Var}(Y_{ij})$, $G_i(\rho)$ is a $J \times J$ "working" correlation matrix that is fully specified by the vector of parameters ρ. Now these estimating equations are not only functions of β, but of ρ and ϕ as well. We often write $U_i(\beta)$ as $U_i(\beta, \rho, \phi)$

to incorporate these parameters. To solve this equation, ρ is often replaced by a $\sqrt{n}$-consistent estimate, $\hat{\rho}(Y, \beta, \phi)$ and ϕ is replaced by a $\sqrt{n}$-consistent estimate, $\hat{\phi}(Y, \beta)$, then plug them into (1.3). Newton-Raphson algorithm is often employed to obtain the solution of (1.3). Specifically, given the initial value $\beta^{(0)}$, we iterate the following two steps until convergence.

1. Given $\beta^{(h)}$, obtain a $\sqrt{n}$-consistent estimate $\phi^{(h)} = \hat{\phi}(Y, \beta^{(h)})$ and a $\sqrt{n}$-consistent estimate $\rho^{(h)} = \hat{\rho}(Y, \beta^{(h)}, \phi^{(h)})$,

2. Then obtain $\beta^{(h+1)}$ as the solution of $\sum_{i=1}^{n} U_i(\beta, \rho^{(h)}, \phi^{(h)}) = 0$ by a Newton-Raphson algorithm, say.

Denoted the limit as $\hat{\beta}$.

Under some regularity conditions and given that

1. $\hat{\rho}$ is $\sqrt{n}$-consistent, given β and ϕ,

2. $\hat{\phi}$ is $\sqrt{n}$-consistent, given β, and

3. $|\partial\hat{\rho}(\beta, \phi)/\partial\phi| \leq H(Y, \beta)$ which is $O_p(1)$,

Liang and Zeger (1986) gave the large-sample properties of $\hat{\beta}$: $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically multivariate Gaussian as $n \to \infty$ with mean zero and covariance matrix $\Sigma = I_0(\beta)^{-1} I_1(\beta) I_0(\beta)^{-1}$, where

$$I_0 = E[-\partial U_i(\beta, \rho, \phi)/\partial\beta]$$

and

$$I_1 = E[U_i(\beta, \rho, \phi) U_i(\beta, \rho, \phi)'].$$

This asymptotic covariance matrix Σ can be consistently estimated by

$$\hat{\Sigma} = \hat{I}_0^{-1}\hat{I}_1\hat{I}_0^{-1},$$

where $\hat{I}_0 = -n^{-1}\sum_{i=1}^{n} \partial U_i(\hat{\beta}, \hat{\rho}, \hat{\phi})/\partial\beta$ and $\hat{I}_1 = n^{-1}\sum_{i=1}^{n} U_i(\hat{\beta}, \hat{\rho}, \hat{\phi})U_i'(\hat{\beta}, \hat{\rho}, \hat{\phi})$.

GEE analysis is valid when the data are complete or the missing data mechanism is MCAR. When data are MAR or MNAR, GEE equations are biased. Rotnitzky and Robins (1995), Robins and Rotnitzky (1995), and Robins et al. (1994, 1995) developed a class of estimators based on Inverse Probability Weighted Generalized Estimating Equations (IPWGEE) in a regression setting when the data are MAR. The IPWGEE are given by

$$\sum_{i=1}^{n} U_i(\beta, \alpha) = \sum_{i=1}^{n} D_i' \cdot V_i^{-1} \cdot \Delta_i(\alpha) \cdot (Y_i - \mu_i(\beta)) = 0, \tag{1.4}$$

where $\Delta_i(\alpha)$ is a diagonal weight matrix that depends on the probabilities of the data having missing. The matrix may be given by $\Delta_i(\alpha) = \text{diag}(I(R_{ij} = 1)/\pi_{ij}(\alpha) : j = 1, 2, \cdots, J)$, where $\pi_{ij}(\alpha) = P(R_{ij} = 1|Y_i, X_i; \alpha)$ and $I(R_{ij} = 1)/\pi_{ij}(\alpha)$ is the so-called occasion-specific weight. Fitzmaurice et al. (1995) proposed a cluster level weight as $\Delta_i(\alpha) = \text{diag}(I(R_{ij} = 1)/\pi_i(\alpha) : j = 1, 2, \cdots, J)$ in the monotone missing data pattern (a monotone missing data process means $R_{ij} = 0$ implies $R_{ik} = 0$ for $k > j$), where $\pi_i(\alpha) = P(R_i = r_i|Y_i, X_i; \alpha)$, is the missing data probability for individual i over the entire observation period. The IPWGEE with occasion-specific level weights is more efficient than an IPWGEE with cluster level weights (Preisser et al., 2002).

Robins et al. (1995) gave the large-sample properties of the solution $\hat{\beta}$ to this equation, which stated that subject to some regularity conditions and given the regression models for the response process and the missing data process are correctly

specified, $\sqrt{n}(\hat{\beta}-\beta)$ is asymptotically multivariate Gaussian with mean 0 and asymptotic covariance matrix $\Gamma^{-1}C\Gamma^{-1'}$, where

$$\begin{aligned}\Gamma &= E\left\{\frac{\partial U_i(\beta,\alpha)}{\partial\beta'}\right\},\\ C &= \mathrm{Var}\{\mathrm{Res}(U_i(\beta,\alpha),S_i(\alpha))\},\end{aligned}$$

in which

$$\mathrm{Res}(M_i,N_i)=M_i-E[M_iN_i']\{E(N_iN_i')\}^{-1}N_i,$$

and $S_i(\alpha)$ is the score function of the missing data model. Furthermore, this asymptotic covariance matrix can be consistently estimated by $\hat{\Gamma}^{-1}\hat{C}\hat{\Gamma}^{-1'}$ with

$$\begin{aligned}\hat{\Gamma} &= n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial U_i(\hat{\beta},\hat{\alpha})}{\partial\beta'}\right\},\\ \hat{C} &= n^{-1}\sum_{i=1}^{n}\left\{\widehat{\mathrm{Res}}[U_i(\hat{\beta},\hat{\alpha}),S_i(\hat{\alpha})]\right\}^{\otimes 2},\\ \widehat{\mathrm{Res}}(M_i,N_i) &= M_i-\left\{\sum_{i=1}^{n}M_iN_i'\right\}\left\{\sum_{i=1}^{n}N_iN_i'\right\}^{-1}N_i,\end{aligned}$$

where $A^{\otimes 2}=AA'$.

1.2.3 Modeling the Missing Data Process

For the joint modeling method and the marginal method, modeling the distribution of the missing data process is required. In the following, we discuss the method of modeling the missing data process. One option is to use the binomial model (Ibrahim et al., 2001)

$$P(R_i=r_i|Y_i,X_i;\alpha) = \prod_{j=1}^{J}\left\{\pi_{ij}^{r_{ij}}(1-\pi_{ij})^{1-r_{ij}}\right\}$$

assuming the conditional independence between the R_{ij}, where $\pi_{ij} = P(R_{ij} = 1|Y_i, X_i; \alpha)$. Often a generalized linear model

$$g(\pi_{ij}) = Z'_{ij}\alpha$$

is specified to link the binomial probabilities to the covariate vector Z_{ij}. Here $g(\cdot)$ is a known link function and α measures the influence of the covariates on these probabilities.

A multinomial missing data model (Ibrahim et al., 2001) specifies the joint distribution of R_i through a sequence of one-dimensional conditional distributions:

$$P(R_i|Y_i, X_i; \alpha) = \prod_{j=2}^{J} P(R_{ij}|H^r_{ij}, Y_i, X_i; \alpha_j) \cdot P(R_{i1}|Y_i, X_i; \alpha_1), \tag{1.5}$$

where $H^r_{ij} = \{R_{i,j-1}, \ldots, R_{i1}\}$ and α_j is a vector of indexing parameters for the jth conditional distribution and $\alpha = (\alpha_1, \ldots, \alpha_J)'$. This accommodates nonmonotone patterns of missing data, and provides a natural way to specify the joint distribution of the missing data indicators when knowledge about the missingness of one response affects the probability of missingness of another. In practice, however, interest often lies in the first order dependence of the serial probability, that is

$$P(R_i|Y_i, X_i; \alpha) = \prod_{j=2}^{J} P(R_{ij}|R_{i,j-1}, Y_i, X_i; \alpha_j) \cdot P(R_{i1}|Y_i, X_i; \alpha_1), \tag{1.6}$$

where the first order Markov property is assumed for the indicator variable R_i. Let $\lambda^*_{ij}(\alpha_j) = P(R_{ij} = 1|R_{i,j-1}, Y_i, X_i; \alpha_j)$ be the conditional probability, which it is often modeled by a logistic regression model

$$\text{logit}(\lambda^*_{ij}(\alpha_j)) = Z'_{ij}\alpha_j, \quad j = 2, \ldots J,$$

where Z_{ij} features the missing data mechanism, which may include the response Y_i, the covariate X_i and the missing indicator $R_{i,j-1}$. If Z_{ij} does not include any

observed or unobserved measurements, it leads to MCAR; if Z_{ij} only includes the observed measurements, it leads to MAR; and if Z_{ij} includes the unobserved response $Y_i^{(m)}$, it leads to MNAR. The joint probability of R_i is

$$P(R_i = r_i|Y_i, X_i; \alpha) = \left[\prod_{j=2}^{J} \left(\lambda_{ij}^*(\alpha_j)\right)^{r_{ij}} \left(1 - \lambda_{ij}^*(\alpha_j)\right)^{1-r_{ij}}\right] \cdot P(R_{i1} = r_{i1}|Y_i, X_i; \alpha_1).$$

1.3 Outline of Thesis

The remaining chapters of this thesis are organized as follows.

Chapter 2

In Chapter 2, likelihood analysis of joint marginal and conditional models are explored for longitudinal categorical data. We develop a Markov model for the analysis of longitudinal categorical data which facilitates modeling marginal and conditional structures. A likelihood formulation is employed for inference, so the resulting estimators enjoy properties such as optimal efficiency and consistency, and remain consistent when data are missing at random. Simulation studies are given, which demonstrate that the proposed method performs well under a variety of situations. Application to data from a smoking prevention study illustrates the utility of the model and interpretation of covariate effects.

Chapter 3

Chapter 3 involves modeling progressive multi-state processes with incomplete observations, including the discrete time progressive process and continuous time progressive process. For the discrete time progressive process, we directly model the conditional transition probability using the generalized linear model, while for the continuous time progressive process, intensity based models are introduced to in-

corporate the covariate effects. Although model formulations are different, the estimation methods are the same, which is maximum likelihood based on the EM algorithm. Louis method is used to calculate the standard errors. Simulations and the asymptotic biases are explored to evaluate the performance of the proposed method.

Chapter 4

In Chapter 4, we consider the inverse probability weighted generalized estimating equations (IPWGEE) to handle longitudinal data with both missing response and missing covariate. The idea behind this is that we incorporate the association between the missing response and missing covariates. The simulations support the assumptions that the proposed method gives consistent estimators and is more efficient than the method that ignores the association when it is present.

Chapter 5

Chapter 5 involves addressing the impact of missing covariates on the association parameters in clustered longitudinal studies. Weighted first and second order estimating equations are constructed to obtain consistent estimates of association parameters. Clustering in the missing data process is addressed to get efficient estimates.

Chapter 6

Chapter 6 briefly summarizes overall findings and outlines the future work.

Chapter 2

Likelihood Analysis of Joint Marginal and Conditional Models for Longitudinal Categorical Data

2.1 Introduction

Marginal methods are commonly used to model longitudinal categorical data through specification of covariate effects on marginal, or "population averaged", attributes via a specified link function. Liang and Zeger (1986) proposed a class of first-order generalized estimating equations (GEE) for longitudinal data when the marginal distributions are in the exponential family. Prentice (1988) and Zhao and Prentice (1990) developed second order generalized estimating equations (GEE2) which facilitate modeling covariate effects on parameters characterizing the association between responses. Methods for regression with longitudinal categorical data were developed by Liang et al. (1992), who again focussed on marginal models

for both the mean and association structures. Inference for semiparametric models based on generalized estimating equations is attractive because it does not require full model specification for such complex processes, however the resulting estimates can be inefficient (e.g., Fitzmaurice et al., 1993).

Alternative approaches for dealing with longitudinal or clustered data include the use of mixed effects models where covariate effects are specified given a latent subject-specific random effect. The most common approach for inference is to base it on the marginal (joint) distribution obtained by integrating the joint distribution of the data and the random effect, with respect to the random effect. Limitations of this approach include the need to specify the random effect distribution, the fact that covariate effects have a subject-specific interpretation, and the computational challenges associated with calculation of the marginal likelihood typically used for estimation and inference.

Transition models are appealing when scientific interest is directed at how responses change over time (Neuhaus, 1992). In transition models, the probability distribution of the response at a particular time is expressed as a function of an individual's past q responses and covariates. While likelihood-based inferences are straightforward with transition models, a limitation is that the interpretation of covariate effects change as the order q changes, and it may therefore be difficult to interpret and compare models on the same dataset or inferences from models in different datasets.

Azzalini (1994) introduced a Markov chain model which incorporated serial dependence and facilitated expression of covariate effects on marginal features. Heagerty and Zeger (2000) and Heagerty (2002) extended this work to a qth order

marginalized transition model. These models are based on binary data, and do not deal with the more general issue of categorical data which arise in many biomedical studies. The objective here is to describe a general approach for modeling longitudinal categorical data based on a Markov model which accommodates regression modeling on marginal moments as well as on association parameters. Likelihood-based inferences are possible since the model is fully specified, so the resulting estimators are consistent and fully efficient.

The remainder of the chapter is organized as follows. In Section 2.2, we present the details of the model formulation and describe a Fisher-scoring algorithm which can be used for parameter estimation to avoid the need to compute the hessian matrix in the spirit of Kalbfleisch and Lawless (1985). Numerical studies are conducted in Section 2.3 which show that the proposed method works well. Adaptations for handling incomplete data, including the EM algorithm (Dempster et al., 1977), are discussed in Section 2.4. Data from the motivating study called the Waterloo Smoking Prevention Project (Cameron et al., 1999) are analyzed in Section 2.5.

2.2 Model Formulation

2.2.1 Marginal and Conditional Models

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ_i})'$ be a categorical response vector of subject i observed at time $t_1, \ldots, t_{J_i}$, and X_{ij} be the covariate vectors recorded for subject i at the jth time point, $j = 1, \ldots, J_i$, $i = 1, \ldots, n$. Each response component Y_{ij} may take a value from the integers $0, 1, \ldots, K$. Here we may also think of those $K+1$ values as $K+1$ distinct states. Denote $X_i = (X_{i1}', X_{i2}', \ldots, X_{iJ_i}')'$. Let $\mu_{ijk}^M = P(Y_{ij} = k|X_i)$

be the marginal probability that subject i is in state k at the jth assessment given the covariates X_i, $k = 0, \ldots, K$. A regression model may be specified as

$$g\left(\frac{\mu_{ijk}^M}{\mu_{ij0}^M}\right) = X_{ijk}'\beta_k, \quad k = 1, \ldots, K, \tag{2.1}$$

by using μ_{ij0}^M as a reference, where $g(\cdot)$ is a monotone link function, X_{ijk} may be a subset of X_{ij}, featuring the influence of the covariates on the average response in state k at the jth assessment, and β_k is the vector of regression coefficients. We note that an implicit assumption $P(Y_{ij} = k|X_i) = P(Y_{ij} = k|X_{ijk})$ is made here. In practice, $g(\cdot)$ is often chosen as a logarithm function. If Y_{ij} is a binary variable, (2.1) reduces to a standard logistic regression model; when Y_{ij} represents more than two categories, (2.1) allows an analogous interpretation to the odds ratio for binary outcomes. Let $\beta = (\beta_1', \beta_2', \ldots, \beta_K')'$ denote the full vector of regression coefficients.

Sometimes, our interest centers on the dependence of Y_{ij} on its history $H_{ij} = \{Y_{i1}, \ldots, Y_{i,j-1}\}$ and covariates and indeed it is necessary to model this for full model specification. Let $\mu_{ijk}^C = P(Y_{ij} = k|H_{ij}, X_i)$ be the conditional probability, where $k = 0, 1, \ldots, K$. We may employ a regression model to feature the dependence of Y_{ij} on its history and the covariates. That is, specify $\mu_{ijk}^C = h^{-1}(H_{ij}, X_{ijk})$ by a known link function $h(\cdot)$. Again an implicit assumption $P(Y_{ij} = k|H_{ij}, X_i) = P(Y_{ij} = k|H_{ij}, X_{ijk})$ is made here. Typically, we consider a first order dependence of Y_{ij} on its history that is of particular interest in practice. Extensions to any order dependence of Y_{ij} on its history is straightforward though more involved notation may be needed.

In the same spirit of (2.1), we may adopt the following model with μ_{ij0}^C regarded as a reference

$$\log\left(\frac{\mu_{ijk}^C}{\mu_{ij0}^C}\right) = \gamma_{ijk} + \sum_{k'=1}^{K} \gamma_{ijk'k} I(Y_{i,j-1} = k'), \quad k = 1, \ldots, K, \tag{2.2}$$

where $I(\cdot)$ denotes the indicator function. Model (2.2) clearly reflects the dependence of the conditional probability μ_{ijk}^C on the response history. Coefficients γ_{ijk} and $\gamma_{ijk'k}$ have clear interpretations of log odds and log odds ratios, respectively. To be more specific, we have

$$\gamma_{ijk} = \log\frac{P(Y_{ij}=k|Y_{i,j-1}=0,X_i)}{P(Y_{ij}=0|Y_{i,j-1}=0,X_i)},$$

which is the log odds for categorical responses, and the log odds ratio

$$\gamma_{ijk'k} = \log\frac{P(Y_{ij}=k|Y_{i,j-1}=k',X_i)P(Y_{ij}=0|Y_{i,j-1}=0,X_i)}{P(Y_{ij}=0|Y_{i,j-1}=k',X_i)P(Y_{ij}=k|Y_{i,j-1}=0,X_i)},$$

which gives a pairwise association between Y_{ij} and $Y_{i,j-1}$.

The dependence of μ_{ijk}^C on the covariates may be incorporated by assuming regression models for the coefficients $\gamma_{ijk'k}$ in (2.2). Specifically, we consider linear regression models

$$\gamma_{ijk'k} = Z'_{ijk'k}\alpha_{k'k}, \quad k',k=1,\ldots,K, \tag{2.3}$$

though in principle, other regression forms may be adopted as well. Here $\alpha_{k'k}$ is the parameter vector, and $Z_{ijk'k}$ may be subsets of X_{ij}, $k',k=1,\ldots,K$, which feature various types of dependence of Y_{ij} on the covariates. For example, if $Z_{ijk'k}$ simply consists of the unit vector, models (2.2) and (2.3) do not contain any interaction terms between the response history and the covariates. By enlarging $Z_{ijk'k}$ we may include interaction terms. Let $\alpha = (\alpha'_{k'k}, \quad k',k=1,2,\ldots,K)'$ denote the full parameter vector for the conditional models and $\theta = (\beta',\alpha')'$ denote the vector of parameters in both marginal and conditional models (2.1) and (2.2).

If Y_{ij} is a binary response, the proposed models reduce to the marginalized transition models discussed in Heagerty (2002). With binary data, Heagerty (2002)

showed that $\gamma_{ijk}'s$ in (2.2) exist uniquely for any given values of α in (2.3). For categorical data we can obtain the analogous result. That is, given mean model (2.1) and dependence model (2.2) along with regression model (2.3), the intercepts $\gamma_{ijk}'s$ are uniquely determined. Therefore, under constraint (2.4) to be discussed, the proposed models enable us to separate the marginal mean models from the specification of the conditional dependence models. This is very attractive because the interpretation of the regression parameter β does not change when we modify assumptions regarding the conditional dependence models. In contrast, classical transition models focus on modeling the response Y_{ij} conditional on the past response outcomes H_{ij} and covariates. These models may be useful for categorical data which exhibit serial dependence, but the interpretation of the covariate parameters is not resistible to the inclusion of the response history. If the order of the transition model changes, the meaning of the associated parameters would change accordingly.

2.2.2 Estimation and Inference

The likelihood is given by $L(\theta) = \prod_{i=1}^{n} L_i(\theta)$, where

$$\begin{aligned} L_i(\theta) &= P(Y_{i1}, \ldots, Y_{iJ_i}|X_i) = \prod_{j=2}^{J_i} P(Y_{ij}|Y_{i,j-1}, X_i) \cdot P(Y_{i1}|X_i) \\ &= \prod_{j=2}^{J_i}\prod_{k=0}^{K}(\mu_{ijk}^{C})^{I(Y_{ij}=k)} \cdot \prod_{k=0}^{K}(\mu_{i1k}^{M})^{I(Y_{i1}=k)}, \end{aligned}$$

where μ_{i1k}^{M} and μ_{ijk}^{C} are determined from (2.1) and (2.2) respectively. Let $S(\theta) = \sum_{i=1}^{n} S_i(\theta)$ be the score vector, where $S_i(\theta)$ is given by

$$S_i(\theta) = \sum_{j=2}^{J_i}\sum_{k=0}^{K} I(Y_{ij}=k)\frac{1}{\mu_{ijk}^{C}}\frac{\partial \mu_{ijk}^{C}}{\partial \theta} + \sum_{k=0}^{K} I(Y_{i1}=k)\frac{1}{\mu_{i1k}^{M}}\frac{\partial \mu_{i1k}^{M}}{\partial \theta}.$$

To calculate the score function, we need to evaluate the derivatives of the conditional probabilities μ_{ijk}^C and the marginal probabilities μ_{i1k}^M. As these probabilities are constrained by the iterative equation

$$\mu_{ijk}^M = \sum_{k'=0}^{K} P(Y_{ij} = k | Y_{i,j-1} = k'; X_i) \times \mu_{i,j-1,k'}^M, \quad k = 1, \ldots, K, \tag{2.4}$$

the required derivatives may be obtained by differentiating both sides of (2.4). Details for the relevant expressions are provided in Section 2.6.

To solve $S(\theta) = 0$ in order to obtain the estimate $\widehat{\theta}$, we may, in principle, apply the Newton-Raphson algorithm. Let $I(\theta)$ be the observed information matrix constructed from the entire dataset, then the Newton-Raphson algorithm iterates

$$\theta^{(h+1)} = \theta^{(h)} + I^{-1}(\theta^{(h)}) S(\theta^{(h)}), \quad h = 0, 1, \cdots$$

until it converges to $\widehat{\theta}$. This requires the availability of the second derivatives of the log-likelihood. Procedures of calculating the second derivatives are provided in Section 2.6. Under the usual regularity conditions for maximum likelihood estimators, $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, J^{-1}(\theta))$ as the sample size n approaches infinity. Here $J(\theta) = E\left[-\partial S_i(\theta)/\partial\theta\right]$, since the underlying assumption is that the responses of all the subjects are independent and identically distributed conditional on the covariates.

Because the marginal and conditional probabilities μ_{ijk}^M and μ_{ijk}^C are constrained by the iterative equation (2.4), the second derivatives of the log-likelihood are tedious to derive and program. Here we develop a quasi-Newton (or Fisher-scoring) method which eliminates the need for the second derivatives when the covariates are discrete, as in Kalbfleisch and Lawless (1985).

Note that the likelihood can be written as $L(\theta) = \prod_{i=1}^{n} L_i(\theta)$ with

$$L_i(\theta) = \prod_{j=2}^{J_i} \prod_{k',k=0}^{K} \left(\mu^C_{ijk'k}\right)^{I(Y_{ij}=k, Y_{i,j-1}=k')} \cdot \prod_{k=0}^{K} \left(\mu^M_{i1k}\right)^{I(Y_{i1}=k)}$$

and $\mu^C_{ijk'k} = P(Y_{ij} = k | Y_{i,j-1} = k', X_i)$. Accordingly, we obtain the score function

$$\begin{aligned} S_u(\theta) &= \frac{\partial \log L(\theta)}{\partial \theta_u} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=2}^{J_i} \sum_{k',k=0}^{K} \frac{I(Y_{ij}=k, Y_{i,j-1}=k')}{\mu^C_{ijk'k}} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial \theta_u} + \sum_{k=0}^{K} \frac{I(Y_{i1}=k)}{\mu^M_{i1k}} \cdot \frac{\partial \mu^M_{i1k}}{\partial \theta_u} \right\}, \end{aligned}$$

and the second derivative

$$\begin{aligned} \frac{\partial^2 \log L(\theta)}{\partial \theta_u \partial \theta_v} &= \sum_{i=1}^{n} \left\{ - \sum_{j=2}^{J_i} \sum_{k',k=0}^{K} \frac{I(Y_{ij}=k, Y_{i,j-1}=k')}{\left(\mu^C_{ijk'k}\right)^2} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial \theta_u} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial \theta_v} \right. \\ &\quad + \sum_{j=2}^{J_i} \sum_{k',k=0}^{K} \frac{I(Y_{ij}=k, Y_{i,j-1}=k')}{\mu^C_{ijk'k}} \cdot \frac{\partial^2 \mu^C_{ijk'k}}{\partial \theta_u \partial \theta_v} \\ &\quad \left. - \sum_{k=0}^{K} \frac{I(Y_{i1}=k)}{\left(\mu^M_{i1k}\right)^2} \cdot \frac{\partial \mu^M_{i1k}}{\partial \theta_u} \cdot \frac{\partial \mu^M_{i1k}}{\partial \theta_v} + \sum_{k=0}^{K} \frac{I(Y_{i1}=k)}{\mu^M_{i1k}} \cdot \frac{\partial^2 \mu^M_{i1k}}{\partial \theta_u \partial \theta_v} \right\}. \end{aligned}$$

Taking the expectation with respect to the conditional distribution of the response vectors given the covariates, we obtain

$$\begin{aligned} E\left\{ - \frac{\partial^2 \log L(\theta)}{\partial \theta_u \partial \theta_v} \right\} &= \sum_{i=1}^{n} \left\{ \sum_{j=2}^{J_i} \sum_{k',k=0}^{K} \frac{P(Y_{i,j-1}=k'|X_i)}{\mu^C_{ijk'k}} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial \theta_u} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial \theta_v} \right. \\ &\quad \left. + \sum_{k=0}^{K} \frac{1}{\mu^M_{i1k}} \cdot \frac{\partial \mu^M_{i1k}}{\partial \theta_u} \cdot \frac{\partial \mu^M_{i1k}}{\partial \theta_v} \right\}, \end{aligned}$$

by noting that

$$E\left\{I(Y_{ij}=k, Y_{i,j-1}=k')\right\} = P(Y_{ij}=k, Y_{i,j-1}=k'|X_i) = P(Y_{i,j-1}=k'|X_i) \cdot \mu^C_{ijk'k},$$

$$E\{I(Y_{i1}=k)\} = P(Y_{i1}=k|X_i) = \mu^M_{i1k},$$

$$\sum_{k=0}^{K} \mu^C_{ijk'k} = 1$$

and

$$\sum_{k=0}^{K} \mu_{i1k}^{M} = 1.$$

This expectation can be estimated by

$$M_{uv}(\theta) = \sum_{i=1}^{n} \left\{ \sum_{j=2}^{J_i} \sum_{k',k=0}^{K} \frac{p_{i,j-1,k'}}{\mu_{ijk'k}^{C}} \cdot \frac{\partial \mu_{ijk'k}^{C}}{\partial \theta_u} \cdot \frac{\partial \mu_{ijk'k}^{C}}{\partial \theta_v} + \sum_{k=0}^{K} \frac{1}{\mu_{i1k}^{M}} \cdot \frac{\partial \mu_{i1k}^{M}}{\partial \theta_u} \cdot \frac{\partial \mu_{i1k}^{M}}{\partial \theta_v} \right\},$$

where $p_{i,j-1,k'}$ is the proportion of the subjects with covariate x_i in state k' at the $(j-1)$th time point. That is, if $N_{i,j-1,k'} = \sum_{i=1}^{n} I(Y_{i,j-1} = k', X_i = x_i)$ is the total number of subjects with covariate vector $X_i = x_i$ and with response $Y_{i,j-1} = k'$, and $N_{i,j-1} = \sum_{k'=0}^{K} N_{i,j-1,k'}$ is the total number of subjects with covariate vector $X_i = x_i$, then $p_{i,j-1,k'} = N_{i,j-1,k'}/N_{i,j-1}$.

Let $S(\theta)$ be the vector of $(S_u(\theta))$, and $M(\theta)$ be the matrix $[M_{uv}(\theta)]$. Then an updated estimate is obtained as

$$\theta^{(h+1)} = \theta^{(h)} + M(\theta^{(h)})^{-1} S(\theta^{(h)}), \quad h = 0, 1, \cdots, \tag{2.5}$$

where $M(\theta^{(h)})$ is assumed nonsingular. The iteration is cycled through until convergence of $\theta^{(h+1)}$. Let $\widehat{\theta}$ denote the corresponding limit.

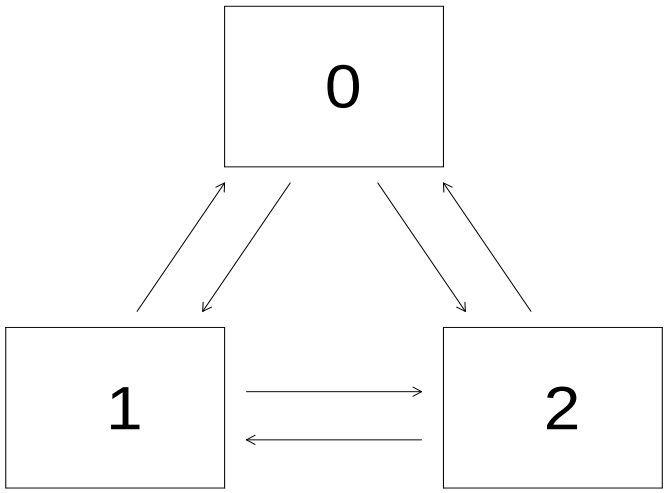
Under the usual regularity conditions for maximum likelihood estimators, $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma^{-1})$ as the sample size n approaches infinity. Here

$$\Sigma = E\left[-\partial^2 \log L_i(\theta)/\partial\theta\partial\theta'\right],$$

and it can be estimated by the observed information matrix given in Section 2.6 or by the expected information matrix given here.

Here we comment that the mean parameter β and the association parameter α are orthogonal if Y_{ij} are binary responses. This property is established by Azzalini

Figure 2.1: Three states diagram of the transitions



(1994) for a restrictive scenario. Heagerty (2002) proved the properties by using the log-linear parametrization of the first order marginalized transition model for binary data. In Section 2.7 we establish an analogous property for categorical data. Finally, we note that model checking can be conducted directly through score tests or likelihood ratio tests as the proposed method is likelihood-based.

2.3 Numerical Studies

2.3.1 Performance of the Proposed Method

In this subsection we conduct simulation studies to assess the performance of the proposed method. We set $n = 500$, $K = 2$ to give three categories and $J_i = 4$ to give four timepoints, for $i = 1, \ldots, n$. Figure 2.1 illustrates the state diagram. Two thousand simulations are run for each parameter configuration.

The response vector is generated from the marginal model

$$\log\left(\frac{\mu_{ijk}^M}{\mu_{ij0}^M}\right) = \beta_{k0} + \beta_{k1}X_{ij1} + \beta_{k2}X_{ij2}, \quad k = 1, 2, \tag{2.6}$$

in combination with the conditional model

$$\log\left(\frac{\mu_{ijk}^C}{\mu_{ij0}^C}\right) = \gamma_{ijk} + \gamma_{ij1k}I(Y_{i,j-1} = 1) + \gamma_{ij2k}I(Y_{i,j-1} = 2), \quad k = 1, 2, \tag{2.7}$$

where $\gamma_{ijk'k}$ are specified as

$$\gamma_{ijk'k} = \alpha_{k'k0} + \alpha_{k'k1}X_{ij2}, \quad k', k = 1, 2. \tag{2.8}$$

Here $X_{ij1} = X_{i1}$ represents the treatment status, generated from the binomial distribution $Bin(1, 0.5)$, and X_{ij2} is specified as $I(j = 3 \text{ or } 4)$, indicating a temporal effect. Let $\beta_k = (\beta_{k0}, \beta_{k1}, \beta_{k2})'$ for $k = 1, 2$, and $X_{ij} = (1, X_{ij1}, X_{ij2})'$ for $j = 1, \ldots, J$. Set $\beta_1 = (-\log(3), \log(0.8), \log(1.2))'$, $\beta_2 = (-\log(3), \log(0.6), \log(1.5))'$, $\alpha_{110} = \log(1.2), \alpha_{210} = \log(1.1), \alpha_{120} = \log(1.5), \alpha_{220} = \log(2.0)$, $\alpha_{111} = \log(1.5), \alpha_{211} = \log(1), \alpha_{121} = \log(1.5)$, and $\alpha_{221} = \log(1.5)$.

Data are generated as follows. Given the covariate vector $X_i = (X_{i1}', X_{i2}', \ldots, X_{iJ}')'$, with parameter θ specified as above, Y_{i1} is generated from a multinomial distribution with probabilities

$$\begin{aligned}
\mu_{i10}^M &= P(Y_{i1} = 0|X_i) = \frac{1}{1 + e^{X_{i1}'\beta_1} + e^{X_{i1}'\beta_2}}, \\
\mu_{i11}^M &= P(Y_{i1} = 1|X_i) = \frac{e^{X_{i1}'\beta_1}}{1 + e^{X_{i1}'\beta_1} + e^{X_{i1}'\beta_2}}, \\
\mu_{i12}^M &= P(Y_{i1} = 2|X_i) = \frac{e^{X_{i1}'\beta_2}}{1 + e^{X_{i1}'\beta_1} + e^{X_{i1}'\beta_2}},
\end{aligned}$$

based on (2.6). Given $Y_{i,j-1}$ then, Y_{ij} is generated from a multinomial distribution

with conditional probabilities

$$\begin{aligned}
\mu_{ij0}^C &= P(Y_{ij}=0|Y_{i,j-1},X_i) \\
&= \frac{1}{1+e^{\gamma_{ij1}+\gamma_{ij11}I(Y_{i,j-1}=1)+\gamma_{ij21}I(Y_{i,j-1}=2)}+e^{\gamma_{ij2}+\gamma_{ij12}I(Y_{i,j-1}=1)+\gamma_{ij22}I(Y_{i,j-1}=2)}},
\end{aligned}$$

$$\begin{aligned}
\mu_{ij1}^C &= P(Y_{ij}=1|Y_{i,j-1},X_i) \\
&= \frac{e^{\gamma_{ij1}+\gamma_{ij11}I(Y_{i,j-1}=1)+\gamma_{ij21}I(Y_{i,j-1}=2)}}{1+e^{\gamma_{ij1}+\gamma_{ij11}I(Y_{i,j-1}=1)+\gamma_{ij21}I(Y_{i,j-1}=2)}+e^{\gamma_{ij2}+\gamma_{ij12}I(Y_{i,j-1}=1)+\gamma_{ij22}I(Y_{i,j-1}=2)}},
\end{aligned}$$

$$\begin{aligned}
\mu_{ij2}^C &= P(Y_{ij}=2|Y_{i,j-1},X_i) \\
&= \frac{e^{\gamma_{ij2}+\gamma_{ij12}I(Y_{i,j-1}=1)+\gamma_{ij22}I(Y_{i,j-1}=2)}}{1+e^{\gamma_{ij1}+\gamma_{ij11}I(Y_{i,j-1}=1)+\gamma_{ij21}I(Y_{i,j-1}=2)}+e^{\gamma_{ij2}+\gamma_{ij12}I(Y_{i,j-1}=1)+\gamma_{ij22}I(Y_{i,j-1}=2)}}.
\end{aligned}$$

To determine these probabilities, we must first determine the values of γ_{ij1} and γ_{ij2} that satisfy the equations

$$\mu_{ijk}^M = P(Y_{ij}=k|X_i)=\sum_{k'=0}^{2}P(Y_{ij}=k|Y_{i,j-1}=k',X_i)\mu_{i,j-1,k'}^M \quad k=1,2.$$

That is, we must solve the nonlinear equations

$$\begin{aligned}
\mu_{ij1}^M &= \frac{e^{\gamma_{ij1}}}{1+e^{\gamma_{ij1}}+e^{\gamma_{ij2}}}\mu_{i,j-1,0}^M+\frac{e^{\gamma_{ij1}+\gamma_{ij11}}}{1+e^{\gamma_{ij1}+\gamma_{ij11}}+e^{\gamma_{ij2}+\gamma_{ij12}}}\mu_{i,j-1,1}^M \\
&\quad +\frac{e^{\gamma_{ij1}+\gamma_{ij21}}}{1+e^{\gamma_{ij1}+\gamma_{ij21}}+e^{\gamma_{ij2}+\gamma_{ij22}}}\mu_{i,j-1,2}^M \\
\mu_{ij2}^M &= \frac{e^{\gamma_{ij2}}}{1+e^{\gamma_{ij1}}+e^{\gamma_{ij2}}}\mu_{i,j-1,0}^M+\frac{e^{\gamma_{ij2}+\gamma_{ij12}}}{1+e^{\gamma_{ij1}+\gamma_{ij11}}+e^{\gamma_{ij2}+\gamma_{ij12}}}\mu_{i,j-1,1}^M \\
&\quad +\frac{e^{\gamma_{ij2}+\gamma_{ij22}}}{1+e^{\gamma_{ij1}+\gamma_{ij21}}+e^{\gamma_{ij2}+\gamma_{ij22}}}\mu_{i,j-1,2}^M
\end{aligned}$$

for γ_{ij1} and γ_{ij2}, for $i=1,2,\ldots,n$ and $j=2,3,4$. Here μ_{ijk}^M is determined by (2.6), and $\gamma_{ijk'k}{}'s$ are given by (2.8). Explicit expressions are typically not available and so numerical methods must be employed to obtain solutions. Specifically, here we use the software R in the numerical implementation. The detailed expression of the first and second derivatives of the log likelihood are included in Appendix A.

In this simulation, we consider the case with *i*) no covariates, *ii*) only a time dependent covariate, and *iii*) the case with both a time dependent and a time independent (treatment) covariate for the marginal model. Both the Newton-Raphson and Fisher-scoring methods are employed. The results are reported in Table 2.1, where ASE represents the average of the standard errors, ESE refers to the empirical standard error, and CP denotes the coverage probability for 95% confidence intervals. For the Fisher-scoring algorithm the confidence intervals were computed using the expected information matrix. It is seen that the two methods give very similar results. The biases for both β and α under the three scenarios are fairly small, indicating that the estimators are consistent. The model-based standard errors agree with the empirical standard errors reasonably well. It is not surprising that standard errors for the estimates of mean parameters $\beta_{k'k}$ are smaller than those for the estimates of association parameters $\alpha_{k'kr}$. The coverage probabilities for all the parameters are in good agreement with the nominal level of 95%, suggesting that the variance estimates obtained from the proposed method are valid.

2.3.2 Comparison of the Proposed Method and GEE

In this subsection we further evaluate the performance of the proposed method, compared to the GEE approach, by focusing on the estimates of the marginal mean parameter β. We consider two scenarios – correct model specification and model misspecification. In particular, with model misspecification we examine the performance of the proposed method only when the conditional model μ^C_{ijk} is misspecified.

For the first scenario that both marginal and conditional models are correctly assumed, we use the same settings as those in Section 2.3.1. That is, a first order

Table 2.1: Simulation results under the three scenarios using Newton-Raphson and Fisher-scoring method

			No Covariate				Trend				Treatment and Trend			
Parameter	Value	Method†	*BIAS	ASE	ESE	CP	*BIAS	ASE	ESE	CP	*BIAS	ASE	ESE	CP
Marginal Model:														
β_{10}	-log(3)	NR	-0.001	0.060	0.061	0.951	-0.002	0.080	0.081	0.949	-0.004	0.104	0.104	0.953
		FS	-0.001	0.061	0.060	0.949	-0.002	0.082	0.080	0.948	-0.005	0.104	0.103	0.951
β_{11}	log(0.8)	NR									-0.005	0.124	0.124	0.954
		FS									-0.004	0.123	0.122	0.952
β_{12}	log(1.2)	NR					0.002	0.116	0.118	0.952	-0.000	0.115	0.115	0.950
		FS					0.002	0.117	0.117	0.953	-0.001	0.114	0.114	0.948
β_{20}	-log(3)	NR	-0.000	0.064	0.064	0.951	-0.006	0.086	0.086	0.951	-0.005	0.105	0.105	0.950
		FS	-0.000	0.065	0.064	0.952	-0.004	0.085	0.086	0.951	-0.006	0.103	0.104	0.951
β_{21}	log(0.6)	NR									-0.000	0.122	0.124	0.950
		FS									-0.001	0.123	0.124	0.952
β_{22}	log(1.5)	NR					0.002	0.113	0.112	0.951	0.002	0.118	0.120	0.952
		FS					0.002	0.112	0.112	0.949	0.002	0.118	0.119	0.953
Association Model:														
α_{110}	log(1.2)	NR	-0.005	0.297	0.297	0.955	-0.007	0.300	0.301	0.950	-0.006	0.326	0.327	0.952
		FS	-0.005	0.302	0.296	0.952	-0.005	0.302	0.300	0.951	-0.005	0.327	0.326	0.951
α_{111}	log(1.5)	NR	0.011	0.361	0.362	0.951	0.021	0.401	0.400	0.953	0.003	0.444	0.445	0.954
		FS	0.013	0.360	0.363	0.950	0.018	0.402	0.401	0.952	0.002	0.450	0.443	0.949
α_{210}	log(1.1)	NR	-0.005	0.313	0.313	0.950	-0.005	0.308	0.307	0.952	-0.004	0.313	0.315	0.954
		FS	-0.005	0.314	0.311	0.952	-0.006	0.310	0.310	0.950	-0.003	0.315	0.312	0.952
α_{211}	log(1.0)	NR	0.016	0.382	0.383	0.949	0.015	0.415	0.414	0.953	0.003	0.422	0.423	0.953
		FS	0.013	0.384	0.384	0.948	0.016	0.412	0.413	0.949	0.006	0.420	0.422	0.949
α_{120}	log(1.5)	NR	-0.001	0.300	0.301	0.952	-0.000	0.297	0.298	0.952	-0.002	0.301	0.301	0.952
		FS	-0.001	0.305	0.301	0.951	-0.002	0.301	0.299	0.950	-0.002	0.302	0.304	0.950
α_{121}	log(1.5)	NR	0.000	0.368	0.368	0.955	0.001	0.363	0.360	0.949	0.004	0.366	0.366	0.953
		FS	0.001	0.367	0.368	0.953	0.001	0.361	0.362	0.947	0.002	0.368	0.366	0.948
α_{220}	log(2.0)	NR	-0.009	0.283	0.282	0.952	-0.008	0.280	0.281	0.949	-0.005	0.282	0.281	0.951
		FS	-0.010	0.281	0.281	0.951	-0.009	0.282	0.282	0.947	-0.007	0.282	0.279	0.949
α_{221}	log(1.5)	NR	0.006	0.345	0.344	0.952	0.007	0.325	0.325	0.952	0.009	0.326	0.327	0.948
		FS	0.007	0.342	0.343	0.955	0.008	0.325	0.324	0.948	0.008	0.324	0.328	0.947

† "NR" represents the Newton-Raphson algorithm, while "FS" denotes the Fisher-scoring method.
* Absolute bias.

dependence is considered as specified in (2.7) and (2.8). With the GEE approach we adopt the working independence correlation matrix as it has been shown to provide fairly efficient estimates in many settings (e.g., Sutradhar and Das, 1999). We particularly fit three marginal models with no covariates, only time trend and both treatment and time trend included, respectively.

Table 2.2 shows the empirical bias (BIAS), the average standard error (ASE), the empirical standard error (ESE) and the empirical coverage probability (CP) of 95% confidence intervals for 2000 samples. For the GEE model the ASE is the average of 2000 robust standard errors based on the sandwich variance formula and for the proposed model the ASE is the average of the 2000 standard errors based on the Fisher information matrix. It is seen that both methods give reasonably comparable estimates with very small finite sample biases. Both methods yield reasonable standard errors as the model based standard errors (ASE) agree very well with the empirical standard errors (ESE). However, the proposed method seems to be more efficient than the GEE method since it tends to produce smaller standard errors and better coverage probabilities.

Now we investigate the performance of the proposed model by examining its sensitivity to the model misspecification. We consider the same marginal model as that in Section 3.1, but use a second order dependence model of the form

$$\log\left(\frac{\mu^C_{ijk}}{\mu^C_{ij0}}\right) = \gamma_{ijk} + \sum_{l=1}^{2}\sum_{k'=1}^{2}\gamma_{ijlk'k}I(Y_{i,j-l}=k'), \quad j=3,4,\ k=1,2,$$

$$\gamma_{ijlk'k} = \alpha_{lk'k0} + \alpha_{lk'k1}X_{ij2}, \quad k,k'=1,2, \quad l=1,2,$$

to accommodate serial correlation for $j = 3$ and 4. Here $\mu^C_{ijk} = P(Y_{ij} = k|Y_{i,j-1}, Y_{i,j-2}, X_i)$ for $j = 3, 4$, are the second order dependence probabilities. When $j = 2$, we assume a first order dependence model for Y_{i2} through the models (2.7) and (2.8) and

take the same parameter values as in Section 2.3.1. For the α parameters in the second order dependence model, we set $\alpha_{lk'k0} = \log(1+((k-1)\cdot 2^2+(k'-1)\cdot 2+(l-1))/8)$ and $\alpha_{lk'k1} = \log(2 + ((k-1)\cdot 2^2 + (k'-1)\cdot 2 + (l-1))/8)$ to feature a weak dependence, and $\alpha_{lk'k0} = \log(5 + ((k-1)\cdot 2^2 + (k'-1)\cdot 2 + (l-1))/8)$ and $\alpha_{lk'k1} = \log(6 + ((k-1)\cdot 2^2 + (k'-1)\cdot 2 + (l-1))/8)$ to represent a strong dependence, $k, k', l = 1, 2$.

Table 2.3 shows that with a weak dependence among $Y'_{ij}s$, both methods produce very small finite sample biases, the model based standard errors (ASE) agree very well with the empirical standard errors (ESE), and the standard errors from the two methods are fairly comparable. The proposed method seems to provide better coverage probabilities than the GEE approach. When the strength of dependence is large, however, it is evident from Table 2.4 that the proposed method may fail to perform satisfactorily. The finite sample biases can be substantial, the standard errors may be inflated, and hence the coverage probabilities deviate from the nominal value considerably.

2.4 Inference with Missing Data

In practice, missing observations arise commonly. This is also the case of the motivating example to be analyzed in Section 2.5. In this section we develop inference methods to handle missing observations. Specifically, we discuss methods based on the observed likelihood and the expectation-maximization (EM) algorithm. Throughout we assume data are missing at random (MAR) (Diggle et al., 2002).

Table 2.2: Comparison of the frequency properties of estimators of regression coefficients by the proposed method and GEE method with correctly specified models

	True	GEE				Proposed Method			
Parameters	Value	†BIAS%	ASE	ESE	CP	†BIAS%	ASE	ESE	CP
No Covariate									
β_{10}	-log(3)	0.00	0.062	0.061	0.947	-0.09	0.061	0.060	0.949
β_{20}	-log(3)	-0.18	0.065	0.065	0.933	-0.00	0.065	0.064	0.952
Trend									
β_{10}	-log(3)	0.09	0.082	0.082	0.955	-0.18	0.080	0.080	0.948
β_{12}	log(1.2)	1.64	0.115	0.116	0.951	1.10	0.113	0.112	0.953
β_{20}	-log(3)	-0.09	0.089	0.089	0.929	-0.36	0.085	0.086	0.951
β_{22}	log(1.5)	0.25	0.116	0.117	0.934	0.49	0.112	0.112	0.949
Treatment and Trend									
β_{10}	-log(3)	0.00	0.106	0.105	0.945	-0.45	0.104	0.103	0.951
β_{11}	log(0.8)	-1.34	0.122	0.123	0.946	-1.79	0.123	0.122	0.952
β_{12}	log(1.2)	1.10	0.118	0.120	0.936	-0.55	0.114	0.114	0.948
β_{20}	-log(3)	-0.55	0.109	0.110	0.933	-0.56	0.103	0.104	0.951
β_{21}	log(0.6)	0.39	0.128	0.128	0.939	-0.20	0.123	0.124	0.952
β_{22}	log(1.5)	1.23	0.125	0.126	0.953	0.49	0.118	0.119	0.953

† Percent relative bias $(\hat{\beta} - \beta_{\text{true}})/\beta_{\text{true}} \times 100$.

Table 2.3: Comparison of the frequency properties of estimators of regression coefficients by the proposed method and GEE method with misspecified conditional model: weak dependence

	True	GEE				Proposed Method			
Parameters	Value	†BIAS%	ASE	ESE	CP	†BIAS%	ASE	ESE	CP
β_{10}	-log(3)	-0.18	0.107	0.108	0.939	-0.64	0.106	0.107	0.959
β_{11}	log(0.8)	0.90	0.140	0.141	0.938	3.13	0.140	0.140	0.957
β_{12}	log(1.2)	0.00	0.133	0.132	0.958	5.48	0.143	0.145	0.953
β_{20}	-log(3)	-0.46	0.115	0.115	0.928	-0.64	0.112	0.113	0.945
β_{21}	log(0.6)	1.19	0.155	0.155	0.923	-0.59	0.147	0.149	0.947
β_{22}	log(1.5)	-1.73	0.134	0.133	0.966	1.97	0.137	0.136	0.965

† Percent relative bias $(\hat{\beta} - \beta_{\text{true}})/\beta_{\text{true}} \times 100$.

Table 2.4: Comparison of the frequency properties of estimators of regression coefficients by the proposed method and GEE method with misspecified conditional model: strong dependence

	True	GEE				Proposed Method			
Parameters	Value	†BIAS%	ASE	ESE	CP	†BIAS%	ASE	ESE	CP
β_{10}	-log(3)	-0.36	0.114	0.115	0.942	-10.56	0.135	0.137	0.806
β_{11}	log(0.8)	-0.45	0.155	0.153	0.919	0.90	0.160	0.164	0.912
β_{12}	log(1.2)	2.19	0.131	0.130	0.966	57.59	0.166	0.168	0.866
β_{20}	-log(3)	-0.72	0.116	0.116	0.926	-15.84	0.175	0.178	0.668
β_{21}	log(0.6)	0.00	0.164	0.162	0.917	-10.76	0.170	0.169	0.886
β_{22}	log(1.5)	1.80	0.131	0.130	0.964	38.47	0.188	0.190	0.730

† Percent relative bias $(\hat{\beta} - \beta_{\text{true}})/\beta_{\text{true}} \times 100$.

2.4.1 A Scoring Method

In this subsection we describe a Fisher-scoring method to handle incomplete data. For a given q, let $\mu_{ijk'k}^{C(q)}$ be the q-step transition probability from state k' at the $(j-q)$th time point to state k at the jth time point, i.e.

$$\begin{aligned}
\mu_{ijk'k}^{C(q)} &= P(Y_{ij}=k|Y_{i,j-q}=k', X_i) \\
&= \sum_{l_1=0}^{K}\cdots\sum_{l_{q-1}=0}^{K} P(Y_{ij}=k, Y_{i,j-1}=l_1,\cdots,Y_{i,j-q+1}=l_{q-1}|Y_{i,j-q}=k', X_i) \\
&= \sum_{l_1=0}^{K}\cdots\sum_{l_{q-1}=0}^{K}\left(\prod_{m=1}^{q}\mu_{i,j-q+m,l_{m-1}l_m}^{C}\right)
\end{aligned}$$

where we denote $l_q=k$ and $l_0=k'$.

Let $j_0<j_1<\cdots<j_{m_i}$ be the ordered observed assessment points for subject i, between two consecutive time points. Here $j_0=1$. The likelihood can be written as

$$L(\theta)=\prod_{i=1}^{n}\left\{\prod_{m=1}^{m_i}\prod_{k',k=0}^{K}\left(\mu_{ij_mk'k}^{C(j_m-j_{m-1})}\right)^{I(Y_{ij_{m-1}}=k',Y_{ij_m}=k)}\cdot\prod_{k=0}^{K}\left(\mu_{i1k}^{M}\right)^{I(Y_{i1}=k)}\right\},$$

leading to the score function

$$\begin{aligned}
S_u(\theta) &= \sum_{i=1}^{n}\left\{\sum_{m=1}^{m_i}\sum_{k',k=0}^{K}\frac{I(Y_{ij_{m-1}}=k', Y_{ij_m}=k)}{\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_u}\right. \\
&\quad \left.+\sum_{k=0}^{K}\frac{I(Y_{i1}=k)}{\mu_{i1k}^{M}}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_u}\right\},
\end{aligned}$$

and the second derivative

$$\begin{aligned}\frac{\partial^2\ell(\theta)}{\partial\theta_u\partial\theta_v} &= \sum_{i=1}^{n}\Bigg\{-\sum_{m=1}^{m_i}\sum_{k',k=0}^{K}\frac{I(Y_{ij_{m-1}}=k',Y_{ij_m}=k)}{\left(\mu_{ij_mk'k}^{C(j_m-j_{m-1})}\right)^2}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_u}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_v}\\ &\quad+\sum_{m=1}^{m_i}\sum_{k',k=0}^{K}\frac{I(Y_{ij_{m-1}}=k',Y_{ij_m}=k)}{\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}\cdot\frac{\partial^2\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_u\partial\theta_v}\\ &\quad-\sum_{k=0}^{K}\frac{I(Y_{i1}=k)}{\left(\mu_{i1k}^{M}\right)^2}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_u}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_v}+\sum_{k=0}^{K}\frac{I(Y_{i1}=k)}{\mu_{i1k}^{M}}\cdot\frac{\partial^2\mu_{i1k}^{M}}{\partial\theta_u\partial\theta_v}\Bigg\},\end{aligned}$$

where $\ell(\theta)$ is the log-likelihood and

$$\frac{\partial\mu_{ijk'k}^{C(q)}}{\partial\theta}=\sum_{l_1=0}^{K}\cdots\sum_{l_{q-1}=0}^{K}\left(\sum_{r=1}^{q}\prod_{m=1,m\neq r}^{q}\mu_{i,j-q+m,l_{m-1}l_m}^{C}\cdot\frac{\partial\mu_{i,j-q+r,l_{r-1}l_r}^{C}}{\partial\theta}\right).$$

Taking expectation yields

$$\begin{aligned}E\left[-\frac{\partial^2\ell(\theta)}{\partial\theta_u\partial\theta_v}\right] &= \sum_{i=1}^{n}\sum_{m=1}^{m_i}\sum_{k',k=0}^{K}\frac{P(Y_{ij_{m-1}}=k')}{\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_u}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_v}\\ &\quad+\sum_{i=1}^{n}\sum_{k=0}^{K}\frac{1}{\mu_{i1k}^{M}}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_u}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_v},\end{aligned}$$

due to the fact that

$$\begin{aligned}E[I(Y_{ij_{m-1}}=k',Y_{ij_m}=k)] &= P(Y_{ij_{m-1}}=k',Y_{ij_m}=k|X_i)\\ &= P(Y_{ij_{m-1}}=k'|X_i)\cdot\mu_{ij_mk'k}^{C(j_m-j_{m-1})},\end{aligned}$$

$$E[Y_{i1}=k]=P(Y_{i1}=k|X_i)=\mu_{i1k}^{M},$$

$$\sum_{k=0}^{K}\mu_{ij_mk'k}^{C(j_m-j_{m-1})}=1$$

and

$$\sum_{k=0}^{K}\mu_{i1k}^{M}=1.$$

Applying the same argument as that in Section 2.2, if the covariates X_i are discrete, $E\left[-\partial^2\ell(\theta)/\partial\theta_u\partial\theta_v\right]$ can be estimated by

$$\begin{aligned} M_{uv}(\theta) &= \sum_{i=1}^{n}\left\{\sum_{m=1}^{m_i}\sum_{k',k=0}^{K}\frac{p_{ij_{m-1}k'}}{\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_u}\cdot\frac{\partial\mu_{ij_mk'k}^{C(j_m-j_{m-1})}}{\partial\theta_v}\right. \\ &\quad \left.+\sum_{k=0}^{K}\frac{1}{\mu_{i1k}^{M}}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_u}\cdot\frac{\partial\mu_{i1k}^{M}}{\partial\theta_v}\right\}, \end{aligned}$$

where $p_{ij_{m-1}k'}$ is the proportion of the subjects with covariate x_i in state k' at the (j_{m-1})th time point. Therefore, the Fisher-scoring method in Section 2.2 may be employed here to obtain the estimate $\widehat{\theta}$. Analogously, we can get the estimate of the asymptotic covariance matrix using the observed information matrix or the expected information given here.

2.4.2 An EM Algorithm

When the missing data proportion is relatively small, the Fisher-scoring method described above works well. However, if the missing proportion is large, the Fisher-scoring method may become computationally burdensome since we need to calculate the q step transition probabilities and their derivatives which will be very computationally intensive. In this subsection we describe an alternative method based on the EM algorithm. The complete data likelihood for subject i is given by

$$L_i(\theta, y_i) = \prod_{j=2}^{J_i}\prod_{k',k=0}^{K}\left(\mu_{ijk'k}^{C}\right)^{I(Y_{ij}=k,Y_{i,j-1}=k')}\cdot\prod_{k=0}^{K}\left(\mu_{i1k}^{M}\right)^{I(Y_{i1}=k)},$$

leading to the complete data log-likelihood

$$\ell_i(\theta, y_i) = \sum_{j=2}^{J_i}\sum_{k',k=0}^{K}I(Y_{ij}=k, Y_{i,j-1}=k')\cdot\log\left(\mu_{ijk'k}^{C}\right)+\sum_{k=0}^{K}I(Y_{i1}=k)\cdot\log\left(\mu_{i1k}^{M}\right).$$

In the E step we construct the conditional expectation

$$Q(\theta;\theta^{(h)}) \;=\; \sum_{i=1}^{n} Q_i(\theta;\theta^{(h)}),$$

where y_i is written as $(y_i^{(m)}, y_i^{(o)})$ to explicitly indicate missing and observed components, $Q_i(\theta;\theta^{(h)}) = E[\ell_i(\theta, y_i)|Y_i^{(o)}, \theta^{(h)}] = \sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \ell_i(\theta, y_i)$, and

$$w_i(y_i;\theta^{(h)}) = \frac{L_i(\theta^{(h)}; y_i^{(m)}, y_i^{(o)})}{\sum_{y_i^{(m)}} L_i(\theta^{(h)}; Y_i^{(m)} = y_i^{(m)}, y_i^{(o)})},$$

may be viewed as a weight.

To maximize $Q(\theta;\theta^{(h)})$, we may use the Newton-Raphson method in the same spirit of Section 2.2. Note that the weighted score function is

$$\begin{aligned} S_u(\theta;\theta^{(h)}) &= \frac{\partial Q(\theta;\theta^{(h)})}{\partial\theta_u} \\ &= \sum_{i=1}^{n}\sum_{j=2}^{J_i}\sum_{k',k=0}^{K}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \frac{I(Y_{ij}=k, Y_{i,j-1}=k')}{\mu^C_{ijk'k}} \cdot \frac{\partial\mu^C_{ijk'k}}{\partial\theta_u} \\ &\quad + \sum_{i=1}^{n}\sum_{k=0}^{K} \frac{I(Y_{i1}=k)}{\mu^M_{i1k}} \cdot \frac{\partial\mu^M_{i1k}}{\partial\theta_u} \end{aligned}$$

and the second derivative is

$$\begin{aligned} I_{uv}(\theta;\theta^{(h)}) &= -\frac{\partial^2 Q(\theta;\theta^{(h)})}{\partial\theta_u\partial\theta_v} \\ &= \sum_{i=1}^{n}\sum_{j=2}^{J_i}\sum_{k',k=0}^{K}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \frac{I(Y_{ij}=k, Y_{i,j-1}=k')}{\left(\mu^C_{ijk'k}\right)^2} \cdot \frac{\partial\mu^C_{ijk'k}}{\partial\theta_u} \cdot \frac{\partial\mu^C_{ijk'k}}{\partial\theta_v} \\ &\quad - \sum_{i=1}^{n}\sum_{j=2}^{J_i}\sum_{k',k=0}^{K}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \frac{I(Y_{ij}=k, Y_{i,j-1}=k')}{\mu^C_{ijk'k}} \cdot \frac{\partial^2\mu^C_{ijk'k}}{\partial\theta_u\partial\theta_v} \\ &\quad + \sum_{i=1}^{n}\sum_{k=0}^{K} \frac{I(Y_{i1}=k)}{\left(\mu^M_{i1k}\right)^2} \cdot \frac{\partial\mu^M_{i1k}}{\partial\theta_u} \cdot \frac{\partial\mu^M_{i1k}}{\partial\theta_v} - \sum_{i=1}^{n}\sum_{k=0}^{K} \frac{I(Y_{i1}=k)}{\mu^M_{i1k}} \cdot \frac{\partial^2\mu^M_{i1k}}{\partial\theta_u\partial\theta_v}. \end{aligned}$$

Let $S(\theta;\theta^{(h)})$ be the score vector $(S_u(\theta;\theta^{(h)}))$ and $I(\theta;\theta^{(h)})$ be the matrix $[I_{uv}(\theta;\theta^{(h)})]$, then given the initial value $\theta^{(h,0)}$, the Newton-Raphson method involves iterating

as

$$\theta^{(h,\ell+1)} = \theta^{(h,\ell)} + I^{-1}(\theta^{(h,\ell)}; \theta^{(h,\ell)}) S(\theta^{(h,\ell)}; \theta^{(h,\ell)})$$

until convergence is achieved at the estimate, say $\theta^{(h+1)}$. We then replace $\theta^{(h)}$ by $\theta^{(h+1)}$, and iterate again using the E and M steps, until convergence is achieved at the estimate $\widehat{\theta}$.

Alternatively, we may employ the Fisher-scoring method discussed in Section 2.2. Specifically, the expectation of the second derivative is

$$\begin{aligned} & E\left\{-\frac{\partial^2 Q(\theta;\theta^{(h)})}{\partial\theta_u \partial\theta_v}\right\} \\ = & \sum_{i=1}^{n}\left\{\sum_{j=2}^{J_i}\sum_{k',k=0}^{K}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \frac{P(Y_{i,j-1}=k'|X_i)}{\mu^C_{ijk'k}} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial\theta_u} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial\theta_v} \right. \\ & \left. + \sum_{k=0}^{K} \frac{1}{\mu^M_{i1k}} \cdot \frac{\partial \mu^M_{i1k}}{\partial\theta_u} \cdot \frac{\partial \mu^M_{i1k}}{\partial\theta_v}\right\}, \end{aligned}$$

where the expectation is taken with respect to $L_i(\theta^{(h)}; y_i)$. Again, this expectation can be estimated by

$$\begin{aligned} M_{uv}(\theta;\theta^{(h)}) = & \sum_{i=1}^{n}\left\{\sum_{j=2}^{J_i}\sum_{k',k=0}^{K}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \frac{p_{i,j-1,k'}}{\mu^C_{ijk'k}} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial\theta_u} \cdot \frac{\partial \mu^C_{ijk'k}}{\partial\theta_v} \right. \\ & \left. + \sum_{k=0}^{K} \frac{1}{\mu^M_{i1k}} \cdot \frac{\partial \mu^M_{i1k}}{\partial\theta_u} \cdot \frac{\partial \mu^M_{i1k}}{\partial\theta_v}\right\}, \end{aligned}$$

where $p_{i,j-1,k'}$ is the proportion of the subjects in state k' at the $(j-1)$th time point with $X_i = x_i$. The iterative equation (2.5) can be used to obtain the estimate $\theta^{(h+1)}$. That is, we replace $\theta^{(h)}$ by $\theta^{(h+1)}$ and iterate between the E and M steps until the estimates converge to the estimate $\widehat{\theta}$.

To obtain the variance estimates for the estimator $\widehat{\theta}$, we may apply the Louis's

method (Louis, 1982). That is, let

$$\Sigma(\widehat{\theta}) = M(\widehat{\theta};\widehat{\theta}) - \sum_{i=1}^{n}\sum_{y_i^{(m)}} w_i(y_i;\widehat{\theta}) \left(\frac{\partial \ell_i(\widehat{\theta})}{\partial \theta}\right)\left(\frac{\partial \ell_i(\widehat{\theta})}{\partial \theta}\right)' + \sum_{i=1}^{n}\left(\frac{\partial Q_i(\widehat{\theta};\widehat{\theta})}{\partial \theta}\right)\left(\frac{\partial Q_i(\widehat{\theta};\widehat{\theta})}{\partial \theta}\right)',$$

then $[\Sigma(\widehat{\theta})]^{-1}$ is the estimate of the asymptotic covariance matrix of $\widehat{\theta}$. Alternatively, one could use the expected information matrix from Section 2.1.

2.5 Application to Waterloo Smoking Prevention Project

The Waterloo Smoking Prevention Project (WSPP) is a randomized longitudinal study designed to investigate smoking behavior among school children (Cameron et al., 1999). A total of 100 schools in seven Ontario school boards were randomized to dispense either the regular health education programmes provided by the school or a more intensive anti-smoking programme delivered by either a specially trained teacher or a public health nurse. Questionnaires regarding smoking attitudes and behavior were administered annually from grade 6 to grade 12. Here we use the subjects who are present at the first assessment. The purposes of this study include evaluating *i*) whether the intensive anti-smoking education programme is more effective than standard school education programme, *ii*) whether students' smoking behavior changes and *iii*) whether other factors have influence on the children's smoking behavior.

The smoking status based on the responses to the questionnaire items can be coded as three states. Children who have never smoked, tried once or quit are classified as 'non-smoker' and are represented by state 0. A child is in state 1 if

Table 2.5: Sample data from two schools participating in the Waterloo Smoking Prevention Project

School A																School B														
Time			1	2	3	4	5	6	7	1	2	3	4	5	6	7	Time			1	2	3	4	5	6	7	1	2	3	4	5	6	7
ID	GENDER	TRT	SMR							STATE							ID	GENDER	TRT	SMR							STATE						
1	0	1	1	1	2	1	2	1	2	1	1	2	1	1	1	1	1	0	0	1	1	2	2	2	2	2	1	1	1	3	1	3	3
2	1	1	2	2	1	1	1	2	1	1	1	.	1	1	3	3	2	1	0	1	1	2	2	2	2	2	1	1	2	3	3	3	3
3	0	1	1	1	1	1	1	1	1	1	2	2	3	3	.	.	3	1	0	2	2	2	3	3	3	3	1	1	1	2	3	3	3
4	0	1	1	1	1	1	1	1	1	1	1	2	2	3	2	3	4	1	0	2	2	2	2	2	2	3	1	1	1	1	1	.	.
5	1	1	2	2	2	2	2	3	3	1	1	1	1	1	.	.	5	0	0	2	2	1	2	2	2	2	1	1	1	1	1	1	1
6	1	1	2	2	2	2	3	2	2	1	1	3	3	3	3	3	6	0	0	1	1	2	2	2	2	2	1	1	1	1	1	1	1
7	0	1	2	2	2	3	3	3	2	1	1	1	1	1	1	1	7	1	0	1	1	3	3	3	3	3	1	1	2	.	.	.	.
8	1	1	2	2	3	3	3	3	3	1	1	2	.	.	.	.	8	1	0	2	2	2	2	3	3	3	1	1	1	.	.	.	.
9	0	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	9	1	0	1	1	1	2	2	2	2	1	1	1	2	3	3	3
10	1	1	2	2	3	2	2	2	3	1	1	3	3	3	3	3	10	1	0	1	1	2	2	2	2	2	1	1	.	.	.	.	.
11	0	1	2	3	3	3	3	3	3	1	2	1	3	1	1	.	11	1	0	2	2	3	3	3	3	3	1	1	1	.	1	1	1
12	0	1	2	3	2	2	2	2	2	1	1	2	2	3	1	2	12	0	0	2	2	2	2	2	2	3	1	1	1	1	1	1	3
13	0	1	1	2	3	3	3	3	3	1	1	1	3	.	.	3	13	0	0	2	2	1	2	2	2	2	1	1	1	1	1	3	3
14	0	1	2	2	3	2	3	3	3	1	1	1	3	1	1	1	14	0	0	1	2	2	2	2	2	2	1	1	1	1	1	1	2

he or she is experimenting with smoking. Children who are regular smokers are classified as in state 2. A three-state diagram is displayed in Figure 2.2 to show possible transitions among the states. Along with the responses, the factors that may influence the children's smoking behavior were recorded. These covariates include gender (coded as GENDER, 0–female, 1–male), treatment group (coded as TRT, 0–control; 1–intervention), social models risk score (coded as SMR, 1–none of parents, siblings or friends smoke; 2–one of parents, siblings or friends smoke; 3–two or more of parents, siblings or friends smoke) and grade indicator (coded as GRADE, 0–secondary school; 1–high school).

There are 3965 subjects in the data set who are present at the first assessment. About 62.6% subjects have missing observations. The missing proportions from grade 7 to grade 12 are 2%, 3%, 7%, 10%, 2% and 1% respectively. In Table 2.5 we display a sample data subset from two participating schools for illustration. We first analyze the complete cases which contain the measurements of 1432 children taken from grade 6 to grade 12.

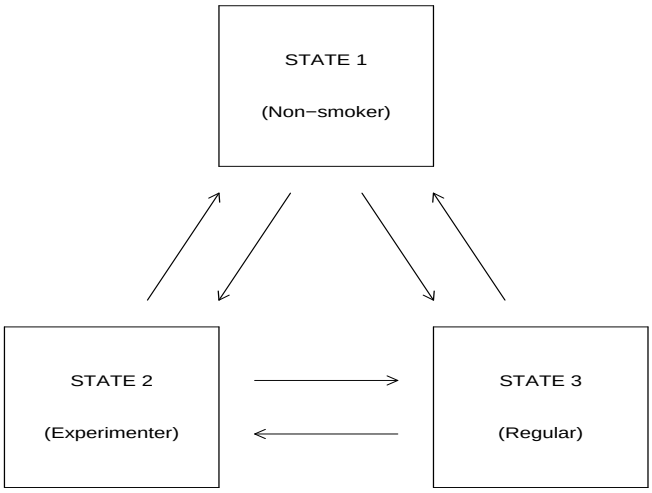
We let Y_{ij} be the state student i was in at time j, i.e., in grade $5+j$, $j = 1, \ldots, 7$, and use the subscripts for covariates in a similar fashion. Consider the model for the marginal probabilities

$$\begin{aligned} \log\left(\frac{\mu^M_{ijk}}{\mu^M_{ij0}}\right) &= \beta_{k0} + \beta_{k1} \cdot \text{GENDER}_i + \beta_{k2} \cdot \text{TRT}_i + \beta_{k3} \cdot \text{SMR2}_{ij} \\ &\quad + \beta_{k4} \cdot \text{SMR3}_{ij} + \beta_{k5} \cdot \text{GRADE}_{ij}, \quad k = 1, 2, \end{aligned}$$

where TRT_i represents the treatment status for subject i, $\text{SMR2}_{ij} = I(\text{SMR}_{ij} = 2)$ and $\text{SMR3}_{ij} = I(\text{SMR}_{ij} = 3)$, along with the model for the conditional probabilities

$$\log\left(\frac{\mu^C_{ijk}}{\mu^C_{ij0}}\right) = \gamma_{ijk} + \gamma_{ij1k} I(Y_{i,j-1} = 1) + \gamma_{ij2k} I(Y_{i,j-1} = 2), \quad k = 1, 2,$$

Figure 2.2: Three-state diagram for the analysis of the Waterloo Smoking Prevention Project Data



where the following regression models are assumed for the coefficients $\gamma_{ijk'k}$:

$$\begin{aligned}\gamma_{ijk'k} &= \alpha_{k'k0} + \alpha_{k'k1} \cdot \mathrm{GENDER}_i + \alpha_{k'k2} \cdot \mathrm{TRT}_i + \alpha_{k'k3} \cdot \mathrm{SMR2}_{ij} \\ &\quad + \alpha_{k'k4} \cdot \mathrm{SMR3}_{ij} + \alpha_{k'k5} \cdot \mathrm{GRADE}_{ij}, \quad k', k = 1, 2.\end{aligned}$$

As the data feature both the longitudinal correlation across subjects and cross-sectional association across schools, here we use robust standard errors by adapting the sandwich type variance formula discussed in Royall (1986) and Cook et al. (2002) to accommodate potential cluster effects in the presence of missing values. Let $S^{(h)}(\theta)$ denote the score vector constructed by means of the formulation in Section 2.4.1, based only on students from school h with the cross-sectional association ignored, $h = 1, 2, \ldots, H$, where H denotes the total number of schools. Modifying the arguments in White (1982) that are applied to the cases without missing data,

we can show that under MAR, the solution $\widehat{\theta}$ to

$$S(\theta) = \sum_{h=1}^{H} S^{(h)}(\theta) = 0$$

converges to θ^* almost surely. Here θ^* solves $E_T[S(\theta)] = 0$ with E_T denoting the expectation taken with respect to the true distribution. Furthermore,

$$\sqrt{H}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, A^{-1}(\theta^*) \cdot B(\theta^*) \cdot A^{-1}(\theta^*)), \quad \text{as } H \longrightarrow \infty,$$

where $A(\theta) = E_T[\partial S^{(h)}(\theta)/\partial\theta]$, and $B(\theta) = E_T\left[S^{(h)}(\theta)[S^{(h)}(\theta)]'\right]$. $A(\theta)$ and $B(\theta)$ may be estimated by

$$\widehat{A}(\widehat{\theta}) = H^{-1} \sum_{h=1}^{H} \partial S^{(h)}(\theta)/\partial\theta|_{\theta=\widehat{\theta}}$$

and

$$\widehat{B}(\widehat{\theta}) = H^{-1} \sum_{h=1}^{H} S^{(h)}(\theta)[S^{(h)}(\theta)]'|_{\theta=\widehat{\theta}},$$

respectively.

Table 2.6 reports on the complete case analysis results with 1432 children contributing complete observations. In the marginal model, both gender and treatment covariates are not statistically significant. However, social model risk score and grade have significant negative effects on the probability of smoking (either experimental or regular). Students are more likely to smoke if their parents, siblings or friends are smokers. Students are more likely to smoke when they are in high school as opposed to being in secondary school. The covariate effects in the conditional model seem to be less striking. Grade and social model risk score have negative effects only for some transitions (see $\alpha_{114}, \alpha_{125}, \alpha_{223}, \alpha_{224}$). The remaining covariate effects are not statistically significant.

Next, we analyze the available data with 3965 subjects contributing complete or partial observations, assuming the missing data mechanism is MAR. Table 2.7

Table 2.6: Complete case analysis of the Waterloo Smoking Prevention Project data

Parameter		Estimate	S.E	R.S.E†	p-value
Marginal Model:					
INTERCEPT1	(β_{10})	-3.126	0.130	0.174	<0.001
GENDER	(β_{11})	0.063	0.077	0.083	0.448
TRT	(β_{12})	0.112	0.081	0.107	0.295
SMR2	(β_{13})	0.569	0.082	0.099	<0.001
SMR3	(β_{14})	1.320	0.096	0.111	<0.001
GRADE	(β_{15})	1.214	0.094	0.135	<0.001
INTERCEPT2	(β_{20})	-5.410	0.171	0.228	<0.001
GENDER	(β_{21})	0.179	0.084	0.104	0.083
TRT	(β_{22})	0.074	0.087	0.117	0.527
SMR2	(β_{23})	1.923	0.136	0.178	<0.001
SMR3	(β_{24})	3.629	0.128	0.180	<0.001
GRADE	(β_{25})	2.114	0.126	0.186	<0.001
Association Model					
INTERCEPT1	(α_{110})	2.712	0.335	0.345	<0.001
GENDER	(α_{111})	0.086	0.168	0.157	0.584
TRT	(α_{112})	-0.108	0.177	0.185	0.559
SMR2	(α_{113})	-0.277	0.207	0.205	0.177
SMR3	(α_{114})	-0.517	0.229	0.168	0.002
GRADE	(α_{115})	-0.309	0.257	0.262	0.238
INTERCEPT2	(α_{210})	-0.503	0.833	0.660	0.446
GENDER	(α_{211})	0.694	0.335	0.367	0.059
TRT	(α_{212})	0.639	0.375	0.333	0.055
SMR2	(α_{213})	0.570	0.457	0.501	0.255
SMR3	(α_{214})	0.376	0.499	0.486	0.439
GRADE	(α_{215})	0.023	0.670	0.585	0.969
INTERCEPT3	(α_{120})	3.460	0.524	0.569	<0.001
GENDER	(α_{121})	-0.077	0.225	0.279	0.783
TRT	(α_{122})	-0.288	0.226	0.222	0.195
SMR2	(α_{123})	0.360	0.369	0.366	0.325
SMR3	(α_{124})	0.244	0.376	0.391	0.533
GRADE	(α_{125})	-1.393	0.358	0.351	<0.001
INTERCEPT4	(α_{220})	2.889	0.635	0.686	<0.001
GENDER	(α_{221})	-0.317	0.234	0.242	0.190
TRT	(α_{222})	0.067	0.235	0.228	0.769
SMR2	(α_{223})	0.985	0.381	0.307	0.001
SMR3	(α_{224})	1.071	0.377	0.360	0.003
GRADE	(α_{225})	-0.022	0.497	0.609	0.971
loglik=-5010.54					

† R.S.E is the robust standard error based on the sandwich variance formula. S.E is the naive standard error that without accommodating the clustering.

reports on the analysis results based on the observed data using the Fisher-scoring method. It can be seen that the results are comparable with those for the complete data analysis. Again, in the marginal model, both gender and treatment covariates are not statistically significant. Social model risk score and grade have significant negative effects on smoking incidence. Students are more likely to smoke if their parents, siblings or friends are smokers. Students are more likely to smoke when they are in high school as opposed to being in secondary school. Finally, we comment on the difference in interpreting the parameters in the marginal and conditional models. In the marginal models, the parameters reflect the covariate at the population level for various time points. Typically, the social model risk score and grade are statistically significant. However, in the conditional models, the parameters feature the covariate effects on transitions among states. The analysis results show that the covariate effects in the conditional model seem to be less striking. Grade and social model risk score have negative effects on some transitions (see $\alpha_{114}, \alpha_{125}, \alpha_{223}, \alpha_{224}$). Other covariate effects are not statistically significant.

Table 2.7: Available data analysis of the Waterloo Smoking Prevention Project data

Parameter		Estimate	S.E	R.S.E[†]	p-value
Marginal Model:					
INTERCEPT1	(β_{10})	-3.181	0.109	0.141	<0.001
GENDER	(β_{11})	-0.119	0.068	0.078	0.127
TRT	(β_{12})	0.095	0.076	0.093	0.307
SMR2	(β_{13})	0.457	0.073	0.088	<0.001
SMR3	(β_{14})	1.293	0.106	0.127	<0.001
GRADE	(β_{15})	0.971	0.071	0.094	<0.001
INTERCEPT2	(β_{20})	-3.642	0.111	0.157	<0.001
GENDER	(β_{21})	0.014	0.059	0.118	0.906
TRT	(β_{22})	-0.041	0.072	0.086	0.634
SMR2	(β_{23})	0.805	0.084	0.127	<0.001
SMR3	(β_{24})	1.803	0.095	0.134	<0.001
GRADE	(β_{25})	2.040	0.088	0.135	<0.001
Association Model					
INTERCEPT1	(α_{110})	2.462	0.298	0.309	<0.001
GENDER	(α_{111})	0.064	0.132	0.138	0.643
TRT	(α_{112})	-0.121	0.144	0.152	0.426
SMR2	(α_{113})	-0.245	0.183	0.191	0.200
SMR3	(α_{114})	-0.499	0.207	0.183	0.006
GRADE	(α_{115})	-0.318	0.224	0.230	0.167
INTERCEPT2	(α_{210})	-0.429	0.679	0.656	0.513
GENDER	(α_{211})	0.438	0.301	0.324	0.176
TRT	(α_{212})	0.547	0.347	0.358	0.127
SMR2	(α_{213})	0.525	0.403	0.437	0.230
SMR3	(α_{214})	0.328	0.456	0.448	0.464
GRADE	(α_{215})	0.030	0.618	0.632	0.962
INTERCEPT3	(α_{120})	3.906	0.483	0.504	<0.001
GENDER	(α_{121})	-0.082	0.211	0.227	0.718
TRT	(α_{122})	-0.297	0.208	0.213	0.163
SMR2	(α_{123})	0.347	0.334	0.362	0.338
SMR3	(α_{124})	0.208	0.338	0.368	0.572
GRADE	(α_{125})	-1.486	0.329	0.320	<0.001
INTERCEPT4	(α_{220})	2.978	0.592	0.633	<0.001
GENDER	(α_{221})	-0.326	0.218	0.235	0.165
TRT	(α_{222})	0.052	0.221	0.241	0.829
SMR2	(α_{223})	0.884	0.337	0.364	0.015
SMR3	(α_{224})	1.009	0.343	0.337	0.003
GRADE	(α_{225})	-0.027	0.446	0.488	0.956
loglik=-13422.53					

[†] R.S.E is the robust standard error based on the sandwich variance formula. S.E is the naive standard error that without accommodating the clustering.

2.6 Derivatives of the Log-Likelihood

In this section, we provide details on calculation of the first and second derivatives of the log-likelihood in Section 2.2.

DERIVATION OF FIRST DERIVATIVES

By constraint (2.4) in the text, we obtain

$$\begin{aligned}\frac{\partial \mu_{ijk}^M}{\partial \beta} &= \sum_{k'=0}^{K}\left\{\sum_{l=1}^{K}\frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial \gamma_{ijl}}\frac{\partial \gamma_{ijl}}{\partial \beta}\mu_{i,j-1,k'}^M \right.\\ &\left. + P(Y_{ij}=k|Y_{i,j-1}=k';X_i)\frac{\partial \mu_{i,j-1,k'}^M}{\partial \beta}\right\}.\end{aligned}$$

Let

$$B_k = \frac{\partial \mu_{ijk}^M}{\partial \beta} - \sum_{k'=0}^{K} P(Y_{ij}=k|Y_{i,j-1}=k';X_i)\frac{\partial \mu_{i,j-1,k'}^M}{\partial \beta},$$

$$A_k = \left(\sum_{k'=0}^{K}\frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial \gamma_{ijl}}\mu_{i,j-1,k'}^M, l=1,\ldots,K\right)'$$

and $D_\beta = (\partial\gamma_{ijl}/\partial\beta, l=1,\ldots,K)'$, $k=1,\ldots,K$, then $B_k = A_k' D_\beta$. In matrix notation, we have $B = AD_\beta$, where $B=(B_1,\ldots,B_K)'$ and $A=(A_1,\ldots,A_K)'$. If A is not singular, then

$$D_\beta = A^{-1}B \tag{2.9}$$

Therefore, the partial derivative $\partial\mu_{ijk}^C/\partial\beta$ in the score functions is given by

$$\frac{\partial \mu_{ijk}^C}{\partial \beta} = \sum_{l=1}^{K}\frac{\partial \mu_{ijk}^C}{\partial \gamma_{ijl}}\frac{\partial \gamma_{ijl}}{\partial \beta},$$

where $\partial\mu_{ijk}^C/\partial\gamma_{ijl}$ is determined by (2.2), and $\partial\gamma_{ijl}/\partial\beta$ is determined by (2.9).

Note that by (2.2), $P(Y_{ij}=k|Y_{i,j-1}=k';X_i)$ is a function of $\gamma_{ijl}, l=1,...,K$ and $\gamma_{ijk'm}, m=1,...,K$, where $\gamma_{ijl}, l=1,...,K$ and $\gamma_{ijk'm}, m=1,...,K$ are functions

of α due to (2.3) and (2.4). So, taking the derivative on both sides of (2.4) with respect to α, we obtain

$$\begin{aligned} 0 &= \frac{\partial \mu_{ijk}^M}{\partial \alpha} \\ &= \sum_{k'=0}^{K} \left\{ \sum_{l=1}^{K} \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial \gamma_{ijl}} \frac{\partial \gamma_{ijl}}{\partial \alpha} \right. \\ &\quad \left. + \sum_{m=1}^{K} \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial \gamma_{ijk'm}} \frac{\partial \gamma_{ijk'm}}{\partial \alpha} \right\} \mu_{i,j-1,k'}^M \\ &= A_k' D_\alpha + M_k \end{aligned}$$

where $D_\alpha = (\partial \gamma_{ijl}/\partial \alpha, l = 1, \ldots, K)'$ and

$$M_k = \sum_{k'=0}^{K} \sum_{m=1}^{K} \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial \gamma_{ijk'm}} \cdot \frac{\partial \gamma_{ijk'm}}{\partial \alpha} \cdot \mu_{i,j-1,k'}^M.$$

In matrix form, we have $AD_\alpha + M = 0$, where $M = (M_1, \ldots, M_K)'$. If A is not singular,

$$D_\alpha = -A^{-1}M. \tag{2.10}$$

Note that μ_{ijk}^C is a function of $\gamma_{ijl}, l = 1, ..., K$ and $\gamma_{ijmn}, m, n = 1, ..., K$, where $\gamma_{ijl}, l = 1, ..., K$ and $\gamma_{ijmn}, m, n = 1, ..., K$ are functions of α. Therefore, by the Chain Rule, the derivative of μ_{ijk}^C with respect to α may be written as

$$\frac{\partial \mu_{ijk}^C}{\partial \alpha} = \sum_{l=1}^{K} \frac{\partial \mu_{ijk}^C}{\partial \gamma_{ijl}} \frac{\partial \gamma_{ijl}}{\partial \alpha} + \sum_{m=1}^{K} \sum_{n=1}^{K} \frac{\partial \mu_{ijk}^C}{\partial \gamma_{ijmn}} \frac{\partial \gamma_{ijmn}}{\partial \alpha}, \quad k = 1, \ldots, K,$$

where $\partial \gamma_{ijl}/\partial \alpha$ is determined by (2.10), $\partial \mu_{ijk}^C/\partial \gamma_{ijl}$ and $\partial \mu_{ijk}^C/\partial \gamma_{ijmn}$ are determined by (2.2) and $\partial \gamma_{ijmn}/\partial \alpha$ is determined by (2.3).

So, the score vector is given by

$$S(\theta) = \sum_{i=1}^{n} \left\{ \sum_{j=2}^{J_i} \sum_{k=0}^{K} I(Y_{ij}=k) \frac{1}{\mu_{ijk}^C} \frac{\partial \mu_{ijk}^C}{\partial \theta} + \sum_{k=0}^{K} I(Y_{i1}=k) \frac{1}{\mu_{i1k}^M} \frac{\partial \mu_{i1k}^M}{\partial \theta} \right\},$$

where $\partial\mu^C_{ijk}/\partial\theta = (\partial\mu^C_{ijk}/\partial\beta', \partial\mu^C_{ijk}/\partial\alpha')'$, $\partial\mu^M_{i1k}/\partial\theta = (\partial\mu^M_{i1k}/\partial\beta', \partial\mu^M_{i1k}/\partial\alpha')'$, and $\partial\mu^M_{i1k}/\partial\beta$ can be obtained from (2.1), $\partial\mu^M_{i1k}/\partial\alpha = 0$.

DERIVATIONS OF SECOND DERIVATIVES

Taking second derivatives in the constraint (2.4), we obtain

$$\begin{aligned}
\frac{\partial^2\mu^M_{ijk}}{\partial\beta_u\partial\beta_v} &= \sum_{k'=0}^{K}\sum_{l=1}^{K}\left\{\sum_{m=1}^{K}\frac{\partial^2 P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}\partial\gamma_{ijm}}\frac{\partial\gamma_{ijl}}{\partial\beta_u}\frac{\partial\gamma_{ijm}}{\partial\beta_v}\mu^M_{i,j-1,k'}\right. \\
&\quad + \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}}\frac{\partial^2\gamma_{ijl}}{\partial\beta_u\partial\beta_v}\mu^M_{i,j-1,k'} \\
&\quad \left. + \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}}\frac{\partial\gamma_{ijl}}{\partial\beta_u}\frac{\partial\mu^M_{i,j-1,k'}}{\partial\beta_v}\right\} \\
&\quad + \sum_{k'=0}^{K}\left\{\sum_{l=1}^{K}\frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}}\frac{\partial\gamma_{ijl}}{\partial\beta_v}\frac{\partial\mu^M_{i,j-1,k'}}{\partial\beta_u}\right. \\
&\quad \left. + P(Y_{ij}=k|Y_{i,j-1}=k';X_i)\frac{\partial^2\mu^M_{i,j-1,k'}}{\partial\beta_u\partial\beta_v}\right\}.
\end{aligned}$$

Let

$$\begin{aligned}
C_{kuv} = \frac{\partial^2\mu^M_{ijk}}{\partial\beta_u\partial\beta_v} &- \sum_{k'=0}^{K}\sum_{l=1}^{K}\left\{\sum_{m=1}^{K}\frac{\partial^2 P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}\partial\gamma_{ijm}}\frac{\partial\gamma_{ijl}}{\partial\beta_u}\frac{\partial\gamma_{ijm}}{\partial\beta_v}\mu^M_{i,j-1,k'}\right. \\
&\quad \left. + \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}}\frac{\partial\gamma_{ijl}}{\partial\beta_u}\frac{\partial\mu^M_{i,j-1,k'}}{\partial\beta_v}\right\} \\
&- \sum_{k'=0}^{K}\left\{\sum_{l=1}^{K}\frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}}\frac{\partial\gamma_{ijl}}{\partial\beta_v}\frac{\partial\mu^M_{i,j-1,k'}}{\partial\beta_u}\right. \\
&\quad \left. + P(Y_{ij}=k|Y_{i,j-1}=k';X_i)\frac{\partial^2\mu^M_{i,j-1,k'}}{\partial\beta_u\partial\beta_v}\right\}
\end{aligned}$$

and

$$D_{\beta_{uv}} = \left(\frac{\partial^2\gamma_{ijl}}{\partial\beta_u\partial\beta_v},\ \ l=1,\ldots,K\right)',$$

we have

$$C_{uv} = AD_{\beta_{uv}},$$

where $C_{uv} = (C_{1uv}, \ldots, C_{Kuv})'$. If A is non-singular, we have

$$D_{\beta_{uv}} = A^{-1} C_{uv}. \tag{2.11}$$

So,

$$\frac{\partial^2 \mu^C_{ijk}}{\partial\beta_u \partial\beta_v} = \sum_{l=1}^{K} \left(\sum_{m=1}^{K} \frac{\partial^2 \mu^C_{ijk}}{\partial\gamma_{ijl}\partial\gamma_{ijm}} \frac{\partial\gamma_{ijl}}{\partial\beta_u} \frac{\partial\gamma_{ijm}}{\partial\beta_v} + \frac{\partial\mu^C_{ijk}}{\partial\gamma_{ijl}} \cdot \frac{\partial^2\gamma_{ijl}}{\partial\beta_u\partial\beta_v} \right), \quad k = 1, \ldots, K,$$

where $\partial^2\gamma_{ijl}/\partial\beta_u\partial\beta_v$ is determined by (2.11), $\partial^2\mu^C_{ijk}/\partial\gamma_{ijl}\partial\gamma_{ijm}$ and $\partial\mu^C_{ijk}/\partial\gamma_{ijl}$ are determined by (2.2), and $\partial\gamma_{ijl}/\partial\beta_u$ and $\partial\gamma_{ijm}/\partial\beta_v$ are determined by (2.9).

Take second derivative on both sides of (2.4) with respect to α_u and α_v, we obtain

$$\begin{aligned}
0 &= \frac{\partial^2 \mu^M_{ijk}}{\partial\alpha_u\partial\alpha_v} \\
&= \sum_{k'=0}^{K}\sum_{l=1}^{K} \left\{ \sum_{m=1}^{K} \frac{\partial^2 P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}\partial\gamma_{ijm}} \frac{\partial\gamma_{ijl}}{\partial\alpha_u}\frac{\partial\gamma_{ijm}}{\partial\alpha_v} \right. \\
&\quad \left. + \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}} \frac{\partial^2\gamma_{ijl}}{\partial\alpha_u\partial\alpha_v} \right\} \mu^M_{i,j-1,k'} \\
&\quad + \sum_{k'=0}^{K}\sum_{m=1}^{K} \left\{ \sum_{n=1}^{K} \frac{\partial^2 P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijk'm}\partial\gamma_{ijk'n}} \frac{\partial\gamma_{ijk'm}}{\partial\alpha_u}\frac{\partial\gamma_{ijk'n}}{\partial\alpha_v} \right. \\
&\quad \left. + \frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijk'm}} \frac{\partial^2\gamma_{ijk'm}}{\partial\alpha_u\partial\alpha_v} \right\} \mu^M_{i,j-1,k'} \\
&= A'_k D_{\alpha_{uv}} + E_{kuv},
\end{aligned}$$

where

$$D_{\alpha_{uv}} = \left(\frac{\partial^2\gamma_{ijl}}{\partial\alpha_u\partial\alpha_v}, l = 1, \ldots, K \right)$$

and

$$\begin{aligned} E_{kuv} &= \sum_{k'=0}^{K}\sum_{l=1}^{K}\sum_{m=1}^{K}\frac{\partial^2 P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijl}\partial\gamma_{ijm}}\frac{\partial\gamma_{ijl}}{\partial\alpha_u}\frac{\partial\gamma_{ijm}}{\partial\alpha_v}\mu^M_{i,j-1,k'} \\ &\quad+\sum_{k'=0}^{K}\sum_{m=1}^{K}\left\{\sum_{n=1}^{K}\frac{\partial^2 P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijk'm}\partial\gamma_{ijk'n}}\frac{\partial\gamma_{ijk'm}}{\partial\alpha_u}\frac{\partial\gamma_{ijk'n}}{\partial\alpha_v}\right. \\ &\quad\left.+\frac{\partial P(Y_{ij}=k|Y_{i,j-1}=k';X_i)}{\partial\gamma_{ijk'm}}\frac{\partial^2\gamma_{ijk'm}}{\partial\alpha_u\partial\alpha_v}\right\}\mu^M_{i,j-1,k'} \end{aligned}$$

In matrix form, we have $AD_{\alpha_{uv}} = E_{uv}$, where $E_{uv} = (E_{1uv},\ldots,E_{Kuv})'$. If A is non-singular, we have

$$D_{\alpha_{uv}} = -A^{-1}E_{uv}. \tag{2.12}$$

So,

$$\begin{aligned} &\frac{\partial^2\mu^C_{ijk}}{\partial\alpha_u\partial\alpha_v} \\ &= \sum_{l=1}^{K}\left(\sum_{m=1}^{K}\frac{\partial^2\mu^C_{ijk}}{\partial\gamma_{ijl}\partial\gamma_{ijm}}\cdot\frac{\partial\gamma_{ijl}}{\partial\alpha_u}\cdot\frac{\partial\gamma_{ijm}}{\partial\alpha_v}+\frac{\partial\mu^C_{ijk}}{\partial\gamma_{ijl}}\frac{\partial^2\gamma_{ijl}}{\partial\alpha_u\partial\alpha_v}\right) \\ &\quad+\sum_{m_1=1}^{K}\sum_{n_1=1}^{K}\left(\sum_{m_2=1}^{K}\sum_{n_2=1}^{K}\frac{\partial^2\mu^C_{ijk}}{\partial\gamma_{ijm_1n_1}\partial\gamma_{ijm_2n_2}}\frac{\partial\gamma_{ijm_1n_1}}{\partial\alpha_u}\frac{\partial\gamma_{ijm_2n_2}}{\partial\alpha_v}+\frac{\partial\mu^C_{ijk}}{\partial\gamma_{ijm_1n_1}}\frac{\partial^2\gamma_{ijm_1n_1}}{\partial\alpha_u\partial\alpha_v}\right), \end{aligned}$$

where $\partial^2\gamma_{ijl}/\partial\alpha_u\partial\alpha_v$ is determined by (2.12), $\partial^2\mu^C_{ijk}/\partial\gamma_{ijl}\partial\gamma_{ijm}$, $\partial\mu^C_{ijk}/\partial\gamma_{ijl}$ and $\partial^2\mu^C_{ijk}/\partial\gamma_{ijm_1n_1}\partial\gamma_{ijm_2n_2}$ are determined by (2.2), $\partial\gamma_{ijl}/\partial\alpha_u$ and $\partial\gamma_{ijm}/\partial\alpha_v$ are determined by (2.10), and $\partial\gamma_{ijm_1n_1}/\partial\alpha_u$, $\partial\gamma_{ijm_2n_2}/\partial\alpha_v$ and $\partial^2\gamma_{ijm_1n_1}/\partial\alpha_u\partial\alpha_v$ are determined by (2.3).

We also note that $\partial^2\mu^C_{ijk}/\partial\alpha\partial\beta = 0$ because α and β are orthogonal based on Appendix B in the following and $\partial^2\mu^M_{i1k}/\partial\alpha\partial\beta = 0$. So, the second derivative of the log-likelihood can be written as

$$\begin{aligned} \frac{\partial^2\log L(\theta)}{\partial\theta_u\partial\theta_v} &= \sum_{i=1}^{n}\sum_{j=2}^{J_i}\sum_{k=0}^{K}I(Y_{ij}=k)\left\{-\frac{1}{\left(\mu^C_{ijk}\right)^2}\frac{\partial\mu^C_{ijk}}{\partial\theta_u}\frac{\partial\mu^C_{ijk}}{\partial\theta_v}+\frac{1}{\mu^C_{ijk}}\frac{\partial^2\mu^C_{ijk}}{\partial\theta_u\partial\theta_v}\right\} \\ &\quad+\sum_{i=1}^{n}\sum_{k=0}^{K}I(Y_{i1}=k)\left\{-\frac{1}{\left(\mu^M_{i1k}\right)^2}\frac{\partial\mu^M_{i1k}}{\partial\theta_u}\frac{\partial\mu^M_{i1k}}{\partial\theta_v}+\frac{1}{\mu^M_{i1k}}\frac{\partial^2\mu^M_{i1k}}{\partial\theta_u\partial\theta_v}\right\}, \end{aligned}$$

and the observed information matrix is given by $[-\partial^2 \log L(\theta)/\partial\theta_u \partial\theta_v]$.

2.7 Log-Linear Parametrization of Categorical Model

For ease of notation we drop the subject index i in the following discussion. Note that

$$\begin{aligned} P(Y_1 = y_1) &= \exp\left(\sum_{k=1}^{K} \eta_{1k} I(y_1 = k)\right) \Big/ \left(1 + \sum_{k=1}^{K} \exp(\eta_{1k})\right) \\ &= \exp\left(\theta_{01} + \sum_{k=1}^{K} \eta_{1k} I(y_1 = k)\right) \end{aligned}$$

where $\eta_{1k} = X'_{1k}\beta_k, k = 1, \ldots, K$ and $\theta_{01} = -\log\left(1 + \sum_{k=1}^{K} \exp(\eta_{1k})\right)$.

$$\begin{aligned} & P(Y_1 = y_1, Y_2 = y_2) \\ = \;& P(Y_2 = y_2 | Y_1 = y_1) P(Y_1 = y_1) \\ = \;& \exp\left(\theta_{02} + \sum_{j=1}^{2}\sum_{k=1}^{K} \theta_{jk} I(y_j = k) + \sum_{j=2}^{2}\sum_{k=1}^{K}\sum_{k'=1}^{K} \gamma_{jk'k} I(y_1 = k', y_2 = k)\right) \end{aligned}$$

where $\theta_{02} = \theta_{01} - \log\left(1 + \sum_{k=1}^{K} \exp(\gamma_{2k})\right)$,

$$\theta_{1k} = \eta_{1k} - \log\left(1 + \sum_{k'=1}^{K} \exp(\gamma_{2k'} + \gamma_{2kk'})\right) + \log\left(1 + \sum_{k'=1}^{K} \exp(\gamma_{2k'})\right)$$

and $\theta_{2k} = \gamma_{2k}$ for $k = 1, \ldots, K$.

In general, we have

$$\begin{aligned} & P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_J = y_J) \\ = \;& \prod_{j=2}^{J} P(Y_j = y_j | Y_{j-1} = y_{j-1}) \cdot P(Y_1 = y_1) \\ = \;& \exp\left(\theta_{0J} + \sum_{j=1}^{J}\sum_{k=1}^{K} \theta_{jk} I(y_j = k) + \sum_{j=2}^{J}\sum_{k=1}^{K}\sum_{k'=1}^{K} \gamma_{jk'k} I(y_{j-1} = k', y_j = k)\right) \end{aligned}$$

where $\theta_{0J} = -\sum_{j=1}^{J} \log\left(1 + \sum_{k=1}^{K} \exp(\gamma_{jk})\right)$ and

$$\theta_{jk} = \eta_{jk} - \log\left(1 + \sum_{k'=1}^{K} \exp(\gamma_{j+1,k'} + \gamma_{j+1,kk'})\right) + \log\left(1 + \sum_{k'=1}^{K} \exp(\gamma_{j+1,k'})\right)$$

for $j < J$, and $\theta_{Jk} = \gamma_{Jk}$ for $k = 1, \ldots, K$, where for simplicity we adopt $\gamma_{1k} = \eta_{1k}$. Here $\eta_{jk} = X'_{jk}\beta_k$.

Therefore, the proposed model is a reparametrization of the canonical log-linear model $(\theta^{(1)}, \gamma^{(1)})$ to $(\mu^M, \gamma^{(1)})$, where $\theta^{(1)} = (\theta_{jk})$ and $\gamma^{(1)} = (\gamma_{jk'k})$ for $j = 1, \ldots, J$ and $k, k' = 1, \ldots, K$. This implies β and α are orthogonal (Barndorff-Nielsen and Cox, 1994).

Chapter 3

Progressive Multi-State Models for Incomplete Longitudinal and Life History Data

3.1 Overview

Multi-state life history data commonly arise in many research areas such as medicine, social sciences and public health. Multi-state models provide a convenient method for characterizing the movement of individuals through a finite set of states. In health research, the most common application of multi-state models is to provide a comprehensive view of a disease process to allow estimation of proportions of individuals who will be in various states at some time in the future, or rates of transitions. Examples of these include illness-death models, competing risk models and progressive models. In continuous time multi-state models, it is often the transition intensities which are of interest. These are the instantaneous conditional probability of transition at some time point given the covariates and the process history. In practice, the intensities are frequently modeled as a function of covariates that are believed to be relevant to the response process.

Sometimes individuals are observed at prespecified assessment times and their

states are to be determined only at these times, so information about transitions between successive observation times is unavailable. This type of data are sometimes referred to as *panel data* in the context of multi-state models (Kalbfleisch and Lawless, 1989), and arise naturally in settings such as clinical trials where patients are examined by physicians periodically and their states are assessed at those visits. In an observational setting the psoriatic arthritis (PsA) study of Gladman et al. (1995), involves radiological assessments of clinic patients at roughly one-year intervals for much of their follow up, but the exact times of events are unknown. Since the assessment times are prespecified, if patients completed their schedule of assessments, only the disease process would need to be modeled (Gruger et al., 1991).

Markov models are widely used in the analysis of multi-state data. These methods have been studied early by Bartholomew (1983), Singer and Spilerman (1976a, b) and Wasserman (1980) among others. Kalbfleisch and Lawless (1985, 1989) proposed a very efficient way of obtaining maximum likelihood estimates for time-homogeneous Markov models with arbitrary transition structures. Gentleman et al. (1994) adapted the method of Kalbfleisch and Lawless (1985, 1989) to incorporate time nonhomogeneous intensities.

Typically, progressive models provide a convenient framework for characterizing the disease processes which arise, for example, when the state represents the degree of the irreversible damage incurred by the disease. For the special case of the progressive time-homogeneous Markov model, Satten (1999) gave a closed form of the transition probability matrix expressed in terms of the transition intensities. Cook et al. (2004) discussed the conditional Markov model for clustered progressive multi-state process under incomplete observation through multivariate random

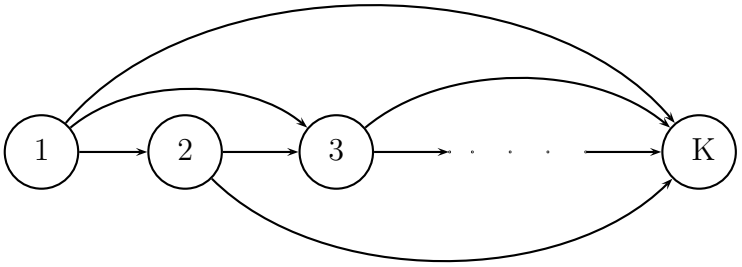
effects.

Relatively little attention has been given to the analysis of progressive models with incomplete assessment data. In general, individuals are observed at prespecified assessment times and their states are determined only at these times. A complete case analysis might lead to the biased results, and so invalid inferences. Likelihood-based methods are commonly used to handle incomplete data problems. It gives valid estimates when the missing data is MCAR and MAR. In this chapter, we provide a general method to handle the incomplete data problem in the progressive model. Maximum likelihood methods of estimation via the EM algorithm are developed to calculate parameter estimates, and variance estimation is based on Louis's method. Section 3.2 is mainly concerned with the discrete time progressive multi-state process, in which the EM algorithm is employed. Simulation studies indicate that this method works well for many settings. Continuous time progressive multi-state processes are considered in Section 3.3. Data from the Waterloo Smoking Prevention Project (Cameron et al., 1999) and the psoriatic arthritis study (Gladman et al., 1995) are analyzed in Section 3.4.

3.2 Modeling Transition Probabilities

Discrete time progressive models are widely used when the data structure is panel data and the observed transition times are not available. In practice, interest often lies in the transition probabilities between different states as well as the association between the transition probabilities and the covariate effects. In this section, we consider discrete time models via the transition probabilities.

Figure 3.1: K-state diagram for progressive process



3.2.1 Notation and Model Formulation

We define an irreversible multi-state process as one in which a given state can be entered at most once, but for the discussion that follows we consider models with a state diagram as in Figure 3.1, which is a K-state transition model. If subjects are to be observed at time points $t_1, t_2, \ldots, t_J$, let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ be the response vector containing the states occupied at the assessment times, and X_{ij} be the vector of covariates recorded for subject i at the jth assessment, $j = 1, \ldots, J$, $i = 1, \ldots, n$. We let $X_i = (X'_{i1}, X'_{i2}, \ldots, X'_{iJ})'$ denote the full covariate vector.

Often, given the covariates, a first order dependence of Y_{ij} on its history is appropriate and we discuss this model in what follows; extensions to models with higher orders of dependence are straightforward. Let $\mu^C_{ijk'k} = P(Y_{ij} = k | Y_{i,j-1} = k', X_i)$ be the transition probability from state k' to state k given X_i, where $k' \leq$

$k \leq K$. The transition probability matrix for subject i at time j is then written as

$$P_{ij} = \begin{pmatrix} \mu^C_{ij11} & \mu^C_{ij12} & \mu^C_{ij13} & \cdots & \mu^C_{ij1,K-1} & \mu^C_{ij1K} \\ 0 & \mu^C_{ij22} & \mu^C_{ij23} & \cdots & \mu^C_{ij2,K-1} & \mu^C_{ij2K} \\ 0 & 0 & \mu^C_{ij33} & \cdots & \mu^C_{ij3,K-1} & \mu^C_{ij3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu^C_{ij,K-1,K-1} & \mu^C_{ij,K-1,K} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

where these transition probabilities satisfy

$$\sum_{k=k'}^{K} \mu^C_{ijk'k} \quad = \quad 1, \quad \text{for } k' = 1, 2, \ldots, K-1.$$

In general, we may adopt the following model with $\mu^C_{ijk'k'}$ regarded as a reference

$$\log\left(\frac{\mu^C_{ijk'k}}{\mu^C_{ijk'k'}}\right) = X'_{ijk'k}\beta_{k'k}, \quad k = k'+1, \ldots, K, \tag{3.1}$$

where $X_{ijk'k}$ may be a subset of X_i, featuring the influence of the covariates on the transition between the responses, and $\beta_{k'k}$ is the vector of regression coefficients. We note that an implicit assumption $P(Y_{ij} = k|Y_{i,j-1}, X_i) = P(Y_{ij} = k|Y_{i,j-1}, X_{ijk'k})$ is made here, but $X_{ijk'k}$ can be easily expanded to ensure all important covariates are included. We let $\beta = (\beta'_{k'k}, k > k', k' = 1, \ldots, K-1)'$.

To model the missing data process, we let R_{ij} be an indicator random variable, which equals 1 if Y_{ij} is observed and 0 otherwise. We then let $R_i = (R_{i1}, R_{i2}, \ldots, R_{iJ})'$, and $r_i = (r_{i1}, r_{i2}, \ldots, r_{iJ})'$ be a realization of R_i. Here we assume all the subjects are observed at the initial enrollment, i.e. $R_{i1} = 1$. For ease of exposition, we write $y_i = (y_i^{(o)}, y_i^{(m)})$ with $y_i^{(o)}$ and $y_i^{(m)}$ denoting the observed and missing components of y_i, respectively. Inference about β is based on the observed data likelihood $L = \prod_{i=1}^n L_i$, where

$$L_i = P(R_i, Y_i^{(o)}|X_i) = \int P(R_i|Y_i^{(o)}, Y_i^{(m)}, X_i) \cdot P(Y_i^{(o)}, Y_i^{(m)}|X_i) dY_i^{(m)}. \tag{3.2}$$

Under a missing completely at random (i.e. $P(R_i|Y_i, X_i) = P(R_i)$) or missing at random (i.e. $P(R_i|Y_i, X_i) = P(R_i|Y_i^{(o)}, X_i)$) mechanism, $P(R_i|Y_i^{(o)}, Y_i^{(m)}, X_i)$ does not depend on the missing components $Y_i^{(m)}$. Then $L_i \propto P(Y_i^{(o)}|X_i)$ can be used in lieu of (3.2), provided $P(R_i|Y_i, X_i)$ does not share any parameters with β. Thus under these settings, inference about β may be directly conducted based on $P(Y_i^{(o)}|X_i)$ and the missing data process needn't be modeled. However, if data are missing not at random (MNAR), $P(R_i|Y_i^{(o)}, Y_i^{(m)}, X_i)$ does depend on $Y_i^{(m)}$, and inferences must be based on (3.2) and a model must be specified for $P(R_i|Y_i, X_i)$.

In applications, incomplete data may arise for a variety of reasons, and it is generally difficult to tell which missing data mechanism is reasonable. Flexible models encompassing various missing data mechanisms are therefore desirable. To this end, here we adopt models that accommodate a nonignorable missing data mechanism. Specifically, let H_{ij}^r denote the history of the missing indicators until the t_{j-1}, and $\lambda_{ij}^* = P(R_{ij} = 1|H_{ij}^r, Y_i, X_i)$. Regression models may be employed to link λ_{ij}^* with functions of Y_i, X_i and H_{ij}^r. Typically, a logistic regression model is commonly used with

$$\text{logit}(\lambda_{ij}^*) = Z_{ij}'\alpha, \tag{3.3}$$

where Z_{ij} is a vector which may include functions of $\{H_{ij}^r, Y_i, X_i\}$. As a typical case, we may write

$$\text{logit}(\lambda_{ij}^*) = \alpha_0 + \alpha_1 \cdot r_{i,j-1} + \alpha_2 \cdot r_{i,j-1}y_{i,j-1} + \alpha_3 \cdot y_{ij} + \alpha_x' \cdot X_{ij}$$

to reflect distinct missing data mechanisms. In this case, for example, $\alpha_2 = \alpha_3 = \alpha_x = 0$ leads to a MCAR mechanism, $\alpha_3 = 0$ and $\alpha_2 \neq 0$ corresponds to a MAR mechanism, and $\alpha_3 \neq 0$ represents a MNAR mechanism.

3.2.2 Model Identifiability

Conditional on the initial state, the complete data likelihood for subject i is written as

$$L_i(\theta; y_i, r_i) = \prod_{j=2}^{J} \left\{ (\lambda_{ij}^*)^{r_{ij}} (1-\lambda_{ij}^*)^{1-r_{ij}} \cdot \prod_{k'=1}^{K} \prod_{k=k'}^{K} \left(\mu_{ijk'k}^C\right)^{I(y_{i,j-1}=k', y_{ij}=k)} \right\}, \quad (3.4)$$

where $\theta = (\alpha', \beta')'$, and hence the observed data likelihood is

$$L_i(\theta; y_i^{(o)}, r_i) = \sum_{y_i^{(m)}} \left\{ \prod_{j=2}^{J} \left[(\lambda_{ij}^*)^{r_{ij}} (1-\lambda_{ij}^*)^{1-r_{ij}} \cdot \prod_{k'=1}^{K} \prod_{k=k'}^{K} \left(\mu_{ijk'k}^C\right)^{I(y_{i,j-1}=k', y_{ij}=k)} \right] \right\}. \quad (3.5)$$

With a MNAR mechanism, parameter identifiability is a central concern. Parameters governing the missing data process may be nonidentifiable due to incomplete information on unobserved responses and sensitivity analyses are therefore often conducted (e.g., Verbeke and Molenberghs, 2000). When the response process follows a progressive model, both response and missing data parameters are identifiable, even under a MNAR mechanism, provided standard conditions for (3.1) and (3.3) are satisfied. That is, if θ and $\tilde{\theta}$ are two values such that $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$ for any $(y_i^{(o)}, r_i)$, then $\theta = \tilde{\theta}$ must hold (Casella and Berger, 2002; Fitzmaurice, Laird and Zahner, 1996). The detailed proof is included in Section 3.5.

3.2.3 EM Algorithm

Here we describe an expectation-maximization (EM) algorithm to maximize (3.2). From (3.4), we obtain the log-likelihood for the complete data contribution

from subject i as

$$\begin{aligned}\ell_i(\theta; y_i, r_i) &= \sum_{j=2}^{J}\{r_{ij}\log\lambda_{ij}^* + (1-r_{ij})\log(1-\lambda_{ij}^*)\} \\ &\quad + \sum_{j=2}^{J}\sum_{k'=1}^{K}\sum_{k=k'}^{K} I(y_{i,j-1}=k', y_{ij}=k)\log\left(\mu_{ijk'k}^C\right), \qquad (3.6)\end{aligned}$$

where y_i is a realization of Y_i. In the expectation step (E-step), we require the conditional expectation

$$Q(\theta;\theta^{(h)}) = \sum_{i=1}^{n} Q_i(\theta;\theta^{(h)}),$$

where $Q_i(\theta;\theta^{(h)}) = E[\ell_i(\theta; y_i, r_i)|y_i^{(o)}, \theta^{(h)}] = \sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) \cdot \ell_i(\theta; y_i, r_i)$, and

$$w_i(y_i;\theta^{(h)}) = \frac{L_i(\theta^{(h)}; y_i^{(m)}, y_i^{(o)}, r_i)}{\sum_{y_i^{(m)}} L_i(\theta^{(h)}; y_i^{(m)}, y_i^{(o)}, r_i)}.$$

The maximization step (M-step) maximizes the function $Q(\theta;\theta^{(h)})$ with respect to the parameter θ, and a Newton-Raphson algorithm can be used for this purpose. Note that from (3.6), we can see that $Q(\theta;\theta^{(h)})$ can be maximized with respect to α and β by treating them separately since

$$Q(\theta;\theta^{(h)}) = Q_1(\alpha;\theta^{(h)}) + Q_2(\beta;\theta^{(h)}),$$

where

$$Q_1(\alpha;\theta^{(h)}) = \sum_{i=1}^{n}\sum_{j=2}^{J}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)})\{r_{ij}\log\lambda_{ij}^* + (1-r_{ij})\log(1-\lambda_{ij}^*)\}$$

and

$$Q_2(\beta;\theta^{(h)}) = \sum_{i=1}^{n}\sum_{j=2}^{J}\sum_{k'=1}^{J}\sum_{k=k'}^{J}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)}) I(y_{i,j-1}=k', y_{ij}=k)\log\left(\mu_{ijk'k}^C\right).$$

Iterating between the E and M steps until convergence leads to the maximimum likelihood estimator $\widehat{\theta}$.

Standard errors for these parameter estimates can be calculated using Louis's method (Louis, 1982), which partitions the complete data information into two parts: the information associated with the observed data and that associated with the missing data. Louis (1982) showed that a consistent estimate of the second derivative matrix could be calculated using

$$I(\theta|Y^{(o)}) = E[I(\theta|Y)|Y^{(o)}] - E[S(\theta|Y)S'(\theta|Y)|Y^{(o)}] + E[S(\theta|Y)|Y^{(o)}]E[S'(\theta|Y)|Y^{(o)}] \,,$$

where $S(\theta) = \sum_{i=1}^{n} \partial \ell_i(\theta; y_i, r_i)/\partial\theta$ is the score vector, and

$$I(\theta) = -\sum_{i=1}^{n} \partial^2 \ell_i(\theta; y_i, r_i)/\partial\theta\partial\theta'$$

is the observed information matrix (Horton and Laird, 1998). Therefore, the estimated observed information matrix of θ based on the observed data is given by

$$\Sigma(\hat{\theta}) = -\frac{\partial^2 Q(\hat{\theta};\hat{\theta})}{\partial\theta\partial\theta'} - \sum_{i=1}^{n}\sum_{y_i^{(m)}} w_i(y_i;\hat{\theta}) S_i(\hat{\theta}) S_i(\hat{\theta})' + \sum_{i=1}^{n} \left(\frac{\partial Q_i(\hat{\theta};\hat{\theta})}{\partial\theta}\right)\left(\frac{\partial Q_i(\hat{\theta};\hat{\theta})}{\partial\theta}\right)' ,$$

where $S_i(\hat{\theta}) = \partial \ell_i(\theta; y_i, r_i)/\partial\theta|_{\theta=\hat{\theta}}$. The estimate of the asymptotic covariance matrix of $\hat{\beta}$ is the lower $p_2 \times p_2$ block of $[\Sigma(\hat{\theta})]^{-1}$, where p_2 is the dimension of β.

3.2.4 Simulation Studies

Now we conduct a simulation study to assess the performance of the proposed method by examining finite sample biases and coverage probabilities for parameter estimates. Our primary aim here is to compare the proposed method with the methods which do not incorporate the missing data mechanism. We show that

large biases and poor coverage probabilities in the parameter estimates can result if the missing data mechanism is not accounted for when it should be.

Set $n = 1000, K = 3$ and $J = 4$. Two thousand simulations are run for each parameter configuration. The response vector is generated from the conditional model

$$\log\left(\frac{\mu^C_{ijk'k}}{\mu^C_{ijk'k'}}\right) = \beta_{k'k0} + \beta_{k'k1}X_{ij1} + \beta_{k'k2}X_{ij2} + \beta_{k'k3}X_{ij3}, \quad k' < k \leq 3,$$

for $j = 2, 3, 4$, where $X_{ij1} = X_{i1}$ represents a time invariant treatment indicator, generated from the Binomial distribution Bin(1,0.5). Here $X_{ij2} = I(j = 3)$ and $X_{ij3} = I(j = 4)$ facilitate the temporal effects. The true values for the parameters are set as $\beta_{120} = \text{logit}(0.5)$, $\beta_{121} = \log(0.5)$, $\beta_{122} = \log(u_\beta)$, $\beta_{123} = 2\log(u_\beta)$, $\beta_{130} = \text{logit}(0.25)$, $\beta_{131} = \log(0.5)$, $\beta_{132} = \log(u_\beta)$, $\beta_{133} = 2\log(u_\beta)$, $\beta_{230} = \text{logit}(0.5)$, $\beta_{231} = \log(0.5)$, and $\beta_{233} = 2\log(u_\beta)$. We take $u_\beta = 1$ or 1.2 to indicate whether the response model is dependent on the temporal effects.

For the missing data process, assume the logistic regression model

$$\begin{aligned}\text{logit}(\lambda^*_{ij}) &= \alpha_0 + \alpha_1(1 - r_{i,j-1}) + \alpha_2 r_{i,j-1}I(Y_{i,j-1} = 2) + \alpha_3 I(Y_{ij} = 2) \\ &\quad + \alpha_4 r_{i,j-1}X_{ij1}I(Y_{i,j-1} = 2) + \alpha_5 X_{ij1}I(Y_{ij} = 2), \end{aligned} \tag{3.7}$$

for $j = 2, 3, \ldots, J$. Set $\alpha_0 = \text{logit}(0.7)$, and $\alpha_1 = \log(0.75)$. When considering an MAR, set $\alpha_3 = \alpha_5 = 0$, $\alpha_2 = \log(u_\alpha)$, and $\alpha_4 = \log(2)$; with an MNAR, set $\alpha_3 = \log(u_\alpha)$, $\alpha_5 = \log(2)$ and $\alpha_2 = \alpha_4 = 0$. To alter the missing data proportion, take $u_\alpha = 0.5$ or 2.

Data generation procedure for the responses is as follows: assume $Y_{i1} = 1$, i.e., all the subjects are in state 1 at the entrance of the study, and given the true parameter vectors $\beta_{k'k} = (\beta_{k'k0}, \beta_{k'k1}, \beta_{k'k2}, \beta_{k'k3})'$,

1. generate the covariate vector $X_{ij} = (1, X_{ij1}, X_{ij2}, X_{ij3})'$, $j = 2, 3, \ldots, J$,

2. given $Y_{i,j-1} = y_{i,j-1}$, calculate the transition probabilities $\mu^C_{ijy_{i,j-1}k}$, $y_{i,j-1} < k$, which is given by

$$\mu^C_{ijy_{i,j-1}k} = \frac{\exp(X'_{ij}\beta_{y_{i,j-1}k})}{1 + \sum_{k=y_{i,j-1}+1}^{K} \exp(X'_{ij}\beta_{y_{i,j-1}k})},$$

3. generate the response Y_{ij} from the discrete distribution $P(Y_{ij} = k|X_{ij}) = \mu^C_{ijy_{i,j-1}k}$ for $k > y_{i,j-1}$.

For the missing data process, assume $R_{i1} = 1$. Given the true parameter α and $R_{i,j-1} = r_{i,j-1}$, $j > 1$, the conditional probabilities λ^*_{ij} can be calculated from (3.7), and we generate the missing indicator R_{ij} via the binomial distribution $\text{Bin}(1, \lambda^*_{ij})$.

Here we conduct two methods, one is the proposed method, and the other is available data analysis, in which we pick up the observations that are consecutively observed and analyze them using the "GENMOD" command in SAS. In the simulation studies, we assume all the models are correctly specified. The simulation results are reported in Tables 3.1-3.8, where SEL denotes the average standard error calculated based on Louis method, ESE is the empirical standard errors and CP represents the coverage probability of the 95% confidence intervals. In the simulation, we considered two missingness proportions – 45% and 30%, corresponding to $u_\alpha = 0.5$ and 2, respectively. It is seen that, as the missingness proportion increases, biases of the parameters increase. It is not surprising that, under MAR, the biases of $\hat{\beta}$ for both the proposed method and the available data analysis are fairly small, the coverage probabilities are very close to 95%, and the empirical standard errors are in good agreement with the standard errors obtained by the Louis method. In contrast, for the missing not at random cases, the biases of both $\hat{\beta}$ and $\hat{\alpha}$ are very

small for the proposed method, but the biases of $\hat{\beta}$ in the available data analysis are remarkable. The coverage probabilities for the proposed method are in good agreement with the nominal level 95%, while the available data analysis produces coverage probabilities that are far from the nominal level.

3.3 Modeling Transition Intensities

Besides the transition probabilities, sometimes interest lies in the transition intensities. In this section, I consider the continuous time progressive model via the transition intensities. In general, one could develop stochastic models for the assessment times. In cohort studies, clinical assessments may be scheduled at roughly equal intervals (e.g. annually), but patients may choose when they want to visit clinics for clinical examination according to their degree of disease activity. This creates a situation somewhat akin to incomplete data in longitudinal studies when data may be missing at random (MAR) if missing status depends on the observed, typically past, responses, or missing not at random (MNAR), where the missing status may depend on the latent disease status. The latter situation is particularly difficult to deal with in general and in most settings analysts must rely on sensitivity analyses to examine the possible effect of this type of observation scheme. We consider, however, progressive models for chronic disease processes, which by their progressive nature, are convenient for jointly modeling the disease and observation processes. We provide a general method to handle this type of data. Maximum likelihood methods are used with parameter estimation carried out via an EM algorithm (Dempster et al., 1977), and variance estimation is performed using Louis' (1982) method.

Table 3.1: Simulation results under MAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with no temporal effect (i.e. $u_\beta = 1.0$)

Transition Model			EM				Available Data		
Transition	Parms	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	0.002	0.101	0.099	0.955	-0.001	0.106	0.948
	β_{121}	log(0.5)	0.005	0.120	0.117	0.954	0.003	0.134	0.951
	β_{122}	log(1.0)	-0.002	0.149	0.155	0.945	-0.003	0.163	0.953
	β_{123}	2log(1.0)	0.002	0.193	0.191	0.956	-0.002	0.209	0.961
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.009	0.145	0.144	0.956	0.0091	0.153	0.953
	β_{131}	log(0.5)	0.002	0.180	0.178	0.955	0.0046	0.203	0.946
	β_{132}	log(1.0)	-0.006	0.232	0.233	0.950	-0.0100	0.242	0.951
	β_{133}	2log(1.0)	0.007	0.313	0.312	0.957	-0.0119	0.325	0.951
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.000	0.162	0.164	0.952	-0.0042	0.194	0.952
	β_{231}	log(0.5)	0.009	0.192	0.192	0.953	0.0194	0.233	0.956
	β_{233}	2log(1.0)	-0.005	0.220	0.220	0.950	-0.0087	0.234	0.956
Missing Data Model									
	α_0	logit(0.7)	0.002	0.053	0.051	0.951			
	α_1	log(0.75)	-0.003	0.104	0.095	0.948			
	α_2	log(0.5)	-0.001	0.130	0.131	0.950			
	α_4	log(2.0)	-0.000	0.191	0.190	0.956			

† Absolute bias.

Table 3.2: Simulation results under MAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with temporal effect (i.e. $u_\beta = 1.2$)

Transition Model			EM				Available Data		
Transition	Parameter	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.002	0.102	0.097	0.957	-0.001	0.105	0.952
	β_{121}	log(0.5)	0.002	0.125	0.121	0.954	0.001	0.140	0.945
	β_{122}	log(1.2)	0.006	0.148	0.150	0.948	0.004	0.160	0.954
	β_{123}	2log(1.2)	0.009	0.196	0.199	0.950	0.006	0.223	0.953
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.006	0.146	0.146	0.953	0.006	0.155	0.948
	β_{131}	log(0.5)	0.002	0.179	0.175	0.955	0.002	0.198	0.954
	β_{132}	log(1.2)	-0.007	0.227	0.225	0.953	-0.003	0.234	0.947
	β_{133}	2log(1.2)	0.016	0.314	0.309	0.957	-0.001	0.318	0.955
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.003	0.163	0.165	0.948	0.004	0.196	0.951
	β_{231}	log(0.5)	0.003	0.190	0.192	0.948	0.001	0.236	0.951
	β_{133}	2log(1.2)	-0.008	0.216	0.218	0.950	-0.008	0.237	0.949
Missing Data Model									
	α_0	logit(0.7)	-0.000	0.054	0.051	0.953			
	α_1	log(0.75)	-0.003	0.105	0.098	0.952			
	α_2	log(0.5)	0.003	0.129	0.129	0.948			
	α_4	log(2.0)	-0.004	0.190	0.185	0.955			

† Absolute bias.

Table 3.3: Simulation results under MAR: about 30% missingness (i.e. $u_\alpha = 2$) with no temporal effect (i.e. $u_\beta = 1.0$)

Transition Model			EM				Available Data		
Transition	Parms	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.000	0.101	0.097	0.956	-0.000	0.104	0.948
	β_{121}	log(0.5)	-0.001	0.120	0.117	0.954	0.001	0.135	0.950
	β_{122}	log(1.0)	0.001	0.149	0.146	0.949	0.001	0.155	0.960
	β_{123}	2log(1.0)	0.001	0.195	0.196	0.948	-0.000	0.217	0.948
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.003	0.145	0.141	0.956	0.003	0.151	0.957
	β_{131}	log(0.5)	-0.001	0.180	0.179	0.952	0.000	0.203	0.948
	β_{132}	log(1.0)	0.003	0.231	0.235	0.956	-0.002	0.242	0.955
	β_{133}	2log(1.0)	0.013	0.302	0.295	0.961	-0.005	0.310	0.960
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.002	0.144	0.146	0.947	0.000	0.161	0.951
	β_{231}	log(0.5)	0.004	0.175	0.176	0.948	0.003	0.194	0.953
	β_{233}	2log(1.0)	-0.002	0.183	0.185	0.948	-0.001	0.193	0.950
Missing Data Model									
	α_0	logit(0.7)	-0.000	0.052	0.051	0.956			
	α_1	log(0.75)	0.001	0.106	0.101	0.951			
	α_2	log(2.0)	-0.006	0.162	0.162	0.947			
	α_4	log(2.0)	-0.000	0.268	0.268	0.958			

† Absolute bias.

Table 3.4: Simulation results under MAR: about 30% missingness (i.e. $u_\alpha = 2.0$) with temporal effect (i.e. $u_\beta = 1.2$)

Transition Model			EM				Available Data		
Transition	Parameter	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.003	0.104	0.096	0.958	-0.002	0.103	0.958
	β_{121}	log(0.5)	0.001	0.118	0.115	0.954	-0.000	0.134	0.952
	β_{122}	log(1.2)	0.002	0.148	0.148	0.947	0.001	0.159	0.952
	β_{123}	2log(1.2)	-0.001	0.197	0.196	0.955	0.002	0.220	0.955
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.001	0.146	0.142	0.957	0.002	0.151	0.952
	β_{131}	log(0.5)	0.006	0.189	0.183	0.961	0.006	0.209	0.938
	β_{132}	log(1.2)	0.009	0.228	0.228	0.951	0.004	0.235	0.955
	β_{133}	2log(1.2)	0.007	0.310	0.307	0.953	-0.007	0.320	0.956
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.002	0.144	0.149	0.943	0.002	0.167	0.943
	β_{231}	log(0.5)	0.003	0.172	0.175	0.945	0.002	0.192	0.951
	β_{233}	2log(1.2)	-0.004	0.180	0.181	0.946	-0.005	0.189	0.953
Missing Data Model									
	α_0	logit(0.7)	0.001	0.053	0.052	0.955			
	α_1	log(0.75)	-0.001	0.107	0.103	0.949			
	α_2	log(2.0)	-0.003	0.161	0.164	0.952			
	α_4	log(2.0)	-0.015	0.265	0.265	0.956			

† Absolute bias.

Table 3.5: Simulation results under MNAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with no temporal effect (i.e. $u_\beta = 1.0$)

Transition Model			EM				Available Data		
Transition	Parms	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.015	0.103	0.110	0.943	0.259	0.111	0.359
	β_{121}	log(0.5)	-0.001	0.125	0.119	0.957	-0.260	0.138	0.547
	β_{122}	log(1.0)	0.026	0.152	0.159	0.944	0.004	0.167	0.956
	β_{123}	2log(1.0)	0.024	0.197	0.199	0.954	-0.005	0.219	0.950
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.006	0.147	0.142	0.959	0.004	0.150	0.950
	β_{131}	log(0.5)	-0.002	0.189	0.182	0.958	0.001	0.197	0.958
	β_{132}	log(1.0)	0.016	0.233	0.239	0.943	0.004	0.244	0.949
	β_{133}	2log(1.0)	0.015	0.313	0.310	0.960	-0.007	0.316	0.960
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.012	0.171	0.178	0.942	-0.266	0.213	0.751
	β_{231}	log(0.5)	0.024	0.198	0.200	0.944	0.273	0.241	0.809
	β_{233}	2log(1.0)	-0.007	0.215	0.218	0.950	0.000	0.249	0.950
Missing Data Model									
	α_0	logit(0.7)	-0.016	0.080	0.081	0.945			
	α_1	log(0.75)	-0.007	0.120	0.109	0.953			
	α_3	log(0.5)	0.025	0.203	0.174	0.944			
	α_5	log(2.0)	0.029	0.148	0.163	0.955			

† Absolute bias.

Table 3.6: Simulation results under MNAR: about 45% missingness (i.e. $u_\alpha = 0.5$) with temporal effect (i.e. $u_\beta = 1.2$)

Transition Model			EM				Available Data		
Transition	Parameter	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.017	0.106	0.110	0.943	0.265	0.113	0.348
	β_{121}	log(0.5)	-0.008	0.129	0.123	0.956	-0.262	0.139	0.535
	β_{122}	log(1.2)	0.025	0.151	0.155	0.946	0.003	0.164	0.961
	β_{123}	2log(1.2)	0.027	0.200	0.205	0.947	-0.000	0.228	0.948
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.010	0.149	0.151	0.955	0.007	0.157	0.946
	β_{131}	log(0.5)	-0.001	0.192	0.186	0.958	0.005	0.202	0.943
	β_{132}	log(1.2)	0.006	0.229	0.228	0.954	-0.007	0.232	0.963
	β_{133}	2log(1.2)	0.015	0.315	0.305	0.958	-0.005	0.315	0.953
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.026	0.168	0.175	0.939	-0.255	0.205	0.772
	β_{231}	log(0.5)	0.019	0.189	0.192	0.943	0.267	0.232	0.813
	β_{233}	2log(1.2)	-0.019	0.211	0.211	0.955	-0.004	0.238	0.957
Missing Data Model									
	α_0	logit(0.7)	-0.021	0.086	0.085	0.952			
	α_1	log(0.75)	-0.009	0.124	0.092	0.956			
	α_3	log(0.5)	0.037	0.224	0.182	0.945			
	α_5	log(2.0)	0.028	0.152	0.174	0.954			

† Absolute bias.

Table 3.7: Simulation results under MNAR: about 30% missingness (i.e. $u_\alpha = 2.0$) with no temporal effect (i.e. $u_\beta = 1.0$)

Transition Model			EM				Available Data		
Transition	Parms	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.007	0.100	0.103	0.941	-0.163	0.100	0.632
	β_{121}	log(0.5)	-0.005	0.119	0.118	0.953	-0.091	0.128	0.882
	β_{122}	log(1.0)	0.007	0.135	0.139	0.945	-0.002	0.152	0.947
	β_{123}	2log(1.0)	0.008	0.173	0.176	0.944	-0.005	0.202	0.945
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.011	0.143	0.141	0.948	0.009	0.152	0.951
	β_{131}	log(0.5)	-0.003	0.181	0.176	0.959	-0.001	0.201	0.954
	β_{132}	log(1.0)	0.012	0.230	0.232	0.950	0.002	0.240	0.954
	β_{133}	2log(1.0)	0.012	0.310	0.313	0.957	-0.007	0.330	0.954
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.000	0.144	0.151	0.940	0.159	0.154	0.821
	β_{231}	log(0.5)	0.018	0.170	0.174	0.943	0.100	0.186	0.928
	β_{233}	2log(1.0)	0.004	0.179	0.181	0.945	0.003	0.188	0.939
Missing Data Model									
	α_0	logit(0.7)	-0.009	0.067	0.065	0.956			
	α_1	log(0.75)	-0.004	0.117	0.114	0.955			
	α_3	log(2.0)	0.003	0.217	0.218	0.952			
	α_5	log(2.0)	0.041	0.303	0.306	0.956			

† Absolute bias.

Table 3.8: Simulation results under MNAR: about 30% missingness (i.e. $u_\alpha = 2.0$) with temporal effect (i.e. $u_\beta = 1.2$)

Transition Model			EM				Available Data		
Transition	Parameter	True	†BIAS	SEL	ESE	CP	†BIAS	ESE	CP
$1 \rightarrow 2$	β_{120}	logit(0.5)	-0.006	0.100	0.108	0.941	-0.163	0.101	0.651
	β_{121}	log(0.5)	-0.009	0.123	0.122	0.951	-0.003	0.130	0.878
	β_{122}	log(1.2)	0.011	0.140	0.141	0.947	0.004	0.155	0.946
	β_{123}	2log(1.2)	0.016	0.177	0.181	0.947	0.001	0.209	0.951
$1 \rightarrow 3$	β_{130}	logit(0.25)	0.001	0.143	0.137	0.958	0.001	0.146	0.963
	β_{131}	log(0.5)	0.006	0.183	0.172	0.959	0.007	0.197	0.957
	β_{132}	log(1.2)	0.015	0.226	0.226	0.954	0.006	0.233	0.956
	β_{133}	2log(1.2)	0.012	0.310	0.305	0.958	-0.009	0.321	0.955
$2 \rightarrow 3$	β_{230}	logit(0.5)	0.005	0.140	0.149	0.942	0.162	0.152	0.833
	β_{231}	log(0.5)	0.013	0.167	0.168	0.948	0.093	0.176	0.935
	β_{233}	2log(1.2)	-0.002	0.176	0.173	0.951	-0.001	0.181	0.954
Missing Data Model									
	α_0	logit(0.7)	-0.012	0.069	0.079	0.947			
	α_1	log(0.75)	-0.006	0.118	0.114	0.954			
	α_3	log(2.0)	-0.003	0.227	0.300	0.951			
	α_5	log(2.0)	0.042	0.317	0.485	0.949			

† Absolute bias.

3.3.1 Continuous Time Progressive Multi-State Models

Suppose there are K states and the transition direction is irreversible. Figure 3.2 is an illustrative diagram of a K-state progressive transition model. Let $Y(t)$ represent the state occupied at time $t \geq 0$, and $\mathcal{H}(t) = \{Y(s), 0 \leq s < t\}$ denote the history of the response process which records the states occupied over the interval $[0, t)$. The transition probability function is written generally as $P(Y(s+t) = k|Y(s) = k', \mathcal{H}(s))$ for $s, t > 0$, and $k \geq k'$, but under a Markov model this simplifies to $P(Y(s+t) = k|Y(s) = k')$, which we denote compactly as $P_{k'k}(s, s+t)$, $k \geq k'$.

Figure 3.2: A diagram of K-state progressive process

$$\boxed{1} \longrightarrow \boxed{2} \longrightarrow \boxed{3} \longrightarrow \cdots \longrightarrow \boxed{\text{K}}$$

The corresponding transition intensity from state k to state $k+1$ at time t is

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(Y(t+\Delta t) = k+1|Y(t) = k)}{\Delta t}, \quad k = 1, \ldots, K-1,$$

(Cox and Miller, 1977). A multi-state progressive model with state space $\{1, 2, \ldots, K\}$ can then be described via the following transition intensity matrix, $Q(t)$:

$$Q(t) = \begin{pmatrix} -\lambda_1(t) & \lambda_1(t) & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2(t) & \lambda_2(t) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{K-1}(t) & \lambda_{K-1}(t) \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Under a time-homogeneous Markov model, let $\lambda_k(t) = \lambda_k$, $k = 1, \ldots, K-1$, and it follows from stationarity that $P_{k'k}(s, s+t) = P_{k'k}(0, t)$, which we may now

write simply as $P_{k'k}(t)$. Let $P(t)$ denote the $K \times K$ matrix with (k', k) element $P_{k'k}(t)$. We assume $\lambda_1, \ldots, \lambda_{K-1}$ are distinct, and let $\lambda = (\lambda_1, \ldots, \lambda_{K-1})'$. For a time-homogeneous model the transition probability from state k' to state k over $[0, t]$ is given by

$$P_{k'k}(t) = \begin{cases} \sum_{j=k'}^{k} C(k', j, k; \lambda) \exp(-\lambda_j t), & k' \leq k, \\ 0, & k' > k, \end{cases}$$

where the coefficients are given by

$$C(k', j, k; \lambda) = \frac{\prod_{h=k'}^{k-1} \lambda_h}{\prod_{h=k', h \neq j}^{k} (\lambda_h - \lambda_j)}$$

for $k' \leq j \leq k$, and $C(j, j, j; \lambda) = 1, j = 1, 2, \ldots, K$ (Satten, 1999). In the simulations and application that follow we focus on time-homogeneous Markov models, but extensions which accommodate nonhomogeneous Markov models can be developed in the same spirit, and so we retain the dependence on t in the following remarks.

To model the dependence of the transition intensities on prognostic variables, we may incorporate covariates in the preceding formulation by expressing the transition intensities as functions of time (in the nonhomogenous case) and the covariates. That is, let $\lambda_k(t) = g_k(t, X)$ for some non-negative known function $g_k(\cdot, \cdot)$, $k = 1, \ldots, K-1$, where X represents the covariate vector. For a given individual i, we often adopt models of the form

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp(X_{ik}' \beta_k), \quad k = 1, \ldots, K-1, \tag{3.8}$$

where the $\lambda_{0k}(t)$ are the baseline transition intensities which may or may not depend on t, and β_k is a vector of regression coefficients associated with the covariates of interest, $X_{ik}, k = 1, \ldots, K-1$. This setup permits the baseline transition intensities

and the regression coefficients to vary across the possible transitions. We let $X_i = (X'_{i1}, \ldots, X'_{i,K-1})'$ and $\beta = (\beta_1, \ldots, \beta_{K-1})'$ denote the vector of all covariates and regression coefficients.

With continuous time models and observation schemes, the response process $\{Y(t), t > 0\}$ may be observed at any time point t over the period of observation. In practice, however, individuals are often observed at random, individual-specific times, which are not necessarily evenly spaced, and their states are determined at these visit times. Let $t_{i1} < t_{i2} < \cdots < t_{iJ_i}$ denote variable assessment times for subject i, $H^a_{ij} = \{t_{ik}, k = 1, \ldots, j-1\}$, $H^y_{ij} = \{Y_i(t_{ik}), k = 1, \ldots, j-1\}$, and $H_{ij} = \{(t_{ik}, Y_i(t_{ik})), k = 1, \ldots, j-1\}$. If we condition on the initial time of assessment and the initial state, the full observed data likelihood contribution from subject i, suppressing dependence on the covariates, is then

$$L_i = \prod_{j=2}^{J_i} P(t_{ij}, Y_i(t_{ij}) | H_{ij}) = \prod_{j=2}^{J_i} P(Y_i(t_{ij}) | H_{ij}) \prod_{j=2}^{J_i} P(t_{ij} | Y_i(t_{ij}), H_{ij}) \,. \tag{3.9}$$

The model for the underlying stochastic process does not typically feature a dependence on the previous assessment times, and so $P(Y_i(t_{ij})|H_{ij}) = P(Y_i(t_{ij})|H^y_{ij})$ is a quite natural assumption. Indeed we usually desire to base inferences strictly on the product of such terms. If $P(t_{ij}|Y_i(t_{ij}), H_{ij})$ does not depend on $Y_i(t_{ij})$ (i.e. the time of the assessment does not depend on the state of the underlying process) then we can treat $\prod_{j=2}^{J_i} P(Y_i(t_{ij})|H^y_{ij})$ as if it were the probability of the observed states, conditional on the assessment times, and this is typically an implicit assumption in standard analyses. If, on the other hand, $P(t_{ij}|Y_i(t_{ij}), H_{ij})$ does depend on $Y_i(t_{ij})$, then we must consider the full likelihood based on (3.9). In this case, one needs to model the conditional distributions of the examination times (i.e. $P(t_{ij}|Y_i(t_{ij}), H_{ij})$) which can be challenging. In this chapter we consider

the problem in which subjects are scheduled to be examined at pre-specified assessment times denoted $a_1, a_2, \ldots, a_J$, and adopt the convenient framework commonly employed to handle incomplete longitudinal data.

Let R_{ij} be the indicator random variable, which equals 1 if response $Y_i(a_j)$ is observed and 0 otherwise. Let $R_i = (R_{i1}, R_{i2}, \ldots, R_{iJ})'$, and $r_i = (r_{i1}, r_{i2}, \ldots, r_{iJ})'$ be a realization of R_i. Here we assume all the subjects are observed at the initial enrollment, i.e., $R_{i1} = 1$. Let $\lambda^*_{ij} = P(R_{ij} = 1|H^r_{ij}, Y_i, X_i)$ be the conditional probability, where H^r_{ij} denotes the history of the missing indicators until the $(j-1)$st time point. When models with a first order dependence are of interest, we write $\lambda^*_{ij} = P(R_{ij} = 1|R_{i,j-1}, Y_i, X_i)$. A logistic regression model is commonly employed to postulate the conditional probability λ^*_{ij}, i.e.,

$$\text{logit}(\lambda^*_{ij}) = Z'_{ij}\alpha,$$

where α is a parameter vector, and Z_{ij} is a covariate vector featuring various missingness.

Let $\theta = (\alpha', \beta')'$. Then the likelihood for the complete data is given by $L(\theta) = \prod_{i=1}^n L_i(\theta; y_i)$, where

$$\begin{aligned} L_i(\theta; y_i) &= P(R_i = r_i|Y_i = y_i, X_i; \alpha)P(Y_i = y_i|X_i; \beta) \\ &\propto \prod_{j=2}^{J} \left\{ (\lambda^*_{ij})^{r_{ij}}(1-\lambda^*_{ij})^{1-r_{ij}} \cdot \prod_{k'=1}^{K}\prod_{k=k'}^{K} \{P_{ik'k}(a_j - a_{j-1})\}^{I(Y_i(a_{j-1})=k', Y_i(a_j)=k)} \right\} \end{aligned} \tag{3.10}$$

or equivalently, its logarithm is

$$\begin{aligned} \ell_i(\theta; y_i) &= \sum_{j=2}^{J}\{r_{ij}\log\lambda^*_{ij} + (1-r_{ij})\log(1-\lambda^*_{ij})\} \\ &+ \sum_{j=2}^{J}\sum_{k'=1}^{K}\sum_{k=k'}^{K} I(Y_i(a_{j-1}) = k', Y_i(a_j) = k)\log P_{ik'k}(a_j - a_{j-1}), \end{aligned} \tag{3.11}$$

with $P_{ik'k}(a_j - a_{j-1}) = P(Y_i(a_j) = k|Y_i(a_{j-1}) = k', X_i)$.

3.3.2 Asymptotic Bias Under Dependent Inspection

In the presence of missing values, we may base inference about θ on using (3.10) or (3.11). The detail will be presented in Section 3.3.3. Here we investigate the impact of ignoring missingness, or the fact of dependent inspection. Specifically, we employ the available data analysis and the complete case analysis that are often used in practice due to their simplicity of implementation. We investigate this problem through application of the theory of misspecified models.

Let $\ell^*(\beta^*) = \sum_{i=1}^n \log P(Y_i^{(*)}|X_i)$ be the naive log-likelihood function where $Y_i^{(*)}$ represents the available data or the complete case data. Here β^* is used to stress that the associated parameter may be different from the parameter of interest β. Solving

$$S^*(\beta^*) = \frac{\partial \ell^*(\beta^*)}{\partial \beta^*} = 0$$

leads to a naive estimate $\hat{\beta}^*$.

White (1982) showed that $\hat{\beta}^*$ converges to β^* almost surely, where β^* solves

$$E_{Y,R,X}[S^*(\beta^*)] = 0.$$

Here $E_{Y,R,X}$ denotes the expectation taken with respect to the joint distribution (3.11) of (Y, R, X) which depends on β and α. In general, it is difficult to obtain an analytical expression for β^* by solving this equation. Instead, to understand the magnitude of the bias, or difference $\beta^* - \beta$, we proceed with a numerical study. Specifically, we solve the equation

$$\sum_{d \in \mathcal{D}} S^*(\beta^*) \cdot P(d; \alpha, \beta) = 0,$$

where $\mathcal{D}$ is the sample space for $D = (R, Y, X)$, and $P(d; \alpha, \beta)$ is the true probability

of observing the realized value d of D. This can be easily solved using the standard software.

We consider the case with $K = 3$ states and $J = 3$ or $J = 5$ time points, and assume that the intensity function for transitions from state k to state $k + 1$ is

$$\lambda_k = \lambda_{0k} e^{\beta_k x}, \quad k = 1, \ldots, K-1, \tag{3.12}$$

where the baseline function is modeled as $\lambda_{0k} = \lambda_0 e^{\gamma(k-1)}$, and x is generated from Bin(1,0.5), representing a treatment indicator, for instance. The true values of the coefficients are taken to be $\lambda_0 = 0.5, \gamma = 0.2$, and $\beta_k = 1/k, k = 1, \ldots, K-1$. The study duration, τ, is selected such that $P(T < \tau) = 0.9$ where T denotes the time to entry of state K. The assessment time points are chosen as $a_j = (j-1)/(J-1) \cdot \tau$, $j = 1, \ldots, J$, equally cutting the interval $[0, \tau]$. We assume that all of the subjects are observed at the first assessment time, which is plausible in settings where the observation process begins upon entry to a clinic. The conditional probabilities λ^*_{ij} are modeled as

$$\text{logit}(\lambda^*_{ij}) = \alpha_0 + \alpha_1(1 - r_{i,j-1}) + \alpha_2 y_i(a_{j-1}) + \alpha_3(y_i(a_j) - y_i(a_{j-1})) + \alpha_4 x_i. \tag{3.13}$$

Note that $\alpha_2 \neq 0$ or $\alpha_3 \neq 0$ represents a nonignorable missing mechanism (Little and Rubin, 1987; Laird, 1988).

In the study considered here we set $\alpha_0 = \log(4)$, $\alpha_1 = \log(0.75)$ and $\alpha_4 = \log(2)$. The parameters α_2 and α_3 are changed from $\log(0.5)$ to $\log(2.0)$ to reflect varying degrees of the missing data proportion and the dependence of the missingness on the previous observation and the present observation. For example, as α_2 and α_3 increase, the missingness proportion reduces, and the dependence on unobserved data becomes weaker (if $\alpha_2 \leq 0$ and $\alpha_3 \leq 0$) or stronger (if $\alpha_2 \geq 0$ and $\alpha_3 \geq 0$).

The results are displayed in Figures 3.3 – 3.5. Not surprisingly, the asymptotic biases for the complete case analysis are bigger than for the available data analysis, and as the absolute values of α_2 and α_3 decrease, the biases become smaller. Note that, for the baseline intensity function, the biases are all negative when $\alpha_2 < 0$ and $\alpha_3 < 0$, which suggests the naive baseline intensity estimates are underestimates. However, when $\alpha_2 > 0$ and $\alpha_3 > 0$, the biases of the baseline intensity function are indicating overestimated results; yet the magnitudes are very small.

3.3.3 Maximum Likelihood Estimate and EM Algorithm

In this subsection we develop an EM algorithm for valid inference. In the E step, we construct the conditional expectation, at the hth iteration,

$$Q(\theta;\theta^{(h)}) \;=\; \sum_{i=1}^{n} Q_i(\theta;\theta^{(h)}),$$

where $Q_i(\theta;\theta^{(h)}) = E[\ell_i(\theta,y_i)|Y_i^{(o)},\theta^{(h)}] = \sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)})\cdot\ell_i(\theta,y_i)$, y_i is written as $(y_i^{(m)},y_i^{(o)})$ to explicitly indicate missing and observed components, and

$$w_i(y_i;\theta^{(h)}) = \frac{L_i(\theta^{(h)};y_i^{(m)},y_i^{(o)})}{\sum_{y_i^{(m)}} L_i(\theta^{(h)};y_i^{(m)},y_i^{(o)})},$$

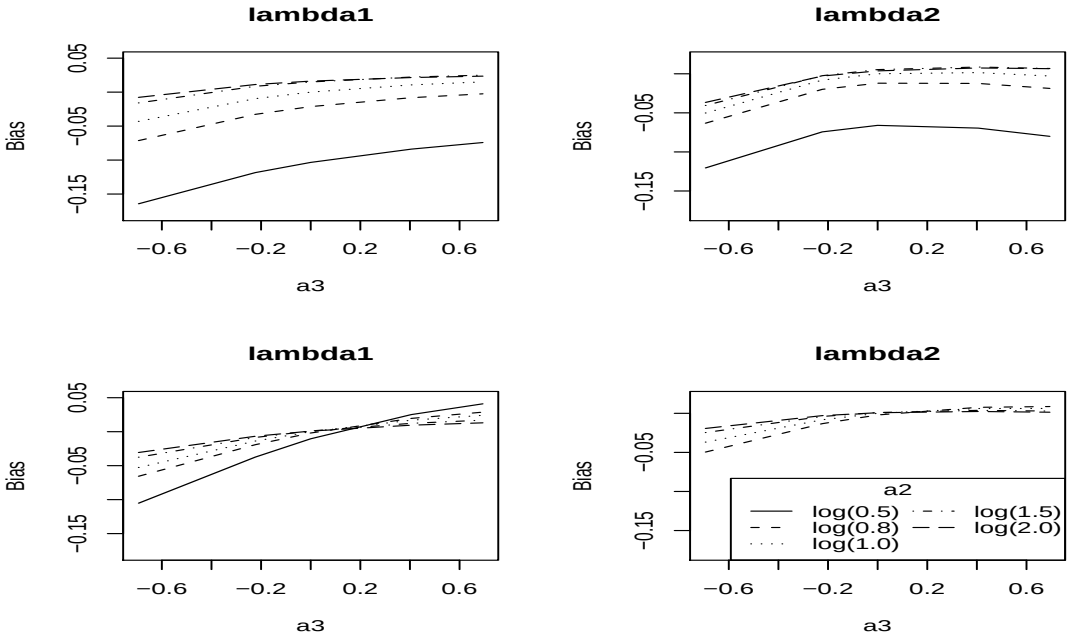
facilitating the conditional probability of the missing data given the observed data. Here $L_i(\cdot)$ and $\ell_i(\cdot)$ are the complete data likelihood and log-likelihood given by (3.10) and (3.11), respectively. The dependence on the covariates is suppressed in the notation.

We note, by (3.11), that parameters α and β can be separated in the conditional expectation $Q(\theta;\theta^{(h)})$ with the form:

$$Q(\theta;\theta^{(h)}) = Q_1(\alpha;\theta^{(h)}) + Q_2(\beta;\theta^{(h)}),$$

Figure 3.3: Asymptotic bias for missing not at random without covariates with 3 states. The first and the third horizonal rows are plots for the complete case analysis and the second and the fourth horizonal rows are plots for the available data analysis, with 3 and 5 observations, respectively.

$K = 3, J = 3$



$K = 3, J = 5$

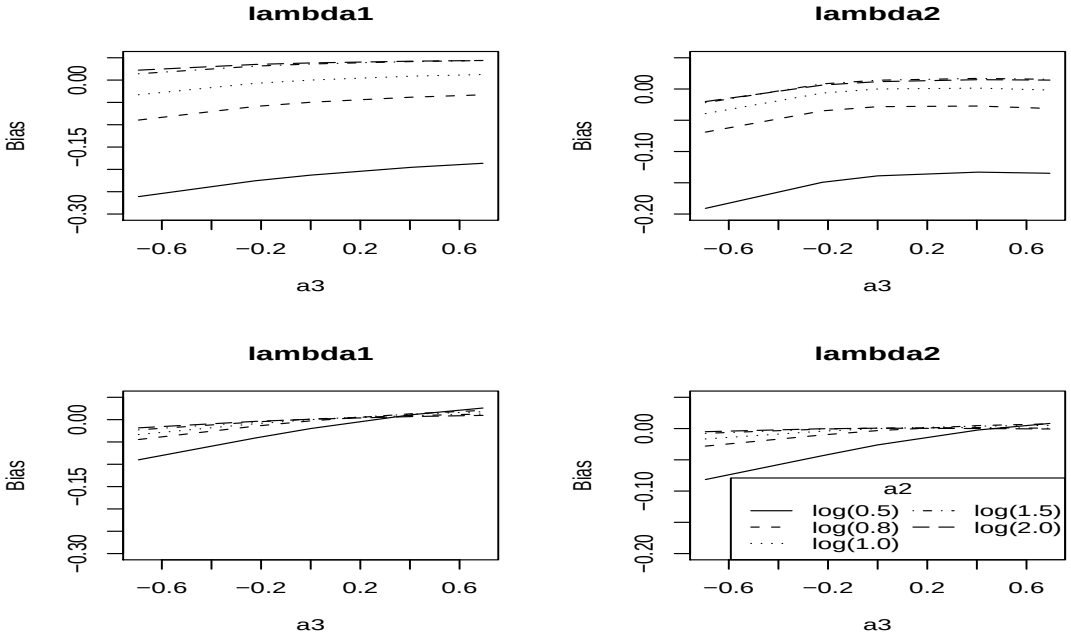
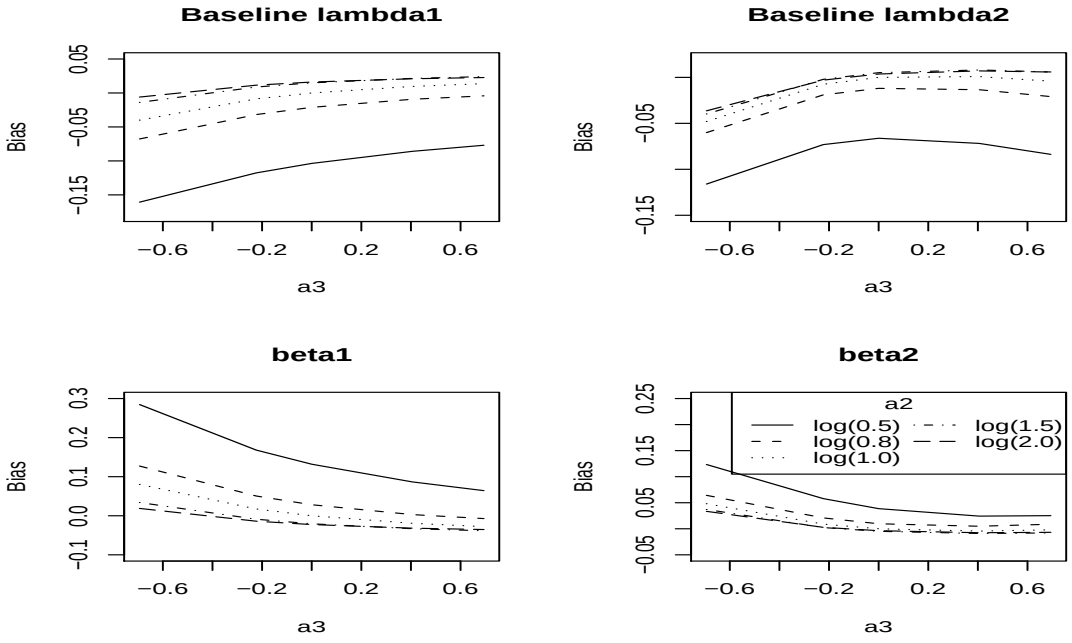


Figure 3.4: Asymptotic bias for missing not at random with one covariate with 3 states 3 observations.

Complete Case



Available Data

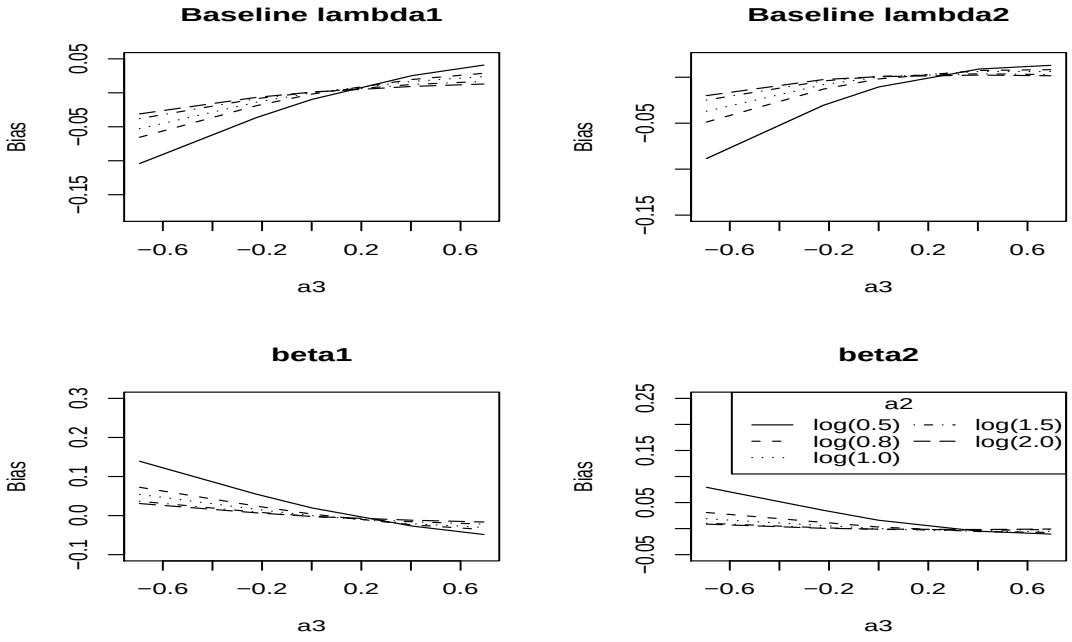
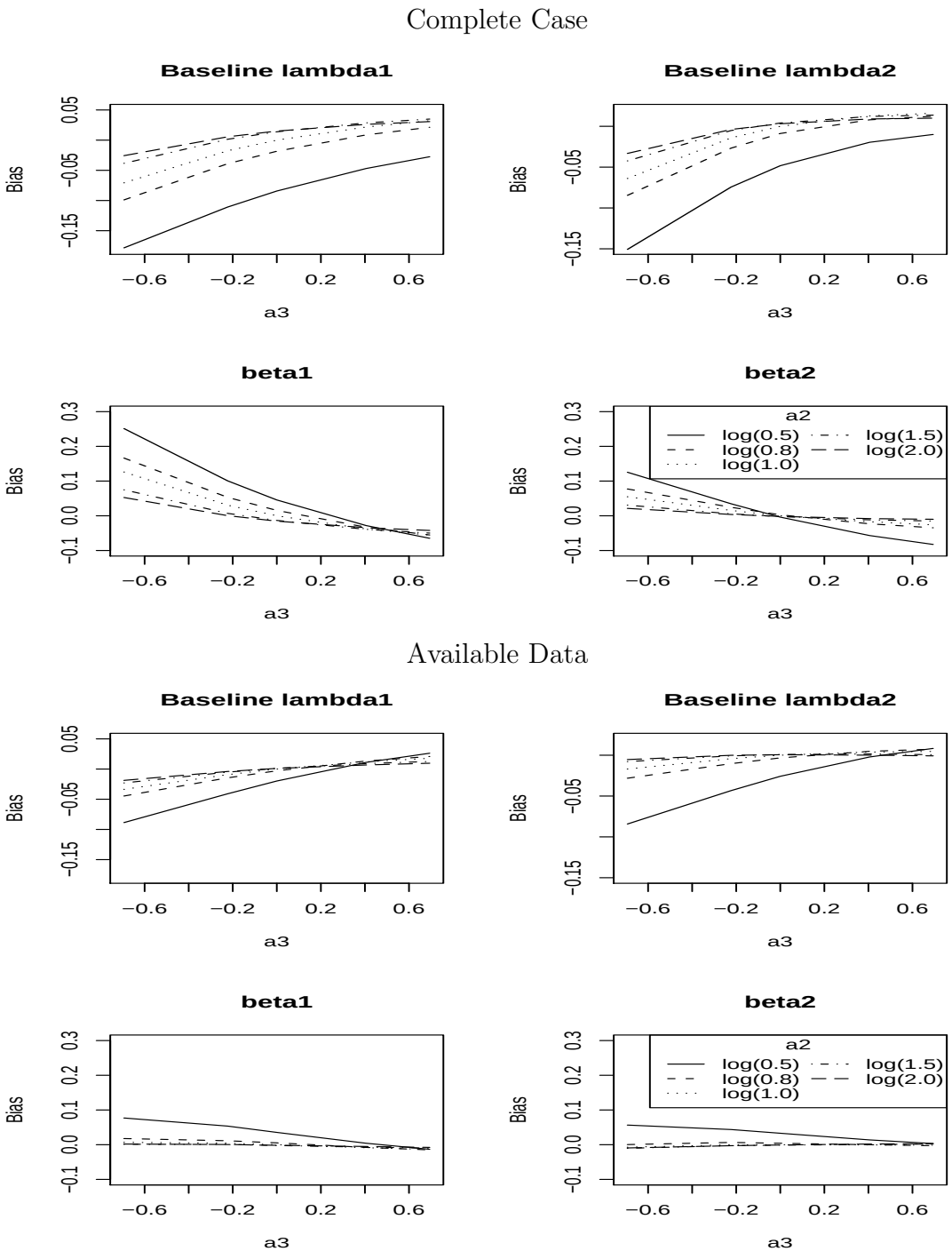


Figure 3.5: Asymptotic bias for missing not at random with one covariate with 3 states 5 observations.



where

$$Q_1(\alpha;\theta^{(h)}) = \sum_{i=1}^{n}\sum_{j=2}^{J}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)})\cdot\{r_{ij}\log\lambda_{ij}^* + (1-r_{ij})\log(1-\lambda_{ij}^*)\}$$

and

$$Q_2(\beta;\theta^{(h)}) = \sum_{i=1}^{n}\sum_{j=2}^{J}\sum_{k'=1}^{K}\sum_{k=k'}^{K}\sum_{y_i^{(m)}} w_i(y_i;\theta^{(h)})\cdot I(Y_i(a_{j-1})=k', Y_i(a_j)=k)\cdot\log P_{ik'k}(a_j-a_{j-1}).$$

Consequently, in the M-step we can separately maximize the progressive model part $Q_2(\beta;\theta^{(h)})$ and the missing data process part $Q_1(\alpha;\theta^{(h)})$. Standard statistical software such as R can be readily adapted to implement this step. Iterate through the E and M steps until $\theta^{(h)}$ converges. Denote the limit as $\hat{\theta}$.

To estimate the variance for $\hat{\theta}$, we use the Louis formula (Louis, 1982), given by

$$\Sigma(\hat{\theta}) = -\frac{\partial^2 Q(\hat{\theta};\hat{\theta})}{\partial\theta\partial\theta'} - \sum_{i=1}^{n}\sum_{y_i^{(m)}} w_i(y_i;\hat{\theta})S_i(\hat{\theta})S_i(\hat{\theta})' + \sum_{i=1}^{n}\left(\frac{\partial Q_i(\hat{\theta};\hat{\theta})}{\partial\theta}\right)\left(\frac{\partial Q_i(\hat{\theta};\hat{\theta})}{\partial\theta}\right)',$$

where $S_i(\hat{\theta}) = \partial\ell_i(\theta;y_i)/\partial\theta|_{\theta=\hat{\theta}}$. The estimate of the asymptotic covariance matrix of $\hat{\beta}$ is the lower $p_2\times p_2$ block of $[\Sigma(\hat{\theta})]^{-1}$, where p_2 is the dimension of β.

3.3.4 Identifiability of the Model

In this subsection, we show that the parameters of both the time homogeneous progressive model and the missing data model are identifiable for general missing data patterns for the panel data form. If we let Y_{ij} denote $Y_i(a_j)$, and let $\mu_{ijk'k}^C = P_{ik'k}(a_j-a_{j-1}) = P(Y_i(a_j)=k|Y_i(a_{j-1})=k', X_i)$ denote the conditional transition probability for subject i, then the likelihood can also be written as

$$\begin{aligned} L_i(\theta;y_i) &= \prod_{j=2}^{J}(\lambda_{ij}^*)^{r_{ij}}(1-\lambda_{ij}^*)^{1-r_{ij}} \\ &\quad\cdot\prod_{j=2}^{J}\prod_{k'=1}^{K}\prod_{k=k'+1}^{K}\left(\mu_{ijk'k}^C\right)^{I(Y_{i,j-1}=k',Y_{ij}=k)}\cdot\left(1-\sum_{l=k'+1}^{K}\mu_{ijk'l}^C\right)^{I(Y_{i,j-1}=k',Y_{ij}=k')}. \end{aligned}$$

Based on the results in Section 3.2.4, λ^*_{ij} and $\mu^C_{ijk'k}$ are identifiable.

From the identifiability of λ^*_{ij}, it is easy to show that α is identifiable. Now we show that β is also identifiable. First we show that the intensity function λ_k is identifiable. Given the time interval t, we need to show that $\lambda_l = \tilde{\lambda}_l$ for $l = 1, \ldots, K-1$ if $P_{k'k}(t) = \tilde{P}_{k'k}(t)$ for all $k' \leq k \leq K$, where $\tilde{P}_{k'k}(t)$ is $P_{k'k}(t)$ evaluated at $\tilde{\lambda}$. Note that for the time-homogeneous model, the expression for the transition probability from state k' to state k over $[0, t]$ is

$$P_{k'k}(t) = \begin{cases} \sum_{j=k'}^{k} C(k', j, k; \lambda) exp(-\lambda_j t), & k' \leq k, \\ 0, & k' > k, \end{cases}$$

where the coefficients are given by

$$C(k', j, k; \lambda) = \frac{\prod_{h=k'}^{k-1} \lambda_h}{\prod_{h=k', h \neq j}^{k} (\lambda_h - \lambda_j)}$$

for $k' \leq j \leq k$, and $C(j, j, j; \lambda) = 1, j = 1, 2, \ldots, K$ (Satten, 1999). If we let $k' = k$ for all $k = 1, \ldots, K-1$, we get $e^{-\lambda_k t} = e^{-\tilde{\lambda}_k t}$, thus $\lambda_k = \tilde{\lambda}_k$ for all $k = 1, \ldots, K-1$. The identifiability of β is easy to show from the identifiability of $\lambda_k, k = 1, \ldots, K-1$.

3.3.5 Simulation Studies

In this subsection we report on a simulation study to assess the performance of the proposed method. We consider the case with $K = 3$ states and $J = 3$ or $J = 5$ time points, and a sample of $n = 500$ individuals. We assume all the subjects are in state 1 when entering the study. Data generation procedures are very similar as in Section 3.2.4. Two thousand samples are simulated for each parameter configuration. The intensity function and the missing data model are the same as (3.12) and (3.13) in Section 3.3.2, respectively.

Here we consider the cases with no covariates (i.e., $\lambda_k = \lambda_{0k}$) and with one covariate (i.e., $\lambda_k = \lambda_{0k}e^{\beta_k x}$, $k = 1, 2$) included in the response model. We conduct three analyses – complete case analysis, available data analysis and the analysis using the proposed method. The results are reported in Tables 3.9 to 3.12, where SEL denotes the average standard error calculated based on the Louis formula, ASE is the average naive standard error calculated using the Hessian matrix, ESE is the empirical standard errors for the 2000 estimates, and CP represents the 95% coverage probability of the parameters. Table 3.9 displays the results for the case without covariates under two scenarios for $J = 3$. In Scenario I the missingness proportions for Y_2 and Y_3 are about 25% and 28%, respectively, while in Scenario II the missingness proportions for Y_2 and Y_3 are about 55% and 65%, respectively. Similarly, Table 3.10 reports the results for the case with one covariate under two scenarios for $J = 3$. In Scenario I the missingness proportions for Y_2 and Y_3 are about 22% and 30%, respectively, while in Scenario II the missingness proportions for Y_2 and Y_3 are about 50% and 60%, respectively. Table 3.11 displays the results for the case without covariates under two scenarios for $J = 5$. In Scenario I the missingness proportions for Y_2 to Y_5 are about 20%, 28%, 30% and 30%, respectively, while in Scenario II the missingness proportions for Y_2 to Y_5 are about 48%, 60%, 65% and 65%, respectively. Similarly, Table 3.12 reports the results for the case with one covariate under two scenarios for $J = 5$. In Scenario I the missingness proportions for Y_2 to Y_5 are about 25%, 28%, 30% and 30%, respectively, while in Scenario II the missingness proportions for Y_2 to Y_5 are about 45%, 60%, 60% and 62%, respectively.

It can be seen that both the complete case analysis and the available data analysis produce biased estimates, whereas the proposed method yields satisfactory

results with considerably smaller finite sample biases. As expected, as the proportion of missing observations increases the biases produced by the complete case and available data analyses become larger, but the proposed method retains small bias. Comparisons between the ASE and ESE suggest that the effect of missing data on variance estimation is not as striking as that on parameter estimation. Variance estimation based on SEL adjusts for missingness and the results agree with the empirical version (ESE) much better than the naive version of the ASE does. Furthermore, the coverage probabilities of the parameters obtained from the proposed method agree well with the nominal level 95% under different settings, but the complete case and available data analyses yield coverage probabilities that are far away from the nominal value, and in some situations they may completely fail to capture the true values of the parameters.

In many settings, prevalence functions, such as the one giving the proportion of subjects in the absorbing state, are of interest. Graphical plots in Figure 3.6 reveal how the three methods differ in estimation of the survival function $S(t) = 1 - P_{1K}(t)$ for the case without covariates. The estimates obtained from the proposed method are almost identical to the true survival functions, however, the survival functions estimated from the available data analysis and the complete case analysis are both above the true curve, revealing a positive bias in the survival probabilities. It is not surprising that the complete case analysis produces a curve that is farther from the true curve than the available data analysis. The differences among the curves become more substantial as time increases.

Table 3.9: Empirical performance of regression estimators by various methods for the case without covariates: $J = 3$

			EM				Complete Case				Available Data			
	Parameters[†]	BIAS%	SEL	ASE	ESE	CP	BIAS%	ASE	ESE	CP	BIAS%	ASE	ESE	CP
Scenario I:	Transition Model													
	λ_1	0.6	0.030	0.026	0.030	0.955	-2.8	0.033	0.033	0.913	-0.004	5.6	0.027	0.952
	λ_2	0.6	0.044	0.038	0.044	0.952	-1.0	0.049	0.049	0.941	-0.003	6.8	0.042	0.943
	Missing Data Model													
	α_0	-0.2	0.172	0.159	0.171	0.949								
	α_1	-0.7	0.214	0.177	0.214	0.952								
	α_2	0.1	0.081	0.070	0.080	0.950								
	α_3	0.1	0.051	0.039	0.050	0.950								
Scenario II:	Transition Model													
	λ_1	0.6	0.029	0.026	0.030	0.953	-32.4	0.042	0.041	0.079	-21.1	0.028	0.029	0.076
	λ_2	0.8	0.065	0.062	0.064	0.948	-20.0	0.075	0.076	0.582	-14.7	0.049	0.047	0.509
	Missing Data Model													
	α_0	0.0	0.119	0.099	0.127	0.952								
	α_1	1.0	0.181	0.143	0.186	0.954								
	α_2	0.1	0.082	0.070	0.080	0.949								
	α_3	0.5	0.050	0.038	0.048	0.947								

[†] $\lambda_1 = 0.500$ and $\lambda_2 = 0.611$

Scenario I: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.85)$, $\alpha_3 = \log(0.95)$

Scenario II: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.5)$, $\alpha_3 = \log(0.5)$

Table 3.10: Empirical performance of regression estimators by various methods for the case with covariates: $J = 3$

		EM					Complete Case				Available Data			
	Parameters[†]	BIAS%	SEL	ASE	ESE	CP	BIAS%	ASE	ESE	CP	BIAS%	ASE	ESE	CP
Scenario I:	Transition Model													
	λ_{01}	0.8	0.044	0.037	0.043	0.953	5.0	0.073	0.061	0.967	2.2	0.059	0.046	0.976
	λ_{02}	1.3	0.064	0.052	0.064	0.948	3.3	0.060	0.069	0.915	1.6	0.051	0.061	0.914
	β_1	-0.4	0.252	0.212	0.252	0.952	6.8	0.253	0.244	0.981	5.2	0.225	0.214	0.977
	β_2	-1.6	0.182	0.169	0.180	0.951	-0.4	0.184	0.185	0.964	-1.0	0.163	0.179	0.954
	Missing Data Model													
	α_0	-0.1	0.258	0.201	0.262	0.949								
	α_1	-1.0	0.221	0.188	0.215	0.951								
	α_2	0.9	0.141	0.112	0.140	0.948								
	α_3	0.2	0.146	0.121	0.147	0.947								
	α_4	-0.1	0.151	0.118	0.150	0.953								
Scenario II:	Transition Model													
	λ_{01}	1.2	0.042	0.037	0.043	0.951	-18.9	0.103	0.112	0.484	-17.0	0.058	0.053	0.528
	λ_{02}	1.1	0.076	0.059	0.077	0.946	-9.0	0.092	0.129	0.756	-12.3	0.055	0.072	0.616
	β_1	-0.4	0.255	0.227	0.261	0.957	34.0	0.429	0.417	0.984	20.4	0.271	0.263	0.992
	β_2	-1.0	0.187	0.175	0.195	0.949	20.1	0.321	0.322	0.964	14.7	0.205	0.199	0.943
	Missing Data Model													
	α_0	-0.1	0.154	0.129	0.161	0.950								
	α_1	2.0	0.180	0.168	0.189	0.948								
	α_2	-0.3	0.092	0.078	0.089	0.955								
	α_3	0.3	0.066	0.060	0.065	0.952								
	α_4	0.4	0.138	0.122	0.141	0.947								

[†] $\lambda_1 = 0.500, \lambda_2 = 0.611, \beta_1 = 1.000$ and $\beta_2 = 0.500$

Scenario I: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.85)$, $\alpha_3 = \log(0.95)$, $\alpha_4 = \log(2.0)$

Scenario II: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.5)$, $\alpha_3 = \log(0.5)$, $\alpha_4 = \log(2.0)$

Table 3.11: Empirical performance of regression estimators by various methods for the case without covariates: $J = 5$

		EM					Complete Case				Available Data			
	Parameters[†]	BIAS%	SEL	ASE	ESE	CP	BIAS%	ASE	ESE	CP	BIAS%	ASE	ESE	CP
Scenario I:	Transition Model													
	λ_1	0.8	0.024	0.024	0.025	0.949	-7.1	0.040	0.037	0.840	-0.8	0.024	0.024	0.942
	λ_2	0.2	0.035	0.033	0.035	0.952	-2.4	0.056	0.056	0.928	-1.6	0.034	0.034	0.945
	Missing Data Model													
	α_0	0.4	0.134	0.115	0.134	0.948								
	α_1	-0.8	0.124	0.114	0.124	0.952								
	α_2	1.2	0.058	0.052	0.058	0.949								
	α_3	1.3	0.051	0.048	0.052	0.947								
Scenario II:	Transition Model													
	λ_1	0.2	0.026	0.023	0.026	0.958	-50.2	0.064	0.062	0.097	-18.0	0.023	0.023	0.044
	λ_2	0.5	0.041	0.040	0.042	0.948	-25.8	0.140	0.156	0.649	-12.9	0.039	0.038	0.448
	Missing Data Model													
	α_0	0.4	0.105	0.092	0.106	0.949								
	α_1	-0.7	0.103	0.090	0.102	0.953								
	α_2	-0.6	0.052	0.049	0.052	0.952								
	α_3	-0.3	0.051	0.049	0.052	0.948								

[†] $\lambda_1 = 0.500$ and $\lambda_2 = 0.611$

Scenario I: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.85)$, $\alpha_3 = \log(0.95)$

Scenario II: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.5)$, $\alpha_3 = \log(0.5)$

Table 3.12: Empirical performance of regression estimators by various methods for the case with covariates: $J = 5$

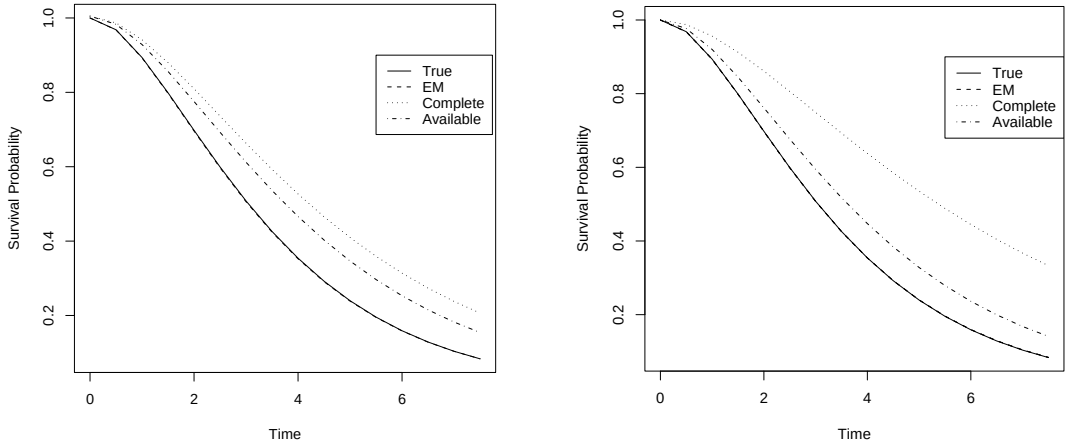
			EM				Complete Case				Available Data			
	Parameters[†]	BIAS%	SEL	ASE	ESE	CP	BIAS%	ASE	ESE	CP	BIAS%	ASE	ESE	CP
Scenario I:	Transition Model													
	λ_{01}	0.2	0.033	0.031	0.033	0.950	7.8	0.045	0.060	0.827	2.4	0.028	0.035	0.887
	λ_{02}	0.8	0.051	0.046	0.050	0.953	5.6	0.054	0.078	0.822	1.8	0.035	0.049	0.852
	β_1	0.3	0.108	0.098	0.108	0.949	6.0	0.167	0.167	0.930	0.4	0.109	0.110	0.954
	β_2	0.2	0.117	0.102	0.116	0.954	3.7	0.173	0.177	0.943	0.8	0.114	0.116	0.949
	Missing Data Model													
	α_0	0.4	0.159	0.137	0.158	0.952								
	α_1	1.4	0.125	0.110	0.125	0.952								
	α_2	-0.8	0.063	0.056	0.064	0.946								
	α_3	-0.9	0.063	0.059	0.062	0.949								
	α_4	0.9	0.106	0.091	0.106	0.953								
Scenario II:	Transition Model													
	λ_{01}	0.4	0.037	0.033	0.037	0.956	-32.4	0.078	0.088	0.397	-14.2	0.027	0.033	0.293
	λ_{02}	0.8	0.057	0.053	0.057	0.952	-14.1	0.124	0.154	0.609	-11.4	0.038	0.055	0.515
	β_1	0.3	0.116	0.105	0.116	0.954	52.9	0.478	0.460	0.861	9.1	0.125	0.119	0.904
	β_2	-0.6	0.126	0.117	0.127	0.947	43.0	0.534	0.533	0.934	9.8	0.141	0.141	0.937
	Missing Data Model													
	α_0	0.5	0.128	0.118	0.128	0.948								
	α_1	-0.3	0.105	0.098	0.098	0.949								
	α_2	-0.4	0.060	0.053	0.060	0.952								
	α_3	-0.8	0.060	0.056	0.060	0.951								
	α_4	0.6	0.099	0.090	0.099	0.953								

[†] $\lambda_1 = 0.500, \lambda_2 = 0.611, \beta_1 = 1.000$ and $\beta_2 = 0.500$

Scenario I: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.85)$, $\alpha_3 = \log(0.95)$, $\alpha_4 = \log(2.0)$

Scenario II: $\alpha_0 = \log(4.0)$, $\alpha_1 = \log(0.75)$, $\alpha_2 = \log(0.5)$, $\alpha_3 = \log(0.5)$, $\alpha_4 = \log(2.0)$

Figure 3.6: Survival functions for missing not at random without covariates with 3 states. The left figure is for 3 observations and the right is for 5 observations.



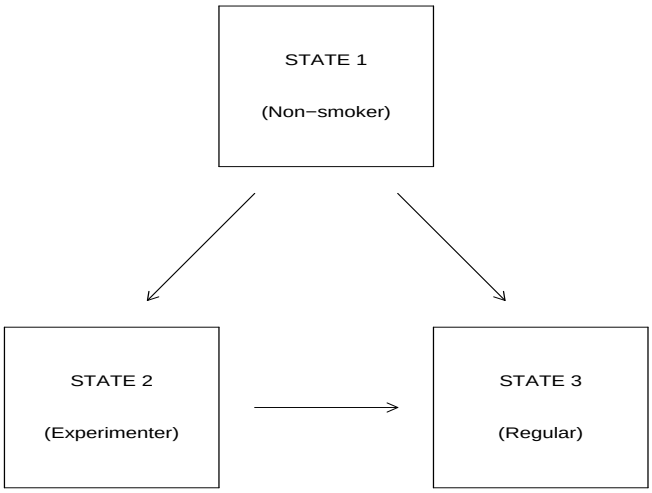
3.4 Applications

3.4.1 Application to a Smoking Prevention Project

Here we reanalyze the Waterloo Smoking Prevention Project data described in Chapter 2. Here we focus on progression of students' smoking behavior. That is, we model the response process with a progressive model. It is often of interest to understand the impact of covariates on the transition probability. Figure 3.7 is an illustrative diagram for the progressive model. The description of the data set and some notations are the same as in Chapter 2. To study this problem, we only select subjects with the progressive transition patterns. There are 3027 subjects in the data set who are present at the first assessment and in state 1. About 42.4% subjects have missing observations. The missingness proportion is about 15.3%. In the complete case analysis, there are 1849 subjects with complete observations.

Let Y_{ij} be the state student i was in at time j, i.e., in grade $5 + j$, $j = 1, \ldots, 7$,

Figure 3.7: Three-state progressive diagram for the analysis of the Waterloo Smoking Prevention Project Data



and use the subscripts for covariates in a similar fashion. Consider the model for the transitional probabilities

$$\begin{aligned}\log\left(\frac{\mu^C_{ijk'k}}{\mu^C_{ijk'k'}}\right) &= \beta_{k'k0} + \beta_{k'k1}\cdot \mathrm{GENDER}_i + \beta_{k'k2}\cdot \mathrm{TRT}_i + \beta_{k'k3}\cdot \mathrm{GRADE}_{ij} \\ &\quad +\beta_{k'k4}\cdot \mathrm{SMR2}_{ij} + \beta_{k'k5}\cdot \mathrm{SMR3}_{ij}, \quad k' = 1, 2,\ k' < k \leq 3,\end{aligned}$$

where TRT_i represents the treatment status for subject i, $\mathrm{SMR2}_{ij} = I(\mathrm{SMR}_{ij} = 2)$, $\mathrm{SMR3}_{ij} = I(\mathrm{SMR}_{ij} = 3)$, GRADE_{ij} is the grade indicator for subject i at grade $5 + j$, taking value 0 at secondary school (grade 6 to grade 8) and 1 at high school (grade 9 to grade 12). For the missing data process R_{ij}, we build the model

$$\begin{aligned}\mathrm{logit}\left(\lambda^*_{ij}\right) &= \alpha_0 + \alpha_1\cdot \mathrm{GENDER}_i + \alpha_2\cdot \mathrm{TRT}_i + \alpha_3\cdot \mathrm{GRADE}_{ij} + \alpha_4\cdot \mathrm{SMR2}_{ij} \\ &\quad +\alpha_5\cdot \mathrm{SMR3}_{ij} + \alpha_6\cdot r_{i,j-1} + \alpha_7\cdot Z_{ij12} + \alpha_8\cdot Z_{ij13} + \alpha_9\cdot Z_{ij23} \\ &\quad +\alpha_{10}\cdot I(y_{i,j-1} = 2) + \alpha_{11}\cdot I(y_{i,j-1} = 3),\end{aligned}$$

where $Z_{ijk'k} = I(y_{i,j-1} = k', y_{ij} = k)$ is the indicator covariate featuring the transitions from state k' to state k at time j, $k' < k$. Here α_7, α_8 and α_9 measure the influence of different transition occurrence on the missing process and α_{10} and α_{11} measure the influence of the previous states on the missing probability of the present responses. It leads to MNAR when at least one of α_7 to α_{11} is not equal to 0.

We analyze the WSPP data with three methods– the proposed method, complete case analysis and available data analysis. The results are reported in Table 3.13. Complete case and available data analyses produce generally agreeable estimates. Although the estimated treatment effects for transitions from states 1 to 2, and 1 to 3 are in opposite directions from both methods, they are not statistically significant. As expected, the standard errors produced from the available data analysis are smaller than those obtained from the complete case analysis. The proposed method reveals the same nature of statistical significance (or non-significance) as that obtained from the complete case and available data analyses for each covariate effect. The proposed method suggests that the gender and treatment effects are not statistically significant in all the transition models. In the transition models from state 1 to state 2 and from state 1 to state 3, social model risk score and grade have significant negative effects on smoking incidence ($\hat{\beta}_{123} = 0.863$, p-value$< 0.001$; $\hat{\beta}_{124} = 0.427$, p-value$< 0.001$; $\hat{\beta}_{125} = 0.658$, p-value$< 0.001$; $\hat{\beta}_{133} = 1.408$, p-value$< 0.001$; $\hat{\beta}_{134} = 0.577$, p-value$< 0.001$; $\hat{\beta}_{135} = 1.097$, p-value$< 0.001$). Students are more likely to smoke if their parents, siblings or friends are smokers. Students are more likely to smoke when they are in high school as opposed to secondary school. In the transition from state 2 to state 3, no covariate is statistically significant.

The results regarding the missing data mechanism are reported on in the bottom of Table 3.13. It is seen that α_7, α_8, α_{10} and α_{11} are statistically significant, suggesting that a missing not at random mechanism is perhaps reasonable. The occurrences of the transitions and the previous observations have negative effects on the probabilities of observing the present observations ($\hat{\alpha}_7 = -3.863$, p-value$<$ 0.001; $\hat{\alpha}_8 = -2.145$, p-value$<$ 0.001; $\hat{\alpha}_{10} = -3.072$, p-value$<$ 0.001; $\hat{\alpha}_{11} = -2.229$, p-value$<$ 0.001). The significance of α_6 ($\hat{\alpha}_6 = 3.708$, p-value$<$ 0.001) indicates that there is a serial dependence in the missingness of consecutive observations. Moreover, if subjects have missing observations at the previous assessment time then they are less likely to be observed at the present assessment. It is seen that both gender and grade are significant, with females being more likely to appear for the assessment ($\hat{\alpha}_2 = -0.183$, p-value$=$ 0.019), and students in public school having a larger probability of being observed compared to those in secondary school ($\hat{\alpha}_3 = -0.580$, p-value$<$ 0.001).

3.4.2 Application to Psoriatic Arthritis Data

Psoriatic arthritis (PsA) is a progressive disease in the sense that without treatment, it can increase in severity causing disability through deformity and destruction of the joints. It is of interest to determine prognostic factors that relate to disease severity and rates of disease progression (Gladman et al., 1995, 1998). Upon entry to the clinic, a comprehensive list of demographic and clinical features are recorded. Covariates include duration of psoriasis at clinic entry (in years) (coded as PSORDUR), sex (coded as SEX, 0–Female, 1–Male), age at onset of PsA (in years) (coded as AGEPSA), family history of psoriasis (coded as FMPS, 0–No, 1–Yes), family history of PsA (coded as FMPSA, 0–No, 1–Yes) and erythrocyte sedimenta-

Table 3.13: Analysis of the Waterloo Smoking Prevention Project data

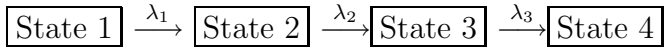
Transition Model		Proposed Method			Complete Case			Available Data		
Transition	Parameter	Estimate	SE†	p-value	Estimate	SE	p-value	Estimate	SE	p-value
$1 \rightarrow 2$	INTERCEPT (β_{120})	-2.985	0.123	<0.001	-2.933	0.111	<0.001	-2.926	0.095	<0.001
	GENDER (β_{121})	-0.082	0.086	0.342	-0.200	0.080	0.012	-0.150	0.068	0.028
	TRT (β_{122})	0.004	0.107	0.971	-0.026	0.098	0.793	0.018	0.084	0.833
	GRADE (β_{123})	0.863	0.088	<0.001	0.955	0.080	<0.001	0.787	0.068	<0.001
	SMR2 (β_{124})	0.427	0.088	<0.001	0.256	0.082	0.002	0.398	0.070	<0.001
	SMR3 (β_{125})	0.658	0.183	<0.001	0.473	0.179	0.008	0.537	0.149	<0.001
$1 \rightarrow 3$	INTERCEPT (β_{130})	-4.697	0.183	<0.001	-4.475	0.180	<0.001	-4.343	0.148	<0.001
	GENDER (β_{131})	0.180	0.118	0.127	0.038	0.112	0.737	0.002	0.097	0.981
	TRT (β_{132})	0.122	0.150	0.416	-0.066	0.138	0.631	0.019	0.120	0.874
	GRADE (β_{133})	1.408	0.125	<0.001	1.823	0.135	<0.001	1.648	0.106	<0.001
	SMR2 (β_{134})	0.577	0.124	<0.001	0.415	0.118	<0.001	0.458	0.102	<0.001
	SMR3 (β_{135})	1.097	0.214	<0.001	1.124	0.209	<0.001	1.150	0.175	<0.001
$2 \rightarrow 3$	INTERCEPT (β_{230})	-1.699	0.263	<0.001	-0.532	0.426	0.211	-0.555	0.307	0.071
	GENDER (β_{231})	0.077	0.131	0.558	0.257	0.174	0.140	0.201	0.157	0.199
	TRT (β_{232})	0.027	0.169	0.873	-0.297	0.214	0.165	-0.202	0.193	0.296
	GRADE (β_{233})	0.926	0.210	0.105	0.246	0.375	0.512	0.186	0.246	0.450
	SMR2 (β_{234})	0.098	0.135	0.469	0.132	0.179	0.459	0.252	0.161	0.118
	SMR3 (β_{235})	0.154	0.265	0.562	0.313	0.399	0.431	0.180	0.342	0.599
Missing Data Model										
	α_0	0.030	0.137	0.828						
	α_1	-0.183	0.078	0.019						
	α_2	0.092	0.096	0.339						
	α_3	0.580	0.101	<0.001						
	α_4	0.009	0.082	0.912						
	α_5	-0.194	0.160	0.226						
	α_6	3.708	0.095	<0.001						
	α_7	-3.863	0.122	<0.001						
	α_8	-2.145	0.245	<0.001						
	α_9	0.191	0.165	0.246						
	α_{10}	-3.072	0.132	<0.001						
	α_{11}	-2.229	0.129	<0.001						

† SE is the standard error based on the Louis formula (Louis, 1982). GENDER: 0–Female, 1–Male; TRT: treatment effect (0–control, 1–intervention); GRADE: 0–secondary school, 1–high school; SMR2: one of parents, siblings or friends smoke; SMR3: two or more of parents, siblings or friends smoke.

tion rate (ESR) which is a continuous variable measuring degree of inflammation. Patients are then scheduled to be assessed annually and at each followup assessment the number of damaged joints, as determined by clinical examination, is recorded. Table 3.14 lists a sample data set. There are 703 subjects with complete covariates, in which 28 subjects have complete observations over the first 10 years of their participation in the clinic registry. That is, there are 675 subjects with missing observations at different assessment time points, leading to a missing proportion about 61.3%.

Here we consider a multi-state Markov model with four states defined by the number of damaged joints determined by clinical assessment, as used by Gladman et al. (1995, 1998). Specifically, 0, 1-4, 5-9 and 10 or more damaged joints correspond to states 1, 2, 3 and 4 representing no damage, mild, moderate and severe damage, respectively. The rationale behind this state structure is that a larger number of damaged joints corresponds to a more severe disease. Figure 3.8 displays the transitions among the four states.

Figure 3.8: Four-state progression diagram for psoriatic arthritis data



Let $Y_i(a_j)$ denote the state subject i was in at time a_j, $j = 0, \ldots, 10$. The transition intensity functions are modeled as

$$\begin{aligned} \lambda_{ik} &= \lambda_{0k} \exp(\beta_1 \cdot \text{PSORDUR}_i + \beta_2 \cdot \text{AGEPSA}_i + \beta_3 \cdot \text{FMPS}_i \\ &\quad + \beta_4 \cdot \text{FMPSA}_i + \beta_5 \cdot \text{ESR}_i + \beta_6 \cdot \text{SEX}_i), \quad k = 1, 2, 3, \end{aligned}$$

Table 3.14: Sample data of the psoriatic arthritis study

ASSESSMENT							0	1	2	3	4	5	6	7	8	9	10
ID	PSORDUR	AGEPSA	FMPS	FMPSA	ESR	SEX	STATE										
1	21.5	33	0	0	6	1	1	.	.	.	.	.	1	.	1	.	1
2	38.3	40	1	0	36	0	1	.	.	.	.	.	.	.	.	.	1
3	15.1	25	0	0	4	1	1	.	.	.	.	.	.	.	.	.	4
4	23.0	24	0	0	25	1	1	.	.	.	.	.	.	.	.	.	4
5	7.1	34	0	0	83	0	1	.	.	1	1	.	1	1	1	.	1
6	11.7	24	0	0	2	1	1	.	.	.	1	.	1	2	.	.	2
7	7.4	28	1	1	16	1	1	.	.	.	2	.	4	4	4	4	4
8	4.2	23	1	1	34	1	1	.	.	2	.	.	3	3	3	4	.
9	11.8	49	1	0	23	0	1	.	.	.	.	.	.	.	.	.	3
10	56.7	31	1	0	47	0	1	.	.	.	.	.	.	.	.	.	4
11	41.1	32	1	0	65	1	1	.	.	.	.	.	.	.	.	.	4
12	10.1	25	1	0	26	0	1	.	.	4	.	4	.	.	.	4	4
13	31.6	70	1	0	25	0	1	1	.	.	.	.	.	.	.	.	.
14	33.0	40	1	0	17	1	1	.	.	1	.	.	1	.	.	1	1
15	35.9	51	0	0	57	1	1	.	.	1	.	2	.	.	.	.	2
16	10.1	36	1	0	12	0	1	.	.	.	.	.	.	.	.	.	1
17	23.6	43	0	0	20	1	1	.	.	.	4	4	.	.	.	.	4
18	16.8	10	1	0	74	0	1	.	.	.	.	.	.	.	.	.	1
19	0.2	43	0	0	24	0	1	.	.	.	2	2	2	2	2	4	4
20	19.9	26	0	0	40	0	1	.	.	.	.	.	.	.	.	.	1
21	29.9	25	0	0	99	0	1	.	.	.	.	.	.	.	.	.	4
22	11.9	21	1	1	15	0	1	.	1	1	.	1	1	1	1	1	1
23	12.3	23	1	0	16	1	1	.	.	.	.	.	.	2	.	.	2
24	31.0	36	1	1	73	0	1	.	.	.	.	.	.	.	.	.	4
25	10.6	36	1	0	8	0	1	1	1	1	.	.	.	.	.	.	.
26	9.1	50	1	0	10	1	1	.	.	.	.	.	1	1	1	.	1
27	30.2	36	0	0	30	1	1	.	.	.	.	.	.	.	.	.	1

PSORDUR: duration of psoriasis at time of clinic entry (years); AGEPSA: age at onset of psoriatic arthritis (years); FMPS: family history of psoriasis (0–No, 1–Yes); FMPSA: family history of psoriatic arthritis (0–No, 1–Yes); ESR: erythrocyte sedimentation rate; SEX: 0–Female, 1–Male.

where the $\lambda_{0k}{}'s$ are the baseline intensities. For the missing data process, we assume

$$\begin{aligned}
\text{logit}(\lambda^*_{ij}) &= \alpha_0 + \alpha_1 \cdot \text{PSORDUR}_i + \alpha_2 \cdot \text{AGEPSA}_i + \alpha_3 \cdot \text{FMPS}_i \\
&+ \alpha_4 \cdot \text{FMPSA}_i + \alpha_5 \cdot \text{ESR}_i + \alpha_6 \cdot \text{SEX}_i + \alpha_7 \cdot Z_{ij12} + \alpha_8 \cdot Z_{ij13} \\
&+ \alpha_9 \cdot Z_{ij14} + \alpha_{10} \cdot Z_{ij23} + \alpha_{11} \cdot Z_{ij24} + \alpha_{12} \cdot Z_{ij34} \\
&+ \alpha_{13} \cdot I(Y_i(a_{j-1}) = 2) + \alpha_{14} \cdot I(Y_i(a_{j-1}) = 3) + \alpha_{15} \cdot I(Y_i(a_{j-1}) = 4) \\
&+ \alpha_{16} \cdot r_{i,j-1},
\end{aligned}$$

where $Z_{ijk'k} = I(Y_i(a_{j-1}) = k', Y_i(a_j) = k)$ is the indicator featuring the transitions from states k' to k at time a_j, $k' < k$.

Table 3.15 reports the results obtained from the proposed method as well as from the complete case and available data analyses that ignore the missing data mechanism. The duration of psoriasis at clinic entry has a significant effect on the rate of PsA progression ($\hat{\beta}_1 = 0.057$ with p-value<0.001); that is, the relative rate of progression increases 5.9% for each additional year since diagnosis, controlling for other factors. The age at onset of PsA is also significantly associated with the rate of transition ($\hat{\beta}_2 = -0.073$; p-value<0.001); that is, the older the age at onset the slower the rate of progression (the risk decreases about 7.0% for each additional year of age at onset of PsA, when controlling other factors). A family history of psoriasis or PsA were not significantly related to the rate of progression ($\hat{\beta}_3 = -0.064$; p-value=0.560 and $\hat{\beta}_4 = -0.103$, p-value=0.423 respectively), but ESR level has an effect on PsA progression ($\hat{\beta}_5 = 0.013$; p-value<0.001) such that those with a higher ESR value have rates of damage (the relative risk increases about 1.3% for one unit of ESR increasing when controlling other factors). The effect of SEX is significant ($\hat{\beta}_6 = 0.177$; p-value=0.030), indicating that males have higher rates of progression than females ($RR = 1.194$).

Table 3.15: Analysis of the psoriatic arthritis data

Transition Model	Proposed Method			Complete Case			Available Data		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Baseline Intensities									
λ_{01}	0.064	0.016	<0.001	0.013	0.014	0.349	0.048	0.013	<0.001
λ_{02}	0.116	0.030	<0.001	0.013	0.014	0.361	0.086	0.024	<0.001
λ_{03}	0.150	0.042	<0.001	0.030	0.036	0.396	0.109	0.036	<0.001
Covariate Effects									
PSORDUR	0.057	0.005	<0.001	0.044	0.022	0.047	0.078	0.004	<0.001
AGEPSA	-0.073	0.007	<0.001	0.025	0.020	0.217	-0.063	0.005	<0.001
FMPS	-0.064	0.110	0.560	0.579	0.371	0.119	-0.050	0.079	0.530
FMPSA	-0.103	0.130	0.423	0.101	0.602	0.866	0.074	0.122	0.543
ESR	0.013	0.002	<0.001	0.012	0.009	0.180	0.007	0.002	<0.001
SEX	0.177	0.083	0.030	0.298	0.554	0.591	0.147	0.079	0.064
Missing Data Model									
α_0	-1.4979	0.1247	<0.001						
α_1	-0.0262	0.0028	<0.001						
α_2	0.0028	0.0024	0.2495						
α_3	0.0754	0.0661	0.2543						
α_4	0.0641	0.1026	0.5318						
α_5	-0.0019	0.0016	0.2119						
α_6	0.1122	0.0636	0.0776						
α_7	-0.4598	0.1489	<0.001						
α_8	1.4775	0.4276	<0.001						
α_9	3.0478	0.5670	<0.001						
α_{10}	-0.4507	0.2087	0.0308						
α_{11}	0.0699	0.4025	0.8620						
α_{12}	-0.6821	0.2674	0.0108						
α_{13}	0.4032	0.0822	<0.001						
α_{14}	0.7547	0.1359	<0.001						
α_{15}	0.7786	0.1174	<0.001						
α_{16}	1.8648	0.0627	<0.001						

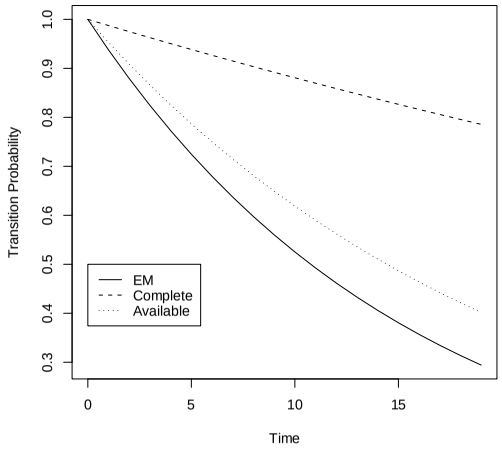
PSORDUR: duration of psoriasis at time of clinic entry (years); AGEPSA: age at onset of psoriatic arthritis (years); FMPS: family history of psoriasis (0–No, 1–Yes); FMPSA: family history of psoriatic arthritis (0–No, 1–Yes); ESR: erythrocyte sedimentation rate; SEX: 0–Female, 1–Male.

In Figure 3.9 we plot the transition probabilities starting from state 1 to other possible states to show the differences of the three analyses. It is seen that the available data analysis tends to yield less different curves from those obtained from the proposed method than the complete case analysis does. The probabilities staying in state 1 decrease as time goes by, while transition probabilities $P_{13}(t)$ and $P_{14}(t)$ have increasing trends with time. However, transition probabilities $P_{12}(t)$ obtained from the three methods are quite different. The proposed method produces a first-increasing-then-decreasing curve, the available data analysis yields a first-increasing-then-stable curve, but the complete case analysis leads to a fairly straight, increasing curve.

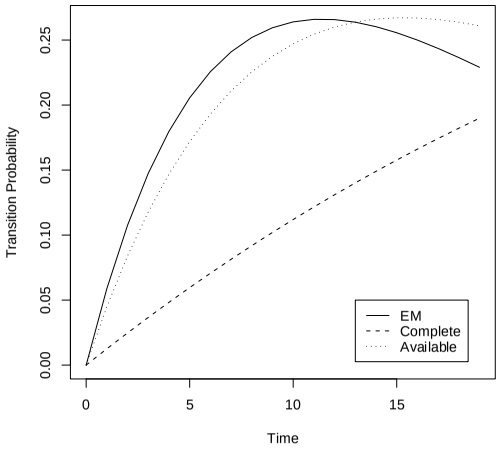
For the missing data process model, we find that the α_j coefficients with $j = 7$, 8, 9, 10, 12, 13, 14 and 15 are all significant, suggesting that nonignorable missing mechanisms are perhaps reasonable. In particular, we report that $\hat{\alpha}_{13} = 0.403$ with p-value$<$0.001, $\hat{\alpha}_{14} = 0.757$ with p-value$<$0.001 and $\hat{\alpha}_{15} = 0.779$ with p-value$<$0.001. It suggests that the more severe the disease at the previous assessment, the more likely he or she would appear for the present assessment. This seems to make intuitive sense since patients may be more likely to attend a clinic when their disease becomes more severe. If subjects are missing at an assessment, they are less likely to be observed at the next assessment because the estimate of α_{16} is 1.865 with p-value$<$0.001. As for the covariates, only the duration of initial psoriasis is significant ($\hat{\alpha}_1 = -0.026$; p-value$<$0.001), indicating the shorter the duration, the more likely for a patient to appear for assessment.

Figure 3.9: Transition probabilities for the analyses of the psoriatic arthritis data

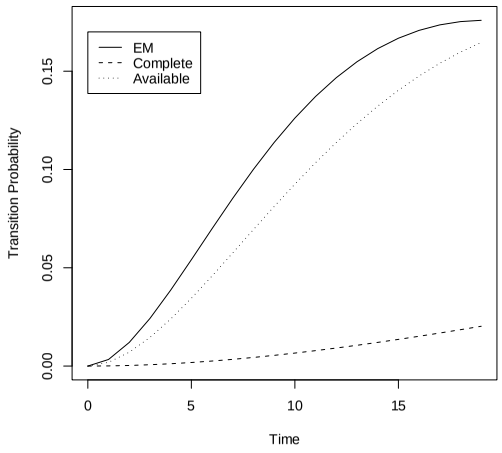
$P_{11}(t)$



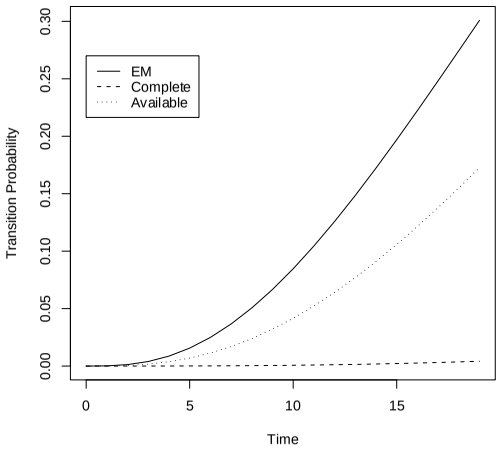
$P_{12}(t)$



$P_{13}(t)$



$P_{14}(t)$



3.5 Proof of the Identifiability of the Model

The parameter θ is identifiable if $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$ if and only if $\theta = \tilde{\theta}$. If $\theta = \tilde{\theta}$, it is easy to show $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$. Now it suffices to show that $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$ implies $\theta = \tilde{\theta}$. If θ and $\tilde{\theta}$ are two parameter values such that $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$ for all $(y_i^{(o)}, r_i)$, we now need to show that $\theta = \tilde{\theta}$. First, we introduce some notation for ease of exposition. Let $\tilde{\mu}^C_{ijk'k}$ and $\mu^C_{ijk'k}$ represent the conditional probability $P(Y_{ij} = k | Y_{i,j-1} = k', X_i)$ evaluated at $\tilde{\theta}$ and θ, respectively, and $\tilde{\lambda}^*_{ij}$ and λ^*_{ij} are defined analogously. We use $\lambda^*_{ij}(y_i)$ to explicitly indicate the dependence of λ^*_{ij} on response y_i whenever needed. Identifiability is established through two steps:

Step 1: We show that

$$\lambda^*_{ij} = \tilde{\lambda}^*_{ij}, \quad \text{for } j = 2, 3, \cdots, J,$$

and the identifiability of parameter α follows from a proper form of (3.3).

Step 2: We show that

$$\mu^C_{ijk'k} = \tilde{\mu}^C_{ijk'k}, \quad \text{for } j = 2, 3, \cdots, J; k' < k,$$

and the identifiability of parameter β follows from a proper form of (3.1).

We now proceed with the first step. Suppose $J \geq 3$. First, take $y_i = (k, k, k, k_4, \ldots, k_J)$, (i.e., subject i was in state k for the first three time points) then from $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$, we have

$$\begin{aligned} &\prod_{j=2}^{J} \lambda^*_{ij}(y_i) \cdot \left\{ \mu^C_{i2kk} \cdot \mu^C_{i3kk} \cdot \mu^C_{i4kk_4} \cdot \prod_{j=5}^{J} \mu^C_{ijk_{j-1}k_j} \right\} \\ =&\prod_{j=2}^{J} \tilde{\lambda}^*_{ij}(y_i) \cdot \left\{ \tilde{\mu}^C_{i2kk} \cdot \tilde{\mu}^C_{i3kk} \cdot \tilde{\mu}^C_{i4kk_4} \cdot \prod_{j=5}^{J} \tilde{\mu}^C_{ijk_{j-1}k_j} \right\}. \end{aligned} \tag{3.14}$$

Now consider $\tilde{y}_i = (k, *, k, k_4, \ldots, k_J)$, i.e. the response of subject i is the same as y_i except that there is a missing value in the second time point. Then the observed likelihood

$$L_i(\theta; \tilde{y}_i^{(o)}, r_i) = \sum_{\tilde{y}_{i2}} \left\{ (1 - \lambda_{i2}^*(\tilde{y}_i)) \cdot \prod_{j=3}^{J} \lambda_{ij}^*(\tilde{y}_i) \cdot \left[\mu_{i2k\tilde{y}_{i2}}^C \cdot \mu_{i3\tilde{y}_{i2}k}^C \cdot \mu_{i4kk_4}^C \cdot \prod_{j=5}^{J} \mu_{ijk_{j-1}k_j}^C \right] \right\}.$$

The feature of progressive process (if $s < t$, then $y(s) \leq y(t)$) implies that the missing observation $*$ should be k, and hence $\tilde{y}_i = y_i$. Then, from $L_i(\theta; \tilde{y}_i^{(o)}, r_i) = L_i(\tilde{\theta}; \tilde{y}_i^{(o)}, r_i)$ with the missingness probability incorporated, we have

$$\begin{aligned} &\left\{ (1 - \lambda_{i2}^*(\tilde{y}_i)) \cdot \prod_{j=3}^{J} \lambda_{ij}^*(\tilde{y}_i) \right\} \cdot \left\{ \mu_{i2kk}^C \cdot \mu_{i3kk}^C \cdot \mu_{i4kk_4}^C \cdot \prod_{j=5}^{J} \mu_{ijk_{j-1}k_j}^C \right\} \\ =&\left\{ (1 - \tilde{\lambda}_{i2}^*(\tilde{y}_i)) \cdot \prod_{j=3}^{J} \tilde{\lambda}_{ij}^*(\tilde{y}_i) \right\} \cdot \left\{ \tilde{\mu}_{i2kk}^C \cdot \tilde{\mu}_{i3kk}^C \cdot \tilde{\mu}_{i4kk_4}^C \cdot \prod_{j=5}^{J} \tilde{\mu}_{ijk_{j-1}k_j}^C \right\}. \end{aligned} \tag{3.15}$$

Comparing (3.14) and (3.15) in combination with that $y_i = \tilde{y}_i$, we obtain that

$$\frac{\lambda_{i2}^*(y_i)}{1 - \lambda_{i2}^*(y_i)} = \frac{\tilde{\lambda}_{i2}^*(y_i)}{1 - \tilde{\lambda}_{i2}^*(y_i)},$$

and thus,

$$\lambda_{i2}^*(y_i) = \tilde{\lambda}_{i2}^*(y_i).$$

In the same spirit, we can get

$$\lambda_{ij}^*(y_i) = \tilde{\lambda}_{ij}^*(y_i)$$

for $j = 2, 3, \ldots, J-1$. At time point J, if we consider $y_i = (y_{i1}, y_{i2}, \ldots, y_{i,J-2}, K, K)$ and $\tilde{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,J-2}, K, *)$, we can obtain $\lambda_{iJ}^*(y_i) = \tilde{\lambda}_{iJ}^*(y_i)$. Thus, we can obtain

$$\lambda_{ij}^*(y_i) = \tilde{\lambda}_{ij}^*(y_i) \tag{3.16}$$

for $j = 2, 3, \ldots, J$ for some patterns of y_i.

We assume, by examining all possible values of these special types of y_i listed above, that all the parameters α are identifiable. We comment that this assumption is not rigorous in practice when the number of observations for each subjects is not very small and the number of parameters in the missing data models is not very large. For example, consider model (3.3) which is given by $\text{logit}\lambda^*_{ij} = \alpha_0 + \alpha_1 x_i + \alpha_2 y_{ij}$, where x_i is a binary covariate, say a treatment indicator. It is easy to obtain that $\lambda^*_{i2}(y_i) = \tilde{\lambda}^*_{i2}(y_i)$ for $y_i = (k, k, k)$ with $k = 1, 2, 3$. The identifiability of α can be obtained from these patterns of y_i. Taking $y_i = (1, 1, 1)$, (3.16) leads to

$$\alpha_0 + \alpha_1 x_i + \alpha_2 = \tilde{\alpha}_0 + \tilde{\alpha}_1 x_i + \tilde{\alpha}_2; \tag{3.17}$$

taking $y_i = (2, 2, 2)$, (3.16) leads to

$$\alpha_0 + \alpha_1 x_i + 2\alpha_2 = \tilde{\alpha}_0 + \tilde{\alpha}_1 x_i + 2\tilde{\alpha}_2. \tag{3.18}$$

From (3.17) and (3.18), we can get $\alpha_0 = \tilde{\alpha}_0$ and $\alpha_2 = \tilde{\alpha}_2$; by evaluating x_i, we can get $\alpha_1 = \tilde{\alpha}_1$. Therefore, we establish the identifiability of the parameters in λ^*_{ij}.

Now it remains to show the second step. Since $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$ holds for any $y_i^{(o)}$ and r_i, we specifically examine those y_i with complete observations. With complete data y_i, the corresponding missing data indicator R_i assumes value 1 at every time point. That is, the identity $L_i(\theta; y_i^{(o)}, r_i) = L_i(\tilde{\theta}; y_i^{(o)}, r_i)$ leads to

$$\begin{aligned}
&\prod_{j=2}^{J} \left\{ \lambda^*_{ij} \cdot \prod_{k'=1}^{K} \prod_{k=k'+1}^{K} \left(\mu^C_{ijk'k}\right)^{I(Y_{i,j-1}=k', Y_{ij}=k)} \cdot \left(1 - \sum_{l=k'+1}^{K} \mu^C_{ijk'l}\right)^{I(Y_{i,j-1}=k', Y_{ij}=k')} \right\} \\
=&\prod_{j=2}^{J} \left\{ \tilde{\lambda}^*_{ij} \cdot \prod_{k'=1}^{K} \prod_{k=k'+1}^{K} \left(\tilde{\mu}^C_{ijk'k}\right)^{I(Y_{i,j-1}=k', Y_{ij}=k)} \cdot \left(1 - \sum_{l=k'+1}^{K} \tilde{\mu}^C_{ijk'l}\right)^{I(Y_{i,j-1}=k', Y_{ij}=k')} \right\}
\end{aligned} \tag{3.19}$$

for all complete data y_i. As it has been shown that $\lambda_{ij}^* = \tilde{\lambda}_{ij}^*$, (3.19) then becomes

$$\begin{aligned}
&\prod_{j=2}^{J}\left\{\prod_{k'=1}^{K}\prod_{k=k'+1}^{K}\left(\mu_{ijk'k}^{C}\right)^{I(Y_{i,j-1}=k',Y_{ij}=k)}\cdot\left(1-\sum_{l=k'+1}^{K}\mu_{ijk'l}^{C}\right)^{I(Y_{i,j-1}=k',Y_{ij}=k')}\right\} \\
=&\prod_{j=2}^{J}\left\{\prod_{k'=1}^{K}\prod_{k=k'+1}^{K}\left(\tilde{\mu}_{ijk'k}^{C}\right)^{I(Y_{i,j-1}=k',Y_{ij}=k)}\cdot\left(1-\sum_{l=k'+1}^{K}\tilde{\mu}_{ijk'l}^{C}\right)^{I(Y_{i,j-1}=k',Y_{ij}=k')}\right\}
\end{aligned} \tag{3.20}$$

for any complete data y_i. It suffices to show that

$$\mu_{ijk'k}^{C} = \tilde{\mu}_{ijk'k}^{C} \quad \text{for any } j, k' < k. \tag{3.21}$$

To this end, we examine (3.20) for different values of y_i for a fixed time point $j = J, J-1, \ldots, 1$.

First, fix $j = J$ and take $y_i = (k, \ldots, k, k, k)$ with $k < K$, then we obtain, by (3.20),

$$\prod_{j=2}^{J}(1-\sum_{l=k+1}^{K}\mu_{ijkl}^{C}) = \prod_{j=2}^{J}(1-\sum_{l=k+1}^{K}\tilde{\mu}_{ijkl}^{C}). \tag{3.22}$$

Now take $y_i = (k, \ldots, k, k, k_0)$, $k_0 \geq k+1$, then we obtain, by (3.20),

$$\prod_{j=2}^{J-1}(1-\sum_{l=k+1}^{K}\mu_{ijkl}^{C})\cdot\mu_{iJkk_0}^{C} = \prod_{j=2}^{J-1}(1-\sum_{l=k+1}^{K}\tilde{\mu}_{ijkl}^{C})\cdot\tilde{\mu}_{iJkk_0}^{C}. \tag{3.23}$$

Comparing (3.22) and (3.23) leads to

$$\frac{1-\sum_{l=k+1}^{K}\mu_{iJkl}^{C}}{\mu_{iJkk_0}^{C}} = \frac{1-\sum_{l=k+1}^{K}\tilde{\mu}_{iJkl}^{C}}{\tilde{\mu}_{iJkk_0}^{C}},\ k = 1, 2, \cdots, K-1, k_0 \geq k+1. \tag{3.24}$$

Repeatedly using this identity for different values of k establishes (3.21) for $j = J$. To be specific, we proceed with the following steps.

(1). Let $k = K-1$, then we obtain

$$\frac{1-\mu_{iJ,K-1,K}^{C}}{\mu_{iJ,K-1,K}^{C}} = \frac{1-\tilde{\mu}_{iJ,K-1,K}^{C}}{\tilde{\mu}_{iJ,K-1,K}^{C}},$$

and hence,

$$\mu_{iJ,K-1,K}^{C} = \tilde{\mu}_{iJ,K-1,K}^{C}.$$

(2). Let $k = K-2$, then $k_0 = K-1$ or K, and hence,

$$\frac{1-\mu^C_{iJ,K-2,K-1}-\mu^C_{iJ,K-2,K}}{\mu^C_{iJ,K-2,K-1}} = \frac{1-\tilde{\mu}^C_{iJ,K-2,K-1}-\tilde{\mu}^C_{iJ,K-2,K}}{\tilde{\mu}^C_{iJ,K-2,K-1}} \tag{3.25}$$

and

$$\frac{1-\mu^C_{iJ,K-2,K-1}-\mu^C_{iJ,K-2,K}}{\mu^C_{iJ,K-2,K}} = \frac{1-\tilde{\mu}^C_{iJ,K-2,K-1}-\tilde{\mu}^C_{iJ,K-2,K}}{\tilde{\mu}^C_{iJ,K-2,K}}, \tag{3.26}$$

which gives

$$\frac{\mu^C_{iJ,K-2,K}}{\mu^C_{iJ,K-2,K-1}} = \frac{\tilde{\mu}^C_{iJ,K-2,K}}{\tilde{\mu}^C_{iJ,K-2,K-1}}. \tag{3.27}$$

Combining (3.27) and (3.25), we obtain

$$\mu^C_{iJ,K-2,K-1} = \tilde{\mu}^C_{iJ,K-2,K-1},$$

and therefore,

$$\mu^C_{iJ,K-2,K} = \tilde{\mu}^C_{iJ,K-2,K},$$

which is from (3.27) and (3.26).

(3). In general, for $k = 1, 2, \cdots, K-3$ and $k_0 = k+1, \cdots, K$, we have

$$\frac{1-\sum_{l=k+1}^K \mu^C_{iJkl}}{\mu^C_{iJk,k+1}} = \frac{1-\sum_{l=k+1}^K \tilde{\mu}^C_{iJkl}}{\tilde{\mu}^C_{iJk,k+1}} \tag{3.28}$$

$$\frac{1-\sum_{l=k+1}^K \mu^C_{iJkl}}{\mu^C_{iJk,k+2}} = \frac{1-\sum_{l=k+1}^K \tilde{\mu}^C_{iJkl}}{\tilde{\mu}^C_{iJk,k+2}} \tag{3.29}$$

$$\frac{1-\sum_{l=k+1}^K \mu^C_{iJkl}}{\mu^C_{iJk,k+3}} = \frac{1-\sum_{l=k+1}^K \tilde{\mu}^C_{iJkl}}{\tilde{\mu}^C_{iJk,k+3}} \tag{3.30}$$

$$\vdots$$

$$\frac{1-\sum_{l=k+1}^K \mu^C_{iJkl}}{\mu^C_{iJkK}} = \frac{1-\sum_{l=k+1}^K \tilde{\mu}^C_{iJkl}}{\tilde{\mu}^C_{iJkK}} \tag{3.31}$$

- Dividing (3.28) by (3.29), ..., (3.31), respectively, we obtain

$$\frac{\mu^C_{iJk,k+2}}{\mu^C_{iJk,k+1}} = \frac{\tilde{\mu}^C_{iJk,k+2}}{\tilde{\mu}^C_{iJk,k+1}},$$
$$\vdots$$
$$\frac{\mu^C_{iJkK}}{\mu^C_{iJk,k+1}} = \frac{\tilde{\mu}^C_{iJkK}}{\tilde{\mu}^C_{iJk,k+1}},$$

  and hence, in combination with (3.28), we obtain

$$\mu^C_{iJk,k+1} = \tilde{\mu}^C_{iJk,k+1}.$$

- Dividing (3.29) by (3.30), ..., (3.31), respectively, and combining those identities with (3.29), we obtain

$$\mu^C_{iJk,k+2} = \tilde{\mu}^C_{iJk,k+2}.$$

- Analogously,

$$\mu^C_{iJkk_0} = \tilde{\mu}^C_{iJkk_0}$$

  for all $k = 1, 2, \cdots, K-3$ and $k_0 = k+1, \ldots, K$. That is

$$\mu^C_{iJkk_0} = \tilde{\mu}^C_{iJkk_0}, \text{ for all } k < k_0.$$

Secondly, fix $j = J-1$ and take $y_i = (k, \ldots, k, k_0, k_0)$, $k_0 \geq k+1$. Then we obtain, by (3.20),

$$\begin{aligned} &\prod_{j=2}^{J-2}(1 - \sum_{l=k+1}^{K} \mu^C_{ijkl}) \cdot \mu^C_{i,J-1,kk_0}(1 - \sum_{l=k_0+1}^{K} \mu^C_{iJk_0l}) \\ =&\prod_{j=2}^{J-2}(1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{ijkl}) \cdot \tilde{\mu}^C_{i,J-1,kk_0}(1 - \sum_{l=k_0+1}^{K} \tilde{\mu}^C_{iJk_0l}). \end{aligned} \tag{3.32}$$

Comparing (3.32) and (3.22) leads to

$$\begin{aligned} &\frac{\mu^C_{i,J-1,kk_0}(1 - \sum_{l=k_0+1}^{K} \mu^C_{iJk_0l})}{(1 - \sum_{l=k+1}^{K} \mu^C_{i,J-1,kl})(1 - \sum_{l=k+1}^{K} \mu^C_{iJkl})} \\ = \quad &\frac{\tilde{\mu}^C_{i,J-1,kk_0}(1 - \sum_{l=k_0+1}^{K} \tilde{\mu}^C_{iJk_0l})}{(1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{i,J-1,kl})(1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{iJkl})}, \end{aligned}$$

for $k = 1, 2, \ldots, K-1; k_0 = k+1, \ldots, K$. As shown before, $\mu^C_{iJkk_0} = \tilde{\mu}^C_{iJkk_0}$ for all $k < k_0$, we then obtain

$$\frac{1 - \sum_{l=k+1}^{K} \mu^C_{i,J-1,kl}}{\mu^C_{i,J-1,kk_0}} = \frac{1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{i,J-1,kl}}{\tilde{\mu}^C_{i,J-1,kk_0}}.$$

This structure is the same as (3.24) and therefore, repeating the same arguments above, we establish

$$\mu^C_{i,J-1,kk_0} = \tilde{\mu}^C_{i,J-1,kk_0}$$

for all $k < k_0$.

Analogously, by the same arguments, we can show (3.21) for $j = 2, \ldots, J-2$. Or more precisely, using mathematical induction we can establish (3.21). To be specific, assume

$$\mu^C_{ijkk_0} \quad = \quad \tilde{\mu}^C_{ijkk_0}$$

are true for all $k < k_0$, $j = j_0 + 1, \cdots, J$. Now we need to show that

$$\mu^C_{ij_0kk_0} \quad = \quad \tilde{\mu}^C_{ij_0kk_0}$$

for all $k < k_0$.

Take $y_i = (k, \ldots, k, k_0, k_0, \ldots, k_0)$, $k_0 \geq k+1$, where the first k_0 starts at time point j_0, then we obtain, by (3.20),

$$\begin{aligned} &\prod_{j=2}^{j_0-1}(1 - \sum_{l=k+1}^{K} \mu^C_{ijkl}) \cdot \mu^C_{ij_0kk_0} \cdot \prod_{j=j_0+1}^{J}(1 - \sum_{l=k_0+1}^{K} \mu^C_{ijk_0l}) \\ =&\prod_{j=2}^{j_0-1}(1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{ijkl}) \cdot \tilde{\mu}^C_{ij_0kk_0} \cdot \prod_{j=j_0+1}^{J}(1 - \sum_{l=k_0+1}^{K} \tilde{\mu}^C_{ijk_0l}). \end{aligned} \tag{3.33}$$

Comparing (3.33) and (3.22), we obtain

$$\begin{aligned} &\frac{\mu^C_{ij_0kk_0} \cdot \prod_{j=j_0+1}^{J}(1 - \sum_{l=k_0+1}^{K} \mu^C_{ijk_0l})}{(1 - \sum_{l=k+1}^{K} \mu^C_{ij_0kl}) \cdot \prod_{j=j_0+1}^{J}(1 - \sum_{l=k+1}^{K} \mu^C_{ijkl})} \\ = \quad &\frac{\tilde{\mu}^C_{ij_0kk_0} \cdot \prod_{j=j_0+1}^{J}(1 - \sum_{l=k_0+1}^{K} \tilde{\mu}^C_{ijk_0l})}{(1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{ij_0kl}) \cdot \prod_{j=j_0+1}^{J}(1 - \sum_{l=k+1}^{K} \tilde{\mu}^C_{ijkl})}. \end{aligned}$$

By the hypothesis of the induction, then we obtain

$$\frac{1-\sum_{l=k+1}^{K}\mu^{C}_{ij_0kl}}{\mu^{C}_{ij_0kk_0}}=\frac{1-\sum_{l=k+1}^{K}\tilde{\mu}^{C}_{ij_0kl}}{\tilde{\mu}^{C}_{ij_0kk_0}}$$

for $k=1,2,\ldots,K-1; k_0=k+1,\ldots,K$. Then by using the arguments above, we obtain

$$\mu^{C}_{ij_0kk_0}=\tilde{\mu}^{C}_{ij_0kk_0}$$

for all $k<k_0$. Therefore, we obtain that

$$\mu^{C}_{ijk'k}=\tilde{\mu}^{C}_{ijk'k}$$

for $j=2,3,\ldots,J$ and $k'<k$.

Identifiability of the β parameters can be established, provided model (3.1) is identifiable. For example, consider $K=3$ and $J=4$, and model (3.1) is given by

$$\log\left(\frac{\mu^{C}_{ijk'k}}{\mu^{C}_{ijk'k'}}\right)=\beta_{k'k0}+\beta_{k'k1}X_{ij1}+\beta_{k'k2}X_{ij2}+\beta_{k'k3}X_{ij3},\quad k'<k\leq 3,$$

for $j=2,3,4$, where $X_{ij1}=X_{i1}$ represents a time invariant treatment indicator, and $X_{ij2}=I(j=3)$ and $X_{ij3}=I(j=4)$ facilitate the temporal effects. Taking $k'=2$, $X_{ij1}=0$, $X_{ij2}=0$ and $X_{ij3}=0$, then (3.21) leads to $\beta_{230}=\tilde{\beta}_{230}$; taking $k'=2$, $X_{ij1}=1$, $X_{ij2}=0$, $X_{ij3}=0$, and using the fact that $\beta_{230}=\tilde{\beta}_{230}$, we obtain $\beta_{231}=\tilde{\beta}_{231}$ from (3.21); taking $k'=2$, $X_{ij1}=0$, $X_{ij2}=1$, $X_{ij3}=0$, and using the fact that $\beta_{230}=\tilde{\beta}_{230}$, we obtain $\beta_{232}=\tilde{\beta}_{232}$ from (3.21); and taking $k'=2$, $X_{ij1}=0$, $X_{ij2}=0$, $X_{ij3}=1$, and using the fact that $\beta_{230}=\tilde{\beta}_{230}$, we obtain $\beta_{233}=\tilde{\beta}_{233}$ from (3.21). Therefore, we show $\beta_{23}=(\beta_{230},\beta_{231},\beta_{232},\beta_{233})'$ is identifiable. Similarly, we can show $\beta_{12}=(\beta_{120},\beta_{121},\beta_{122},\beta_{123})'$ and $\beta_{13}=(\beta_{130},\beta_{131},\beta_{132},\beta_{133})'$ are identifiable.

Chapter 4

Marginal Methods for Longitudinal Data Analysis with Missing Response and Missing Covariates

4.1 Introduction

Incomplete longitudinal data often arise in clinical trials due to missing responses, partially filled out forms or questionnaires yielding missing covariate data, or study subjects failing to attend a scheduled clinic visit. Problems arise if the mechanism leading to the missing data is related to the values of response or covariates. For example, analyses based on individuals with complete data can lead to invalid inferences. Under a missing completely at random (MCAR) mechanism (Little and Rubin, 1987), analyses based on generalized estimating equations (GEE) (Liang and Zeger, 1986) yields consistent estimates for the regression parameters. When the data are missing at random (MAR) or missing not at random (MNAR) (Little and Rubin, 1987), analyses based on GEE give inconsistent param-

eter estimates. Robins and Rotnitzky (1995) and Robins et al. (1995) developed a class of estimators based on an inverse probability weighted generalized estimating equations (IPWGEE) approach in a regression setting when data are MAR. This approach involves modeling the missing data process and weighting the estimating equations by the inverse of a probability that is calculated based on the models for the missing data process. If the models for both the marginal mean of the response and the missing data process are correctly formulated, the IPWGEE approach corrects the bias and gives consistent estimates under the MAR mechanism.

The growing literature on methods for missing data has primarily dealt with either missing response or missing covariates data (Horton and Laird, 1998; Molenberghs et al., 1997; Lipsitz et al., 1999; Ibrahim et al., 2001; Zhao et al., 1996), but not both. In practice, of course, data are often unavailable for both responses and covariates, and sometimes there is an association between the missingness of the response and covariates. Valid analysis of this type of data therefore requires taking this association into consideration. Ignoring such correlation can bias the statistical inference. Chen et al. (2008) give theoretical investigation for inference with missing response and covariate data for general regression models using the likelihood method via EM algorithm. Shardell and Miller (2008) propose a marginal modeling approach to estimate the association between a time-dependent covariate and an outcome in longitudinal studies with missing response and missing covariate, but they focus on methods with an assumption that responses are independent. The purpose of this manuscript is to describe a general approach to the construction of estimating equations for parameters of marginal models for longitudinal data with incomplete response and covariate data. The approach is based on inverse probability weighted estimating equations for which the association between the

missingness of the response and covariate is addressed. We also highlight the poor properties of estimators when ignoring the correlations between the missingness of responses and covariate.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce notation and models. In Section 4.3, we provide details on estimation and inference. In Section 4.4, we provide a method which gives more efficient estimates. Numerical studies concerning asymptotic bias are given in Section 4.5. Data arising from the Waterloo Smoking Prevention Project (Cameron et al., 1999) and a study of bone metastases are also analyzed for illustration in Section 4.5. In Section 4.6, we extend the proposed methods to accommodate multiple missing covariates in conjunction with possibly missing responses. Section 4.7 and 4.8 are appendices.

4.2 Notation and Model Formulation

Consider a trial comprised of n individuals, each with J visits planned. Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ denote the response vector for subject i, some elements of which may be unobserved. Let X_{ij} denote a scalar time-dependent covariate for subject i at the jth visit which may or may not be observed, and let Z_{ij} denote a covariate vector for subject i at the jth visit which is fully observed. The case where multiple covariates may be missing will be considered in Section 4.6. For convenience we let $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})'$ and $Z_i = (Z_{i1}', Z_{i2}', \ldots, Z_{iJ}')'$. The conditional mean of Y_{ij} is denoted $\mu_{ij} = E(Y_{ij}|X_i, Z_i)$, and we let $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iJ})'$ denote the full vector of means. We suppose the mean structure of Y_{ij} depends on the covariate vector for subject i at time j (e.g., Pepe and Anderson, 1994; Robins, Greenland

and Hu, 1999), and consider a model for the mean of the form

$$g(\mu_{ij}) = X_{ij}\beta_x + Z'_{ij}\beta_z$$

for $j = 1, \ldots, J, i = 1, \ldots, n$, where $g(\cdot)$ is a monotone differentiable link function and $\beta = (\beta_x, \beta'_z)'$ is a $p \times 1$ vector of regression coefficients that is of interest here. We assume the variance is $v_{ij} = \text{var}(Y_{ij}|X_i, Z_i) = \kappa h(\mu_{ij})$, where $h(\cdot)$ is the variance function and κ is the dispersion parameter that is not of primary interest. It is usually estimated from an additional source before performing estimation of parameter β.

Let $R^y_{ij} = 1$ if Y_{ij} is observed and $R^y_{ij} = 0$ otherwise, $R^x_{ij} = 1$ if X_{ij} is observed and $R^x_{ij} = 0$ otherwise, $R^y_i = (R^y_{i1}, R^y_{i2}, \ldots, R^y_{iJ})'$ and $R^x_i = (R^x_{i1}, R^x_{i2}, \ldots, R^x_{iJ})'$. We assume the response and covariate are always observed at the first assessment so $R^y_{i1} = R^x_{i1} = 1$, and let future realizations be governed by the conditional probability $\lambda^y_{ij} = P(R^y_{ij} = 1|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, X_i, Z_i)$, where $\bar{R}^y_{ij} = \{r^y_{i1}, \ldots, r^y_{i,j-1}\}$, and $\bar{R}^x_{ij} = \{r^x_{i1}, \ldots, r^x_{i,j-1}\}$; λ^x_{ij} is defined similarly. We model λ^y_{ij} and λ^x_{ij} via logistic regression and specify

$$\text{logit}(\lambda^y_{ij}) = u'_{ij}\alpha_y \ ,$$

and

$$\text{logit}(\lambda^x_{ij}) = v'_{ij}\alpha_x \ ,$$

where u_{ij} and v_{ij} contain functions of $\{\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, X_i, Z_i\}$, $j = 2, 3, \ldots, J$, and α_y and α_x are regression parameters; let $\alpha_{xy} = (\alpha'_y, \alpha'_x)'$.

At each time point j, the observation status of the response and covariate may be associated within subjects because of common factors affecting the marginal observation processes. To model this association we define the conditional odds

ratio

$$\psi_{ij} = \frac{P(R^y_{ij}=1, R^x_{ij}=1|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, Z_i, X_i) \cdot P(R^y_{ij}=0, R^x_{ij}=0|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, Z_i, X_i)}{P(R^y_{ij}=1, R^x_{ij}=0|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, Z_i, X_i) \cdot P(R^y_{ij}=0, R^x_{ij}=1|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, Z_i, X_i)},$$

where the covariate and response variables appear symmetrically in this measure. The parameter ψ_{ij} can be viewed as the relative odds that Y_{ij} is observed (e.g. $R^y_{ij}=1$) when X_{ij} is observed versus when X_{ij} is missing.

We let $\lambda^{xy}_{ij} = P(R^y_{ij}=1, R^x_{ij}=1|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, Z_i, X_i)$ be the joint probability for the pair $R_{ij} = (R^y_{ij}, R^x_{ij})'$, conditional on the histories of the indicator variables and the entire vector of response and covariates. By noting that

$$\psi_{ij} = \frac{\lambda^{xy}_{ij}[1-\lambda^x_{ij}-\lambda^y_{ij}+\lambda^{xy}_{ij}]}{(\lambda^x_{ij}-\lambda^{xy}_{ij})(\lambda^y_{ij}-\lambda^{xy}_{ij})},$$

we write

$$\lambda^{xy}_{ij} = \begin{cases} \frac{a_{ij}-[a^2_{ij}-4\psi_{ij}(\psi_{ij}-1)\lambda^x_{ij}\lambda^y_{ij}]^{1/2}}{2(\psi_{ij}-1)}, & \text{if } \psi_{ij} \neq 1, \\ \lambda^x_{ij} \cdot \lambda^y_{ij}, & \text{if } \psi_{ij} = 1, \end{cases}$$

where $a_{ij} = 1-(1-\psi_{ij})(\lambda^x_{ij}+\lambda^y_{ij})$ (e.g. Lipsitz et al., 1991). Regression models may be used to allow the odds ratio ψ_{ij} to change with time-varying covariates. We may specify, for example,

$$\log(\psi_{ij}) = u^{*\prime}_{ij} \cdot \phi_j,$$

where u^*_{ij} is a covariate vector and ϕ_j is a vector of regression coefficients. Let $\phi = (\phi'_2, \phi'_3, \ldots, \phi'_J)'$, and $\alpha = (\alpha'_{xy}, \phi')'$ be of dimension q.

Here we consider a missing at random mechanism which assumes

$$P(R^y_{ij}=r^y_{ij}, R^x_{ij}=r^x_{ij}|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y_i, X_i, Z_i) = P(R^y_{ij}=r^y_{ij}, R^x_{ij}=r^x_{ij}|\bar{R}^y_{ij}, \bar{R}^x_{ij}, Y^{(o)}_i, X^{(o)}_i, Z_i).$$

Informally, we write $X_i = (X^{(o)}_i, X^{(m)}_i)$ where $X^{(o)}_i$ and $X^{(m)}_i$ denote the observed and missing components of X_i, respectively. For subject i, let $\pi^{xy}_{ij} = P(R^y_{ij} =$

$1, R_{ij}^x = 1|Y_i, Z_i, X_i)$ be the conditional probability of complete data for subject i at time j given the response vector Y_i and covariates Z_i and X_i, $j \geq 2$; $\pi_{i1}^{xy} = 1$ is assumed. The joint probability π_{ij}^{xy} can then be written as

$$\begin{aligned}
\pi_{ij}^{xy} &= \sum_{\bar{R}_{ij}^y} \sum_{\bar{R}_{ij}^x} \left\{ P(R_{ij}^y = 1, R_{ij}^x = 1 | \bar{R}_{ij}^y, \bar{R}_{ij}^x, Y_i, X_i, Z_i) \cdot \prod_{l=2}^{j-1} P(R_{il}^y = r_{il}^y, R_{il}^x = r_{il}^x | \bar{R}_{il}^y, \bar{R}_{il}^x, Y_i, X_i, Z_i) \right\} \\
&= \sum_{\bar{R}_{ij}^y} \sum_{\bar{R}_{ij}^x} \left\{ \lambda_{ij}^{xy} \cdot [\prod_{l=2}^{j-1} (\lambda_{il}^{xy})^{r_{il}^y r_{il}^x} (\lambda_{il}^x - \lambda_{il}^{xy})^{(1-r_{il}^y) r_{il}^x} (\lambda_{il}^y - \lambda_{il}^{xy})^{(1-r_{il}^x) r_{il}^y} \right. \\
&\quad \left. \cdot (1 - \lambda_{il}^x - \lambda_{il}^y + \lambda_{il}^{xy})^{(1-r_{il}^y)(1-r_{il}^x)}] \right\} \qquad (4.1)
\end{aligned}$$

for $j \geq 2$, where the summation is taken over all the possible values of the histories $\bar{R}_{ij}^y$ and $\bar{R}_{ij}^x$.

4.3 Estimation and Inference

4.3.1 Estimating Equations for Response Parameters

Following the same spirit of IPWGEE advocated by Robins et al. (1995), we may include a weight matrix $\Delta_i^*(\alpha)$ to the usual GEE to adjust for the effects of missingness occurring in both the response and covariate variables. That is, let

$$\Delta_i^*(\alpha) = \text{diag}(I(R_{ij}^y = 1, R_{ij}^x = 1)/\pi_{ij}^{xy}, 1 \leq j \leq J),$$

then the product $\Delta_i^*(Y_i - \mu_i)$ yields an adjusted contribution from subject i which involves the observed data alone yet retains the unbiasedness property, and hence estimating equations for β can be given by

$$U^*(\beta, \alpha) = \sum_{i=1}^{n} U_i^*(\beta, \alpha) = 0, \qquad (4.2)$$

where $U_i^*(\beta, \alpha) = D_i V_i^{-1} \Delta_i^*(\alpha)(Y_i - \mu_i)$ with $D_i = \partial \mu_i' / \partial \beta$ being a $p \times J$ derivative matrix, and V_i the working covariance matrix for the response Y_i.

In practice, the covariance matrix V_i is often expressed as

$$V_i = \kappa F_i^{1/2} R_i(\rho) F_i^{1/2},$$

where $R_i(\rho)$ is a working correlation matrix which may contain parameter ρ that is distinct from the β parameter, and $F_i = \text{diag}(h(\mu_{ij}), j = 1, 2, \ldots, J)$. When the working correlation matrix $R_i(\rho)$ is the identity matrix, (4.2) is computable. However, when a working independence assumption is not adopted, (4.2) may not be computable since elements of $D_i V_i^{-1}$ associated with those observed paired response Y_{ij} and covariate X_{ij} may still be unknown because of the involvement of other missing covariates X_{ik}'s ($k \neq j$). Here we modify (4.2) to incorporate the general working correlation matrices. We define

$$\Delta_i(\alpha) = \begin{pmatrix} \frac{I(R_{i1}^y=1, R_{i1}^x=1)}{\pi_{i1}^{xy}} & \frac{I(R_{i1}^x=1, R_{i2}^y=1, R_{i2}^x=1)}{\pi_{i12}^{xy}} & \cdots & \frac{I(R_{i1}^x=1, R_{iJ}^y=1, R_{iJ}^x=1)}{\pi_{i1J}^{xy}} \\ \frac{I(R_{i2}^x=1, R_{i1}^y=1, R_{i1}^x=1)}{\pi_{i21}^{xy}} & \frac{I(R_{i2}^y=1, R_{i2}^x=1)}{\pi_{i2}^{xy}} & \cdots & \frac{I(R_{i2}^x=1, R_{iJ}^y=1, R_{iJ}^x=1)}{\pi_{i2J}^{xy}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{I(R_{iJ}^x=1, R_{i1}^y=1, R_{i1}^x=1)}{\pi_{iJ1}^{xy}} & \frac{I(R_{iJ}^x=1, R_{i2}^y=1, R_{i2}^x=1)}{\pi_{iJ2}^{xy}} & \cdots & \frac{I(R_{iJ}^y=1, R_{iJ}^x=1)}{\pi_{iJ}^{xy}} \end{pmatrix}_{J \times J},$$

where $\pi_{ijk}^{xy} = P(R_{ij}^x = 1, R_{ik}^y = 1, R_{ik}^x = 1 | Y_i, X_i, Z_i)$ for $j \neq k$, and denote

$$M_i = \kappa^{-1} F_i^{-1/2} [R_i^{-1}(\rho) \bullet \Delta_i(\alpha)] F_i^{-1/2},$$

where $A \bullet B = [a_{ij} \cdot b_{ij}]$ denotes the Hadamard product of $J \times J$ matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ (e.g., Horn and Johnson, 1994). By introducing the condition that X_{ij} must be observed for elements in row j of $\Delta_i(\alpha)$, we ensure that all required elements of $D_i[V_i^{-1} \bullet \Delta_i(\alpha)](Y_i - \mu_i)$ can be computed.

The generalized estimating functions for β are given by

$$U(\beta, \alpha) = \sum_{i=1}^{n} U_i(\beta, \alpha) = 0, \tag{4.3}$$

where $U_i(\beta,\alpha) = D_i M_i(Y_i - \mu_i)$, and this yields consistent estimators since

$$E_{(R_i^y,R_i^x)|(Y_i,X_i,Z_i)}[R_i^{-1}(\rho) \bullet \Delta_i(\alpha)] = R_i^{-1}(\rho)$$

and hence

$$\begin{aligned} E_{(R_i^y,R_i^x,Y_i,X_i,Z_i)}[U_i(\beta,\alpha)] &= E_{(Y_i,X_i,Z_i)}E_{(R_i^y,R_i^x)|(Y_i,X_i,Z_i)}[U_i(\beta,\alpha)] \\ &= E_{(Y_i,X_i,Z_i)}[D_i V_i^{-1}(Y_i - \mu_i)] \\ &= E_{(X_i,Z_i)}E_{Y_i|(X_i,Z_i)}[D_i V_i^{-1}(Y_i - \mu_i)] \\ &= 0. \end{aligned}$$

It is easy to see that estimating function (4.3) depends on the observed data and the parameters only, and hence is computable. To employ (4.3) to estimate β, one needs to evaluate the joint probability π_{ijk}^{xy} which can be written as, for example, for $j < k$

$$\begin{aligned} \pi_{ijk}^{xy} = \sum_{r_{i,k-1}^y, r_{i,k-1}^x} \cdots \sum_{r_{i,j+1}^y, r_{i,j+1}^x} \sum_{r_{ij}^y} \Bigg\{ \lambda_{ik}^{xy} \cdot \prod_{\ell=j+1}^{k-1} [(\lambda_{i\ell}^{xy})^{r_{i\ell}^y r_{i\ell}^x} (\lambda_{i\ell}^x - \lambda_{i\ell}^{xy})^{(1-r_{i\ell}^y) r_{i\ell}^x} \\ \cdot (\lambda_{i\ell}^y - \lambda_{i\ell}^{xy})^{(1-r_{i\ell}^x) r_{i\ell}^y} (1 - \lambda_{i\ell}^x - \lambda_{i\ell}^y + \lambda_{i\ell}^{xy})^{(1-r_{i\ell}^y)(1-r_{i\ell}^x)}] \cdot (\pi_{ij}^{xy})^{r_{ij}^y} (\pi_{ij}^x - \pi_{ij}^{xy})^{1-r_{ij}^y} \Bigg\}, \end{aligned}$$

where $\pi_{ij}^x = P(R_{ij}^x = 1 | Y_i, X_i, Z_i)$, and it can be expressed in terms of $\lambda_{ij'}^x$, $\lambda_{ij'}^y$ and $\lambda_{ij'}^{xy}$, $j' \le j$. Similar notation applies to π_{ij}^y.

The working correlation matrix $R_i(\rho)$ is usually unknown and must be estimated. It is estimated in the iterative fitting process using the current value of β to compute the appropriate function of the Pearson residual

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{h(\mu_{ij})}} \cdot \delta_{ij},$$

where $\delta_{ij} = r_{ij}^y r_{ij}^x / \pi_{ij}^{xy}$. The estimator of the parameter ρ is different for different correlation structures. For example, for the unstructured correlation matrix that $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk}$ for $j \neq k$, we estimate ρ_{jk} by

$$\hat{\rho}_{jk} = \frac{1}{(n-p)\kappa} \sum_{i=1}^{n} e_{ij} e_{ik} \cdot \pi_{ij}^{xy} \pi_{ik}^{xy} / \pi_{ijk}^{*},$$

where $\pi_{ijk}^{*} = P(R_{ij}^x = 1, R_{ij}^y = 1, R_{ik}^x = 1, R_{ik}^y = 1 | Y_i, X_i, Z_i)$ which can be calculated in the same spirit of calculation of π_{ijk}^{xy} above, and the dispersion parameter κ is estimated by

$$\hat{\kappa} = \frac{1}{nJ - p} \sum_{i=1}^{n} \sum_{j=1}^{J} e_{ij}^2 \cdot \pi_{ij}^{xy}.$$

If α were known, then the estimate of β can be obtained by solving $U(\beta, \alpha) = 0$. In practice α is unknown and one must replace α in (3) with a consistent estimate which may be obtained as we describe in the next subsection.

4.3.2 Estimation of Parameters for the Missing Data Processes

Let $\Lambda_{ij} = (\lambda_{ij}^y, \lambda_{ij}^x)'$, $R_i = ({R_{i2}}', {R_{i3}}', \ldots, {R_{iJ}}')'$, $\Lambda_i = (\Lambda_{i2}', \Lambda_{i3}', \ldots, \Lambda_{iJ}')'$, and let $V_i^* = \text{diag}(V_{i2}^*, V_{i3}^*, \ldots, V_{iJ}^*)$ be the covariance matrix of R_i, where

$$V_{ij}^* = \begin{pmatrix} \lambda_{ij}^y (1 - \lambda_{ij}^y) & \lambda_{ij}^{xy} - \lambda_{ij}^y \lambda_{ij}^x \\ \lambda_{ij}^{xy} - \lambda_{ij}^y \lambda_{ij}^x & \lambda_{ij}^x (1 - \lambda_{ij}^x) \end{pmatrix}$$

is the covariance matrix of R_{ij}. If $D_i^* = \partial \Lambda_i' / \partial \alpha_{xy}$, then the estimating functions for α_{xy} are given by $\sum_{i=1}^{n} S_{1i}(\alpha)$, where $S_{1i}(\alpha) = D_i^* [V_i^*]^{-1} (R_i - \Lambda_i)$.

We use second order estimating equations for estimation of the association parameter ϕ. To construct these we define the pairwise product $R_{ij}^* = R_{ij}^y R_{ij}^x$ and the vector $R_i^* = (R_{i2}^*, R_{i3}^*, \ldots, R_{iJ}^*)'$, and let $\lambda_i^{xy} = (\lambda_{i2}^{xy}, \lambda_{i3}^{xy}, \ldots, \lambda_{iJ}^{xy})'$, $C_i^* = \partial [\lambda_i^{xy}]' / \partial \phi$,

and $W_i^* = \text{diag}(\lambda_{ij}^{xy} \cdot (1 - \lambda_{ij}^{xy}),\ j = 2, 3, \ldots, J)$. The estimating functions for ϕ are then given by $\sum_{i=1}^n S_{2i}(\alpha)$, where $S_{2i}(\alpha) = C_i^*[W_i^*]^{-1} \cdot (R_i^* - \lambda_i^{xy})$. Then the estimating equation for α are

$$\sum_{i=1}^{n} S_i(\alpha) = 0. \tag{4.4}$$

where $S_i(\alpha) = (S_{1i}'(\alpha), S_{2i}'(\alpha))'$.

4.3.3 Estimation and Inference

We may employ a Fisher-scoring algorithm for estimation of $\theta = (\alpha', \beta')'$. To do this we let

$$H_i(\theta) = \begin{pmatrix} S_i(\alpha) \\ U_i(\beta, \alpha) \end{pmatrix}, \qquad M^*(\alpha) = \begin{pmatrix} -\sum_{i=1}^n D_i^*[V_i^*]^{-1}[D_i^*]' \\ -\sum_{i=1}^n C_i^*[W_i^*]^{-1}[C_i^*]' \end{pmatrix},$$

and $M(\theta) = -\sum_{i=1}^n D_i M_i D_i'$. As the estimating functions $S_i(\alpha)$ are free of the β parameters, the derivative matrix $\partial H_i(\theta)/\partial\theta'$ is lower triangular, i.e.,

$$\frac{\partial H_i(\theta)}{\partial \theta'} = \begin{pmatrix} \frac{\partial S_i(\alpha)}{\partial \alpha'} & 0 \\ \frac{\partial U_i(\beta,\alpha)}{\partial \alpha'} & \frac{\partial U_i(\beta,\alpha)}{\partial \beta'} \end{pmatrix},$$

and therefore, given an initial value $\theta^{(0)}$, an updated estimates are obtained with the iterative equation

$$\theta^{(t+1)} = \theta^{(t)} - \begin{pmatrix} M^*(\alpha^{(t)}) & 0 \\ \sum_{i=1}^n [\partial U_i(\theta)/\partial\alpha']|_{\theta^{(t)}} & M(\theta^{(t)}) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n S_i(\alpha^{(t)}) \\ \sum_{i=1}^n U_i(\theta^{(t)}) \end{pmatrix}, \tag{4.5}$$

$t = 0, 1, \cdots$, until $\theta^{(t+1)}$ converges to the solution $\widehat{\theta}$.

Alternatively, one can invoke a two-stage estimation procedure. Under this scheme an estimate of α is obtained as the solution to $\sum_{i=1}^n S_i(\alpha) = 0$ by Fisher-scoring, and then a Fisher-scoring algorithm is employed to solve $\sum_{i=1}^n U_i(\beta, \hat{\alpha}) = 0$

where $\hat{\alpha}$ is used in place of α in (3). This two-stage estimation procedure employs the iterative equations

$$\alpha^{(s+1)} = \alpha^{(s)} - [M^*(\alpha^{(s)})]^{-1} \cdot \sum_{i=1}^{n} S_i(\alpha^{(s)}) \quad s = 0, 1, \cdots \tag{4.6}$$

and

$$\beta^{(t+1)} = \beta^{(t)} - [M(\beta^{(t)}, \hat{\alpha})]^{-1} \cdot \sum_{i=1}^{n} U_i(\beta^{(t)}, \hat{\alpha}) \quad t = 0, 1, \cdots . \tag{4.7}$$

The two-stage iterative equations (4.6) and (4.7) differ from joint iterative equation (4.5). Even under the special situation that the components of the left lower corner are zero in the inverse matrix of (4.5), (4.5) does not necessarily yield the same updated values as those from (4.6) and (4.7). However, the updated values from these two procedures converge to the same limit under mild regularity conditions (e.g., Prentice, 1988; Lipsitz et al., 1991).

While the two-stage procedure based on (4.6) and (4.7) is much easier to use for estimation of $\hat{\theta}$, the joint formulation based on $H_i(\theta)$ is more useful for developing the asymptotic distribution for $\hat{\theta}$. Note that since $E[H_i(\theta)] = 0$ and by Theorem 3.4 of Newey and McFadden (1993), under standard regularity conditions there is a unique solution $\hat{\theta}$ to the equation $\sum_{i=1}^{n} H_i(\theta) = 0$ with probability approaching 1, that satisfies

$$n^{1/2}(\hat{\theta} - \theta) = -\{E[\partial H_i(\theta)/\partial\theta']\}^{-1} \cdot n^{-1/2} \sum_{i=1}^{n} H_i(\theta) + o_p(1) .$$

For the estimator $\hat{\beta}$ of central interest, we have

$$n^{1/2}(\hat{\beta} - \beta) \;\; = \;\; -\Gamma^{-1} n^{-1/2} \cdot \sum_{i=1}^{n} Q_i(\beta, \alpha) + o_p(1),$$

where $\Gamma = E[\partial U_i(\beta, \alpha)/\partial\beta']$, and

$$Q_i(\beta, \alpha) = U_i(\beta, \alpha) - E[\partial U_i(\beta, \alpha)/\partial\alpha'] \cdot [E(\partial S_i(\alpha)/\partial\alpha')]^{-1} \cdot S_i(\alpha).$$

The central limit theorem then leads to the asymptotic distribution for $n^{1/2}(\hat{\beta}-\beta)$, which is normal with mean 0 and asymptotic variance $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$, where $\Sigma = E[Q_i(\beta,\alpha)Q_i'(\beta,\alpha)]$. A discussion on the variance estimate is included in Section 4.7.

4.4 More Efficient Estimation via Augmented IP-WGEE

Note that the estimating functions in (4.2) include merely the measurements collected at those time points j when both Y_{ij} and X_{ij} are observed, together with an observed covariate X_{ik}. There may be some information loss relative to the methods that may include all the available measurements. Under the missing at random mechanism, Robins et al. (1994, 1995), Robins and Rotnitzky (1995) and Scharfstein et al. (1999) proposed methods to improve the efficiency of the inverse probability weighted estimates. The notion is that adding a function with zero expectation to the estimating function maintains an unbiased estimating function but with suitable choice of this second function, efficiency may be improved. This approach has, to our knowledge, only been investigated to address missingness in either the response or covariates processes. In this section, we describe an efficient method that applies to the case either response or covariate data may be missing at any assessment time, or both.

Corresponding to each missingness pattern, we consider a vector A_{ir} $(r = 1, 2, 3)$ that picks up available measurements that may not be included in (4.3). For example, we take

$$A_{i1} = \left(\left[\frac{I(R_{ij}^y = 1, R_{ij}^x = 0)}{\pi_{ij}^y - \pi_{ij}^{xy}} \cdot \pi_{ij}^y - 1 \right] \cdot R_{ij}^y Y_{ij},\ j = 1, 2, \ldots, J \right)',$$

$$A_{i2} = \left(\left[\frac{I(R_{ij}^y = 0, R_{ij}^x = 1)}{\pi_{ij}^x - \pi_{ij}^{xy}} \cdot \pi_{ij}^x - 1 \right] \cdot R_{ij}^x X_{ij},\ j = 1, 2, \ldots, J \right)',$$

$$A_{i3} = \left(\left[\frac{I(R_{ij}^y = 0, R_{ij}^x = 0)}{1 - \pi_{ij}^x - \pi_{ij}^y + \pi_{ij}^{xy}} - 1 \right] \cdot Z_{ij}',\ j = 1, 2, \ldots, J \right)',$$

and $A_i = (A_{i1}', A_{i2}', A_{i3}')'$. The key point here is to make A_i have zero mean and be expressed in terms of the observed data. For ease of implementation, A_i is often chosen to be free of the unknown β parameter, but it may depend on the α parameter. We now explicitly denote it $A_i(\alpha)$.

Let $\text{Res}\{A, B\} = A - E[AB']\{E[BB']\}^{-1}B$ denote the residual obtained by regressing A on B. Let

$$\eta = E[\text{Res}\{U_i(\beta, \alpha), S_i(\alpha)\}\text{Res}\{A_i(\alpha), S_i(\alpha)\}'][\text{var}(\text{Res}\{A_i(\alpha), S_i(\alpha)\})]^{-1},$$

and $U_i^\dagger(\beta, \alpha) = U_i(\beta, \alpha) - \eta A_i(\alpha)$. Then, if α is known, the estimator $\tilde{\beta}^\dagger$ obtained from solving

$$\sum_{i=1}^{n} U_i^\dagger(\beta, \alpha) = 0 \tag{4.8}$$

is consistent for β since $U_i^\dagger(\beta, \alpha)$ is unbiased.

Under regularity conditions of Robins et al. (1995), $n^{1/2}(\tilde{\beta}^\dagger - \beta)$ has an asymptotic distribution $N(0, \Gamma^{-1}\Sigma^\dagger[\Gamma^{-1}]')$ with $\Sigma^\dagger = \text{var}\{\text{Res}(U_i(\beta, \alpha), H_i^*)\}$, where $H_i^* = (A_i'(\alpha), S_i'(\alpha))'$, and when $\eta \neq 0$, $\tilde{\beta}^\dagger$ is more efficient than $\hat{\beta}$; the proof is given in Section 4.8. We note that the efficiency of $\tilde{\beta}^\dagger$ relies on the choice of function $A_i(\alpha)$, and there is no universal way to specify an optimal $A_i(\alpha)$ function to produce the most efficient estimator $\tilde{\beta}^\dagger$. However, as long as that $A_i(\alpha)$ is correlated with $U_i(\beta, \alpha)$ some improvement in efficiency will be realized.

In practice it is usually not possible to solve (4.8) since η will typically be unknown. A modified version of (4.8) may be solvable, however, by replacing η

with a $n^{1/2}$-consistent estimate $\hat{\eta} = \hat{\eta}_1 \hat{\eta}_2^{-1}$, where

$$\hat{\eta}_1 = n^{-1} \sum_{i=1}^{n} \widehat{\text{Res}}[U_i(\hat{\beta}, \hat{\alpha}), S_i(\hat{\alpha})] \widehat{\text{Res}}[A_i(\hat{\alpha}), S_i(\hat{\alpha})]',$$

$$\hat{\eta}_2 = n^{-1} \sum_{i=1}^{n} \widehat{\text{Res}}[A_i(\hat{\alpha}), S_i(\hat{\alpha})] \widehat{\text{Res}}[A_i(\hat{\alpha}), S_i(\hat{\alpha})]',$$

and

$$\widehat{\text{Res}}(A_i, B_i) = A_i - \sum_{i=1}^{n} [A_i B_i'] [\sum_{i=1}^{n} B_i B_i']^{-1} B_i.$$

Under regularity conditions of Robins et al. (1995), the resultant estimator has the same asymptotic distribution as $\tilde{\beta}^{\dagger}$, and variance matrix $\Gamma^{-1}\Sigma^{\dagger}[\Gamma^{-1}]'$ can be consistently estimated by $\hat{\Gamma}^{-1}\hat{\Sigma}^{\dagger}[\hat{\Gamma}^{-1}]'$ with

$$\hat{\Sigma}^{\dagger} = n^{-1} \sum_{i=1}^{n} \{\widehat{\text{Res}}[U_i(\hat{\beta}, \hat{\alpha}), \{A_i(\hat{\alpha}), S_i(\hat{\alpha})\}]\} \widehat{\text{Res}}\{[U_i(\hat{\beta}, \hat{\alpha}), \{A_i(\hat{\alpha}), S_i(\hat{\alpha})\}]\} \ .$$

4.5 Empirical Studies and Applications

4.5.1 Simulation Studies for Comparison of Procedures

In this section we assess the empirical performance of the methods through simulation studies. We consider a setting with $J = 3$ and $n = 500$, and simulate the longitudinal binary responses from a model with

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$$

where x_{ij} is a time-dependent binary covariate generated independently from Bin(1,0.5) which may be missing at some time points. We set $\text{expit}(\beta_0) = 0.6$ and $\exp(\beta_1) = 0.5$, where $\text{expit}(t) = \exp(t)/(1+\exp(t))$. The association between the responses is specified as exchangeable with correlation coefficient ρ, which is specified as 0, 0.3 and 0.6. The data generation procedures follow Preisser et al. (2002).

For the missing response process, we take

$$\text{logit}(\lambda_{ij}^{y}) = \alpha_{y0} + \alpha_{y1} r_{i,j-1}^{y} + \alpha_{y2} r_{i,j-1}^{y} y_{i,j-1}, \quad j = 2, 3 \, ,$$

and for the missing covariate process, we take

$$\text{logit}(\lambda_{ij}^{x}) = \alpha_{x0} + \alpha_{x1} r_{i,j-1}^{x} + \alpha_{x2} r_{i,j-1}^{x} x_{i,j-1}, \quad j = 2, 3 \, .$$

We assume the response and covariates are available at the first assessment time, so $r_{i1}^{y} = r_{i1}^{x} = 1$. The association between R_{ij}^{x} and R_{ij}^{y} is assumed constant over time with values of $\psi_{i2} = \psi_{i3} = \psi = 8$, 4, 2 or 1. The true values for the regression parameters of the missing data processes are set to $\text{expit}(\alpha_{y0}) = \text{expit}(\alpha_{x0}) = 0.5$, $\exp(\alpha_{y1}) = \exp(\alpha_{x1}) = 1.5$, and $\exp(\alpha_{y2}) = \exp(\alpha_{x2}) = 0.1$, 0.5 or 2.0. Five hundred simulations are run for each parameter configuration.

Here we assess the performance of the proposed method along with other methods which might be used in practice using different models for the formulation of the weight. The first method, labeled "GEE" in the tables, is based on generalized estimating equations obtained by setting π_{ij}^{xy} and π_{ijk}^{xy} to be 1 in (4.3), for $j = 1, 2, \ldots, J$. The second and third methods, labeled "IPWGEE-M1" and "IPWGEE-M2" respectively, use marginal weights in the generalized estimating equation (4.3) based on a single missing data model for R_{ij}^{*} where $R_{ij}^{*} = 1$ if both Y_{ij} and X_{ij} are observed and $R_{ij}^{*} = 0$ otherwise. Then $\lambda_{ij}^{*} = P(R_{ij}^{*} = 1 | R_{i,j-1}^{*}, Y_{i}^{(o)}, X_{i}^{(o)})$ and a logistic model is formed as

$$\text{logit } \lambda_{ij}^{*} = w_{ij}' \alpha, \quad j = 2, 3 \, . \tag{4.9}$$

The second and third methods employ

$$\{1, r_{i,j-1}^{*}, r_{i,j-1}^{*} y_{i,j-1}, r_{i,j-1}^{*} x_{i,j-1}\}$$

and

$$\{1, r^*_{i,j-1}, r^y_{i,j-1}y_{i,j-1}, r^x_{i,j-1}x_{i,j-1}\}$$

for w_{ij} in (4.9) respectively, accommodating different covariate dependencies of the marginal missing data processes for the responses and covariates. The weight matrix now is

$$\Delta_i(\alpha) = \begin{pmatrix} \frac{I(R^*_{i1}=1)}{\pi^{xy}_{i1}} & \frac{I(R^*_{i1}=1,R^*_{i2}=1)}{\pi^*_{i12}} & \dots & \frac{I(R^*_{i1}=1,R^*_{iJ}=1)}{\pi^*_{i1J}} \\ \frac{I(R^*_{i2}=1,R^*_{i1}=1)}{\pi^*_{i12}} & \frac{I(R^*_{i2}=1)}{\pi^{xy}_{i2}} & \dots & \frac{I(R^*_{i2}=1,R^*_{iJ}=1)}{\pi^*_{i2J}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{I(R^*_{iJ}=1,R^*_{i1}=1)}{\pi^*_{i1J}} & \frac{I(R^*_{iJ}=1,R^*_{i2}=1)}{\pi^*_{i2J}} & \dots & \frac{I(R^*_{iJ}=1)}{\pi^{xy}_{iJ}} \end{pmatrix}_{J\times J},$$

where the probabilities π^{xy}_{ij} in (4.3) are therefore determined by

$$\pi^{xy}_{ij} \quad = \quad P(R^*_{ij}=1|Y^{(o)}_i, X^{(o)}_i) = \sum_{r^*_{i2}, r^*_{i3}, \dots, r^*_{i,j-1}} \left\{ \lambda^*_{ij} \cdot \prod_{l=2}^{j-1} (\lambda^*_{il})^{r^*_{il}} (1-\lambda^*_{il})^{1-r^*_{il}} \right\}, \quad (4.10)$$

and $\pi^*_{ijk} = P(R^*_{ij}=1, R^*_{ik}=1|Y_i, X_i, Z_i)$ can be expressed in terms of λ^*_{ij}. Instead of modeling R^y_{ij} and R^x_{ij} with a single indicator $R^*_{ij} = R^y_{ij}R^x_{ij}$, in the fourth and fifth methods we use separate models described in Section 4.2 to characterize R^y_{ij} and R^x_{ij}. The fourth method, labeled "IPWGEE-I", constrains ψ_{ij} to be 1, while the fifth method, labeled "IPWGEE-J", accommodates the association structure through ψ_{ij}. The sixth method, labeled "AIPWGEE-J", is the augmented IPWGEE accommodating the association structure through ψ_{ij}, where we specify $A_i(\alpha) = (A'_{i1}, A'_{i2})'$ as

$$A_{i1} = \left(\left[\frac{I(R^y_{ij}=1, R^x_{ij}=0)}{\pi^y_{ij} - \pi^{xy}_{ij}} \cdot \pi^y_{ij} - 1 \right] \cdot R^y_{ij}Y_{ij},\ j = 1, 2, \dots, J \right)',$$

and

$$A_{i2} = \left(\left[\frac{I(R^y_{ij}=0, R^x_{ij}=1)}{\pi^x_{ij} - \pi^{xy}_{ij}} \cdot \pi^x_{ij} - 1 \right] \cdot R^x_{ij}X_{ij},\ j = 1, 2, \dots, J \right)'.$$

The correlation parameter ρ is estimated by $1/(N^*-p)\sum_{i=1}^{n}\sum_{j<k} e_{ij}e_{ik}\cdot\pi_{ij}^{xy}\pi_{ik}^{xy}/\pi_{ijk}^{*}$ with $N^* = 1/2 \cdot nJ(J-1)$ as discussed in Section 4.3.1.

The results are reported in Tables 4.1 to 4.3, where ESE is the empirical standard error, and CP represents the empirical coverage probability for 95% confidence intervals. It is seen that the "GEE", "IPWGEE-M1", "IPWGEE-M2" and "IPWGEE-I" approaches yield larger biases than the "IPWGEE-J" and "AIPWGEE-J" methods. As the missing proportion increases, the bias increases. When a high percentage of data are missing, the "GEE", "IPWGEE-M1", "IPWGEE-M2" and "IPWGEE-I" methods provide confidence intervals with poor coverage probabilities, while the "IPWGEE-J" and "AIPWGEE-J" methods give reliable ones. As the association parameter ψ increases, performances of the "GEE", "IPWGEE-M1", "IPWGEE-M2" and "IPWGEE-I" approaches become worse; biases are more substantial, and coverage probabilities are far from these nominal levels. Their performances also deteriorate as the longitudinal association ρ increases. However, under a wide range of scenarios, the "IPWGEE-J" and "AIPWGEE-J" methods perform satisfactorily, but the "AIPWGEE-J" method gives more efficient estimates than those obtained from the "IPWGEE-J" method. When the missing proportion increases, the efficiency gain increases; when the association between the missingness increases, the improvement becomes more considerable. Also note that when the correlation ρ between the responses becomes stronger, the efficiency gain increases.

Table 4.1: Empirical bias, standard errors and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.6$)

		$\alpha_2 = 0.1$						$\alpha_2 = 0.5$						$\alpha_2 = 2.0$					
		$\beta_0^\ddagger$			$\beta_1^\ddagger$			$\beta_0^\ddagger$			$\beta_1^\ddagger$			$\beta_0^\ddagger$			$\beta_1^\ddagger$		
ψ	Method	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%
8	GEE	-15.8	0.107	85.5	5.8	0.146	90.5	-8.6	0.109	85.5	0.1	0.123	89.0	7.0	0.110	86.0	0.7	0.107	92.5
	IPWGEE-M1	91.2	0.577	88.0	11.5	0.336	90.6	6.2	0.292	90.2	-0.2	0.145	92.1	0.5	0.114	93.0	-0.2	0.103	93.0
	IPWGEE-M2	50.1	0.582	94.8	11.2	0.326	92.4	-11.3	0.280	91.5	-1.9	0.147	93.8	-0.3	0.144	93.3	2.7	0.112	93.5
	IPWGEE-I	75.6	0.561	93.1	18.1	0.550	95.8	21.8	0.181	91.2	1.0	0.165	93.8	-16.5	0.123	90.4	-0.9	0.121	93.0
	IPWGEE-J	0.7	0.433	94.4	0.3	0.432	94.5	0.8	0.172	93.8	0.5	0.150	94.4	-0.8	0.118	94.7	-1.2	0.115	94.2
	AIPWGEE-J	0.5	0.426	94.5	0.7	0.425	94.1	-0.6	0.166	94.9	-0.7	0.147	94.2	-0.9	0.116	94.4	-1.0	0.113	94.0
4	GEE	-15.6	0.120	82.5	1.5	0.130	91.0	-8.1	0.110	84.0	-0.5	0.119	93.5	10.7	0.120	89.4	-0.5	0.111	93.8
	IPWGEE-M1	90.2	0.585	89.3	7.7	0.378	91.7	5.4	0.395	91.4	-1.5	0.157	92.2	2.3	0.190	93.5	-0.7	0.124	93.5
	IPWGEE-M2	46.4	0.514	92.5	7.2	0.335	87.6	3.1	0.305	92.1	-1.0	0.173	93.1	-3.0	0.130	93.5	-0.9	0.111	94.0
	IPWGEE-I	64.1	0.608	93.4	10.8	0.551	96.9	20.4	0.189	91.2	1.4	0.180	92.6	-15.8	0.135	89.4	-1.2	0.127	94.0
	IPWGEE-J	0.6	0.428	94.6	0.8	0.434	94.8	0.6	0.184	94.0	-1.3	0.178	93.9	-0.7	0.131	94.4	-0.1	0.114	95.1
	AIPWGEE-J	0.8	0.420	94.8	0.7	0.429	94.7	-0.2	0.178	94.9	-0.7	0.175	95.0	-0.6	0.130	94.6	-0.3	0.113	94.3
2	GEE	-15.3	0.117	85.5	3.6	0.134	91.5	-10.7	0.108	90.0	-0.2	0.133	93.5	10.9	0.110	91.0	-0.3	0.109	94.0
	IPWGEE-M1	85.9	0.563	90.5	4.6	0.349	92.1	8.5	0.295	91.7	-0.8	0.168	93.1	-4.2	0.139	93.5	-0.6	0.121	94.0
	IPWGEE-M2	43.4	0.469	93.0	17.7	0.403	89.0	-3.1	0.303	92.1	-0.4	0.157	93.4	1.7	0.152	93.9	-1.5	0.116	93.9
	IPWGEE-I	46.0	0.552	96.8	6.7	0.544	94.8	16.4	0.188	94.0	0.2	0.182	93.6	-9.4	0.133	93.4	-0.6	0.129	94.6
	IPWGEE-J	0.8	0.423	94.8	0.3	0.377	94.7	0.9	0.180	94.7	0.1	0.178	95.5	-0.4	0.130	94.4	-1.3	0.125	94.0
	AIPWGEE-J	-0.3	0.416	95.1	0.4	0.372	94.5	-0.6	0.176	94.6	-0.7	0.174	94.4	-0.5	0.128	94.3	-0.3	0.124	95.4
1	GEE	-14.8	0.111	92.0	3.2	0.154	95.0	-11.3	0.111	92.5	-0.0	0.140	93.0	9.9	0.111	82.5	-2.0	0.113	93.5
	IPWGEE-M1	83.7	0.609	90.8	6.3	0.360	94.4	-4.4	0.347	93.6	5.8	0.208	96.0	-7.5	0.154	95.4	0.0	0.117	94.4
	IPWGEE-M2	42.8	0.421	93.7	1.1	0.359	89.6	11.8	0.302	93.0	-0.3	0.208	94.0	2.0	0.142	94.9	-0.8	0.139	94.8
	IPWGEE-I	2.0	0.554	97.3	1.8	0.524	95.0	1.8	0.202	93.0	0.8	0.195	92.2	-1.4	0.124	96.0	-0.7	0.123	95.6
	IPWGEE-J	1.2	0.421	95.1	1.2	0.477	95.2	0.1	0.205	94.6	-1.2	0.193	94.8	-0.8	0.130	95.0	-0.7	0.124	94.4
	AIPWGEE-J	0.9	0.418	94.8	0.8	0.473	94.2	-0.4	0.203	94.7	-0.5	0.192	94.3	-0.3	0.131	94.4	-1.0	0.124	94.2

‡ The true values are $\beta_0 = \log(1.5)$ and $\beta_1 = \log(0.5)$.

* Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

Table 4.2: Empirical bias, standard errors and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.3$)

		$\alpha_2 = 0.1$						$\alpha_2 = 0.5$						$\alpha_2 = 2.0$					
		$\beta_0^\ddagger$			$\beta_1^\ddagger$			$\beta_0^\ddagger$			$\beta_1^\ddagger$			$\beta_0^\ddagger$			$\beta_1^\ddagger$		
ψ	Method	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%
8	GEE	-13.4	0.109	92.5	1.6	0.160	92.5	-5.7	0.091	97.0	0.8	0.124	96.0	5.3	0.101	93.5	2.8	0.118	93.5
	IPWGEE-M1	15.7	0.283	88.8	1.7	0.328	90.4	0.5	0.126	91.5	1.3	0.149	94.5	-0.1	0.097	96.5	-0.1	0.137	91.0
	IPWGEE-M2	15.4	0.374	87.1	2.8	0.336	90.5	-0.4	0.121	94.0	0.9	0.154	94.0	1.8	0.100	92.5	1.2	0.125	94.5
	IPWGEE-I	28.3	0.472	93.0	6.7	0.579	96.0	9.8	0.156	96.3	3.8	0.178	94.3	-2.1	0.113	91.5	3.8	0.133	96.0
	IPWGEE-J	1.2	0.239	94.7	0.5	0.299	94.8	-0.0	0.112	95.4	0.4	0.155	94.9	-1.0	0.095	95.4	-0.6	0.125	95.4
	AIPWGEE-J	-0.5	0.229	94.6	0.6	0.293	94.9	-0.4	0.109	94.4	-0.9	0.151	94.2	-0.3	0.093	95.1	0.3	0.123	94.8
4	GEE	-12.3	0.117	90.5	-0.9	0.144	95.0	-3.0	0.107	95.0	1.7	0.134	93.5	4.3	0.097	93.0	-0.5	0.124	91.0
	IPWGEE-M1	-14.9	0.265	92.6	-1.1	0.318	94.1	-2.7	0.135	94.5	1.9	0.154	96.5	0.5	0.097	97.5	3.5	0.128	93.0
	IPWGEE-M2	14.5	0.338	87.9	2.6	0.370	85.4	-2.4	0.134	92.9	2.8	0.174	93.4	-3.1	0.111	94.0	-1.9	0.132	94.5
	IPWGEE-I	10.5	0.492	97.4	5.0	0.410	96.5	6.6	0.175	91.8	0.5	0.183	95.4	-1.3	0.120	92.0	2.5	0.147	90.5
	IPWGEE-J	0.5	0.254	95.4	0.9	0.319	94.4	0.3	0.121	95.0	0.4	0.166	95.0	-0.5	0.100	94.9	-0.4	0.126	95.4
	AIPWGEE-J	0.4	0.246	94.6	0.7	0.314	94.7	0.7	0.118	95.2	0.7	0.163	94.6	-0.2	0.098	94.7	0.1	0.125	95.0
2	GEE	-11.6	0.112	92.0	0.2	0.154	93.5	-1.4	0.107	94.5	1.1	0.123	95.5	5.1	0.101	93.0	-0.2	0.125	93.5
	IPWGEE-M1	14.2	0.326	86.2	4.3	0.373	88.7	-2.2	0.149	94.0	-0.7	0.173	93.5	-2.7	0.111	92.0	-1.7	0.138	93.5
	IPWGEE-M2	11.5	0.324	82.7	0.8	0.390	82.7	-2.5	0.139	95.0	0.1	0.183	94.5	2.4	0.105	95.0	2.0	0.134	94.5
	IPWGEE-I	8.8	0.399	96.0	5.3	0.467	96.0	5.5	0.153	94.5	-1.2	0.194	94.0	-2.8	0.108	94.5	-0.6	0.132	94.0
	IPWGEE-J	-1.1	0.330	95.2	2.5	0.360	94.4	3.6	0.144	95.0	2.2	0.175	95.5	-0.7	0.105	95.5	-0.5	0.125	95.0
	AIPWGEE-J	-0.1	0.326	94.8	-0.6	0.355	94.5	0.6	0.140	94.1	0.3	0.173	95.5	-0.4	0.104	95.0	0.2	0.125	94.9
1	GEE	-10.8	0.123	91.0	-0.2	0.149	96.0	-6.0	0.109	94.0	-0.6	0.148	94.5	5.7	0.097	93.5	-0.1	0.139	89.5
	IPWGEE-M1	13.6	0.319	88.1	9.6	0.369	84.6	2.3	0.155	93.9	3.6	0.196	93.9	0.9	0.110	95.5	1.3	0.139	94.0
	IPWGEE-M2	7.1	0.389	80.4	-0.2	0.407	86.4	-5.9	0.149	96.4	0.5	0.193	94.9	2.9	0.121	94.5	0.2	0.162	93.5
	IPWGEE-I	1.1	0.307	97.1	1.6	0.373	93.0	2.3	0.143	95.5	0.5	0.182	95.5	0.8	0.104	94.5	1.1	0.145	94.0
	IPWGEE-J	-0.1	0.319	94.2	0.9	0.376	94.2	0.1	0.157	95.0	1.0	0.201	94.5	-0.3	0.110	94.5	0.0	0.141	94.0
	AIPWGEE-J	0.7	0.318	94.6	-0.9	0.374	94.9	0.6	0.155	94.8	0.7	0.200	94.5	0.9	0.110	94.7	1.0	0.140	94.1

‡ The true values are $\beta_0 = \log(1.5)$ and $\beta_1 = \log(0.5)$.

* Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

Table 4.3: Empirical bias, standard errors and coverage probabilities for six approaches to estimation and inference with incomplete covariate and response data ($\rho = 0.0$)

		$\alpha_2 = 0.1$						$\alpha_2 = 0.5$						$\alpha_2 = 2.0$					
		$\beta_0^\ddagger$			$\beta_1^\ddagger$			$\beta_0^\ddagger$			$\beta_1^\ddagger$			$\beta_0^\ddagger$			$\beta_1^\ddagger$		
ψ	Method	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%	Bias*	ESE	CP%
8	GEE	1.5	0.110	95.0	1.6	0.134	94.6	1.6	0.103	95.4	1.1	0.145	94.3	0.6	0.096	94.2	-0.0	0.134	94.2
	IPWGEE-M1	1.5	0.357	93.7	-0.5	0.616	88.0	3.3	0.126	91.5	3.0	0.179	90.0	-1.4	0.097	96.0	-0.5	0.137	94.0
	IPWGEE-M2	- 5.0	0.308	91.7	-11.5	0.530	93.6	1.3	0.116	95.0	1.7	0.160	96.5	-1.2	0.106	91.0	-2.0	0.138	95.0
	IPWGEE-I	- 1.9	0.413	87.8	-1.6	0.595	89.8	0.0	0.115	94.0	1.3	0.175	94.5	0.6	0.093	95.0	1.5	0.141	94.5
	IPWGEE-J	-1.1	0.197	95.9	0.2	0.275	94.8	-0.5	0.109	94.2	0.3	0.161	93.9	-0.6	0.086	95.9	-0.7	0.133	95.9
	AIPWGEE-J	0.7	0.191	94.9	-0.6	0.271	94.2	-0.4	0.106	94.6	-0.5	0.156	94.4	-0.5	0.083	95.0	-0.1	0.132	94.8
4	GEE	-0.6	0.101	95.2	-1.2	0.142	94.2	-1.1	0.101	94.5	-0.8	0.142	94.7	0.9	0.093	93.4	0.7	0.133	94.2
	IPWGEE-M1	- 8.3	0.337	92.8	- 4.1	0.572	90.4	1.1	0.118	94.5	0.8	0.170	96.5	1.2	0.100	96.0	-0.4	0.132	95.0
	IPWGEE-M2	-0.3	0.282	90.0	-1.5	0.466	90.5	-2.1	0.114	95.5	-0.3	0.166	95.0	-0.4	0.096	95.0	-0.1	0.133	97.0
	IPWGEE-I	-1.4	0.371	93.9	2.2	0.550	89.6	-2.2	0.124	95.0	-0.4	0.185	94.0	-1.9	0.096	96.0	-2.3	0.133	95.0
	IPWGEE-J	0.5	0.196	94.7	-0.7	0.284	94.5	0.2	0.127	94.5	1.4	0.165	93.5	-0.2	0.095	95.4	-0.0	0.129	94.9
	AIPWGEE-J	-0.3	0.192	94.3	0.3	0.280	94.6	0.0	0.125	94.9	0.4	0.163	94.2	-0.9	0.094	94.6	-0.2	0.128	95.1
2	GEE	-2.7	0.111	93.8	-2.8	0.157	93.2	1.5	0.110	94.8	1.4	0.148	93.7	-1.3	0.096	94.6	-1.1	0.135	96.2
	IPWGEE-M1	-7.2	0.304	86.1	-0.3	0.449	93.0	0.6	0.117	96.5	0.8	0.175	93.0	-2.1	0.105	93.5	-0.0	0.150	95.0
	IPWGEE-M2	-3.3	0.334	86.2	-0.0	0.501	86.2	-0.2	0.116	95.5	3.1	0.175	95.0	2.4	0.101	94.0	0.3	0.151	93.0
	IPWGEE-I	-1.7	0.390	94.1	11.9	0.473	92.9	0.3	0.134	93.5	0.4	0.175	97.5	0.1	0.108	92.5	-1.1	0.152	92.5
	IPWGEE-J	-0.7	0.236	94.9	0.8	0.324	94.1	0.8	0.124	94.5	-0.5	0.176	94.5	-0.1	0.096	95.5	-0.4	0.127	97.5
	AIPWGEE-J	-0.7	0.233	94.5	1.0	0.320	95.1	0.6	0.121	94.4	0.7	0.174	95.0	0.5	0.095	94.5	0.2	0.127	95.3
1	GEE	0.8	0.114	92.0	1.4	0.161	93.6	2.0	0.114	95.1	1.2	0.151	94.0	0.5	0.099	94.9	0.2	0.138	95.7
	IPWGEE-M1	1.2	0.318	84.3	10.3	0.453	83.1	4.4	0.145	93.5	3.9	0.196	95.5	-2.2	0.111	93.5	-0.3	0.169	92.0
	IPWGEE-M2	2.4	0.430	83.9	2.4	0.531	84.5	0.0	0.129	96.0	0.3	0.172	96.0	-0.5	0.103	95.0	0.6	0.136	96.0
	IPWGEE-I	-3.3	0.278	92.4	2.7	0.388	92.9	-4.3	0.123	98.0	-3.9	0.189	93.5	-1.2	0.101	95.5	-0.3	0.148	96.5
	IPWGEE-J	0.7	0.287	95.3	0.7	0.422	94.4	0.9	0.136	95.5	0.4	0.197	94.0	-0.0	0.103	95.0	-0.1	0.142	95.0
	AIPWGEE-J	0.3	0.285	94.2	0.7	0.419	95.3	-0.5	0.134	94.7	0.5	0.195	94.5	-0.7	0.103	94.9	0.6	0.142	95.4

‡ The true values are $\beta_0 = \log(1.5)$ and $\beta_1 = \log(0.5)$.

* Relative bias defined by $(\hat{\beta} - \beta_{true})/\beta_{true} \times 100$.

4.5.2 Study of Asymptotic Bias under Misspecification of Association Structure for Missing Data Procedures

We now focus on evaluating the asymptotic biases induced by misspecifying the association structure between the missing data indicators of the response and covariate. Specifically, we consider the scenario that R_{ij}^y and R_{ij}^x are regarded as independent when they are actually correlated. Let $\hat{\beta}^\dagger$ denote the resultant estimator for the response model.

To characterize the asymptotic bias of $\hat{\beta}^\dagger$, we use the methods of White (1982) to find the value to which $\hat{\beta}^\dagger$ converges. In the spirit of Rotnitzky and Wypij (1994), Fitzmaurice et al. (1995) and Cook et al. (2004), we take the expectation of $U(\beta, \alpha)$ with respect to the true distribution of $G = (R_i^y, R_i^x, Y_i, X_i, Z_i)$ and set it equal to zero. The solution to this equation, denoted β^*, is the value to which $\hat{\beta}^\dagger$ converges in probability. If $\mathcal{G}$ is the sample space for G, we must therefore solve the equation

$$\sum_{g \in \mathcal{G}} D_i M_i (Y_i - \mu_i) \cdot P(g; \alpha, \beta) = 0, \tag{4.11}$$

where $P(g; \alpha, \beta)$ is the true probability of observing the realized value g of G.

The asymptotic covariance matrix of $n^{1/2}(\hat{\beta}^\dagger - \beta^*)$ is given by

$$\text{ascov}(\sqrt{n}(\hat{\beta}^\dagger - \beta^*)) = A^{-1}(\beta^*) B(\beta^*) A^{-1}(\beta^*), \tag{4.12}$$

where $A(\beta) = \sum_{g \in \mathcal{G}} \partial U_i(\beta, \alpha)/\partial \beta' \cdot P(g; \alpha, \beta)$, $B(\beta) = \sum_{g \in \mathcal{G}} U_i(\beta, \alpha) U_i'(\beta, \alpha) \cdot P(g; \alpha, \beta)$, and the dependence on α is suppressed in the notation. To investigate the asymptotic bias of the IPWGEE estimators under misspecification, we evaluate the expectation (4.11), and solve for β_k^*, $k = 0, 1$. We consider $\rho = 0.6$, 0.3 and 0 to consider a decreasing strength of association among the response components.

In the missing data model, we set $\alpha_{y2} = \alpha_{x2} = \alpha_2$ and for each setting, we take $\psi_{i2} = \psi_{i3} = \psi$ and alter it from 1 to 8 to represent different magnitudes of the associations between the missing response and missing covariate indicators.

We report the asymptotic percent relative bias, defined by $100 \times (\beta_k^* - \beta_k)/\beta_k$, $k = 0, 1$, in Figure 4.1 for different values of ρ. It can be seen that, as the missing proportion increases, the relative bias increases if other conditions are held fixed. Moreover, when the association between the missing data indicators increases, the asymptotic relative bias increases. It is also interesting to note that the stronger the correlation between responses the larger the relative bias.

4.5.3 Application to a Smoking Prevention Project

We now reanalyze the Waterloo Smoking Prevention Project data introduced in Chapter 2. The smoking status can be represented by a binary variable. $Y_{ij} = 1$ indicates subject i is a smoker in grade $j + 5$, and 0 otherwise, $j = 1, 2, 3$. The covariates that may influence the children's smoking behavior include gender (coded as GENDER, 0–female; 1–male), treatment indicator (coded as TRT, 0–control; 1–intervention), and social models risk score (coded as SMR, 0–none of parents, siblings or friends smoke; 1– otherwise). There are 4400 subjects in the data set who enter the study in grade 6. About 15.5% subjects have incomplete data; 13.7% of the students have no observations either in grade 7 or grade 8; 15.2% of the students have no social models risk score either in grade 7 or grade 8; and 5.1% of the students have no social models risk score and response either in grade 7 or grade 8. Table 4.4 lists a sample of the dataset.

Figure 4.1: Asymptotic relative bias of regression coefficients under a misspecified models of the association structures for the missing covariate and response processes

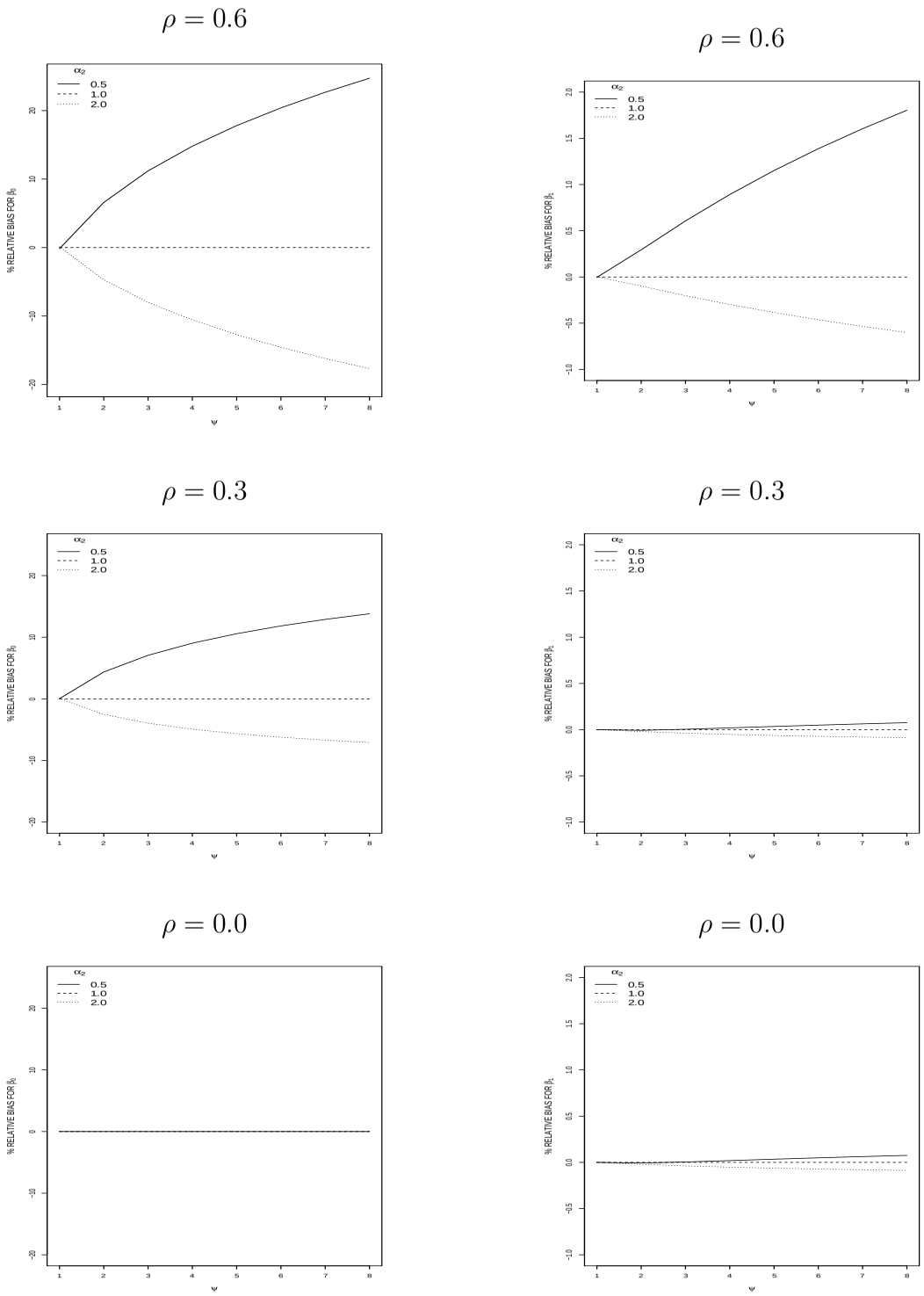


Table 4.4: Sample data from the Waterloo Smoking Prevention Project

ID	GENDER	TRT	SMR†			RESPONSE†		
			1	2	3	1	2	3
1	1	1	0	0	.	0	1	1
2	1	1	1	.	1	1	1	1
3	0	1	0	.	1	1	.	1
4	0	1	0	0	1	1	.	1
5	1	1	1	.	.	0	.	.
6	1	1	1	.	1	0	1	1
7	0	1	1	1	1	1	1	1
8	1	1	1	.	1	0	.	1
9	0	1	1	1	1	1	1	1
10	1	1	1	1	1	1	0	1
11	0	1	1	1	.	0	1	.
12	0	1	1	.	1	1	.	1
13	0	1	0	1	.	0	.	1
14	1	1	0	.	1	1	1	.

† Missing data are denoted by ·.

Consider the regression model for the response process

$$\begin{aligned}\text{logit}(\mu_{ij}) &= \beta_0 + \beta_1 \cdot \text{GENDER}_i + \beta_2 \cdot \text{TRT}_i + \beta_3 \cdot \text{GRADE7}_{ij} \\ &\quad + \beta_4 \cdot \text{GRADE8}_{ij} + \beta_5 \cdot \text{SMR}_{ij}, \quad j = 1, 2, 3,\end{aligned}$$

where GRADE7_{ij} is an indicator that student i is in grade 7 at time j, and GRADE8_{ij} is an indicator that student i is in grade 8 at time j.

For the missingness indicators, we assume models

$$\text{logit}(\lambda^y_{i2}) = \alpha_{y20} + \alpha_{y21} \cdot y_{i1} + \alpha_{y22} \cdot \text{GENDER}_i + \alpha_{y23} \cdot \text{TRT}_i + \alpha_{y24} \cdot \text{SMR}_{i1}, \quad (4.13)$$

$$\text{logit}(\lambda^y_{i3}) = \alpha_{y30} + \alpha_{y31} \cdot r^y_{i2} y_{i2} + \alpha_{y32} \cdot \text{GENDER}_i + \alpha_{y33} \cdot \text{TRT}_i + \alpha_{y34} \cdot r^x_{i2} \text{SMR}_{i2} + \alpha_{y35} \cdot r^y_{i2}, \quad (4.14)$$

and

$$\text{logit}(\lambda^x_{i2}) = \alpha_{x20} + \alpha_{x21} \cdot y_{i1} + \alpha_{x22} \cdot \text{GENDER}_i + \alpha_{x23} \cdot \text{TRT}_i + \alpha_{x24} \cdot \text{SMR}_{i1}, \quad (4.15)$$

$$\text{logit}(\lambda_{i3}^{x}) = \alpha_{x30}+\alpha_{x31}\cdot r_{i2}^{y}y_{i2}+\alpha_{x32}\cdot\text{GENDER}_i+\alpha_{x33}\cdot\text{TRT}_i+\alpha_{x34}\cdot r_{i2}^{x}\text{SMR}_{i2}+\alpha_{x35}\cdot r_{i2}^{x}, \tag{4.16}$$

respectively.

In line with the simulation studies, here we undertake five methods to analyze the data. The first analysis, labeled "GEE" , is an unweighted analysis based on generalized estimating equations. The second analysis, labeled "IPWGEE-M2", is based on a weighted version of the generalized estimating equations in which the weights are determined by fitting logistic models:

$$\text{logit}(\lambda_{i2}^{*}) = \alpha_{20} + \alpha_{21}\cdot y_{i1} + \alpha_{22}\cdot\text{GENDER}_i + \alpha_{23}\cdot\text{TRT}_i + \alpha_{24}\cdot\text{SMR}_{i1}$$

and

$$\text{logit}(\lambda_{i3}^{*}) = \alpha_{30}+\alpha_{31}\cdot r_{i2}^{y}y_{i2}+\alpha_{32}\cdot\text{GENDER}_i+\alpha_{33}\cdot\text{TRT}_i+\alpha_{34}\cdot r_{i2}^{x}\text{SMR}_{i2}+\alpha_{35}\cdot r_{i2}^{*},$$

where $\lambda_{ij}^{*} = P(R_{ij}^{*} = 1|R_{i,j-1}^{*}, Y_i^{(o)}, X_i^{(o)})$, $j = 2, 3$. The third analysis is the "IPWGEE-I" method, in which the weights are determined from standard logistic regression models given by (4.13), (4.14), (4.15) and (4.16) with the assumption $\psi_{ij}^{xy} = 1$. The fourth analysis, labeled "IPWGEE-J", is based on a weighted generalized estimating equations given by (4.13), (4.14), (4.15) and (4.16) by accommodating the association between the missingness indicators of the response and the covariate, as described in Section 4.3. Namely, assume the model $\log(\psi_{ij}) = \phi_j$ for $j = 2, 3$. The last method, entitled "AIPWGEE-J", is the method described in Section 4. 4, where we choose $A_i(\alpha) = (A_{i1}', A_{i2}', A_{i3}')'$ with

$$A_{i1} = \left(\left[\frac{I(R_{ij}^{y}=1, R_{ij}^{x}=0)}{\pi_{ij}^{y}-\pi_{ij}^{xy}}\cdot\pi_{ij}^{y}-1\right]\cdot R_{ij}^{y}\cdot Y_{ij},\ j=2,3\right)',$$

$$A_{i2} = \left(\left[\frac{I(R_{ij}^{y}=0, R_{ij}^{x}=1)}{\pi_{ij}^{x}-\pi_{ij}^{xy}}\cdot\pi_{ij}^{x}-1\right]\cdot R_{ij}^{x}\cdot \text{SMR}_{ij},\ j=2,3\right)',$$

$$A_{i3} = \left(\left[\frac{I(R_{ij}^y = 0, R_{ij}^x = 0)}{1 - \pi_{ij}^x - \pi_{ij}^y + \pi_{ij}^{xy}} - 1 \right] \cdot Z'_{ij},\ j = 2, 3 \right)',$$

and $Z_{ij} = (\text{GENDER}_i, \text{TRT}_i, \text{GRADE7}_{ij}, \text{GRADE8}_{ij})'$, $j = 2, 3$.

To understand how the estimation of the mean parameter β may be influenced by different specifications of the covariance structure, here we consider various association structures for the response process, and the results are reported in Table 4.5. Under each association structure specification, these five methods produce fairly comparable results, although the estimates obtained from the "IPWGEE-J" and "AIPWGEE-J" methods tend to be closer than the estimates obtained from other methods. The "AIPWGEE-J" method also gives smaller standard errors, supporting the expectation that this method is more efficient than the "IPWGEE-J" method. All the five methods reveal that both gender and treatment covariates are not statistically significant, while social model risk score and grade have significant negative effects on smoking incidence. Students are more likely to smoke if their parents, siblings or friends are smokers, and they are more likely to smoke when in higher grades.

Table 4.6 records the results for the missing data processes obtained from the "IPWGEE-I", "IPWGEE-J" and "IPWGEE-M2" methods. The "IPWGEE-I" and "IPWGEE-J" methods lead to fairly comparable estimates for the marginal mean parameters associated with both the missing response and missing covariate processes in grade 7. The results for grade 8 differ more noticeably. However, both methods reveal the same nature of the missing data mechanism. Specifically, α_{y21}, α_{y24}, α_{y31}, α_{y34} and α_{y35} are statistically significant, suggesting that a missing at random mechanism is perhaps reasonable for the missing response process. Similarly, significance of α_{x21}, α_{x31} and α_{x34} suggests that a missing at random mech-

Table 4.5: Results of estimation based on unweighted and weighted GEE when analyzing data from the Waterloo Smoking Prevention Project: response models

	GEE			IPWGEE-M2			IPWGEE-I			IPWGEE-J			AIPWGEE-J		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value
							Exchangeable								
β_0	-4.093	0.146	<0.001	-3.986	0.123	<0.001	-3.977	0.151	<0.001	-3.993	0.123	<0.001	-3.992	0.121	<0.001
β_1	0.042	0.081	0.600	-0.015	0.071	0.829	-0.007	0.088	0.940	-0.017	0.070	0.810	-0.017	0.068	0.809
β_2	-0.091	0.096	0.346	-0.098	0.084	0.243	-0.098	0.105	0.352	-0.096	0.084	0.253	-0.096	0.083	0.239
β_3	0.747	0.085	<0.001	0.730	0.068	<0.001	0.762	0.082	<0.001	0.728	0.069	<0.001	0.727	0.067	<0.001
β_4	1.545	0.080	<0.001	1.493	0.065	<0.001	1.518	0.079	<0.001	1.498	0.065	<0.001	1.498	0.064	<0.001
β_5	1.745	0.100	<0.001	1.731	0.087	<0.001	1.720	0.106	<0.001	1.734	0.088	<0.001	1.736	0.086	<0.001
							AR(1)								
β_0	-4.104	0.178	<0.001	-3.969	0.111	<0.001	-3.961	0.123	<0.001	-3.978	0.113	<0.001	-3.977	0.110	<0.001
β_1	0.093	0.096	0.339	0.008	0.065	0.897	0.011	0.072	0.878	0.004	0.065	0.956	0.004	0.064	0.955
β_2	-0.131	0.116	0.260	-0.122	0.077	0.115	-0.119	0.086	0.167	-0.120	0.077	0.120	-0.120	0.075	0.109
β_3	0.753	0.102	<0.001	0.738	0.061	<0.001	0.767	0.066	<0.001	0.737	0.063	<0.001	0.736	0.061	<0.001
β_4	1.536	0.108	<0.001	1.494	0.060	<0.001	1.522	0.065	<0.001	1.502	0.061	<0.001	1.503	0.060	<0.001
β_5	1.760	0.122	<0.001	1.717	0.077	<0.001	1.705	0.084	<0.001	1.720	0.080	<0.001	1.721	0.077	<0.001
							Unstructured								
β_0	-4.094	0.146	<0.001	-3.956	0.135	<0.001	-3.943	0.124	<0.001	-3.968	0.114	<0.001	-3.969	0.111	<0.001
β_1	0.086	0.081	0.278	0.013	0.077	0.863	0.017	0.072	0.818	0.008	0.064	0.900	0.008	0.063	0.899
β_2	-0.119	0.097	0.215	-0.116	0.092	0.207	-0.111	0.086	0.197	-0.113	0.077	0.142	-0.113	0.075	0.131
β_3	0.752	0.087	<0.001	0.740	0.073	<0.001	0.772	0.065	<0.001	0.737	0.061	<0.001	0.738	0.060	<0.001
β_4	1.547	0.083	<0.001	1.492	0.072	<0.001	1.517	0.065	<0.001	1.499	0.059	<0.001	1.501	0.057	<0.001
β_5	1.742	0.100	<0.001	1.701	0.096	<0.001	1.685	0.089	<0.001	1.706	0.082	<0.001	1.707	0.080	<0.001

Table 4.6: Results of estimation based on unweighted and weighted GEE when analyzing data from the Waterloo Smoking Prevention Project: missing data models

		IPWGEE-I			IPWGEE-J		
Parameters		Estimate	S.E.	p-value	Estimate	S.E.	p-value
Marginal missing-response models:							
Grade 7:							
INTERCEPT	α_{y20}	3.073	0.165	< 0.001	3.041	0.163	< 0.001
PREV. RES	α_{y21}	-1.059	0.181	< 0.001	-1.060	0.181	< 0.001
GENDER	α_{y22}	-0.295	0.116	0.011	-0.286	0.115	0.013
TRT	α_{y23}	-0.164	0.144	0.255	-0.143	0.142	0.313
PREV. SMR	α_{y24}	-0.343	0.122	0.005	-0.325	0.121	0.007
Grade 8:							
INTERCEPT	α_{y30}	-0.643	0.172	< 0.001	-0.318	0.162	0.050
PREV. RES	α_{y31}	-0.570	0.174	0.001	-0.503	0.175	0.004
GENDER	α_{y32}	-0.240	0.112	0.033	-0.289	0.109	0.008
TRT	α_{y33}	0.132	0.133	0.322	0.151	0.130	0.245
PREV. SMR	α_{y34}	-0.475	0.138	0.001	-0.461	0.137	0.001
PREV. MIS. IND.	α_{y35}	3.693	0.159	< 0.001	3.352	0.150	< 0.001
Marginal missing-covariate models:							
Grade 7:							
INTERCEPT	α_{x20}	2.882	0.157	< 0.001	2.881	0.156	< 0.001
PREV. RES	α_{x21}	-1.012	0.179	< 0.001	-1.020	0.178	< 0.001
GENDER	α_{x22}	-0.178	0.109	0.103	-0.179	0.109	0.102
TRT	α_{x23}	-0.273	0.140	0.052	-0.271	0.140	0.053
PREV. SMR	α_{x24}	-0.212	0.114	0.062	-0.212	0.114	0.063
Grade 8:							
INTERCEPT	α_{x30}	-0.316	0.160	0.048	-0.264	0.159	0.097
PREV. RES	α_{x31}	-0.385	0.173	0.026	-0.399	0.173	0.021
GENDER	α_{x32}	-0.332	0.106	0.002	-0.321	0.106	0.002
TRT	α_{x33}	0.119	0.127	0.351	0.099	0.127	0.435
PREV. SMR	α_{x34}	-0.440	0.132	0.001	-0.370	0.129	0.004
PREV. MIS. IND	α_{x35}	3.266	0.149	< 0.001	3.170	0.144	< 0.001
Association:							
ϕ_2					8.860	4.586	0.053
ϕ_3					6.877	1.252	<0.001
Estimates for IPWGEE-M2 Analysis:							
Grade 7:							
INTERCEPT	α_{20}	2.835	0.154	< 0.001			
PREV. RES	α_{21}	-1.020	0.178	< 0.001			
GENDER	α_{22}	-0.173	0.109	0.110			
TRT	α_{23}	-0.237	0.138	0.086			
PREV. SMR	α_{24}	-0.216	0.113	0.057			
Grade 8:							
INTERCEPT	α_{30}	-0.292	0.158	0.065			
PREV. RES	α_{31}	-0.408	0.172	0.017			
GENDER	α_{32}	-0.320	0.105	0.002			
TRT	α_{33}	0.122	0.126	0.331			
PREV. SMR	α_{34}	-0.367	0.129	0.004			
PREV. MIS. IND	α_{35}	3.164	0.145	< 0.001			

anism is perhaps reasonable for the missing covariate process. Significance of ϕ_2 and ϕ_3 in the "IPWGEE-J" model indicates that there is association between missingness of the response and covariate, and this association should be taken into account for the inference. Significance of α_{y35} and α_{x35} indicates there is a serial dependence among consecutive observations. Moreover, if subjects have missing observations at the previous assessment time, they are less likely to be observed at the present assessment. Significance of α_{y21}, α_{y31}, α_{x21} and α_{x31} indicates that the previously observed smoking status has a negative effect on observing the present assessment. Significance of GENDER in the missing response model suggests that female students are more likely to participate in the study compared to male students. However, it is not significant in the missing covariate models. Treatment has no significant effects on the missingness of response or covariate. Significance of α_{y24}, α_{y34}, α_{x24} and α_{x34} suggests that the previously observed social models risk score has a negative effect on the missingness of the assessment. Students are more likely to participate in the study when none of their parents, siblings or friends smoke. The estimates based on the "IPWGEE-M2" method are not compatible with those from the "IPWGEE-I" and "IPWGEE-J" methods. However, it appears that the "IPWGEE-M2" modeling method also detects evidence for a missing at random mechanism, indicating by the nature of the estimates for α_{21}, α_{24}, α_{31} and α_{34}.

4.5.4 Application to a Study of Patients with Skeletal Metastases

In this subsection, we apply the proposed methods to study a bone metastases data set (Hortobagyi et al., 1998). Women with advanced breast cancer often experience bone metastases. From January 1991 to March 1994, the Protocol 19

Aredia Breast Cancer Study Group of Novartis Pharmaceuticals Inc. conducted a randomized clinical trial at 97 sites in the United States, Canada, Australia and New Zealand. The osteoclast activating factors released by tumor cells cause destruction of bone, which in turn leads to the occurrence of the aforementioned skeletal complications. Radiographic surveys of bone lesions were performed and new bone lesions were recorded. The objective of this study is to evaluate covariate effects on the occurrence of bone lesions for patients with breast cancer. The response is the lesion code. Covariates of interests include age at study entry (coded as AGE: 1 for age $\geq$ 50, 0 for age $<$ 50), ECOG score at study entry (coded as ECOG: 1 for two or more, 0 otherwise), the number of fractures at baseline (coded as FRACT: 1 for one or more, 0 for none), pain score at study entry (coded as PSCORE) which is coded as four levels based on the 25%, 50% and 75% quantiles, and urinary hydroxyproline/creatinine ratio (coded as HYCRR). Table 4.7 represents a sample dataset.

Two hundred and twenty patients entered the study and were intended to be assessed at baseline, 6 months and 12 months from the baseline. However, the collected measurements are incomplete. Proportions of various patterns of the missingness $(R_2^y, R_2^x) = (1,1)$, $(0,0)$, $(0,1)$ and $(1,0)$ are 70.0%, 14.5%, 2.3% and 13.2%, respectively, and $(R_3^y, R_3^x) = (1,1)$, $(0,0)$, $(0,1)$ and $(1,0)$ are 70.0%, 9.1%, 1.8% and 19.1%, respectively.

Let $Y_{ij} = 1$ if patient i at time j has a new lesion, and 0 otherwise, $j = 1, 2, 3$. Consider the model for the marginal probabilities

$$\begin{aligned} \text{logit}\, \mu_{ij} &= \beta_0 + \beta_1 \cdot \text{AGE}_i + \beta_2 \cdot \text{ECOG}_i + \beta_3 \cdot \text{FRACT}_i + \beta_4 \cdot \text{PSCORE1}_i \\ &\quad + \beta_5 \cdot \text{PSCORE2}_i + \beta_6 \cdot \text{PSCORE3}_i + \beta_7 \cdot \text{HYCRR}_{ij}, \quad j = 1, 2, 3, \end{aligned}$$

Table 4.7: Sample data from a bone metastases study

							HYCRR†			LESION†		
ID	AGE	ECOG	FRACT	PSCORE1	PSCORE2	PSCORE3	1	2	3	1	2	3
1	1	1	0	0	0	1	0.076	0.064	0.042	1	0	0
2	1	0	0	0	1	0	0.039	0.006	0.011	0	0	1
3	1	1	1	0	0	0	0.094	·	0.103	1	1	0
4	1	0	1	0	0	1	0.050	0.034	0.030	0	0	1
5	1	0	0	0	1	0	0.027	0.027	0.046	1	0	0
6	1	1	1	1	0	0	0.103	0.071	0.127	0	1	0
7	1	0	0	0	0	0	0.067	0.052	0.029	1	1	1
8	1	1	1	0	0	0	0.175	0.147	0.177	0	1	0
9	1	1	0	1	0	0	0.044	0.077	0.040	0	0	1
10	1	0	1	0	1	0	0.068	0.080	0.057	1	·	1
11	1	0	0	1	0	0	0.012	0.011	0.008	0	0	1
12	1	0	0	0	1	0	0.059	0.061	·	0	·	1
13	1	1	1	0	1	0	0.040	0.028	0.040	0	0	1
14	1	0	0	0	0	0	0.026	0.028	0.020	0	0	1
15	1	0	0	1	0	0	0.026	·	0.026	0	0	1
16	1	0	1	0	0	1	0.033	0.043	·	0	·	1
17	1	0	0	1	0	0	0.051	0.026	0.025	0	0	1
18	1	0	1	0	0	1	0.048	0.069	0.038	0	0	0
19	1	0	0	0	0	0	0.027	0.014	0.018	0	1	1
20	1	0	0	0	0	1	0.051	0.013	·	0	·	1
21	1	0	1	0	0	0	0.044	0.004	0.018	0	0	1
22	1	0	0	0	1	0	0.018	·	0.012	0	0	0
23	1	0	0	0	1	0	0.041	·	·	0	0	1
24	1	0	0	1	0	0	0.039	0.028	0.032	1	·	0
25	1	1	0	1	0	0	0.077	0.051	·	0	·	1
26	1	1	0	0	0	0	0.173	0.152	0.106	0	0	1
27	1	0	0	0	0	0	0.201	0.104	0.063	0	0	1
28	1	0	0	0	1	0	0.048	0.022	0.019	0	0	1

† Missing data are denoted by ·.

where $\text{PSCORE1}_i = 1$ if the pain score at study entry is between the 25% and 50% quantiles, and 0 otherwise; $\text{PSCORE2}_i = 1$ if the pain score at study entry is between the 50% and 75% quantiles, and 0 otherwise; and $\text{PSCORE3}_i = 1$ if the pain score at study entry is higher than the 75% quantile, and 0 otherwise.

For the missing response and covariate indicators, we specify the models

$$\begin{aligned}\text{logit } \lambda_{ij}^{y} &= \alpha_{y0} + \alpha_{y1} \cdot \text{AGE}_i + \alpha_{y2} \cdot \text{ECOG}_i + \alpha_{y3} \cdot \text{FRACT}_i + \alpha_{y4} \cdot \text{PSCORE1}_i \\ &\quad + \alpha_{y5} \cdot \text{PSCORE2}_i + \alpha_{y6} \cdot \text{PSCORE3}_i + \alpha_{y7} \cdot r_{i,j-1}^{x} \text{HYCRR}_{i,j-1} \\ &\quad + \alpha_{y8} \cdot r_{i,j-1}^{y} y_{i,j-1}, \quad j = 2, 3, \end{aligned} \tag{4.17}$$

and

$$\begin{aligned}\text{logit } \lambda_{ij}^{x} &= \alpha_{x0} + \alpha_{x1} \cdot r_{i,j-1}^{x} + \alpha_{x2} \cdot \text{AGE}_i + \alpha_{x3} \cdot \text{ECOG}_i + \alpha_{x4} \cdot \text{FRACT}_i \\ &\quad + \alpha_{x5} \cdot \text{PSCORE1}_i + \alpha_{x6} \cdot \text{PSCORE2}_i + \alpha_{x7} \cdot \text{PSCORE3}_i \\ &\quad + \alpha_{x8} \cdot r_{i,j-1}^{x} \text{HYCRR}_{i,j-1} + \alpha_{x9} \cdot r_{i,j-1}^{y} y_{i,j-1}, \quad j = 2, 3. \end{aligned} \tag{4.18}$$

For the association between the missingness of the response and covariate, we assume a common odds ratio at each assessment, i.e., $\log(\psi_{ij}) = \phi$ for $j = 2, 3$.

Analogous to Section 4.5.3, here we undertake five methods to analyze the data. Specifically, in the "IPWGEE-M2" analysis we use the weights determined from the model

$$\begin{aligned}\text{logit}\lambda_{ij}^{*} &= \alpha_0 + \alpha_1 \cdot r_{i,j-1}^{*} + \alpha_2 \cdot \text{AGE}_i + \alpha_3 \cdot \text{ECOG}_i + \alpha_4 \cdot \text{FRACT}_i \\ &\quad + \alpha_5 \cdot \text{PSCORE1}_i + \alpha_6 \cdot \text{PSCORE2}_i + \alpha_7 \cdot \text{PSCORE3}_i \\ &\quad + \alpha_8 \cdot r_{i,j-1}^{x} \text{HYCRR}_{i,j-1} + \alpha_9 \cdot r_{i,j-1}^{y} y_{i,j-1}, \end{aligned}$$

where $\lambda_{ij}^{*} = P(R_{ij}^{*} = 1 | R_{i,j-1}^{*}, Y_i^{(o)})$. For the "AIPWGEE-J" method we choose

$A_i(\alpha) = (A'_{i1}, A'_{i2}, A'_{i3})'$ where

$$A_{i1} = \left(\left[\frac{I(R^y_{ij} = 1, R^x_{ij} = 0)}{\pi^y_{ij} - \pi^{xy}_{ij}} \cdot \pi^y_{ij} - 1 \right] \cdot R^y_{ij} \cdot Y_{ij},\ j = 2, 3 \right)',$$

$$A_{i2} = \left(\left[\frac{I(R^y_{ij} = 0, R^x_{ij} = 1)}{\pi^x_{ij} - \pi^{xy}_{ij}} \cdot \pi^x_{ij} - 1 \right] \cdot R^x_{ij} \cdot \text{HYCRR}_{ij},\ j = 2, 3 \right)',$$

$$A_{i3} = \left(\left[\frac{I(R^y_{ij} = 0, R^x_{ij} = 0)}{1 - \pi^x_{ij} - \pi^y_{ij} + \pi^{xy}_{ij}} - 1 \right] \cdot Z'_{ij},\ j = 2, 3 \right)',$$

and $Z_{ij} = (\text{AGE}_i, \text{ECOG}_i, \text{FRACT}_i, \text{PSCORE1}_i, \text{PSCORE2}_i, \text{PSCORE3}_i)'$, $j = 2, 3$.

Table 4.8 reports on the results for the response model. The "GEE", "IPWGEE-M2" and "IPWGEE-I" methods give the estimates that are much different from those obtained from those obtained from the "IPWGEE-J" and "AIPWGEE-J" methods. Again, it is seen that the "AIPWGEE-J" method leads to smaller standard errors than the "IPWGEE-J" method, which agrees with our expectation that the "AIPWGEE-J" method is more efficient. All these methods suggest that only the HYCRR is statistically significant. The "GEE" method provides strongest evidence of the HYCRR effect, while the "IPWGEE-M2" method tends to reveal weakest evidence for that. It even fails to support an HYCRR effect when the association among the response components is assumed unstructured. The "AIPWGEE-J" method seems to provide stronger evidence for the HYCRR effect than the "IPWGEE-J" method.

Table 4.9 records the results for the missing data processes obtained from the "IPWGEE-I", "IPWGEE-J" and "IPWGEE-M2" methods. The "IPWGEE-I" and "IPWGEE-J" methods lead to fairly comparable estimates for the parameters associated with both the missing response and missing covariate processes. Both methods reveal the same nature of the missing data mechanism. Specifically, α_{y8}

Table 4.8: Results of estimation based on unweighted and weighted GEE when analyzing data from a bone metastases study: response models

	Unweighted			IPWGEE-M2			IPWGEE-I			IPWGEE-J			AIPWGEE-J		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value
							Exchangeable								
β_0	-0.332	0.245	0.175	-0.572	0.357	0.109	-0.846	0.432	0.050	-0.806	0.413	0.051	-0.805	0.394	0.041
β_1	-0.112	0.160	0.484	-0.111	0.234	0.635	-0.071	0.284	0.803	-0.075	0.272	0.783	-0.074	0.268	0.782
β_2	-0.197	0.189	0.297	-0.226	0.270	0.403	-0.280	0.345	0.417	-0.266	0.332	0.423	-0.267	0.317	0.400
β_3	-0.221	0.182	0.225	-0.128	0.231	0.580	-0.048	0.280	0.864	-0.044	0.273	0.872	-0.045	0.263	0.864
β_4	0.076	0.224	0.734	0.090	0.323	0.781	0.110	0.391	0.778	0.110	0.376	0.770	0.109	0.368	0.767
β_5	0.151	0.212	0.476	0.181	0.310	0.559	0.244	0.369	0.508	0.235	0.355	0.508	0.234	0.344	0.496
β_6	0.188	0.201	0.350	0.239	0.335	0.476	0.202	0.412	0.624	0.196	0.397	0.622	0.195	0.389	0.616
β_7	8.430	2.625	0.001	9.155	3.863	0.018	10.846	4.932	0.028	10.201	4.713	0.030	10.199	4.594	0.026
							AR(1)								
β_0	-0.639	0.286	0.025	-0.969	0.339	0.004	-1.158	0.398	0.004	-0.785	0.395	0.047	-0.784	0.372	0.035
β_1	-0.087	0.189	0.645	-0.090	0.221	0.684	-0.048	0.260	0.854	-0.041	0.257	0.873	-0.041	0.254	0.872
β_2	-0.257	0.226	0.255	-0.315	0.262	0.229	-0.325	0.321	0.311	-0.295	0.317	0.352	-0.294	0.306	0.337
β_3	-0.323	0.223	0.147	-0.182	0.213	0.393	-0.109	0.248	0.660	-0.102	0.253	0.687	-0.103	0.242	0.670
β_4	0.086	0.273	0.753	0.123	0.308	0.690	0.125	0.356	0.725	0.107	0.355	0.763	0.107	0.349	0.759
β_5	0.260	0.252	0.302	0.309	0.295	0.295	0.340	0.334	0.309	0.285	0.334	0.393	0.286	0.325	0.379
β_6	0.271	0.240	0.259	0.328	0.325	0.313	0.287	0.376	0.445	0.217	0.380	0.568	0.219	0.375	0.559
β_7	12.410	3.018	<0.001	13.465	3.718	<0.001	14.186	4.617	0.002	12.843	4.736	0.007	12.840	4.539	0.005
							Unstructured								
β_0	0.037	0.120	0.758	-0.118	0.324	0.716	-0.372	0.278	0.181	-0.297	0.263	0.259	-0.298	0.248	0.230
β_1	-0.153	0.075	0.041	-0.130	0.215	0.545	-0.078	0.186	0.675	-0.087	0.177	0.623	-0.086	0.174	0.621
β_2	-0.096	0.087	0.270	-0.107	0.242	0.658	-0.179	0.217	0.409	-0.153	0.207	0.460	-0.152	0.198	0.443
β_3	-0.197	0.086	0.022	-0.154	0.198	0.437	-0.075	0.175	0.668	-0.081	0.168	0.630	-0.082	0.162	0.613
β_4	0.020	0.106	0.850	0.018	0.280	0.949	0.027	0.243	0.912	0.010	0.229	0.965	0.009	0.223	0.968
β_5	0.066	0.100	0.509	0.141	0.276	0.609	0.180	0.233	0.440	0.168	0.221	0.447	0.167	0.213	0.433
β_6	0.143	0.093	0.124	0.185	0.301	0.539	0.149	0.259	0.565	0.137	0.247	0.579	0.137	0.243	0.573
β_7	3.269	1.279	0.011	5.766	3.398	0.090	7.698	3.061	0.012	6.716	2.871	0.019	6.719	2.755	0.015

Table 4.9: Results of estimation based on unweighted and weighted GEE when analyzing data from a bone metastases study: missing data models

		IPWGEE-I			IPWGEE-J		
Parameters		Estimate	S.E.	p-value	Estimate	S.E.	p-value
Marginal missing-response models:							
INTERC.	α_{y0}	0.837	0.343	0.015	0.802	0.338	0.018
AGE	α_{y1}	0.357	0.248	0.149	0.332	0.245	0.176
ECOG	α_{y2}	0.094	0.297	0.752	0.009	0.292	0.975
FRACT	α_{y3}	0.574	0.349	0.099	0.603	0.345	0.081
PSCORE1	α_{y4}	-0.335	0.378	0.375	-0.263	0.372	0.480
PSCORE2	α_{y5}	-0.127	0.356	0.721	-0.056	0.350	0.873
PSCORE3	α_{y6}	-0.476	0.330	0.149	-0.394	0.322	0.221
PREV. HYCRR	α_{y7}	-0.422	3.340	0.899	0.285	3.310	0.931
PREV. RESP	α_{y8}	0.846	0.253	0.001	0.800	0.249	0.001
Marginal missing-covariate models:							
INTERC.	α_{x0}	-0.841	0.353	0.017	-0.610	0.335	0.069
PREV. MIS. IND	α_{x1}	1.650	0.340	<0.001	0.915	0.281	0.001
AGE	α_{x2}	0.127	0.226	0.572	0.210	0.220	0.340
ECOG	α_{x3}	-0.133	0.268	0.619	-0.285	0.260	0.274
FRACT	α_{x4}	0.693	0.300	0.021	0.712	0.295	0.016
PSCORE1	α_{x5}	-0.082	0.340	0.809	-0.065	0.330	0.845
PSCORE2	α_{x6}	-0.123	0.311	0.693	-0.023	0.303	0.940
PSCORE3	α_{x7}	-0.357	0.288	0.216	-0.318	0.281	0.257
PREV. HYCRR	α_{x8}	-2.450	3.393	0.470	2.814	3.295	0.393
PREV. RESP	α_{x9}	0.508	0.223	0.022	0.583	0.217	0.007
Association:							
ϕ					3.929	1.547	<0.001
Estimates for IPWGEE-M2 Analysis:							
INTERC.	α_0	-0.851	0.338	0.012			
PREV. MIS. IND	α_1	1.668	0.317	<0.001			
AGE	α_2	0.137	0.221	0.534			
ECOG	α_3	-0.034	0.262	0.898			
FRACT	α_4	0.651	0.291	0.025			
PSCORE1	α_5	-0.086	0.332	0.796			
PSCORE2	α_6	-0.160	0.302	0.597			
PSCORE3	α_7	-0.285	0.283	0.313			
PREV. HYCRR	α_8	-2.916	3.192	0.361			
PREV. RESP	α_9	0.279	0.221	0.206			

is statistically significant, suggesting that a missing at random mechanism is perhaps reasonable for the missing response process. Similarly, little significance of α_{x4} and α_{x9} suggests that a missing at random mechanism is perhaps reasonable for the missing covariate process. Significance of the association parameter ϕ in the "IPWGEE-J" method suggests there is association between missingness of the response and covariate. Significance of α_{x1} indicates there is a serial dependence among consecutive observations of covariate HYCRR. Moreover, if subjects have missing covariate HYCRR at the previous assessment time, then they are more likely to miss the present assessment. Significance of α_{y8} and α_{x9} indicates that the previously observed new lesion has a positive effect on observing the present assessment. FRACT in the missing response model is moderately significant, and it is significant in the missing covariate model. The more number of fractures, the larger probability to observe the responses and covariate. The estimates based on the "IPWGEE-M2" method are not compatible with those from the "IPWGEE-I" and "IPWGEE-J" methods. However, it appears that the "IPWGEE-M2" modeling method also detects evidence for a missing at random mechanism, indicating by the nature of the estimates for α_4.

4.6 Extension to Accommodate Multiple Missing Covariates

In the preceding sections we focus on the case that only a single covariate, along with the response, may be missing. In this section, we extend the proposed methods to accommodate circumstances that multiple covariates could be missing. Slightly different notation is used in this section.

Let R_{ijk} be the missing indicator for covariate X_{ijk}, $k = 1, 2, \ldots, p_1$ with $p_1 = \dim(X_{ij})$ where $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp_1})'$, and R_{ij0} be the missing indicator for the response Y_{ij}. Denote $R_{ij} = (R_{ij0}, R_{ij1}, \ldots, R_{ijp_1})'$. Assume $R_{i1} = \mathbb{1}$, where $\mathbb{1}$ is the $(p_1+1) \times 1$ vector of element 1. Let $\pi_{ij} = P(R_{ij} = \mathbb{1}|Y_i, X_i, Z_i)$. Under a MAR mechanism with

$$P(R_{ij} = r_{ij}|\bar{R}_{i,j-1}, Y_i, X_i, Z_i) = P(R_{ij} = r_{ij}|\bar{R}_{i,j-1}, Y_i^{(o)}, X_i^{(o)}, Z_i) \ ,$$

where $\bar{R}_{i,j-1} = \{r_{i1}, r_{i2}, \ldots, r_{i,j-1}\}$ with r_{ik} being a realization of R_{ik}, we write

$$\pi_{ij} \quad = \quad \sum_{\bar{R}_{i,j-1}} \{P(R_{ij} = \mathbb{1}|\bar{R}_{i,j-1}, Y_i, X_i, Z_i) \cdot \prod_{\ell=1}^{j-1} P(R_{i\ell} = r_{i\ell}|\bar{R}_{i,\ell-1}, Y_i, X_i, Z_i)\}, \quad (4.19)$$

for $j \geq 2$, where $\pi_{i1} = 1$ is assumed. To determine π_{ij}, we further model the joint probability $P(R_{i\ell} = r_{i\ell}|\bar{R}_{i,\ell-1}, Y_i, X_i, Z_i)$. More specifically, let $\delta_{ijk} = P(R_{ijk} = 1|\bar{R}_{i,j-1}, Y_i, X_i, Z_i)$ for $k = 0, 1, \ldots, p_1$. Let $R^*_{ijk} = (R_{ijk} - \delta_{ijk})/\sqrt{\delta_{ijk}(1-\delta_{ijk})}$, $\rho_{ijst} = E(R^*_{ijs}R^*_{ijt})$, and $\rho_{ijs_1s_2\cdots s_K} = E(R^*_{ijs_1}R^*_{ijs_2} \cdots R^*_{ijs_K})$ be the Kth-order correlation among components $R_{ijs_1}, R_{ijs_2}, \ldots, R_{ijs_K}$ of R_{ij}, where $\rho = (\rho_{01}, \rho_{12}, \ldots, \rho_{01\cdots p_1})'$. For given time point j, we employ the Bahadur representation (Bahadur, 1961; Cox, 1972) to express the joint probability

$$\begin{aligned} & P(R_{ij} = r_{ij}|\bar{R}_{i,j-1}, Y_i, X_i, Z_i) \\ = \ & \prod_{k=0}^{p} \left\{\delta_{ijk}^{r_{ijk}}(1-\delta_{ijk})^{1-r_{ijk}}\right\} \cdot \Big\{1 + \sum_{s<t} \rho_{ijst} r^*_{ijs} r^*_{ijt} \\ & + \sum_{u<s<t} \rho_{ijust} r^*_{iju} r^*_{ijs} r^*_{ijt} + \cdots + \rho_{01\cdots p} r^*_{ij0} r^*_{ij1} \cdots r^*_{ijp_1}\Big\}. \end{aligned} \quad (4.20)$$

This strategy requires modeling the correlation structures of all orders. In practice, it is often the case that the second order dominates the association structure while the third and higher order association is null or nearly null. Under such circumstances, we may typically perform estimation along the lines of Sections 4.2

and 4.3. That is, given j, for $k' > k = 0, 1, \ldots, p_1 - 1$, let

$$\psi_{ijkk'} = \frac{P(R_{ijk} = 1, R_{ijk'} = 1|\bar{R}_{i,j-1}, Y_i, X_i, Z_i)P(R_{ijk} = 0, R_{ijk'} = 0|\bar{R}_{i,j-1}, Y_i, X_i, Z_i)}{P(R_{ijk} = 0, R_{ijk'} = 1|\bar{R}_{i,j-1}, Y_i, X_i, Z_i)P(R_{ijk} = 1, R_{ijk'} = 0|\bar{R}_{i,j-1}, Y_i, X_i, Z_i)}$$

be the odds ratio featuring the association between R_{ijk} and $R_{ijk'}$. Regression models may be invoked to characterize the building blocks δ_{ijk} and $\rho_{ijkk'}$ (or equivalently, $\psi_{ijkk'}$) in (4.20) where the third or higher order correlations are constrained to be zero. Let α^* and ϕ be the parameters associated with the models δ_{ijk} and $\psi_{ijkk'}$ respectively, and denote $\alpha = (\alpha^{*\prime}, \phi')'$. Parameters α^* and ϕ for the missing data processes can be estimated by solving

$$S_1(\alpha) = \sum_{i=1}^{n} S_{1i}(\alpha) = 0$$

and

$$S_2(\alpha) = \sum_{i=1}^{n} S_{2i}(\alpha) = 0$$

where $S_{1i}(\alpha) = [\partial\delta_i/\partial\alpha^{*\prime}]W_i^{-1}(R_i - \delta_i)$, $S_{2i}(\alpha) = [\partial\delta_i^*/\partial\phi']W_i^{*-1}(R_i^* - \delta_i^*)$, $R_i = (R_{ij}', j = 2, 3, \ldots, J)'$, $\delta_i = (\delta_{ij}', j = 2, 3, \ldots, J)'$, $\delta_{ij} = (\delta_{ij0}, \delta_{ij1}, \ldots, \delta_{ijp_1})'$, $W_i = \text{diag}(W_{ij}, j = 2, 3, \ldots, J)$, W_{ij} is the $(p_1 + 1) \times (p_1 + 1)$ matrix with (k, k) element $\delta_{ijk}(1 - \delta_{ijk})$ and (k, k') element $\delta_{ijkk'} - \delta_{ijk}\delta_{ijk'}$, $R_i^* = (R_{ij}^{*\prime}, j = 2, 3, \ldots, J)'$, $R_{ij}^* = (R_{ijk} \cdot R_{ijk'}, k < k')'$, $\delta_i^* = (\delta_{ij}^{*\prime}, j = 2, 3, \ldots, J)'$, $\delta_{ij}^* = (\delta_{ijkk'}, k < k')'$, and $W_i^* = \text{diag}(\delta_i^*(1 - \delta_i^*))$.

If the third or higher order correlation is not zero, we need to calculate π_{ij} using (4.20) for which we may use an ad hoc way (e.g., Lipsitz et al., 1995) to replace the ℓth order correlation $\rho_{ijk_1k_2\cdots k_\ell}$ with

$$\hat{\rho}_{ijk_1k_2\cdots k_\ell} = n^{-1}\sum_{i=1}^{n} \hat{R}_{ijk_1}^* \hat{R}_{ijk_2}^* \cdots \hat{R}_{ijk_\ell}^*,$$

where $\hat{R}^*_{ijk} = (R_{ijk} - \hat{\delta}_{ijk})/\sqrt{\hat{\delta}_{ijk}(1-\hat{\delta}_{ijk})}$. Consequently, estimation of the response parameter β can be performed by solving the estimating equations

$$\sum_{i=1}^{n}[U_i(\beta,\hat{\alpha}) - \hat{\eta}A_i(\hat{\alpha})] = 0,$$

where $U_i(\beta,\alpha) = D_i M_i(Y_i - \mu_i)$ with $M_i = \kappa^{-1}F_i^{-1/2}[R_i^{-1}(\rho) \bullet \Delta_i(\alpha)]F_i^{-1/2}$,

$$\Delta_i(\alpha) = \begin{pmatrix} \frac{I(R_{i1}=\mathbb{1})}{\pi_{i1}} & \frac{I(R_{i11}=1,\cdots,R_{i1p_1}=1,R_{i2}=\mathbb{1})}{\pi_{i12}} & \cdots & \frac{I(R_{i11}=1,\cdots,R_{i1p_1}=1,R_{iJ}=\mathbb{1})}{\pi_{i1J}} \\ \frac{I(R_{i21}=1,\cdots,R_{i2p_1}=1,R_{i1}=\mathbb{1})}{\pi_{i12}} & \frac{I(R_{i2}=\mathbb{1})}{\pi_{i2}} & \cdots & \frac{I(R_{i21}=1,\cdots,R_{i2p_1}=1,R_{iJ}=\mathbb{1})}{\pi_{i2J}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{I(R_{iJ1}=1,\cdots,R_{iJp_1}=1,R_{i1}=\mathbb{1})}{\pi_{i1J}} & \frac{I(R_{iJ1}=1,\cdots,R_{iJp_1}=1,R_{i2}=\mathbb{1})}{\pi_{i2J}} & \cdots & \frac{I(R_{iJ}=\mathbb{1})}{\pi_{iJ}} \end{pmatrix}_{J\times J},$$

$\pi_{ijk} = P(R_{ij1}=1,\cdots,R_{ijp_1}=1,R_{ik}=\mathbb{1}|Y_i,X_i,Z_i)$, and $\hat{\eta} = \hat{\eta}_1\hat{\eta}_2^{-1}$ with

$$\hat{\eta}_1 = n^{-1}\sum_{i=1}^{n}\widehat{\mathrm{Res}}[U_i(\hat{\beta},\hat{\alpha}),S_i(\hat{\alpha})]\widehat{\mathrm{Res}}[A_i(\hat{\alpha}),S_i(\hat{\alpha})]'$$

and

$$\hat{\eta}_2 = n^{-1}\sum_{i=1}^{n}\widehat{\mathrm{Res}}[A_i(\hat{\alpha}),S_i(\hat{\alpha})]\widehat{\mathrm{Res}}[A_i(\hat{\alpha}),S_i(\hat{\alpha})]'.$$

$A_i(\alpha)$ is a $m\times 1$ vector, typically chosen by the investigator, that does not involve unobserved data but satisfies $E[A_i(\alpha)] = 0$. For example, we could choose $A_i(\alpha) = (A'_{i0},\cdots,A'_{ip_1})'$ with

$$A_{ik} = \left(\left[\frac{I(R_{ij}=r_{ij})}{\pi_{ij}(r_{ij})}\cdot\pi_{ijk} - 1\right]\cdot R_{ijk}\cdot X_{ijk},\ j=0,1,\ldots,J\right)',$$

where $r_{ij} = (r_{ij0},\ldots,r_{ijp_1})'$ is a $(p_1+1)\times 1$ vector with $r_{ijk} = 1$, and $\pi_{ij}(r_{ij}) = P(R_{ij}=r_{ij}|Y_i,Z_i,X_i)$. The choice of the $A_i(\alpha)$ functions is not unique. It is often chosen for convenience, and in practice, a wide range of choices can lead to the improvement in efficiency. The asymptotic distribution of the resulting estimator can be established analogously to that in Section 4.4.

4.7 Estimation of the Asymptotic Covariance Matrix

For each component α_ℓ of α, $\ell = 1, 2, \ldots, q$, we define

$$G_\ell(\beta, \alpha) = \sum_{i=1}^{n} D_i \cdot (\partial M_i(\alpha)/\partial \alpha_\ell) \cdot (Y_i - \mu_i).$$

Then $E(\partial U_i(\beta, \alpha)/\partial \alpha')$ is consistently estimated by, as $n \to \infty$,

$$G(\hat{\beta}, \hat{\alpha}) = n^{-1} \left(G_1(\hat{\beta}, \hat{\alpha}), G_2(\hat{\beta}, \hat{\alpha}), \ldots, G_q(\hat{\beta}, \hat{\alpha}) \right).$$

If we let

$$M_{21}^*(\alpha) = -\sum_{i=1}^{n} C_i^* W_i^{*-1} \cdot (\partial \lambda_i^{xy}/\partial \alpha_{xy}'),$$

then $E(\partial S_i(\alpha)/\partial \alpha')$ is consistently estimated by, as $n \to \infty$,

$$M^*(\hat{\alpha}) = n^{-1} \begin{pmatrix} -\sum_{i=1}^{n} D_i^* V_i^{*-1} D_i^{*'} & 0 \\ M_{21}^*(\hat{\alpha}) & -\sum_{i=1}^{n} C_i^* W_i^{*-1} C_i^{*'} \end{pmatrix}.$$

The matrix Σ is consistently estimated by, as $n \to \infty$,

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} Q_i(\hat{\beta}, \hat{\alpha}) Q_i'(\hat{\beta}, \hat{\alpha}),$$

where $Q_i(\hat{\beta}, \hat{\alpha}) = U_i(\hat{\beta}, \hat{\alpha}) - G(\hat{\beta}, \hat{\alpha}) \cdot [M^*(\hat{\alpha})]^{-1} \cdot S_i(\hat{\alpha})$, and the matrix Γ is consistently estimated by, as $n \to \infty$, $\hat{\Gamma} = n^{-1} M(\hat{\beta}, \hat{\alpha})$. Inference about β is conducted by replacing Σ and Γ with these consistent estimates in the expression of the asymptotic covariance matrix.

4.8 Some Proof for the Efficient Estimate via Augmented IPWGEE

We first show that η is actually equal to the regression coefficient of $A_i(\alpha)$ in the population regression of $U_i(\beta,\alpha)$ on H_i^*. To see this, let η^* be the regression coefficient of H_i^* in the population regression of $U_i(\beta,\alpha)$ on H_i^*, that is, $\eta^* = E[U_i H_i^{*\prime}][E(H_i^* H_i^{*\prime})]^{-1}$, then by $H_i^* = (A_i'(\alpha), S_i'(\alpha))'$, we have

$$\begin{aligned}
\eta^* &= E[U_i A_i', U_i S_i'] \begin{pmatrix} E[A_i A_i'] & E[A_i S_i'] \\ E[S_i A_i'] & E[S_i S_i'] \end{pmatrix}^{-1} \\
&= (E[U_i A_i'], E[U_i S_i']) \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix} \\
&= \begin{pmatrix} E[U_i A_i']A^{11} + E[U_i S_i']A^{21} & E[U_i A_i']A^{12} + E[U_i S_i']A^{22} \end{pmatrix}
\end{aligned}$$

where

$$A^{11} = (E[A_i A_i'] - E[A_i S_i'] E[S_i S_i']^{-1} E[S_i A_i'])^{-1},$$

$$A^{21} = -E[S_i S_i']^{-1} E[S_i A_i'](E[A_i A_i'] - E[A_i S_i'] E[S_i S_i']^{-1} E[S_i A_i'])^{-1},$$

$$A^{12} = -E[A_i A_i']^{-1} E[A_i S_i'](E[S_i S_i'] - E[S_i A_i'] E[A_i A_i']^{-1} E[A_i S_i'])^{-1},$$

and

$$A^{22} = (E[S_i S_i'] - E[S_i A_i'] E[A_i A_i']^{-1} E[A_i S_i'])^{-1}.$$

Thus, the regression coefficient of A_i is $E[U_i A_i']A^{11} + E[U_i S_i']A^{21}$, which is equal to η after some algebra.

Subject to regularity conditions and that η is chosen as above, we obtain, using the same arguments in Section 4.3.3,

$$n^{1/2}(\tilde{\beta}^\dagger - \beta) = -\Gamma^{-1} n^{-1/2} \sum_{i=1}^{n} Q_i + o_p(1),$$

where $Q_i = U_i^\dagger - E[\partial U_i^\dagger/\partial\alpha'][E(\partial S_i/\partial\alpha')]^{-1}\cdot S_i$. Now we show that $Q_i = \text{Res}(U_i(\beta,\alpha), H_i^*)$. Noting that $E[\partial U_i^\dagger/\partial\alpha'] = -E[U_i^\dagger S_i']$ and $E(\partial S_i/\partial\alpha') = -E[S_i S_i']$, we write

$$\begin{aligned} Q_i &= U_i^\dagger - E[\partial U_i^\dagger/\partial\alpha'][E(\partial S_i/\partial\alpha')]^{-1}\cdot S_i \\ &= U_i - \eta A_i - E[U_i S_i' - \eta A_i S_i'][E(S_i S_i')]^{-1} S_i \\ &= U_i - \eta A_i - \eta_1 S_i, \end{aligned}$$

where $\eta_1 = E[U_i S_i' - \eta A_i S_i'][E(S_i S_i')]^{-1}$. Analogous to the preceding calculation, we can show that η_1 is the regression coefficient of S_i in the regression of U_i on H_i^*, so we have $Q_i = U_i - \eta^* H_i^* = \text{Res}(U_i(\beta,\alpha), H_i^*)$ with $\eta^* = (\eta, \eta_1)$. Thus, by the Central Limit Theorem, $n^{1/2}(\tilde\beta^\dagger - \beta)$ has the asymptotic covariance $\Gamma^{-1}\Sigma^\dagger[\Gamma^{-1}]'$ where $\Sigma^\dagger = \text{var}\{\text{Res}(U_i(\beta,\alpha), H_i^*)\}$.

Now it remains to show that $\tilde\beta^\dagger$ is more efficient than $\hat\beta$. Note that $n^{1/2}(\hat\beta-\beta)$ has the asymptotic covariance $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$ where Σ can be written as $\text{var}\{\text{Res}(U_i(\beta,\alpha), S_i(\alpha))\}$. If letting η_2 be the regression coefficient of U_i on S_i, then

$$\begin{aligned} \text{var}\{\text{Res}(U_i, S_i)\} &= \text{var}[U_i - \eta_2 S_i] \\ &= E[U_i - \eta_2 S_i][U_i - \eta_2 S_i]' \\ &= E[(U_i - \eta^* H_i^*) + (\eta^* H_i^* - \eta_2 S_i)][(U_i - \eta^* H_i^*) + (\eta^* H_i^* - \eta_2 S_i)]' \\ &= E[[U_i - \eta^* H_i^*][U_i - \eta^* H_i^*]' + E[\eta^* H_i^* - \eta_2 S_i][\eta^* H_i^* - \eta_2 S_i]' \\ &\geq E[[U_i - \eta^* H_i^*][U_i - \eta^* H_i^*]' \\ &= \text{var}\{\text{Res}(U_i, H_i^*)\}. \end{aligned}$$

The inequality is strict unless $\eta^* = 0$. The third last step uses the fact that $E[\eta^* H_i^* - \eta_2 S_i][U_i - \eta^* H_i^*]' = 0$ for the residual $U_i - \eta^* H_i^*$ of the projection of U_i on the expanded space of H_i^*. Therefore, $\tilde\beta^\dagger$ is more efficient than $\hat\beta$ when $\eta \neq 0$.

Chapter 5

Association Studies for Longitudinal Data Arising in Clusters with Missing Covariates

5.1 Introduction

Many analyses for longitudinal incomplete data focus on studying the impact of covariates on the mean responses. Fitzmaurice et al. (2001) considered the case with missing responses for longitudinal binary data. A number of estimating equations approaches are considered for cases where drop-out cannot be assumed to be missing completely at random. These approaches include first-order generalized estimating equations (GEE) (Liang and Zeger, 1986), GEE based on conditional residuals, GEE based on multivariate normal estimating equations for the covariance matrix, and second-order generalized estimating equations (GEE2) that feature association structures among repeated measurements. Bias analyses may be performed for estimation of both the association parameters and mean parameters.

However, in clinical trials and observational studies, complete covariate data are often not available for every subject. Missing data may arise in many circumstances,

including the unavailability of covariate measurements, survey nonresponse, study subjects failing to report to a clinic for monthly evaluations, respondents refusing to answer certain items on a questionnaire, and loss of data. Problems arise if the mechanism leading to the missing data is related to these covariates. Complete case analysis can give invalid inference. Under the missing completely at random (MCAR) mechanism, the first-order GEE approach yields consistent estimates for the regression parameters. When the data are missing at random (MAR) or missing not at random (MNAR), an analysis based on first-order GEE gives inconsistent estimates of parameters for the regression model. Robins and Rotnitzky (1995), and Robins et al. (1994, 1995) developed a class of estimators based on an inverse probability weighted generalized estimating equations (IPWGEE) in a regression setting when data are MAR. This approach involves modeling the missing data process and weighting the estimating equations by the inverse of a probability that is calculated based on the models for the missing data process. If the models for both the marginal mean of the response and the missing data process are correctly formulated, the IPWGEE approach corrects the bias and gives consistent estimates under the MAR mechanism.

In many situations, longitudinal data arise in clusters. Common examples include longitudinal community intervention studies (e.g., Perry et al., 1989), family studies involving repeated assessments of individual members over time (Payment et al., 1991), and longitudinal school-based studies in which individual schools are randomized to receive an experimental or control intervention (Cameron et al., 1999). Clustered longitudinal data feature both a cross-sectional and a longitudinal correlation structure, and interest often resides in the strength of both types of association. When the association parameters are of central importance, second-

order GEEs can be constructed to facilitate their more efficient estimation. Prentice (1988) developed such equations and emphasized estimation of correlation parameters. Fitzmaurice et al. (1993) proposed a model that parameterizes the association in terms of conditional odds ratios. Lipsitz, Laird and Harrington (1991), Liang, Zeger and Qaqish (1992), Carey, Zeger and Diggle (1993), Molenberghs and Lesaffre (1994), Lang and Agresti (1994), and Fitzmaurice and Lipsitz (1995) have proposed models that parameterize the association in terms of marginal odds ratios. Yi and Cook (2002) discussed marginal methods for incomplete responses in longitudinal data arising in clusters, where the inverse probability weighted second-order estimating equations are developed. Under MAR, this method facilitates consistent estimation of the marginal mean parameters and association parameters as well.

However, little attention has been directed to address the impact of missing covariates on the association parameters in clustered longitudinal studies. This chapter mainly addresses this problem. Weighted first and second order estimating equations may be constructed to obtain consistent estimates of association parameters. In cross-sectionally clustered longitudinal data, clustering in the missing data process may need to be addressed to get efficient estimates (Yi and Cook, 2002).

This chapter is organized as follows. Section 5.2 gives a special case by addressing the cross-sectional studies arising in clusters with missing covariates. Section 5.3 addresses the more general case of association studies for incomplete longitudinal data.

5.2 Cross-Sectional Studies

5.2.1 Notation and Model Assumptions

Response Process

Suppose that there are n clusters and J_i individuals within cluster i, $i = 1, 2, \ldots, n$. Let $Y_i = (Y_{i1}, \ldots, Y_{iJ_i})'$, where Y_{ij} denotes the binary response for subject j in cluster i. Let X_{ij} be a scalar covariate that may be missing and $X_i = (X_{i1}, \ldots, X_{iJ_i})'$. Let $Z_{ij} = (1, Z_{ij1}, Z_{ij2}, \ldots, Z_{ij,p-2})'$ be the covariate vector that are always observed, and $Z_i = (Z'_{i1}, \ldots, Z'_{iJ_i})'$.

Define $\mu_{ij} = E(Y_{ij}|X_i, Z_i) = P(Y_{ij} = 1|X_i, Z_i)$, and let $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ_i})'$. Provided that the mean structure of Y_{ij} depends only on the covariate vector for subject j in cluster i, we may consider logistic regression models for the mean of the form

$$\text{logit}(\mu_{ij}) = X_{ij}\beta_x + Z'_{ij}\beta_z$$

for $j = 1, \ldots, J_i$. Let $\beta = (\beta_x, \beta'_z)'$ be a vector of regression parameters. The variance for the response Y_{ij} is specified as

$$v_{ij} = \text{Var}(Y_{ij}|X_i, Z_i) = \mu_{ij}(1 - \mu_{ij}),$$

which depends on the regression parameter vector β.

The joint probability for any pair of binary responses

$$\mu_{ijj'} = E(Y_{ij}Y_{ij'}|X_i, Z_i) = P(Y_{ij} = 1, Y_{ij'} = 1|X_i, Z_i)$$

can be modeled in terms of the two marginal probabilities $\mu_{ij}(\beta)$ and $\mu_{ij'}(\beta)$ in combination with an association parameter vector. One approach is to use the

conditional correlation between Y_{ij} and $Y_{ij'}$, given Z_i and X_i, where

$$\phi_{ijj'} = \text{Corr}(Y_{ij}, Y_{ij'}|X_i, Z_i) = \frac{\mu_{ijj'} - \mu_{ij}\mu_{ij'}}{[\mu_{ij}(1-\mu_{ij})\mu_{ij'}(1-\mu_{ij'})]^{1/2}}.$$

In terms of the correlation coefficient, the joint probability $\mu_{ijj'}$ can then be expressed as

$$\mu_{ijj'} = \mu_{ij}\mu_{ij'} + \phi_{ijj'} \cdot [\mu_{ij}(1-\mu_{ij})\mu_{ij'}(1-\mu_{ij'})]^{1/2}.$$

One may alternatively use odds ratio to characterize the association among responses. Let $\psi_{ijj'}$ be the odds ratio between Y_{ij} and $Y_{ij'}$, which is defined by

$$\psi_{ijj'} = \frac{P(Y_{ij}=1, Y_{ij'}=1|X_i, Z_i)P(Y_{ij}=0, Y_{ij'}=0|X_i, Z_i)}{P(Y_{ij}=1, Y_{ij'}=0|X_i, Z_i)P(Y_{ij}=0, Y_{ij'}=1|X_i, Z_i)}. \tag{5.1}$$

Regression models for the association are typically specified as

$$\log(\psi_{ijj'}) = u'_{ijj'} \cdot \phi,$$

where $u_{ijj'}$ is a vector of covariates which specifies the form of the association between Y_{ij} and $Y_{ij'}$, and ϕ is a vector of regression parameters. Letting $u_{ijj'}$ be the scalar 1, for example, leads to the exchangeable association between responses (Yi and Cook, 2002).

The joint probability $\mu_{ijj'}$ is determined by the marginal means and the odds ratio. Note that

$$\psi_{ijj'} = \frac{\mu_{ijj'}(1-\mu_{ij}-\mu_{ij'}+\mu_{ijj'})}{(\mu_{ij}-\mu_{ijj'})(\mu_{ij'}-\mu_{ijj'})}.$$

Using the quadratic formula, we can solve for $\mu_{ijj'}$ given by

$$\mu_{ijj'} = \begin{cases} \frac{a_{ijj'} - [a_{ijj'}^2 - 4\psi_{ijj'}(\psi_{ijj'}-1)\mu_{ij}\mu_{ij'}]^{1/2}}{2(\psi_{ijj'}-1)}, & \text{if } \psi_{ijj'} \neq 1, \\ \mu_{ij} \cdot \mu_{ij'}, & \text{if } \psi_{ijj'} = 1, \end{cases}$$

where $a_{ijj'} = 1 - (1-\psi_{ijj'})(\mu_{ij} + \mu_{ij'})$ (Lipsitz et al., 1991). Given this, the correlation $\phi_{ijj'}$ can be written in terms of the marginal means and the odds ratio $\psi_{ijj'}$.

Missing Data Models

Let R_{ij} be the missing data indicator for covariate X_{ij}, where $R_{ij} = 1$ if X_{ij} is observed and 0 otherwise. Let r_{ij} be a realization of R_{ij}.

Let $\lambda_{ij} = P(R_{ij} = 1|Y_i, X_i, Z_i)$, known up to a vector of unknown parameters γ. Typically, a logistic link may relate λ_{ij} to a linear function of Y_i, X_i and Z_i, i.e.

$$\text{logit}(\lambda_{ij}) = u_{ij}^{*\prime} \cdot \gamma$$

where u_{ij}^* may be a function of $\{Y_i, X_i, Z_i\}$.

We define the odds ratio for subject j and j' in cluster i as

$$\psi_{ijj'}^* = \frac{P(R_{ij} = 1, R_{ij'} = 1|Y_i, Z_i, X_i) \cdot P(R_{ij} = 0, R_{ij'} = 0|Y_i, Z_i, X_i)}{P(R_{ij} = 1, R_{ij'} = 0|Y_i, Z_i, X_i) \cdot P(R_{ij} = 0, R_{ij'} = 1|Y_i, Z_i, X_i)}.$$

Let ϕ^* be the regression parameters linking the odds ratios $\psi_{ijj'}^*$ to the related covariates, $u_{ijj'}^*$, say. For example,

$$\log \psi_{ijj'}^* = u_{ijj'}^{*\prime} \cdot \phi^*.$$

Denote $\alpha = (\gamma', \phi^{*\prime})'$ to be the q vector of parameters associated with the missing-data process.

Let $\lambda_{ijj'} = P(R_{ij} = 1, R_{ij'} = 1|Y_i, Z_i, X_i)$ be the joint probability for $(R_{ij}, R_{ij'})$. From

$$\psi_{ijj'}^* = \frac{\lambda_{ijj'}[1 - \lambda_{ij} - \lambda_{ij'} + \lambda_{ijj'}]}{(\lambda_{ij} - \lambda_{ijj'})(\lambda_{ij'} - \lambda_{ijj'})}$$

we can get

$$\lambda_{ijj'} = \begin{cases} \frac{a_{ijj'}^* - [a_{ijj'}^{*2} - 4\psi_{ijj'}^*(\psi_{ijj'}^* - 1)\lambda_{ij}\lambda_{ij'}]^{1/2}}{2(\psi_{ijj'}^* - 1)}, & \text{if } \psi_{ijj'}^* \neq 1, \\ \lambda_{ij} \cdot \lambda_{ij'}, & \text{if } \psi_{ijj'}^* = 1, \end{cases}$$

where $a_{ijj'}^* = 1 - (1 - \psi_{ijj'}^*)(\lambda_{ij} + \lambda_{ij'})$ (e.g. Lipsitz et al., 1991).

Here we focus on dealing with the MAR mechanism, where we assume that

$$P(R_{ij} = 1|Y_i, X_i, Z_i) = P(R_{ij} = 1|Y_i, X_i^{(o)}, Z_i)$$

$$\psi^*_{ijj'} = \frac{P(R_{ij} = 1, R_{ij'} = 1|Y_i, Z_i, X_i^{(o)}) \cdot P(R_{ij} = 0, R_{ij'} = 0|Y_i, Z_i, X_i^{(o)})}{P(R_{ij} = 1, R_{ij'} = 0|Y_i, Z_i, X_i^{(o)}) \cdot P(R_{ij} = 0, R_{ij'} = 1|Y_i, Z_i, X_i^{(o)})},$$

and hence $\lambda_{ijj'}$ does not depend on the unobserved $X_i^{(m)}$.

5.2.2 Estimation Procedures

Our primary interest lies in estimating parameters β associated with the mean responses as well as association parameters ϕ. Let $\theta = (\beta', \phi')'$.

Estimating Equations for Mean Parameters

Let $D_i = \partial \mu_i'/\partial \beta$ be the $p \times J_i$ derivative matrix of the mean vector μ_i with respect to β. Let

$$\Delta_i(\alpha) = \text{diag}(I(R_{ij} = 1)/\lambda_{ij}, 1 \le j \le J_i)$$

be the $J_i \times J_i$ weight matrix, and $V_i = \text{diag}(\mu_{ij}(1 - \mu_{ij}), 1 \le j \le J_i)$.

The GEE for β are given by

$$U_1(\theta, \alpha) = \sum_{i=1}^{n} U_{1i}(\theta, \alpha) = 0, \tag{5.2}$$

where $U_{1i}(\theta, \alpha) = D_i V_i^{-1} \Delta_i(\alpha)(Y_i - \mu_i)$.

Estimating Equations for Association Parameters

Let

$$\Delta_i^*(\alpha) = \text{diag}(I(R_{ij} = 1, R_{ij'} = 1)/\lambda_{ijj'}, \quad j < j').$$

The GEEs for ϕ are of the form

$$U_2(\theta, \alpha) = \sum_{i=1}^{n} U_{2i}(\theta, \alpha) = 0, \tag{5.3}$$

where $U_{2i}(\theta, \alpha) = C_i W_i^{-1} \cdot \Delta_i^*(\alpha) \cdot (\mathcal{Y}_i - \xi_i)$, $\mathcal{Y}_i = (Y_{ij} Y_{ij'}, j < j')'$, $\xi_i = E(\mathcal{Y}_i | X_i, Z_i) = (\mu_{ijj'}, j < j')'$, $C_i = \partial \xi_i' / \partial \phi$ is the derivative matrix of the mean vector ξ_i with respect to ϕ, W_i is a working covariance matrix. The covariance matrix of $\mathcal{Y}_i$ involves third and forth moments of the responses, which we would rather not estimate here. The independence working covariance matrix

$$W_i = \text{diag}(\mu_{ijj'} \cdot (1 - \mu_{ijj'}),\ j < j')$$

is often used.

Here we remark that only the data with fully observed covariates are used in the estimating equations (5.2) and (5.3). Therefore, the resulting estimators may lose efficiency. As suggested in Chapter 4, we may employ the improved inverse probability weighted estimates which are theoretically more efficient under the assumption of data missing at random. Or, we can also develop doubly robust estimators, which are robust under certain conditions to misspecification of the model for the probability of response. Note that we only use the independence working correlation matrix in (5.2) and (5.3). We can adapt the idea of Chapter 4 to incorporate the general correlation matrix, and this is the future research work.

Estimation for Parameters of Missing-Data Process

Let $V_i^* = (v_{ijj'}^*)$ with $v_{ijj'}^* = \lambda_{ij}(1 - \lambda_{ij})$ if $j = j'$ and $\lambda_{ijj'} - \lambda_{ij} \cdot \lambda_{ij'}$ otherwise. Let $R_i = (R_{i1}, R_{i2}, \ldots, R_{iJ_i})'$, $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{iJ_i})'$, and $D_i^* = \partial \lambda_i' / \partial \gamma$, then the

estimating equations for γ are of the form

$$\sum_{i=1}^{n} S_{1i}(\alpha) = 0, \tag{5.4}$$

where $S_{1i}(\alpha) = D_i^* V_i^{*-1}(R_i - \lambda_i)$.

For second order estimating equations for the association parameter ϕ^*, we define $\mathcal{R}_i = (R_{i1}R_{i2}, \ldots, R_{i,J_i-1}R_{iJ_i})'$, $\Lambda_i = E(\mathcal{R}_i|Y_i, X_i, Z_i)$, $C_i^* = \partial\Lambda_i'/\partial\phi^*$ and

$$W_i^* = \text{diag}(\lambda_{ijj'} \cdot (1 - \lambda_{ijj'}), \quad j < j').$$

Then estimating equations for ϕ^* are given by

$$\sum_{i=1}^{n} S_{2i}(\alpha) = 0, \tag{5.5}$$

where $S_{2i}(\alpha) = C_i^* W_i^{*-1} \cdot (\mathcal{R}_i - \Lambda_i)$. Let $S_i(\alpha) = (S_{1i}'(\alpha), S_{2i}'(\alpha))'$.

5.2.3 Estimation and Inference

We estimate the parameters based on the following two stages in the same spirit of Chapter 4:

Stage 1: Solve (5.4) and (5.5) for the missing data parameter α using Fisher-scoring algorithm as follows. Define

$$M_1^*(\alpha) = -\sum_{i=1}^{n} D_i^* V_i^{*-1} D_i^{*'}$$

and

$$M_2^*(\alpha) = -\sum_{i=1}^{n} C_i^* W_i^{*-1} C_i^{*'}.$$

For any initial values $\alpha = \alpha^{(0)}$, simultaneously update α using

$$\alpha^{(t)} = \alpha^{(t-1)} - \begin{pmatrix} [M_1^*(\alpha^{(t-1)})]^{-1} \\ [M_2^*(\alpha^{(t-1)})]^{-1} \end{pmatrix} \cdot \begin{pmatrix} \sum_{i=1}^{n} S_{1i}(\alpha^{(t-1)}) \\ \sum_{i=1}^{n} S_{2i}(\alpha^{(t-1)}) \end{pmatrix}$$

until $\alpha^{(t)}$ converges to $\hat{\alpha}$, say.

State 2: Replace α with the estimate $\hat{\alpha}$ and solve (5.2) and (5.3) for θ via Fisher-scoring algorithm as follows:

Let

$$M_1(\theta, \hat{\alpha}) = -\sum_{i=1}^{n} D_i V_i^{-1} \cdot \Delta_i(\hat{\alpha}) \cdot D_i'$$

and

$$M_2(\theta, \hat{\alpha}) = -\sum_{i=1}^{n} C_i W_i^{-1} \cdot \Delta_i^*(\hat{\alpha}) \cdot C_i'.$$

For any initial values $\theta = \theta^{(0)}$, simultaneously update β and ϕ by the iterative equations

$$\theta^{(t)} = \theta^{(t-1)} - \begin{pmatrix} [M_1(\theta^{(t-1)}, \hat{\alpha})]^{-1} \\ [M_2(\theta^{(t-1)}, \hat{\alpha})]^{-1} \end{pmatrix} \cdot \begin{pmatrix} U_1(\theta^{(t-1)}, \hat{\alpha}) \\ U_2(\theta^{(t-1)}, \hat{\alpha}) \end{pmatrix}$$

until $\theta^{(t)}$ converges to $\hat{\theta}$, say.

We conclude this section with a discussion of the asymptotic distribution of the estimate $\hat{\theta}$ and inferential issues. Let $U_i(\theta, \alpha) = (U_{1i}'(\theta, \alpha), U_{2i}'(\theta, \alpha))'$. When α is specified to be α_0, under standard regularity conditions for estimating functions,

$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Gamma_0^{-1} E(U_i(\theta, \alpha_0) U_i'(\theta, \alpha_0))[\Gamma_0^{-1}]'), \quad \text{as } n \to \infty$$

where $\Gamma_0 = E(\partial U_i(\theta, \alpha_0)/\partial \theta')$. When α is unspecified and estimated, the variation in the estimator $\hat{\alpha}$ must be taken into account, and under the regularity conditions stated by Robins et al. (1995), $n^{1/2}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and asymptotic variance $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$, where $\Gamma = E[\partial U_i(\theta, \alpha)/\partial \theta']$, $\Sigma = E[Q_i(\theta, \alpha) Q_i'(\theta, \alpha)]$, and $Q_i(\theta, \alpha) = U_i(\theta, \alpha) - E(\partial U_i(\theta, \alpha)/\partial \alpha') \cdot [E(\partial S_i(\alpha)/\partial \alpha')]^{-1} \cdot S_i(\alpha)$.

For each component α_ℓ of α, $\ell = 1, 2, \ldots, q$, we define

$$G_{1\ell}(\theta, \alpha) = \sum_{i=1}^{n} D_i V_i^{-1} \cdot (\partial \Delta_i(\alpha)/\partial \alpha_\ell) \cdot (Y_i - \mu_i)$$

and

$$G_{2\ell}(\theta, \alpha) = \sum_{i=1}^{n} C_i W_i^{-1} \cdot (\partial \Delta_i^*(\alpha)/\partial \alpha_\ell) \cdot (\mathcal{Y}_i - \xi_i).$$

Then $E(\partial U_i(\theta, \alpha)/\partial \alpha')$ is consistently estimated by, as $n \to \infty$,

$$G(\hat{\theta}, \hat{\alpha}) = n^{-1} \begin{pmatrix} G_{11}(\hat{\theta}, \hat{\alpha}) & G_{12}(\hat{\theta}, \hat{\alpha}) & \cdots & G_{1q}(\hat{\theta}, \hat{\alpha}) \\ G_{21}(\hat{\theta}, \hat{\alpha}) & G_{22}(\hat{\theta}, \hat{\alpha}) & \cdots & G_{2q}(\hat{\theta}, \hat{\alpha}) \end{pmatrix}.$$

If we let

$$M_{21}^*(\alpha) = -\sum_{i=1}^{n} C_i^* W_i^{*-1} \cdot (\partial \Lambda_i/\partial \gamma),$$

then $E(\partial S_i(\alpha)/\partial \alpha')$ is consistently estimated by, as $n \to \infty$,

$$M(\hat{\alpha}) = n^{-1} \cdot \begin{pmatrix} M_1^*(\hat{\alpha}) & 0 \\ M_{21}^*(\hat{\alpha}) & M_2^*(\hat{\alpha}) \end{pmatrix}.$$

The matrix Σ is consistently estimated by, as $n \to \infty$,

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} Q_i(\hat{\theta}, \hat{\alpha}) Q_i'(\hat{\theta}, \hat{\alpha}),$$

where $Q_i(\hat{\theta}, \hat{\alpha}) = U_i(\hat{\theta}, \hat{\alpha}) - G(\hat{\theta}, \hat{\alpha}) \cdot [M(\hat{\alpha})]^{-1} \cdot S_i(\alpha)$, and the matrix Γ is consistently estimated by, as $n \to \infty$,

$$\hat{\Gamma} = n^{-1} \begin{pmatrix} M_1(\hat{\theta}, \hat{\alpha}) & 0 \\ M_{21}(\hat{\theta}, \hat{\alpha}) & M_2(\hat{\theta}, \hat{\alpha}) \end{pmatrix},$$

where $M_{21}(\hat{\theta}, \hat{\alpha}) = -\sum_{i=1}^{n} C_i W_i^{-1} \cdot \Delta_i^*(\alpha) \cdot (\partial \xi_i/\partial \beta')$. Inferences about θ are conducted by replacing Σ and Γ with these consistent estimates in the expression of the asymptotic covariance matrix.

5.2.4 Simulation Studies

In the simulation study, we focus on a setting where $J_i = 3$, $i = 1, 2, \ldots, n$, and $n = 500$. We simulate the longitudinal binary responses from a model with

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2},$$

where x_{ij1} is a time varying binary covariate which is independently generated from Bin(1,0.5) and it may be missing at some time points, x_{ij2} is another time varying covariate generated from Bin(1,0.5), and it is always observed. We take $\text{expit}(\beta_0) = 0.4$, $\exp(\beta_1) = 0.5$ and $\exp(\beta_2) = 0.8$. The association between the responses is specified through odds ratios given by (5.1) as $\psi_{ijj'} = 2$.

For the missing data process, we take

$$\text{logit}(\lambda_{ij}) = \alpha_0 + \alpha_1 y_{ij} + \alpha_2 x_{ij2}.$$

We specify an exchangeable association structure with $\psi^*_{ijj'} = 1, 2$ or 4 for $j \neq j'$. The true values are taken as $\text{expit}(\alpha_0) = 0.7$, $\exp(\alpha_1) = 2.0$, $\exp(\alpha_2) = 2.0$ (leading to 20% missingness) and $\exp(\alpha_1) = 0.5$, $\exp(\alpha_2) = 0.5$ (leading to 50% missingness).

Table 5.1 reports the simulation results. We compare the two methods. One is to consider the analysis with the independence weights (assuming no clustering in the missing data process). The second is the analysis with the clustered weights (assuming there is a clustering association). ASE is the average standard error based on the robust variance estimators in Section 5.2.3, ESE is the empirical standard error and CP is the 95% coverage probability. RE is the relative efficiency defined by the the empirical variance of clustered weights estimators over the empirical variance of independence weights estimators.

It is seen that both methods give consistent results for the mean parameters, because the equations for β in (5.2) have weights that are not functions of the association parameters for the missing-data process; ASE is very close to ESE, and the CP agrees well with the nominal level 95%. For the association parameters $\psi_{ijj'}$, independence weights analysis gives biased estimates when $\psi^*_{ijj'} \neq 1$, while estimates using clustered weights give very small finite sample biases.

For the regression coefficient β, the relative efficiency is very close to 1. It is not surprising that there is not much difference in efficiency for inferences about the regression coefficients, because the equations for β have weights that are not functions of the association parameters for the missing-data process.

5.2.5 Asymptotic Studies

Interest here lies in studying the asymptotic biases due to assuming the association of missingness of the covariate in the same cluster is independent while the association should be considered.

In the spirit of Rotnitzky and Wypij (1994) and Fitzmaurice, Molenberghs, and Lipsitz (1995), to identify the probabilistic limit of $\hat{\theta}$, we need to take the expectation of $U_i(\theta, \alpha)$ with respect to the joint distribution of $D = (R_i, Y_i, X_i^{(o)}, Z_i)$ and set it equal to zero. The solution to this equation, which we denote $\theta^\dagger$, is the parameter to which $\hat{\theta}$ converges in probability. If $\mathcal{D}$ is the sample space for D, we must solve the equation

$$E[U_i(\theta, \alpha)] = \sum_{d \in \mathcal{D}} D_i V_i^{-1} \Delta_i(\alpha)(Y_i - \mu_i) \cdot P(d; \alpha, \theta) = 0, \tag{5.6}$$

where d denotes a realized value for $D \in \mathcal{D}$ and $P(d; \alpha, \theta)$ is the true probability of observing the realized value d. Equation (5.6) can be solved using standard software

Table 5.1: Simulation results for the association study with missing covariates

Percent		True	Independence Weights				Clustered Weights				
Missing	Parameters	Value	Bias%	ASE	ESE	CP	Bias%	ASE	ESE	CP	RE
50%						$\psi^*_{ijj'} = 4$					
	β_0	log(1.5)	0.00	0.114	0.114	0.952	-0.97	0.114	0.114	0.950	1.002
	β_1	log(0.5)	-0.28	0.159	0.159	0.947	-0.14	0.158	0.158	0.948	0.987
	β_2	log(0.8)	3.14	0.151	0.151	0.946	-1.74	0.149	0.150	0.947	0.987
	$\psi_{ijj'}$	2	58.75	1.455	1.472	0.814	4.08	0.779	0.786	0.939	–
						$\psi^*_{ijj'} = 2$					
	β_0	log(1.5)	1.97	0.120	0.120	0.951	-0.74	0.121	0.121	0.951	1.016
	β_1	log(0.5)	2.02	0.162	0.162	0.943	-1.01	0.161	0.162	0.948	1.003
	β_2	log(0.8)	1.34	0.153	0.153	0.945	-1.74	0.152	0.152	0.944	0.987
	$\psi_{ijj'}$	2	36.85	1.219	1.250	0.906	3.75	0.816	0.819	0.942	–
						$\psi^*_{ijj'} = 1$					
	β_0	log(1.5)	0.94	0.116	0.116	0.950	-0.49	0.116	0.116	0.950	1.002
	β_1	log(0.5)	1.18	0.147	0.147	0.951	-0.28	0.146	0.146	0.952	0.993
	β_2	log(0.8)	1.34	0.148	0.148	0.949	0.00	0.148	0.148	0.950	1.001
	$\psi_{ijj'}$	2	2.85	0.905	0.907	0.946	2.87	0.923	0.925	0.948	–
20%						$\psi^*_{ijj'} = 4$					
	β_0	log(1.5)	0.00	0.106	0.106	0.950	1.97	0.105	0.106	0.945	0.998
	β_1	log(0.5)	-0.14	0.126	0.126	0.950	-2.01	0.126	0.126	0.942	1.004
	β_2	log(0.8)	1.34	0.115	0.115	0.948	0.44	0.115	0.115	0.951	1.002
	$\psi_{ijj'}$	2	-3.08	0.287	0.291	0.912	1.85	0.276	0.273	0.942	–
						$\psi^*_{ijj'} = 2$					
	β_0	log(1.5)	1.97	0.100	0.100	0.943	-1.47	0.100	0.100	0.946	1.002
	β_1	log(0.5)	-2.47	0.118	0.118	0.942	-0.29	0.119	0.118	0.950	1.001
	β_2	log(0.8)	1.34	0.118	0.118	0.945	2.25	0.117	0.118	0.946	0.994
	$\psi_{ijj'}$	2	-3.06	0.310	0.307	0.936	0.65	0.297	0.299	0.945	–
						$\psi^*_{ijj'} = 1$					
	β_0	log(1.5)	-1.47	0.098	0.098	0.948	0.24	0.100	0.098	0.949	1.002
	β_1	log(0.5)	0.29	0.113	0.113	0.951	-0.57	0.113	0.113	0.948	1.000
	β_2	log(0.8)	-2.25	0.118	0.118	0.946	-0.44	0.117	0.117	0.953	0.998
	$\psi_{ijj'}$	2	0.67	0.351	0.352	0.948	0.05	0.359	0.357	0.947	–

for generalized estimating equations by constructing a data set consisting of one entry for each unique element of $\mathcal{D}$ and solving the corresponding set of equations with a weight for outcome d given by $\Delta_i(\alpha) \cdot P(d; \alpha, \theta)$.

The asymptotic covariance matrix of $n^{1/2}(\hat{\theta} - \theta^{\dagger})$ is given by

$$\text{ascov}(\sqrt{n}(\hat{\theta} - \theta^{\dagger})) = A^{-1}(\theta^{\dagger})B(\theta^{\dagger})A^{-1}(\theta^{\dagger}), \quad \text{as } n \to \infty \tag{5.7}$$

where

$$A(\theta) = E(\partial U_i(\theta, \alpha)/\partial\theta) = \sum_{d \in \mathcal{D}} \partial U_i(\theta, \alpha)/\partial\theta \cdot P(d; \alpha, \theta)$$

and

$$B(\theta) = E(U_i(\theta, \alpha)U_i'(\theta, \alpha)) = \sum_{d \in \mathcal{D}} U_i(\theta, \alpha)U_i'(\theta, \alpha) \cdot P(d; \alpha, \theta).$$

In this study, we assume the same physical settings as those in the simulation studies. In the missing data model, we change α_2 to adjust the missing proportion. For each setting, we assume $\psi^*_{ijj'} = \psi^*$ are the same for different subjects, and change it from 1 to 10 to indicate the magnitude of the associations among the missing covariate in the cluster.

It is easy to see that the estimates of the mean parameters β for independence weights analysis (assuming there is no association) and clustered weights analysis (considering the missing association) give the same results because the estimating equations for the mean parameters do not depend on the association parameters. So, here we focus on the association parameter $\psi_{ijj'}$. We study the asymptotic relative bias of independence weights analysis, where the relative asymptotic bias is defined by $(\psi^{\dagger}_{ijj'} - \psi_{ijj'})/\psi_{ijj'}$. Figures 5.1, 5.2 and 5.3 report the results. It is seen that as the missing proportion increases, the relative bias increases if controlling other conditions; also, as the missing association increases, the bias increases; the relative bias increases as the association $\psi_{ijj'}$ increases.

Figure 5.1: Asymptotic relative bias of association parameter $\psi_{ijj'}$ in independence weights analysis with $\psi_{ijj'} = 4$

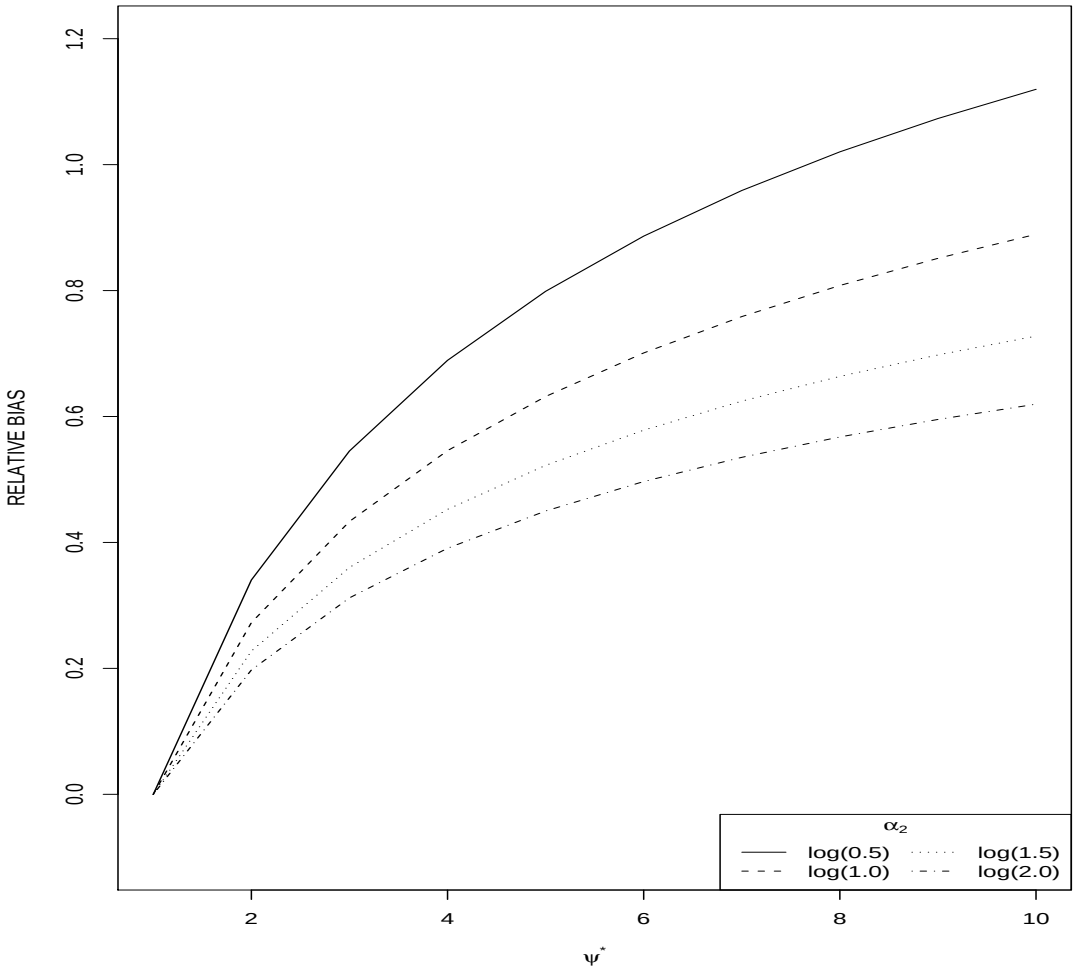


Figure 5.2: Asymptotic relative bias of association parameter $\psi_{ijj'}$ in independence weights analysis with $\psi_{ijj'} = 2$

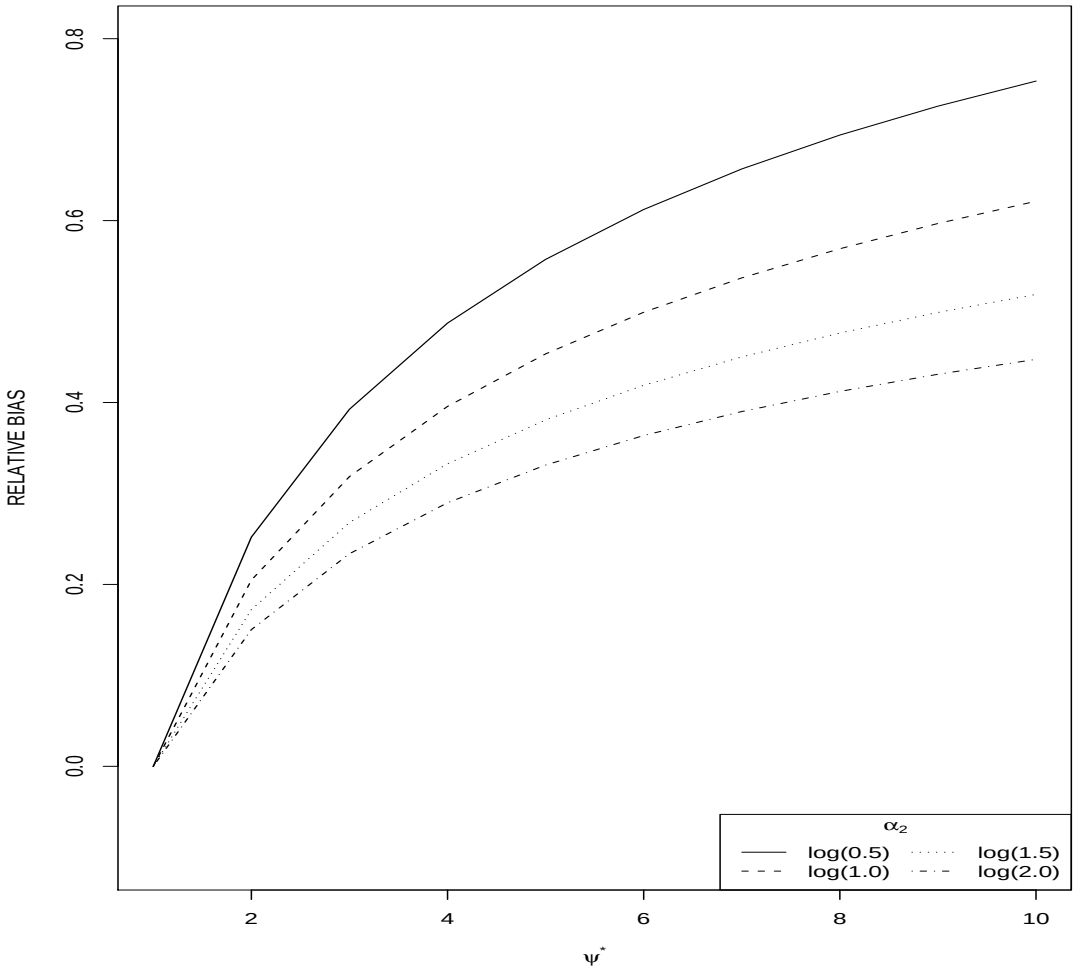
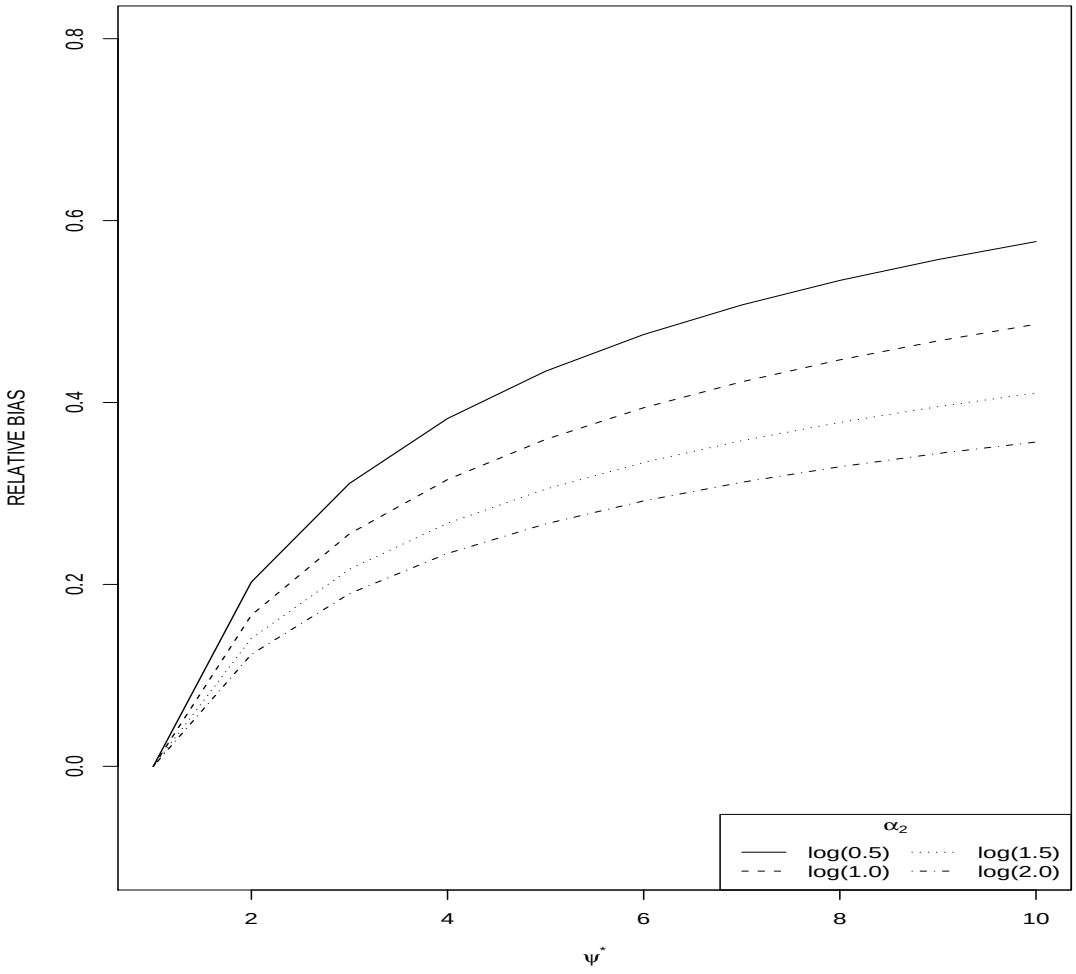


Figure 5.3: Asymptotic relative bias of association parameter $\psi_{ijj'}$ in independence weights analysis with $\psi_{ijj'} = 1$



5.3 Clustered Longitudinal Data

5.3.1 Notation and Model Assumptions

Response Process

Suppose that there are n clusters and J_i individuals within cluster i, $i = 1, 2, \ldots, n$. Furthermore suppose that there are K visits planned. Let $Y_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijK})'$ denote the K response vector for subject j which we assume that it is always observed. Let $Y_i = (Y_{i1}', \ldots, Y_{iJ_i}')'$. Let $X_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijK})'$ be the covariate vector subject to missingness that may have missing values, and let $X_i = (X_{i1}', \ldots, X_{iJ_i}')'$. Let $Z_{ijk} = (1, Z_{ijk1}, Z_{ijk2}, \ldots, Z_{ijk,p-2})'$ be the covariate vector that are always observed, $Z_{ij} = (Z_{ij1}', \ldots, Z_{ijK}')'$, and $Z_i = (Z_{i1}', \ldots, Z_{iJ_i}')'$.

Define $\mu_{ijk} = E(Y_{ijk}|X_i, Z_i) = P(Y_{ijk} = 1|X_i, Z_i)$, and let $\mu_{ij} = (\mu_{ij1}, \mu_{ij2}, \ldots, \mu_{ijK})'$, $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$. Let $\mu_i = (\mu_{i1}', \ldots, \mu_{iJ_i}')'$. Provided that the mean structure of Y_{ijk} depends only on the covariate vector for subject j at time k in cluster i (e.g., Pepe and Anderson, 1994; Robins, Greenland and Hu, 1999), we may consider logistic regression models for the mean of the form

$$\text{logit}\mu_{ijk} = X_{ijk}\beta_x + Z_{ijk}'\beta_z$$

for $k = 1, \ldots, K, j = 1, \ldots, J_i$. Let $\beta = (\beta_x, \beta_z')'$ be the vector of regression parameters. The variance for the response Y_{ijk} is specified as

$$v_{ijk} = \text{Var}(Y_{ijk}|X_i, Z_i) = \mu_{ijk}(1 - \mu_{ijk}),$$

which depends on the regression parameter vector β.

The joint probability for any pair of binary responses

$$\mu_{i;jk;j'k'} = E(Y_{ijk}Y_{ij'k'}|X_i, Z_i) = P(Y_{ijk} = 1, Y_{ij'k'} = 1|X_i, Z_i)$$

can be modeled in terms of the two marginal probabilities $\mu_{ijk}(\beta)$ and $\mu_{ij'k'}(\beta)$ in addition to an association parameter vector. One approach is to use the correlation between Y_{ijk} and $Y_{ij'k'}$ given Z_i and X_i, where

$$\phi_{i;jk;j'k'} = \text{Corr}(Y_{ijk}, Y_{ij'k'}|X_i, Z_i) = \frac{\mu_{i;jk;j'k'} - \mu_{ijk}\mu_{ij'k'}}{[\mu_{ijk}(1-\mu_{ijk})\mu_{ij'k'}(1-\mu_{ij'k'})]^{1/2}}.$$

In terms of the correlation coefficient, the joint probability $\mu_{i;jk;j'k'}$ can then be expressed as

$$\mu_{i;jk;j'k'} = \mu_{ijk}\mu_{ij'k'} + \phi_{i;jk;j'k'} \cdot [\mu_{ijk}(1-\mu_{ijk})\mu_{ij'k'}(1-\mu_{ij'k'})]^{1/2}.$$

One may alternatively use odds ratio to characterize the association among responses. Let $\psi_{i;jk;j'k'}$ be the odds ratio between Y_{ijk} and $Y_{ij'k'}$, which is defined by

$$\psi_{i;jk;j'k'} = \frac{P(Y_{ijk}=1, Y_{ij'k'}=1|X_i, Z_i)P(Y_{ijk}=0, Y_{ij'k'}=0|X_i, Z_i)}{P(Y_{ijk}=1, Y_{ij'k'}=0|X_i, Z_i)P(Y_{ijk}=0, Y_{ij'k'}=1|X_i, Z_i)}. \tag{5.8}$$

Regression models for the association are typically specified as

$$\log \psi_{i;jk;j'k'} = u'_{i;jk;j'k'} \cdot \phi,$$

where $u_{i;jk;j'k'}$ is a vector of covariates which specifies the form of the association between Y_{ijk} and $Y_{ij'k'}$, and ϕ is a vector of regression parameters. Letting $u_{i;jk;j'k'}$ be the scalar 1, for example, leads to the exchangeable association between responses (Yi and Cook, 2002). Specifically, we can adopt the following structure

$$\log \psi_{i;jk;j'k'} = \phi_0 + \phi_1 \cdot I(j=j') + \phi_2 \cdot I(k=k').$$

Let $\phi = (\phi_0, \phi_1, \phi_2)'$.

The joint probability $\mu_{i;jk;j'k'}$ can be determined by the marginal means and the odds ratio. Note that

$$\psi_{i;jk;j'k'} = \frac{\mu_{i;jk;j'k'}(1-\mu_{ijk}-\mu_{ij'k'}+\mu_{i;jk;j'k'})}{(\mu_{ijk}-\mu_{i;jk;j'k'})(\mu_{ij'k'}-\mu_{i;jk;j'k'})}.$$

Using the quadratic formula, we can solve for $\mu_{i;jk;j'k'}$ given by

$$\mu_{i;jk;j'k'} = \begin{cases} \frac{a_{i;jk;j'k'} - [a^2_{i;jk;j'k'} - 4\psi_{i;jk;j'k'}(\psi_{i;jk;j'k'}-1)\mu_{ijk}\mu_{ij'k'}]^{1/2}}{2(\psi_{i;jk;j'k'}-1)}, & \psi_{i;jk;j'k'} \neq 1 \\ \mu_{ijk} \cdot \mu_{ij'k'}, & \psi_{i;jk;j'k'} = 1, \end{cases}$$

where $a_{i;jk;j'k'} = 1 - (1 - \psi_{i;jk;j'k'})(\mu_{ijk} + \mu_{ij'k'})$ (e.g. Lipsitz et al., 1991). Given this, the correlation $\phi_{i;jk;j'k'}$ can be written in terms of the marginal means and the odds ratio $\psi_{i;jk;j'k'}$.

Missing Data Models

Let $R_{ij} = (R_{ij1}, R_{ij2}, \ldots, R_{ijK})'$ be the missing data indicator vector for covariate vector X_{ij}, where $R_{ijk} = 1$ if X_{ijk} is observed and 0 otherwise. Let $r_{ij} = (r_{ij1}, \ldots, r_{ijK})'$ be a realization of R_{ij}. Let $H^r_{ijk} = \{r_{ij1}, \ldots, r_{ij,k-1}\}$ denote the history of the missing data indicators for subject j up to but not include visit k, $k = 2, 3, \ldots, K$, $j = 1, \ldots, J_i$. We shall focus on the monotone missing-data patterns, that is, $R_{ijk} = 0$ implies $R_{ijk'} = 0$ for $k' > k$, in which case H^r_{ijk} consists of a sequence of consecutive $1's$ or $0's$.

Here we focus on dealing with MAR mechanism in marginal models, where we assume that

$$P(R_{ijk} = 1 | R_{ij,k-1} = 1, Y_i, X_i, Z_i) = P(R_{ijk} = 1 | R_{ij,k-1} = 1, Y_i, H^x_{ijk}, Z_i)$$

where $H^x_{ijk} = \{x_{ij1}, \ldots, x_{ij,k-1}\}$. We also assume, for $j \neq j'$

$$\begin{aligned} & P(R_{ijk} = 1, R_{ij'k} = 1 | R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, X_i, Z_i) \\ = \;& P(R_{ijk} = 1, R_{ij'k} = 1 | R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, Z_i, H^x_{ijk}, H^x_{ij'k}). \quad (5.9) \end{aligned}$$

Let $\lambda_{ijk} = P(R_{ijk} = 1 | R_{ij,k-1} = 1, Y_i, X_i, Z_i)$, known up to a vector of unknown parameters γ_k, where $R_{ij,k-1} = 1$ represents the history H^r_{ijk} of the indicator variables. Typically, a logistic link may relate a linear function of Y_i, H^x_{ijk} and Z_i,

i.e.

$$\text{logit}(\lambda_{ijk}) = {u^*_{ijk}}' \cdot \gamma_k$$

where u^*_{ijk} may be a subset of $\{Y_i, H^x_{ijk}, Z_i\}$.

Let $\pi_{ijk} = P(R_{ijk} = 1|Y_i, X_i, Z_i)$ be the marginal probability of observing subject j at time k in cluster i, given the entire vectors of responses and covariates; it is given by $\pi_{ijk} = \prod_{t=2}^{k} \lambda_{ijt}$ for $k \geq 2$, and we assume $\pi_{ij1} = 1$.

In some situations subjects within the same cluster may have substantial influence on each other when assessed at the same time point in the dropout process. We model the association of the dropout process at each fixed time point k, where we assume that

$$\begin{aligned} & P(R_{ijk} = 1|R_{ij,k-1} = 1, R_{ij'k'} = 1, Y_i, X_i, Z_i) \\ = \;& P(R_{ijk} = 1|R_{ij,k-1} = 1, Y_i, Z_i, X_i) \quad \text{if } j \neq j' \text{ and } k' < k \end{aligned}$$

which states that the probability of observing subject j at time k does not depend on the missingness of other subjects at earlier observation times, given that subject j is present at time $k-1$. At time k, we define the odds ratio for subjects j and j' in cluster i as

$$\begin{aligned} \psi^*_{i;jk;j'k} = \;& (P(R_{ijk} = 1, R_{ij'k} = 1|R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, Z_i, X_i) \\ & \cdot P(R_{ijk} = 0, R_{ij'k} = 0|R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, Z_i, X_i)) \\ & /(P(R_{ijk} = 1, R_{ij'k} = 0|R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, Z_i, X_i) \\ & \cdot P(R_{ijk} = 0, R_{ij'k} = 1|R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, Z_i, X_i)). \end{aligned}$$

Let ϕ^*_k be the regression parameters linking the odds ratios $\psi^*_{i;jk;j'k}$ to the related covariates, $u^*_{i;jk;j'k}$, say. For example,

$$\log(\psi^*_{i;jk;j'k}) = {u^*}'_{i;jk;j'k} \cdot \phi^*_k.$$

Denote $\gamma = (\gamma_2', \gamma_3', \ldots, \gamma_K')'$, $\phi^* = (\phi_2^{*\prime}, \phi_3^{*\prime}, \ldots, \phi_K^{*\prime})'$, and let $\alpha = (\gamma', \phi^{*\prime})'$ be the q vector of parameters associated with the missing-data process.

Let $\lambda_{i;jk;j'k} = P(R_{ijk} = 1, R_{ij'k} = 1 | R_{ij,k-1} = 1, R_{ij',k-1} = 1, Y_i, Z_i, X_i)$ be the joint probability for $(R_{ijk}, R_{ij'k})$. From

$$\psi^*_{i;jk;j'k} = \frac{\lambda_{i;jk;j'k}[1 - \lambda_{ijk} - \lambda_{ij'k} + \lambda_{i;jk;j'k}]}{(\lambda_{ijk} - \lambda_{i;jk;j'k})(\lambda_{ij'k} - \lambda_{i;jk;j'k})}$$

we can get

$$\lambda_{i;jk;j'k} = \begin{cases} \frac{a^*_{i;jk;j'k} - [{a^*_{i;jk;j'k}}^2 - 4\psi^*_{i;jk;j'k}(\psi^*_{i;jk;j'k} - 1)\lambda_{ijk}\lambda_{ij'k}]^{1/2}}{2(\psi^*_{i;jk;j'k} - 1)}, & \text{for } \psi^*_{i;jk;j'k} \neq 1 \\ \lambda_{ijk} \cdot \lambda_{ij'k}, & \text{for } \psi^*_{i;jk;j'k} = 1, \end{cases}$$

where $a^*_{i;jk;j'k} = 1 - (1 - \psi^*_{i;jk;j'k})(\lambda_{ijk} + \lambda_{ij'k})$ (e.g. Lipsitz et al., 1991).

Let $\pi_{i;jk;j'k'} = P(R_{ijk} = 1, R_{ij'k'} = 1 | Y_i, Z_i, X_i)$ be the marginal probability, which is given by

$$\pi_{i;jk;j'k'} = \begin{cases} \prod_{t=2}^{k'} \lambda_{ijt} & j = j', k < k' \\ \prod_{t=2}^{k} \lambda_{ijt} & j = j', k > k' \\ \prod_{t=2}^{k} \lambda_{i;jt;j't} & j \neq j', k = k' \\ \prod_{t=2}^{k} \lambda_{i;jt;j't} \cdot \prod_{t=k+1}^{k'} \lambda_{ij't} & j \neq j', k < k' \\ \prod_{t=2}^{k'} \lambda_{i;jt;j't} \cdot \prod_{t=k'+1}^{k} \lambda_{ijt} & j \neq j', k > k' \end{cases}$$

with $\pi_{i;j1;j'1} = 1$ for $j \neq j'$ (Yi and Cook, 2002).

5.3.2 Methods of Estimation

Our primary interest lies in estimating parameters β associated with mean responses as well as association parameters ϕ. Let $\theta = (\beta', \phi')'$.

Estimating Equations for Mean Parameters

Let $D_i = \partial\mu_i'/\partial\beta$ be the $p \times J_iK$ derivative matrix of the mean vector μ_i with respect to β. Let

$$\Delta_i(\alpha) = \text{diag}(I(R_{ijk} = 1)/\pi_{ijk}, 1 \le j \le J_i, 1 \le k \le K)$$

be the $J_iK \times J_iK$ weight matrix. Let $V_i = \text{diag}(\mu_{ijk}(1 - \mu_{ijk}), 1 \le j \le J_i, 1 \le k \le K)$.

The GEE for β are given by

$$U_1(\theta, \alpha) = \sum_{i=1}^{n} U_{1i}(\theta, \alpha) = 0, \tag{5.10}$$

where $U_{1i}(\theta, \alpha) = D_i V_i^{-1} \Delta_i(\alpha)(Y_i - \mu_i)$.

Estimating Equations for Association Parameters

Define $(j, k) < (j', k')$ if $j < j'$ or $j = j', k < k'$. Let

$$\Delta_i^*(\alpha) = \text{diag}(I(R_{ijk} = 1, R_{ij'k'} = 1)/\pi_{i;jk;j'k'}, \quad (j, k) < (j', k'))$$

The GEEs for ϕ are of the form

$$U_2(\theta, \alpha) = \sum_{i=1}^{n} U_{2i}(\theta, \alpha) = 0, \tag{5.11}$$

where $U_{2i}(\theta, \alpha) = C_i W_i^{-1} \cdot \Delta_i^*(\alpha) \cdot (\mathcal{Y}_i - \xi_i)$, $\mathcal{Y}_i = (Y_{ijk} Y_{ij'k'}, (j, k) < (j', k'))'$, $\xi_i = E(\mathcal{Y}_i | X_i, Z_i) = (\mu_{i;jk;j'k'}, (j, k) < (j', k'))'$, $C_i = \partial\xi_i'/\partial\phi$ is the derivative matrix of the mean vector ξ_i with respect to ϕ, W_i is a working covariance matrix. The covariance matrix of $\mathcal{Y}_j$ involves third and forth moments of the responses, which we would rather not estimate. The independence working covariance matrix

$$W_i = \text{diag}(\mu_{i;jk;j'k'} \cdot (1 - \mu_{i;jk;j'k'}), \ (j, k) < (j', k'))$$

is often used.

Estimation for Parameters of Missing-Data Process

Let $R^*_{ik} = (R_{i1k}, \ldots, R_{iJ_ik})'$ indicate the missingness for subjects in cluster i at time k, and let $R^*_i = (R'_{i2}, R'_{i3}, \ldots, R'_{iK})'$. Let $\Lambda_{ik} = (\lambda_{i1k}, \ldots, \lambda_{iJ_ik})'$ be the vector of conditional expectations of R^*_{ik}, and let $\Lambda_i = (\Lambda'_{i2}, \Lambda'_{i3}, \ldots, \Lambda'_{iK})'$. Because the R^*_{ik} terms are binary, we specify the covariance matrix as $V^*_{ik} = (v^*_{i;jk;j'k})$ where $v^*_{i;jk;j'k} = \lambda_{ijk}(1 - \lambda_{ijk})$ if $j = j'$ and $\lambda_{i;jk;j'k} - \lambda_{ijk} \cdot \lambda_{ij'k}$ otherwise. Let $V_i = \text{diag}(V_{ik},\ k = 2, 3, \ldots, K)$ be the covariance matrix for R^*_i. Let $D^*_i = \partial\Lambda'_i/\partial\gamma$, then the estimating equations for γ are of the form

$$\sum_{i=1}^{n} S_{1i}(\alpha) = 0, \tag{5.12}$$

where $S_{1i}(\alpha) = D^*_i V^{*-1}_i (R^*_i - \Lambda_i)$.

For second order estimating equations for the association parameter ϕ^*, we define $\mathcal{R}_{ik} = (R_{i1k}R_{i2k}, \ldots, R_{i,J_i-1,k}R_{iJ_ik})'$, and $\mathcal{R}_i = (\mathcal{R}_{i2}{}', \mathcal{R}_{i3}{}', \ldots, \mathcal{R}_{iK}{}')'$. Let $\Lambda^*_i = E(\mathcal{R}_i|Y_i, X_i, Z_i)$, $C^*_i = \partial\Lambda^{*\prime}_i/\partial\phi^*$ and

$$W^*_i = \text{diag}(\Lambda_i \cdot (1 - \Lambda_i)).$$

Then estimating equations for ϕ^* are given by

$$\sum_{i=1}^{n} S_{2i}(\alpha) = 0, \tag{5.13}$$

where $S_{2i}(\alpha) = C^*_i W^{*-1}_i \cdot (\mathcal{R}_i - \Lambda^*_i)$. Let $S_i(\alpha) = (S'_{1i}(\alpha), S'_{2i}(\alpha))'$.

5.3.3 Estimation and Inference

We also provide a two-stage estimate procedure here.

Stage 1: Solve (5.12) and (5.13) for the missing data parameter α, using Fisher-scoring algorithm as follows. Define

$$M_1^*(\alpha) = -\sum_{i=1}^{n} D_i^* V_i^{*-1} D_i^{*'}$$

and

$$M_2^*(\alpha) = -\sum_{i=1}^{n} C_i^* W_i^{*-1} C_i^{*'}.$$

For any initial values $\alpha = \alpha^{(0)}$, simultaneously update γ and ϕ^* using

$$\alpha^{(t)} = \alpha^{(t-1)} - \begin{pmatrix} [M_1^*(\alpha^{(t-1)})]^{-1} \\ [M_2^*(\alpha^{(t-1)})]^{-1} \end{pmatrix} \cdot \begin{pmatrix} \sum_{i=1}^{n} S_{1i}(\alpha^{(t-1)}) \\ \sum_{i=1}^{n} S_{2i}(\alpha^{(t-1)}) \end{pmatrix}$$

until $\alpha^{(t)}$ converges to $\hat{\alpha}$, say.

State 2: Replace α with the estimate $\hat{\alpha}$ and solve (5.10) and (5.11) for θ via Fisher-scoring algorithm as follows:

Let

$$M_1(\theta, \hat{\alpha}) = -\sum_{i=1}^{n} D_i V_i^{-1} \cdot \Delta_i(\hat{\alpha}) \cdot D_i'$$

and

$$M_2(\theta, \hat{\alpha}) = -\sum_{i=1}^{n} C_i W_i^{-1} \cdot \Delta_i^*(\hat{\alpha}) \cdot C_i'.$$

For any initial values $\theta = \theta^{(0)}$, simultaneously update β and ϕ by the iterative equations

$$\theta^{(t)} = \theta^{(t-1)} - \begin{pmatrix} [M_1(\theta^{(t-1)}, \hat{\alpha})]^{-1} \\ [M_2(\theta^{(t-1)}, \hat{\alpha})]^{-1} \end{pmatrix} \cdot \begin{pmatrix} U_1(\theta^{(t-1)}, \hat{\alpha}) \\ U_2(\theta^{(t-1)}, \hat{\alpha}) \end{pmatrix}$$

until $\theta^{(t)}$ converges to $\hat{\theta}$, say.

We conclude this section with a discussion of the asymptotic distribution of the estimate $\hat{\theta}$ and inferential issues. Let $U_i(\theta, \alpha) = (U_{1i}'(\theta, \alpha), U_{2i}'(\theta, \alpha))'$. When α is

specified to be α_0, under standard regularity conditions for estimating functions,

$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Gamma_0^{-1} E(U_i(\theta, \alpha_0) U_i'(\theta, \alpha_0))[\Gamma_0^{-1}]'), \quad \text{as } n \to \infty$$

where $\Gamma_0 = E(\partial U_i(\theta, \alpha_0)/\partial\theta')$. When α is unspecified and estimated, the variation in the estimator $\hat{\alpha}$ must be taken into account, and under the regularity conditions stated by Robins et al. (1995), $n^{1/2}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and asymptotic variance $\Gamma^{-1}\Sigma[\Gamma^{-1}]'$, where $\Gamma = E[\partial U_i(\theta, \alpha)/\partial\theta']$, $\Sigma = E[Q_i(\theta, \alpha)Q_i'(\theta, \alpha)]$, and $Q_i(\theta, \alpha) = U_i(\theta, \alpha) - E(\partial U_i(\theta, \alpha)/\partial\alpha') \cdot [E(\partial S_i(\alpha)/\partial\alpha')]^{-1} \cdot S_i(\alpha)$.

For each component α_ℓ of α, $\ell = 1, 2, \ldots, q$, we define

$$G_{1\ell}(\theta, \alpha) = \sum_{i=1}^{n} D_i V_i^{-1} \cdot (\partial \Delta_i(\alpha)/\partial\alpha_\ell) \cdot (Y_i - \mu_i)$$

and

$$G_{2\ell}(\theta, \alpha) = \sum_{i=1}^{n} C_i W_i^{-1} \cdot (\partial \Delta_i^*(\alpha)/\partial\alpha_\ell) \cdot (\mathcal{Y}_i - \xi_i).$$

Then $E(\partial U_i(\theta, \alpha)/\partial\alpha')$ is consistently estimated by, as $n \to \infty$,

$$G(\hat{\theta}, \hat{\alpha}) = n^{-1} \begin{pmatrix} G_{11}(\hat{\theta}, \hat{\alpha}) & G_{12}(\hat{\theta}, \hat{\alpha}) & \cdots & G_{1q}(\hat{\theta}, \hat{\alpha}) \\ G_{21}(\hat{\theta}, \hat{\alpha}) & G_{22}(\hat{\theta}, \hat{\alpha}) & \cdots & G_{2q}(\hat{\theta}, \hat{\alpha}) \end{pmatrix}.$$

If we let

$$M_{21}^*(\alpha) = -\sum_{i=1}^{n} C_i^* W_i^{*-1} \cdot (\partial \xi_i^*/\partial\gamma),$$

then $E(\partial S_i(\alpha)/\partial\alpha')$ is consistently estimated by, as $n \to \infty$,

$$M(\hat{\alpha}) = n^{-1} \begin{pmatrix} M_1^*(\hat{\alpha}) & 0 \\ M_{21}^*(\hat{\alpha}) & M_2^*(\hat{\alpha}) \end{pmatrix}.$$

The matrix Σ is consistently estimated by, as $n \to \infty$,

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} Q_i(\hat{\theta}, \hat{\alpha}) Q_i'(\hat{\theta}, \hat{\alpha}),$$

where $Q_i(\hat{\theta}, \hat{\alpha}) = U_i(\hat{\theta}, \hat{\alpha}) - G(\hat{\theta}, \hat{\alpha}) \cdot [M(\hat{\alpha})]^{-1} \cdot S_i(\alpha)$, and the matrix Γ is consistently estimated by, as $n \to \infty$,

$$\hat{\Gamma} = n^{-1} \begin{pmatrix} M_1(\hat{\theta}, \hat{\alpha}) & 0 \\ M_{21}(\hat{\theta}, \hat{\alpha}) & M_2(\hat{\theta}, \hat{\alpha}) \end{pmatrix},$$

where $M_{21}(\hat{\theta}, \hat{\alpha}) = -\sum_{i=1}^{n} C_i W_i^{-1} \cdot \Delta_i^*(\alpha) \cdot (\partial \xi_i / \partial \beta')$. Inferences about θ are conducted by replacing Σ and Γ with these consistent estimates in the expression of the asymptotic covariance matrix.

5.3.4 Simulation Studies

In the simulation study, we focus on a setting with $K = 3$ and $J_i = 3$, $i = 1, 2, \ldots, n$, and $n = 500$. We simulate the longitudinal binary responses from a model with

$$\text{logit}(\mu_{ijk}) = \beta_0 + \beta_1 x_{ijk1} + \beta_2 x_{ijk2} + \beta_3 x_{ijk3},$$

where x_{ijk1} is time varying binary covariate which is independently generated from Bin(1,0.5) and it may be missing at some time points, $x_{ijk2} = I(k = 2)$, and $x_{ijk3} = I(k = 3)$. We take $\text{expit}(\beta_0) = 0.4$, $\text{expit}(\beta_0 + \beta_2) = 0.5$, $\text{expit}(\beta_0 + \beta_3) = 0.6$, $\exp(\beta_1) = 0.5$. The association between the responses is specified through odds ratios given by (5.8) as $\psi_{i;jk;j'k'} = 1.2$ if $j \neq j', k \neq k'$, $\psi_{i;jk;j'k} = 1.5$ if $j \neq j'$, and $\psi_{i;jk;jk'} = 2.0$ if $k \neq k'$.

For the missing data process, we take

$$\text{logit}(\lambda_{ijk}) = \alpha_0 + \alpha_1 y_{ij,k-1} + \alpha_2 y_{ijk} + \alpha_3 x_{ij,k-1,1}, \quad k = 2, 3.$$

We specify an exchangeable association structure with $\psi^*_{i;jk;j'k} = 1$, 2 or 4 for $j \neq j'$. The true values are take as $\text{expit}(\alpha_0) = 0.7$, $\exp(\alpha_1) = 0.75$. Here α_2 and α_3 are used to adjust the missing proportion.

Tables 5.2 to 5.5 report the results. Here we compare two methods. One is the analysis that uses independence weights in the missing-data process based on standard logistic regression models; the empirical variance is denoted var_1. The second is the analysis that uses clustered weights based on the second order estimating equations accommodating cross-sectional association within clusters; the empirical variance is denoted var_2. Empirical relative efficiency (RE) is defined as $\text{var}_2/\text{var}_1 \times 100$. It is seen that both methods give consistent results for the regression coefficients and there is not much difference in efficiency for inference about the regression coefficients because the equation for β have weights that are not functions of the association parameters for the missing-data process.

However, for the association parameters, independence weights analysis gives bigger bias than that for clustered weights analysis when there is cross-sectional association within clusters. As the missing proportion increases, the bias increases. When there is no cross-sectional association within clusters, i.e. $\psi^*_{i;j2;j'2} = 1$ and $\psi^*_{i;j3;j'3} = 1$, both methods give consistent estimators, and the relative efficiency is close to one with a minor loss for the estimators based on the association weights approach.

5.3.5 Intermittently Missing Data

Monotone missing-data patterns have been the focus of much work in the analysis of longitudinal incomplete data. In practice, however, subjects may miss one or more visits before returning for a subsequent visit, creating what is termed intermittently missing data, where $R_{ijk} = 0$ does not necessarily imply $R_{ijk'} = 0$ for $k < k'$.

In this section we investigate estimation of response parameters with inter-

Table 5.2: Simulation results for the association study with missing covariates: about 20% missing (i.e. $\exp(\alpha_2) = 2.0, \exp(\alpha_3) = 1.5$)

	Independence Weights				Clustered Weights				
Parameters	Bias%	ASE	ESE	CP	Bias%	ASE	ESE	CP	RE
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,4)$						
β_0	0.74	0.106	0.105	94.3	1.84	0.105	0.105	94.8	99.9
β_1	0.14	0.114	0.114	94.6	0.72	0.114	0.114	95.4	99.9
β_2	0.73	0.124	0.125	95.2	0.42	0.125	0.125	94.2	100.1
β_3	0.25	0.140	0.138	94.4	-0.76	0.138	0.138	95.4	100.1
ϕ_0	2.09	0.129	0.130	93.5	-0.51	0.126	0.127	95.4	–
ϕ_1	2.68	0.167	0.166	93.0	0.76	0.161	0.161	94.5	–
ϕ_2	-5.41	0.150	0.149	93.6	0.70	0.142	0.142	94.2	–
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,2)$						
β_0	-2.71	0.111	0.111	94.3	-0.94	0.111	0.111	95.3	100.1
β_1	0.54	0.114	0.114	94.5	-0.14	0.115	0.114	94.9	100.1
β_2	0.49	0.127	0.126	94.4	0.24	0.126	0.126	94.6	99.8
β_3	0.65	0.127	0.128	95.5	0.98	0.128	0.128	94.8	99.9
ϕ_0	-1.63	0.132	0.132	93.8	-1.02	0.129	0.129	95.1	–
ϕ_1	1.02	0.171	0.171	93.4	0.08	0.167	0.167	94.6	–
ϕ_2	-4.33	0.156	0.154	93.6	0.060	0.147	0.146	94.4	–
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (2,2)$						
β_0	-0.49	0.104	0.104	94.2	-0.24	0.104	0.104	95.1	99.8
β_1	0.86	0.114	0.114	94.8	-0.54	0.114	0.114	94.6	100.1
β_2	0.00	0.118	0.120	94.2	0.49	0.120	0.120	95.0	99.9
β_3	0.36	0.130	0.131	94.9	0.49	0.131	0.131	95.1	100.0
ϕ_0	-6.58	0.127	0.127	93.6	-1.24	0.125	0.125	94.5	–
ϕ_1	1.69	0.172	0.171	93.4	0.23	0.170	0.170	93.4	–
ϕ_2	-3.31	0.149	0.151	93.9	0.20	0.145	0.147	94.4	–
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (1,1)$						
β_0	-2.19	0.106	0.106	94.4	-1.03	0.109	0.106	94.6	100.0
β_1	-0.86	0.112	0.111	94.4	0.79	0.110	0.111	94.5	100.0
β_2	2.71	0.126	0.126	94.4	0.73	0.124	0.126	94.5	100.0
β_3	1.23	0.129	0.131	94.3	1.02	0.131	0.131	94.6	100.0
ϕ_0	-4.38	0.129	0.128	94.4	-1.09	0.128	0.128	94.4	100.1
ϕ_1	0.56	0.156	0.156	94.6	0.23	0.157	0.156	94.5	100.3
ϕ_2	-0.40	0.152	0.152	94.6	0.71	0.152	0.152	94.5	100.2

Table 5.3: Simulation results for the association study with missing covariates: about 25% missing (i.e. $\exp(\alpha_2) = 2.0, \exp(\alpha_3) = 1.0$)

	Independence Weights				Clustered Weights				
Parameters	Bias%	ASE	ESE	CP	Bias%	ASE	ESE	CP	RE
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,4)$					
β_0	-0.94	0.113	0.114	94.4	-0.42	0.115	0.114	94.0	99.6
β_1	0.43	0.115	0.115	94.9	-1.00	0.115	0.115	94.5	99.9
β_2	-0.94	0.124	0.124	94.8	1.06	0.124	0.124	94.7	100.0
β_3	0.49	0.141	0.141	94.5	0.24	0.141	0.141	95.3	100.0
ϕ_0	-6.24	0.144	0.144	94.0	-1.05	0.137	0.138	94.2	–
ϕ_1	1.22	0.175	0.175	93.9	0.42	0.171	0.170	94.2	–
ϕ_2	-5.72	0.162	0.162	93.1	0.90	0.155	0.152	93.9	–
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,2)$					
β_0	0.24	0.109	0.109	94.9	-0.49	0.108	0.109	95.1	100.2
β_1	-0.13	0.117	0.117	94.4	-0.72	0.117	0.117	95.4	100.2
β_2	-0.24	0.124	0.124	95.0	0.49	0.124	0.124	95.6	99.9
β_3	1.42	0.140	0.139	94.9	0.36	0.139	0.139	94.6	100.1
ϕ_0	-5.75	0.148	0.149	94.3	-0.54	0.144	0.143	94.3	–
ϕ_1	1.68	0.185	0.186	93.3	0.23	0.180	0.181	94.1	–
ϕ_2	-5.52	0.162	0.163	93.1	0.60	0.155	0.157	94.0	–
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (2,2)$					
β_0	-1.72	0.113	0.112	94.0	0.49	0.112	0.112	94.2	100.2
β_1	0.00	0.115	0.114	94.4	-0.86	0.114	0.114	94.2	99.8
β_2	2.09	0.131	0.132	94.2	0.73	0.132	0.132	94.6	100.1
β_3	0.87	0.142	0.143	95.0	0.12	0.143	0.143	94.5	100.0
ϕ_0	-3.12	0.134	0.133	94.2	-1.08	0.128	0.130	94.4	–
ϕ_1	1.34	0.162	0.163	93.7	0.77	0.160	0.160	95.0	–
ϕ_2	-3.34	0.166	0.166	93.7	0.60	0.161	0.161	94.1	–
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (1,1)$					
β_0	-1.08	0.112	0.112	94.4	0.00	0.111	0.112	94.3	100.0
β_1	0.00	0.114	0.115	94.4	-1.00	0.113	0.115	94.9	100.1
β_2	2.09	0.135	0.132	94.3	0.95	0.132	0.132	94.4	100.1
β_3	0.98	0.141	0.142	94.5	0.36	0.140	0.142	94.8	100.0
ϕ_0	-2.14	0.134	0.133	94.5	-1.00	0.135	0.133	94.5	100.2
ϕ_1	0.34	0.159	0.160	94.7	0.08	0.159	0.160	94.6	100.3
ϕ_2	-0.60	0.167	0.164	94.3	0.60	0.164	0.164	94.3	100.3

Table 5.4: Simulation results for the association study with missing covariates: about 30% missing (i.e. $\exp(\alpha_2) = 0.5, \exp(\alpha_3) = 1.5$)

	Independence Weights				Clustered Weights				
Parameters	Bias%	ASE	ESE	CP	Bias%	ASE	ESE	CP	RE
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,4)$					
β_0	-0.98	0.111	0.111	95.5	0.49	0.111	0.111	95.0	100.2
β_1	-0.28	0.125	0.125	95.2	0.28	0.126	0.125	94.4	100.1
β_2	1.09	0.129	0.130	94.2	-1.01	0.130	0.130	95.2	100.0
β_3	0.12	0.146	0.148	94.1	0.36	0.149	0.148	95.4	99.9
ϕ_0	10.97	0.207	0.206	93.0	1.26	0.190	0.190	93.9	–
ϕ_1	-2.03	0.262	0.263	93.2	-0.45	0.249	0.249	93.5	–
ϕ_2	5.02	0.260	0.261	92.5	-0.20	0.240	0.240	94.3	–
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,2)$					
β_0	-1.46	0.113	0.113	94.3	-0.73	0.113	0.113	95.1	100.0
β_1	0.54	0.133	0.132	95.2	0.72	0.132	0.131	94.4	99.8
β_2	-0.55	0.124	0.125	95.5	-0.24	0.125	0.125	94.4	100.1
β_3	-0.23	0.152	0.152	94.5	-0.12	0.152	0.152	94.7	99.9
ϕ_0	6.01	0.197	0.198	92.6	1.04	0.185	0.188	93.9	–
ϕ_1	-2.47	0.250	0.250	93.7	-0.68	0.239	0.240	94.2	–
ϕ_2	3.73	0.272	0.272	92.6	-0.20	0.252	0.250	94.4	–
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (2,2)$					
β_0	-1.47	0.114	0.114	94.8	-0.84	0.114	0.114	94.5	100.0
β_1	0.52	0.125	0.126	94.5	0.28	0.126	0.126	94.7	100.4
β_2	0.42	0.132	0.131	94.7	-0.95	0.131	0.131	94.3	99.9
β_3	-0.23	0.152	0.151	94.3	0.77	0.151	0.151	95.4	99.8
ϕ_0	-5.51	0.222	0.222	94.0	1.04	0.212	0.213	93.2	–
ϕ_1	-1.88	0.275	0.278	92.9	-0.42	0.272	0.274	94.2	–
ϕ_2	2.82	0.280	0.281	93.2	-0.30	0.266	0.267	94.2	–
				$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (1,1)$					
β_0	-1.25	0.105	0.104	94.5	-0.15	0.103	0.104	94.5	100.0
β_1	0.86	0.129	0.128	94.4	0.75	0.129	0.128	94.3	100.1
β_2	-0.24	0.133	0.132	95.2	-0.25	0.130	0.132	94.6	100.0
β_3	0.98	0.150	0.149	94.7	0.65	0.150	0.149	94.4	100.0
ϕ_0	3.02	0.226	0.224	94.2	1.08	0.222	0.224	94.4	100.4
ϕ_1	-0.33	0.275	0.273	94.6	-0.79	0.272	0.273	94.4	100.2
ϕ_2	0.80	0.280	0.278	94.4	-0.20	0.278	0.278	94.6	100.4

Table 5.5: Simulation results for the association study with missing covariates: about 35% missing (i.e. $\exp(\alpha_2) = 0.5, \exp(\alpha_3) = 1.0$)

	Independence Weights				Clustered Weights				
Parameters	Bias%	ASE	ESE	CP	Bias%	ASE	ESE	CP	RE
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,4)$						
β_0	-0.98	0.112	0.113	96.2	-0.96	0.113	0.113	95.4	99.9
β_1	-0.29	0.137	0.134	94.9	-0.31	0.135	0.134	94.9	100.0
β_2	0.73	0.130	0.130	93.4	0.79	0.130	0.130	94.8	100.2
β_3	-0.76	0.152	0.152	93.7	0.64	0.152	0.152	95.1	99.9
ϕ_0	10.97	0.234	0.234	93.4	1.38	0.213	0.214	95.2	–
ϕ_1	-2.34	0.320	0.317	92.9	-0.42	0.296	0.296	94.3	–
ϕ_2	6.041	0.309	0.308	93.8	-0.70	0.279	0.278	94.1	–
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (4,2)$						
β_0	0.24	0.118	0.117	94.6	-0.98	0.117	0.117	95.3	99.9
β_1	-1.00	0.141	0.142	95.9	-0.54	0.141	0.142	95.5	99.9
β_2	1.07	0.128	0.127	94.1	0.91	0.127	0.127	93.9	99.7
β_3	-1.09	0.150	0.149	96.1	0.74	0.149	0.149	95.2	99.9
ϕ_0	7.62	0.249	0.248	92.8	1.31	0.233	0.232	94.4	–
ϕ_1	-3.18	0.295	0.293	92.6	-0.33	0.279	0.280	96.3	–
ϕ_2	6.76	0.304	0.305	93.3	-0.70	0.280	0.283	94.2	–
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (2,2)$						
β_0	0.98	0.116	0.116	94.4	-1.00	0.116	0.116	94.4	100.2
β_1	-0.86	0.134	0.135	94.7	-0.43	0.135	0.135	93.2	100.0
β_2	0.12	0.124	0.124	94.5	0.37	0.124	0.124	94.1	99.8
β_3	-0.12	0.154	0.152	94.9	0.73	0.152	0.152	94.0	99.9
ϕ_0	6.36	0.236	0.236	93.1	1.02	0.223	0.225	93.7	–
ϕ_1	-3.27	0.287	0.287	93.3	-0.45	0.278	0.276	94.1	–
ϕ_2	3.62	0.297	0.296	92.9	-0.80	0.276	0.275	94.4	–
			$(\psi^*_{i;j2;j'2}, \psi^*_{i;j3;j'3}) = (1,1)$						
β_0	-1.08	0.118	0.117	94.6	-1.00	0.117	0.117	94.4	100.0
β_1	0.20	0.139	0.137	95.4	0.28	0.135	0.137	94.5	100.0
β_2	0.49	0.133	0.134	94.5	0.95	0.134	0.134	94.3	100.0
β_3	0.49	0.153	0.153	94.4	0.12	0.153	0.153	94.4	100.0
ϕ_0	3.26	0.242	0.241	94.7	1.63	0.241	0.241	94.5	100.5
ϕ_1	-0.34	0.301	0.301	95.2	-0.45	0.301	0.301	94.6	100.3
ϕ_2	1.60	0.306	0.307	94.5	-0.80	0.307	0.307	94.5	100.4

mittently missing patterns along the same line as the preceding discussion with monotone missing-data patterns. We let $H_{ijk}^{(o)}$ denote the history of the observed components in H_{ijk}^{x}, $k = 2, \ldots, K$, $j = 1, \ldots, J_i$ and $i = 1, \ldots, n$. We assume that

$$P(R_{ijk} = 1 | H_{ijk}^{r}, Y_i, X_i, Z_i) = P(R_{ijk} = 1 | H_{ijk}^{r}, Y_i, H_{ijk}^{(o)}, Z_i)$$

and

$$\begin{aligned} & P(R_{ijk} = 1, R_{ij'k} = 1 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, X_i, Z_i) \\ = \; & P(R_{ijk} = 1, R_{ij'k} = 1 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, H_{ijk}^{(o)}, H_{ij'k}^{(o)}, Z_i) \quad \text{for } j \neq j'. \quad (5.14) \end{aligned}$$

Assumption (5.14) reduces to (5.9) when the missing-data patterns are monotone, but it facilitates the derivations that follow for the intermittently missing-data patterns.

For cluster i and time k, let $\lambda_{ijk} = P(R_{ijk} = 1 | H_{ijk}^{r}, Y_i, X_i, Z_i)$ be the conditional probability for subject j being observed at time k, given the history of the indicator variable and entire vectors of responses and covariates. For assessment on subject j in cluster i at times k, we assume that

$$P(R_{ijk} = 1 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, X_i, Z_i) = P(R_{ijk} = 1 | H_{ijk}^{r}, Y_i, X_i, Z_i) \quad \text{for } j \neq j'.$$

This states that the probability of observing subject j at time k does not depend on the history of missingness of other subjects at time k, given the history of missingness of subject j at time k and the entire vector of response and covariates. For two subjects j and j' in cluster i, define the odds ratio at time k,

$$\begin{aligned} \psi_{i;jk;j'k}^{*} \; = \; & (P(R_{ijk} = 1, R_{ij'k} = 1 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, X_i, Z_i) \\ & \cdot P(R_{ijk} = 0, R_{ij'k} = 0 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, X_i, Z_i)) \\ & / (P(R_{ijk} = 1, R_{ij'k} = 0 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, X_i, Z_i) \\ & \cdot P(R_{ijk} = 0, R_{ij'k} = 1 | H_{ijk}^{r}, H_{ij'k}^{r}, Y_i, X_i, Z_i)), \end{aligned}$$

and let $\lambda_{i;jk;j'k} = P(R_{ijk} = 1, R_{ij'k} = 1|H^r_{ijk}, H^r_{ij'k}, Y_i, X_i, Z_i)$ be the joint probability for the pair $(R_{ijk}, R_{ij'k})$, conditional on the histories the indicator variables and the entire vectors of response and covariates.

Regression models may be used to characterize the probability λ_{ijk} and the odds ratio $\psi^*_{i;jk;j'k}$ for each fixed time point k as in Section 5.3.2, and the resulting parameters may be estimated as in Section 5.3.3. For cluster i, let $\pi_{ijk} = P(R_{ijk} = 1|Y_i, X_i, Z_i)$ be the conditional probability of the missingness for subject j at time k, and let $\pi_{i;jk;j'k'} = P(R_{ijk} = 1, R_{ij'k'} = 1|Y_i, X_i, Z_i)$ be the conditional probability of the missingness for subject j and j' at time k and k', respectively.

The weight matrices in the estimation equations are then given by $\Delta_i = \text{diag}(I(R_{ijk} = 1)/\pi_{ijk}, 1 \leq j \leq J_i, 1 \leq k \leq K)$ and $\Delta^*_i = \text{diag}(I(R_{ijk} = 1, R_{ij'k'} = 1)/\pi_{i;jk;j'k'}, (j,k) < (j',k'))$.

Chapter 6

Discussion and Future Research

6.1 Likelihood Analysis of Joint Marginal and Conditional Models for Longitudinal Categorical Data

In Chapter 2 we proposed a likelihood-based inference method for categorical longitudinal data. The proposed method allows modeling marginal and conditional structures separately, and this is a particular appealing property for longitudinal data analysis. As the likelihood formulation is employed for inferential procedures, the resulting estimators enjoy nice properties of maximum likelihood estimators such as high efficiency; the simulation results suggest that the proposed method performs well in a wide range of settings. A further advantage of a likelihood based procedure is that model checking can be carried out through score tests or likelihood ratio tests of null and expanded models.

In Chapter 2 we focus on modeling the conditional probability μ_{ijk}^{C} by the first order dependence of Y_{ij} on its history. Generalizations to accommodate any qth

order dependence may proceed in the same manner. For example, we may specify

$$\begin{aligned} \log\left(\frac{\mu^C_{ijk}}{\mu^C_{ij0}}\right) &= \gamma_{ijk} + \sum_{l=1}^{q}\sum_{k'=1}^{K}\gamma_{ijlk'k}I(Y_{i,j-l}=k'), \quad k=1,\ldots,K, \\ \gamma_{ijlk'k} &= Z'_{ijlk'k}\alpha_{lk'k}, \quad k,k'=1,\ldots,K, \quad l=1,\ldots,q. \end{aligned}$$

The likelihood can then be factored as the product of the distribution of the first q response variables $P(Y_{i,j-1}=y_{i,j-1},\ldots,Y_{i,j-q}=y_{i,j-q})$ and the subsequent likelihood contributions with parameters μ^C_{ijk}.

Note that computational complexity of a model with first order dependence included increases linearly with the length of the observation series, J_i. However, with qth order dependence modeled, computational complexity of evaluating the resultant likelihood for subject i increases linearly with $J_i(K+1)^q$. This is because, calculations required to compute and update the $q-$dimension history increase linearly with $(K+1)^q$, and each observation requires such calculations. This computation becomes intensive as J_i increases. How to find a feasible way to handle larger observation times J_i and reduce the computational complexity will be a further research direction.

In Chapter 2, we also develop inference procedures to handle incomplete data. One can proceed based on the observed data likelihood when little data are missing, but the described EM algorithm can be particularly useful if more data are missing. The development here rests on the assumption that the data are missing at random (or missing completely at random) (Diggle et al., 2002). As it is generally not possible to verify missing data mechanisms, it is also desirable to develop estimation procedures for data arising from missing not at random mechanisms (MNAR). A Monte Carlo EM algorithm could be developed in the spirit of Ibrahim et al. (2001), where the missing data process must be modeled. Sensitivity analysis may

be conducted as parameter nonidentifiability may become an issue with MNAR mechanisms (Rotnitzky et al., 1998).

6.2 Progressive Multi-State Models for Incomplete Longitudinal and Life History Data

In Chapter 3 we first proposed a likelihood-based method for the analysis of progressive processes with missing observations. In typical analyses of missing data, parameter nonidentifiability is an issue for MNAR mechanisms. With progressive models, however, we have shown that the model parameters are identifiable for all missing mechanisms. This property is very appealing because it allows us to use a large class of progressive models to analyze incomplete longitudinal data with various missing data mechanisms. Under this setup, the likelihood formulation is easily implemented and the resulting estimators enjoy good properties. The simulation demonstrates that the proposed method performs well under various situations.

A number of important questions can be posed. We note that the WSPP data analyzed in Section 3.4.1 are clustered by school. One can use the same idea as employed in the data analysis section of Chapter 2 to incorporate the cluster effects in the calculation of standard errors. Alternatively, a natural way of addressing this clustering is to develop a random effect model, but this would require high dimensional integration. Or, one could also adapt the idea of Zeng and Cook (2007) to explicitly model the cross-sectional association structure at a particular time point, given the history of the process. This could be achieved using more elaborate fully specified models for maximum likelihood estimation, or by adopting

an estimating function approach for the cross-sectional association parameters. A third approach, which is receiving increasing attention in recent years, would be to apply methods based on composite likelihoods (Cox and Reid, 2004, Fieuws and Verbeke, 2006). Composite likelihood methods have been shown to provide estimators with good properties for a range of settings including in the context of longitudinal data.

Second, as it is generally not possible to verify missing data mechanisms, therefore it is useful to conduct sensitivity analysis (Rotnitzky et al., 1998) for the missing not at random models. Also, our proposed method gives consistent estimates when all the models are correctly specified. However, in practice, we do not know the true models. Therefore, model checking methods in general are important.

Third, in Chapter 3 we focus the discussion on incomplete response data, but in practice data often feature missing covariates. In principle, the proposed method can be adapted to accommodate missing covariate data, or missing covariate and response data. The joint likelihood of the two types of missing data indicators, the response and the covariates that may be missing, need to be formulated for complete data, and an EM algorithm can be used again for estimation in the spirit discussed here.

Fourth, a number of important questions can be posed using the covariate information provided at clinic entry. However, in other settings interest may lie in the effect of time-varying covariates. Relatively little work has been done on fitting regression models with interval censored time-dependent covariates. In the special case of a single interval censored covariate that indicates the development of a particular condition, Goggins et al. (1999) develop methods for Cox regression for a

right censored event time. Cook et al. (2008) consider an extension to the bivariate setting where both the covariate and failure times are interval censored.

Fifth, We have focussed on the time-homogeneous Markov model in Chapter 3. This assumption can be easily relaxed to increase the flexibility of the model. Weakly parametric (e.g. piecewise constant intensities) models may be adopted to model $\lambda_{0k}(t)$ in model (3.8) along the lines of Gentleman et al. (1994). Alternatively, one can use splines to obtain smoother estimates of transition intensities if desired (Staniswalis et al., 1997), or local likelihood methods (Loader, 1996, 1999). Nonparametric methods such as those of Turnbull (1976) can in principle be adapted for the setting of dependent observation schemes when models are progressive. Interval censored recurrent event data (e.g., Thall and Lachin, 1988; Wellner and Zhang, 2000) arise from progressive models, and further work in this area is warranted.

6.3 Longitudinal Data Analysis with Incomplete Response and Covariates

The impact of attrition in longitudinal studies depends on the correlation between the missing response and missing covariate. Ignoring this correlation can induce bias and loss of efficiency to statistical inferences. We have developed and studied a method that incorporates the association between the missing response and missing covariate. The simulations demonstrate that the proposed method gives consistent and reliable estimators.

However, a number of important questions can be posed. First, note that we only considered the estimation and inference for the mean model parameters; one of

the future directions is to consider the estimation and inference for the association parameters, in which the second order estimating equations may be used to estimate the association parameters by adapting the idea of Yi and Cook (2002). Also, following the idea of Yi and Cook (2002), we can further extend the proposed method to incorporate the clustered longitudinal data.

Second, the development here rests on the assumption that the data are at most missing at random (MAR) (Diggle et al., 2002). In practice, we often face data missing not at random (MNAR) and generalizations of the proposed method to deal with this type of data would be worthwhile. One may adapt the method of Rotnitzky et al. (1998) and Scharfstein et al. (1999) to deal with missing response and missing covariates problem through semiparametric methods when the mechanism is MNAR. Alternatively, a Monte Carlo EM algorithm could be developed in the spirit of Ibrahim et al. (2001). As it is generally not possible to check the nature of missing data mechanisms, sensitivity analysis may be warranted to assess the effect of MNAR mechanisms (Rotnitzky et al., 1998).

The third research direction is to develop doubly robust, or doubly protected, estimators (Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Scharfstein et al., 1999), which are robust under certain situations of model misspecification. This method is a refinement of a weighted estimating equation approach proposed by Robins et al. (1995) and Rotnitzky et al. (1998). Further explanation and evaluation of doubly robust estimators have been given by Lunceford and Davidian (2004), Carpenter et al. (2006), Davidian et al. (2005), Bang and Robins (2005), and Kang and Schafer (2007). With increasingly complex models for the missing data process, the double robustness is increasingly important.

Bibliography

[1] Albert, P. S. (2000) A transition model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, **56**, 602–608.

[2] Albert, P. S. and Waclawiw, M. A. (1998) A two-state Markov chain for heterogeneous transitional data: A quasi-likelihood approach. *Statistics in Medicine*, **17**, 1481–1493.

[3] Anderson, D. A. and Aitkin, M. (1985) Variance components models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203–210.

[4] Azzalini, A. (1994) Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, **81**, 767–775.

[5] Bahadur, R. R. (1961) A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, H. Solomon (ed), 158–168. Stanford Mathematical Studies in the Social Sciences VI. Stanford, California: Stanford University Press.

[6] Bang, H. and Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–972.

[7] Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics*. New York: Chapman and Hall.

[8] Bartholomew, D. J. (1983) Some recent developments in social statistics. *International Statistical Review*, **51**, 1–9.

[9] Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 125–134.

[10] Cameron, R., Brown, K. S., Best, J. A., Pelkman, C. L., Madill, C. L., Manske, S. R. and Payne, M. E. (1999) Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and social risk. *American Journal of Public Health*, **89**, 1827–1831.

[11] Carey, V., Zeger, S. L. and Diggle, P. J. (1993) Modeling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517–526.

[12] Carpenter, J., Kenward, M. and Vansteelandt, S. (2006) A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, **169**, 571–584.

[13] Casella, G. and Berger, R. L. (2002) *Statistical Inference*. Duxbury, 2nd ed.

[14] Chen, Q., Ibrahim, J. G., Chen, M. and Senchaudhuri, P. (2008) Theory and inference for regression models with missing responses and covariates. *Journal of Multivariate Analysis*, **99**, 1302–1331.

[15] Cnaan, A., Laird, N. M. and Slasor, P. (1997) Using the general linear mixed

model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, **16**, 2349–2380.

[16] Cook, R. J., Yi, G. Y., Lee, K. A. and Gladman, D. D. (2004) A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, **60**, 436–443.

[17] Cook, R. J., Kalbfleisch, J. D. and Yi, G. Y. (2002) A generalized mover-stayer model for panel data. *Biostatistics*, **3**, 407–420.

[18] Cook, R. J., Zeng, L. and Lee, K-A. (2008) A multi-state model for bivariate interval censored failure time data. *Biometrics*. In press.

[19] Cook R.J., Zeng, L. and Yi, G. Y. (2004) Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, **60**, 820–828.

[20] Cox, D. R. (1972) The analysis of multivariate binary data. *Applied Statistics*, **21**, 113–120.

[21] Cox, D. R. and Miller, H. D. (1977) *The Theory of Stochastic Processes*. London: Chapman and hall.

[22] Cox, D. R. and Reid, N. (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**, 729–737.

[23] Crowder, M. J. (2001) On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society, Series B*, **63**, 55–62.

[24] Davidian, M., Tsiatis, A. A. and Leon, S. (2005) Semiparametric estimation of treatment effect in a pretest-posttest study without missing data. *Statististical Science*, **20**, 261–301.

[25] Dempster, A. P., Laird, N. M. and Rubins, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

[26] Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data.* Oxford University Press, 2nd ed.

[27] Fieuws, S. and Verbeke, G. (2006) Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, **62**, 424-431.

[28] Fitzmaurice, G. M., Laird, N. M. and Rotnitzky, A. (1993) Regression models for discrete longitudinal responses. *Statistical Science*, **8**, 284–299.

[29] Fitzmaurice, G. M., Laird, N. M. and Zahner, G. E. P. (1996) Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, **91**, 99–108.

[30] Fitzmaurice, G. M. and Lipsitz, S. R. (1995) A model for binary time series data with serial odds ratio patterns. *Applied Statistics*, **44**, 51–61.

[31] Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. and Ibrahim J. G. (2001) Bias in estimating association parameters for longitudinal binary responses with drop-outs. *Biometrics*, **57**, 15–21.

[32] Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995) Regression

models for longitudinal binary data responses with informative drop-outs. *Journal of the Royal Statistical Society, Series B*, **57**, 691–704.

[33] Gentleman, R. C., Lawless, J. F., Lindsey, J. C. and Pan, P. (1994) Multistate Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, **13**, 805–821.

[34] Gladman, D. D., Farewell, V. T. and Nadeau, C. (1995) Clinical indicators of progression in psoriatic arthritis (PSA): multivariate relative risk model. *Journal of Rheumatology*, **22**, 675–679.

[35] Gladman, D. D., Farewell, V. T., Kopciuk, K. A. and Cook, R. J. (1998) HLA Markers and progression in psoriatic arthritis. *Journal of Rheumatology*, **25**, 730–733.

[36] Glynn, R. J., Laird, N. M. and Rubin, K. B. (1986) Selection model versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (ed. H. Wainer), 115–142. New York: Springer Verlag.

[37] Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208–1212.

[38] Goggins, W. B., Finkelstein, D. M. and Zaslavsky, A. M. (1999) Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval censored. *Biometrics*, **55**, 445–451.

[39] Gruger, J., Kay, R. and Schumacher, M. (1991) The validity of inferences based on incomplete observations in disease state models. *Biometrics*, **47**, 595–605.

[40] Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338.

[41] Heagerty, P. J. (2002) Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, **58**, 342–351.

[42] Heagerty, P. J. and Zeger, S. L. (2000) Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**, 1–19.

[43] Horn, R. and Johnson, C. (1994) Topics in matrix analysis. Cambridge.

[44] Hortobagyi, G. N., Theriault, R. L., Lipton, A., Porter, L., Blayney, D., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J. J., Knight, R. D., Heffernan, M., Mellars, K. and Reitsma, D. J. (1998) Long-term prevention of skeletal complications of metastatic breast cancer with Pamidronate, *Journal of Clinical Oncology*, **16**, 2038–2044.

[45] Horton, H. J. and Laird, N. M. (1998) Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, **8**, 37–50.

[46] Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2001) Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**, 551–564.

[47] Ibrahim, J. G., Lipsitz, S. R. and Horton N. (2001) Using auxiliary data for parameter estimation with nonignorable missing outcomes. *Applied statistics*, **50**, 361–373.

[48] Kalbfleisch, J. D. and Lawless, J. F. (1985) The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863–871.

[49] Kalbfleish, J. D. and Lawless, J. F. (1989) Some statistical methods for panel life history data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time*, 185–192. Ottawa, Ontario: Statistics Canada.

[50] Kang, J. D. Y. and Schafer J. L. (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**, 523–539.

[51] Laird, N. M. (1988) Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305–315.

[52] Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[53] Lang, J. and Agresti, A. A. (1994) Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, **89**, 625–632.

[54] Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

[55] Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal statistical Society, Series B*, **54**, 3–40.

[56] Lipsitz, S. R, Fitzmaurice, G. M., Sleeper, L. and Zhao, L. P. (1995) Estimation methods for the joint distribution of repeated binary observations. *Biometrics*, 51, 562–570.

[57] Lipsitz, S. R, Ibrahim, J. G. and Zhao, L. P. (1999) A new weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, **94**, 1147–1160.

[58] Lipsitz S. R., Laird N. M. and Harrington D. P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.

[59] Little, R. J. A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 1001–1012.

[60] Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc., 1st ed.

[61] Loader, C. R. (1996) Local likelihood density estimation. *The Annals of Statistics*, **24**, 1602–1618.

[62] Loader, C. (1999) *Local Regression and Likelihood*. New York, NY: Springer-Verlag.

[63] Longford, N. (1993) *Random Coefficient Models*. Oxford, UK: Oxford University Press.

[64] Louis, T. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

[65] Lunceford, J. K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, **23**, 2937–2960.

[66] McLachlan, G. J and Krishnan, T. (1996) *The EM Algorithm and Extensions*. Wiley.

[67] Molenberghs, G. and Lesaffre (1994) Marginal modeling of correlated ordinal data using an n-way plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

[68] Molenberghs, G. M., Kenward, G. and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, **84**, 33–44.

[69] Neuhaus, J. M. (1992) Statistical methods for longitudinal and clustered designs with binary response. *Statistical Methods in Medical Research*, **1**, 249–273.

[70] Newey, W. K. and Mcfadden, D. (1993) Estimation in large samples. In *Handbook of Econometrics*, Vol. 4, eds. D. McFadden and R. Engler, Amsterdam: North-Holland.

[71] Payment, P., Richardson, L., Siemiatycki, J., Dewar, R., Edwards, M. and Franco, E. (1991) A randomized trial to evaluate the risk of gastrointestinal disease due to consumption of drinking water meeting current microbiological standards. *American Journal of Public Health*, **81**, 703–708.

[72] Pepe, M. S. and Anderson, G. L. (1994) A cautionary note on inference for marginal regression models with longitudinal data and general correlated

response data. *Communications in Statistics, Simulation and Computation*, **23**, 939–951.

[73] Perry, C. L., Kelder, S. H., Murray, D. M. and Klepp, K. L. (1989) Community-wide smoking prevention: long-term outcomes of the minnesota heart health program and the class of 1989 study. *American Journal of Public Health*, **82**, 1210–1216.

[74] Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002) Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, **21**, 3035–3054.

[75] Prentice, R. L. (1988) Correlated binary regression with covariate specific to each binary observation. *Biometrics*, **44**, 1022–1048.

[76] Robins, J. M., Greenland, S. and Hu, F. C. (1999) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, **94**, 687–712.

[77] Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

[78] Robins, J. M. and Rotnitzky, A. (2001) Comment on " Inference for semiparametric models: some questions and answer." by P. J. Bickel and J. Kwon. *Statistica Sinica*, **11**, 920–936.

[79] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regres-

sion coefficients when some regressor are not always observed. *Journal of the American Statistical Association*, **89**, 846–866.

[80] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

[81] Rotnitzky, A. and Robins, J. M. (1995) Semiparametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics*, **22**, 323–333.

[82] Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93**, 1321–1339.

[83] Rotnitzky, A. and Wypij, D. (1994) A note on the bias of the estimators with missing data. *Biometrics*, **50**, 1163–1170.

[84] Royall, R. M. (1986) Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, **54**, 221–226.

[85] Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

[86] Satten, G. A. (1999) Estimating the extent of tracking in interval-censored chain-of events data. *Biometrics*, **55**, 1228–1231.

[87] Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, **94**, 1096–1120.

[88] Shardell, M. and Miller, R. (2008) Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine*, **27**, 1008–1025.

[89] Singer, B. and Spilerman, S. (1976a) The representation of social processes by Markov models. *American Journal of Sociology*, **82**, 1–54.

[90] Singer, B. and Spilerman, S. (1976b) Some methodological issues in the analysis of longitudinal surveys. *Annals of Economic and Sociological Measurement*, **5**, 447–474.

[91] Staniswalis, J. G., Thall, P. F. and Salch, J. (1997) Semiparametric regression analysis for recurrent event interval counts. *Biometrics*, **53**, 1334–1353.

[92] Stiratelli, R., Laird, N. M. and Ware, J. H. (1984) Random effects models for serial observations with binary response. *Biometrics*, **40**, 961–971.

[93] Sutradhar, B. C. and Das, K. (1999) On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, **86**, 459–465.

[94] Thall, P. F. and Lachin, J. M. (1988) Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, **83**, 339–347.

[95] Turnbull, B. W. (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290–295.

[96] Van der Laan, M. J. and Robins, J. M. (2003) Unified methods for censored longitudinal data and causality. Springer, New York.

[97] Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer.

[98] Wasserman, S. (1980) Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, **75**, 280–294.

[99] Wellner, J. A. and Zhang, Y. (2000) Two estimators of the mean of a counting process with panel count data. *The Annals of Statistics*, **28**, 779–814.

[100] White, H. (1982) Maximum likelihood estimation under misspecified models. *Econometrica*, **50**, 1–26.

[101] Yi, G. Y. and Cook, R. J. (2002) Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, **97**, 1071–1080.

[102] Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random e!ects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–95.

[103] Zeger, S. L., Liang, K. Y. and Albert, P. S (1988) Models for longitudinal data: a generalized estimating approach. *Biometrics*, **44**, 1049–1060.

[104] Zeng, L. and Cook, R. J. (2007) Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association*, **102**, 211–223.

[105] Zhao, L. P., Lipsitz, S. R. and Lew, D. (1996) Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165–1182.

[106] Zhao, L. P. and Prentice, R. L. (1990) Correlated binary regression using a quadratic exponential model. *Biomatrika*, **77**, 642–648.