2016

Multivariate Binary Longitudinal Data Analysis

Hissah Alzahrani

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

MULTIVARIATE BINARY LONGITUDINAL DATA ANALYSIS

By

HISSAH ALZAHRANI

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2016

Hissah Alzahrani defended this dissertation on December 9, 2016.
The members of the supervisory committee were:

Elizabeth H. Slate

Professor Directing Dissertation

Amy M. Wetherby

University Representative

Daniel L. McGee

Committee Member

Debajyoti Sinha

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To my parents, husband and kids who are always with me to be here

ACKNOWLEDGMENTS

I will always be grateful for Dr. Slate's support, patience and guidance. Thank you so much for the other members of my committee for their diligence and great comments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The longitudinal data analysis plays an important role in a lot of applications today. It is defined by many measurements are obtained over many times. These measurements has complicated correlation structure because they are obtained from the same subjects over the time. In multivariate longitudinal data, there is an additional source of correlation which is "outcomes", the data are obtained over the time for many outcomes for the same subjects. This application could happens in many medical, financial and psychological studies. For example, the patients measurements for some variables are measured over some occasions in order to study the mean changes of these patients. How we can generate and analyze this type of data for complete and incomplete cases is the main goal of this dissertation. It consists of three main studies about the analysis of multivariate binary longitudinal data. The first study is a method to generate correlated binary data for a multivariate longitudinal model with specified correlation structure. This specified structure allows the correlation to be induced over the outcomes or occasions. Second study is a comparison of three methods for analyzing multivariate binary longitudinal data; each one can be beneficial for determined aims. Also, we investigated the difference among the parameter estimations of the three methods. The third study is an investigation of missing data analysis via GEE models, controlling the correlation over the occasions and outcomes via simulation study. However, several methods for handling missing data are used to reduce the bias of the parameter estimations for the incomplete data. these three studies are presented in separated chapters of this dissertation.

CHAPTER 1

INTRODUCTION

Longitudinal data are prevalent in many recent health studies. For example, Verbeke et al. (2014) examined the association between hearing thresholds measurements on both ears of a set of subjects in longitudinal study; Slate et al. (2000) and Lourdes et al. (2004) sought to elucidate the natural growth of longitudinal prostate-specific antigen (PSA); Towers et al. (2010) studied longitudinal CD4 cell counts for people with HIV-negative pregnant patients; and Gilbert et al. (1997) conducted longitudinal study of oral health and dental service utilization at 24 months for three binary outcomes. The longitudinal pattern of key health indicators can play an important role in population epidemiology as well as individual risk assessment, disease diagnosis and treatment evaluation. Hence statistical models for longitudinal data analysis are essential for distinguishing the longitudinal profiles and guiding accurate inference for health monitoring and treatment.

This dissertation focuses on analysis of multivariate longitudinal data for binary responses. Our context is one where a multivariate set of dependent binary responses is observed at many occasions in time for each of multiple individuals under study. We will use the term *outcome* to refer to the series of longitudinal readings for one binary dependent variable for an individual, and the term *response* for any reading from any dependent variable. An individual contributes the multivariate binary vector at each occasion, the collection of which is thought of as the *cluster* of observations associated with that individual. If, for example, investigators are interested to study changes in blood pressure over time, they may record each of systolic blood pressure, diastolic blood pressure and pulse pressure over many occasions. Thus there are three longitudinal outcomes, and statistical modeling must accommodate the dependencies among these outcomes at each occasion and serially over the occasions. The investigators may also want to quantify the effects of covariates on the mean of these outcomes over time.

For univariate longitudinal data, in which one response is recorded serially, statistical analysis must account for the correlation induced over time within subject. The many statistical approaches can be divided into three major models: marginal models, mixed effects models and

transition models. Each of these three models has advantages, e.g., mixed effects models provide a flexible specification of the full distribution of the longitudinal outcome; transition models enable a state-space interpretation, often capitalizing on a Markov property, and marginal models avoid specification of the full outcome distribution but nonetheless permit unbiased inference for regression effects. Because of this substantial advantage of avoiding full distributional specification, we adopt a marginal model perspective in this dissertation. For our focus on multivariate longitudinal data, the statistical model must address the more complicated correlation structure of dependencies among outcomes both within and between time occasions. The effects of missing data, i.e., unobserved responses for some individuals at one or more occasions, also becomes more complicated in the multivariate context. A primary focus of our work is the effect that the complicated correlation structure has on estimates of covariate effects for both complete and incomplete data.

In this dissertation, we present three contributions in order to elucidate aspects of the correlation complications and its effects on the covariate estimations in the analysis of multivariate binary longitudinal data. First, in chapter two we address the question of generating artificial longitudinal binary data, controlling the correlation between the outcomes and occasions. We describe a method for generating correlated binary outcomes with an arbitrary specified correlation structure among outcomes and over time. The ability to simulated multivariate binary longitudinal data with specified correlation enables study of the correlation effects on regression parameter estimates. We illustrate this utility with an application to design of a clinical trial studying two outcomes over three occasions.

The second contribution is addressed in chapter three, in which we compare three marginal models for analyzing binary multivariate longitudinal data. The first model estimates a fixed set of covariate parameters for all binary outcomes while accounting for their multivariate structure. The second model allows the covariate parameters to vary by outcome, again accommodating the full multivariate correlation structure. The third model replaces the multivariate response by a univariate encoding of the response pattern and provides for estimation of separate covariate effects for each response pattern. Chapter three investigates the differences among the parameter estimates for these three models and the consequences for inference concerning covariate effects.

Chapter four describes the third contribution, an investigation of the analysis of incomplete multivariate longitudinal data for binary outcomes. Missingness could affect the bias and precision

of parameter estimates and, for multivariate longitudinal data, these effects may depend on the degree of correlation both among outcomes and over occasions. Chapter four presents and in-depth simulation study of the effect of different missing data mechanisms on regression parameter estimation. The simulation study uses the method of Chapter two for generating multivariate binary longitudinal data with specified correlation structure. In addition to the effects of the missing data mechanism, we study the performance of four methods for handling missing data in analysis, namely Completed Cases (CC), Mean Substitution (MS), Last Observation Carried Forward (LOCF) and Regression Imputation (RI).

Now we discuss potential limitations of this dissertation research:

1. Measuring the association between binary responses could determined using the odds ratio or the Pearson correlation metric. As a measure of association for binary responses, the correlation has the limitation that it is restricted due to its functional relationship with the mean. In this dissertation, however, we use the correlation because the method for generating binary responses developed in Chapter two is consistent with the correlation metric. This method uses the correlation among continuous variables to induce the correlation among the binary responses. The fact that the correlation among the continuous variables must be nonnegative definite led to some restrictions in the correlation scenarios that we investigated. In the future, we would recommended using the odds ratios as association measures, which is expected to provide more flexibility in specification of the dependence structure.

2. We limited our studies to binary outcomes. The method of generating correlated binary data described in Chapter two is limited to binary responses. However, our focus on marginal models, specifically the generalized estimating equitation (GEE) model of Shelton et al. (2004), for the multivariate longitudinal data applies also for count and continuous responses.

3. We limited our focus to within-subject correlation, especially among the outcomes and over the occasions. There are other sources of correlation, however, such as clustering of subjects from the same study center or induced by sharing of other covariates.

4. We limited the multivariate structure by assuming that all outcomes were obtained at the same times. In some circumstances, some outcomes may be obtained according to a different schedule than other outcomes. This leads to an unbalanced design in which subjects have a different number of responses recorded at each occasion, and these number of responses may differ among subjects. In this dissertation, we used the GEE approach for marginal model estimation together with the "sandwich estimator" for standard error estimation, which is considered problematic with a severely unbalanced design. Thus our results may not be applicable when the imbalance is great.

CHAPTER 2

GENERATING MULTIVARIATE LONGITUDINAL BINARY RANDOM VARIABLES FOR MARGINAL MODELS USING BRIDGE DISTRIBUTION

2.1 Summary

Generalized estimating equations (GEE) models are often used to analyze the longitudinal data. It accounts for the within-subject associations through specification of working correlation matrix R. In multivariate longitudinal data, the within-subject correlation is computed by many outcomes are measured over many occasions. Then, the correlation is the main problem in the multivariate longitudinal data. This complicated correlation may affect the parameter estimations precision when it is increased over the outcomes or occasions. Designing a simulation method to investigate the correlation effects on the parameter estimations for the marginal models could be good statistical tool in the longitudinal data analysis. In this paper, we utilize a method to generate correlated binary data for a multivariate longitudinal model with specified R correlation matrix. This specified structure allows the correlation to be induced over the outcomes or occasions. We utilized the methods of Wang and Louis (2003) and Parzen et al. (2011) to use the generalized linear mixed models via a bridge distribution to generate multivariate binary longitudinal data for marginal models. In addition, we conducted a clinical trial simulation study for analyzing multiple and correlated binary outcomes based on control the correlation over the outcomes and occasions, and estimate the effect sample size. This approach could be a good method in simulating the correlated binary data. We include an explanation of some constraints to achieving the best simulation results.

2.2 Introduction

In this paper, our interest is generate multivariate binary longitudinal data for marginal models. It is a simulation study for many longitudinal outcomes. The longitudinal data feature is measuring

the responses over many occasions. Then, the measurements within each subject are supposed to be correlated. In the multivariate longitudinal data, there are many longitudinal outcomes are obtained in many occasions. The main two factors to build up the within subject correlation in multivariate longitudinal data are outcomes and occasions. Because the multivariate longitudinal data has a complicated correlation structure R, there is not a lot of correlation or covariance patterns are defined for the multivariate longitudinal data. This simulation study will be helpful to build up a correlation pattern for the correlated binary response form the artificial data. Generating the data based on the advantage of controlling the correlation over the outcomes and occasions is the main goal in this simulation study. Then, we can study the changes in the responses means over the time based on controlling the correlation.

Generating correlated binary data under the marginal model requires specification of the marginal means or pairwise correlation in R. Different methods are used based on different structures of R and equal or unequal marginal means. In the case of generate the artificial correlated binary data, Lee (1993) developed a method using Copula to generate correlated binary data, but contains only one parameter for R matrix. Lunn and Davies (1998) and Kang and Jung (2001) improved methods for exchangeable patterns and equal means correlations. Qaqish (2003) introduced the conditional linear family of correlated binary distribution for patterned R under equals and unequal means, or unpatterned R and large sample size. The method of Qaqish (2003) is based on a conditional linear family of multivariate binary distributions. Emrich and Piedmonte (1991) proposed a method based on the multivariate probit model using correlated standard normal variables by solving nonlinear equations. Since our goal is build a desired correlation pattern to adopt the multivariate longitudinal data, the method we use should generate the data for unstructured correlation matrix which means no constraints and the maximum parameters to estimate.

The two most practical and applicable methods for unpatterned R are those described by Emrich and Piedmonte (1991) and Qaqish (2003). Generally, Preisser Jr and Qaqish (2012) compared the two method and showed they have good estimations unless in some patterned structures. In Addition to the two methods of Emrich and Piedmonte (1991) and Qaqish (2003) to generate correlated binary data for unstructured pattern of R, we describe a third method to generate the multivariate binary data using bridge distribution. It is a method considered to be for unstructured R or building up a desired pattern. In section 2, we preview some aspects of the multivariate

longitudinal data structure specifically for the binary data. Then, we explained the proposed method and its constraints in sections 3 and section 4. Section 5 is the application to the study design for multivariate binary data. Sections 6 and 7 contain the results and conclusion, respectively.

2.3 Multivariate Binary Longitudinal Data

The multivariate longitudinal data model is an extension of the univariate longitudinal model, but for more than one outcome. Each individual i has a vector of responses for different outcomes, $k = 1, 2, ..K$. Also, each individual is measured at different times or occasions, $j = 1, 2 \dots J_i$, and has cluster size $n_i = J_i K$. Let us model K vectors of outcomes measured corresponding to a vector of times. Then, the structure is in the following figure:

$$\begin{matrix} ID \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ \vdots \\ N \\ N \\ \vdots \\ N \end{pmatrix} \end{matrix} \quad \begin{matrix} Y_{ij1} & Y_{ij2} & \cdots & Y_{ijK} \\ \begin{pmatrix} y_{111} & y_{112} & \cdots & y_{11K} \\ y_{121} & y_{122} & \cdots & y_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1J1} & y_{1J2} & \cdots & y_{1JK} \\ y_{211} & y_{212} & \cdots & y_{21K} \\ y_{221} & y_{222} & \cdots & y_{22K} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_{N11} & y_{N12} & \cdots & y_{N1K} \\ y_{N21} & y_{N22} & \cdots & y_{N2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{NJ1} & y_{NJ2} & \cdots & y_{NJK} \end{pmatrix} \end{matrix} \quad \begin{matrix} time \\ \begin{pmatrix} 1 \\ 2 \\ \vdots \\ J \\ 1 \\ 2 \\ \vdots \\ \vdots \\ 1 \\ 2 \\ \vdots \\ J \end{pmatrix} \end{matrix}$$

Figure 2.1: Multivariate longitudinal data structure

To simplify these notations, we will refer to J_i as J which is the number of occasions or visit numbers over all the observations, the vector of responses for subject i is :

$$Y_i = [Y_{i11}, Y_{i21}, ...Y_{iJ1}, Y_{i12}, Y_{i22}, Y_{i32}, ..., Y_{iJ2}, ..., Y_{i1K}, Y_{i2K}, ..., Y_{iJK}]^T$$

To illustrate aspects of the multivariate longitudinal data structure, lets assume the simple case where there are two longitudinal outcomes, $k = 1, 2$, are measured over three occasions, $j = 1, 2, 3$. for observation i, $i = 1, 2, 3...N$. The correlation matrix R consists of the within subject correlation

parameters. Then, correlation matrix $R(\gamma)$ is a function of γ, where γ represents a vector of within subject association parameters, $\gamma = [\gamma_1, \gamma_2, ..., \gamma_{15}]^T$.

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{21} & Y_{31} & Y_{12} & Y_{22} & Y_{32} \end{array} \\ \begin{pmatrix} 1 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 \\ - & 1 & \gamma_6 & \gamma_7 & \gamma_8 & \gamma_9 \\ - & - & 1 & \gamma_{10} & \gamma_{11} & \gamma_{12} \\ - & - & - & 1 & \gamma_{13} & \gamma_{14} \\ - & - & - & - & 1 & \gamma_{15} \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

Let γ be a vector of size $\binom{JK}{2}$ of all non-redundant pairwise correlation parameters in R. We will use the idea of modeling the correlation matrix to reduce the length of the vector γ. O'Brien and Fitzmaurice (2004) fit a regression model for marginal pairwise odds ratio to estimate less parameters in the binary multivariate longitudinal data structure correlation for GEE model. We will build a model for pairwise correlation parameters to induce the correlation over the outcomes or the occasions for many scenarios. Consider the correlation matrix R consists of three correlation parameters types:

1- Let $\alpha_{jk,j'k}$ be the inter-outcome correlation parameter which compares the outcome k with the outcome $k^{'}$at time j:

$$\alpha_{jk,j'k} = \frac{P(Y_{jk}=1, Y_{jk'}=1) - P(Y_{jk}=1)P(Y_{jk'}=1)}{\sqrt{P(Y_{jk}=1)P(Y_{jk'}=1)(1-P(Y_{jk}=1))(1-P(Y_{jk'}=1))}} \tag{2.1}$$

2- Let $\upsilon_{jk,jk'}$ be the intra-outcome correlation parameter which compares outcome k at time j with the same outcome at time $j^{'}$:

$$\upsilon_{jk,j'k} = \frac{P(Y_{jk}=1, Y_{j'k}=1) - P(Y_{jk}=1)P(Y_{j'k}=1)}{\sqrt{P(Y_{jk}=1)P(Y_{j'k}=1)(1-P(Y_{jk}=1))(1-P(Y_{j'k}=1))}} \tag{2.2}$$

3- Let $\tau_{jk,j'k'}$ be the cross correlation parameter which compares the outcome k at time j with outcome $k^{'}$ at time $j^{'}$:

$$\tau_{jk,j'k'} = \frac{P(Y_{jk}=1, Y_{j'k'}=1) - P(Y_{jk}=1)P(Y_{j'k'}=1)}{\sqrt{P(Y_{jk}=1)P(Y_{j'k'}=1)(1-P(Y_{jk}=1))(1-P(Y_{j'k'}=1))}} \tag{2.3}$$

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array} \\ \begin{pmatrix} 1 & \upsilon_1 & \upsilon_2 & \alpha_1 & \tau_1 & \tau_2 \\ - & 1 & \upsilon_3 & \tau_3 & \alpha_2 & \tau_4 \\ - & - & 1 & \tau_5 & \tau_6 & \alpha_3 \\ - & - & - & 1 & \upsilon_4 & \upsilon_5 \\ - & - & - & - & 1 & \upsilon_6 \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

Then, correlation matrix R is for the response Y_{jk}, $k = 1, 2$ (outcomes), $J = 1, 2, 3$ (occasions). Here we can specify any pattern or build up a parsimonious model to reduce the number of estimated parameter less than $\binom{JK}{2} = 15$ in R. For example, we could assume the time correlation parameters $\alpha^{'}s$ to be exponential and assume the outcomes correlation parameters $\upsilon^{'}s$ to be compound symmetry . To simplify the parameter estimations size, we conducted the simulation under the assumption of exchangeability for each correlation type υ, α *and*, τ and $\tau = 0$ for five correlation scenarios. The following table shows the values for each correlation structure and scenario in R matrix:

Table 2.1: The scenarios of simulation study

	$\alpha = 0.00$	$\alpha = 0.60$	$\alpha = 0.90$
$\upsilon = 0.00$	scenario1	scenario2	scenario3
$\upsilon = 0.60$	scenario4	-	-
$\upsilon = 0.90$	scenario5	-	-

2.4 The Simulation Method

2.4.1 Generating correlated binary data using bridge distribution

The goal in this study is to generate correlated binary data for marginal model. We used a regular generalized liner mixed model using bridge distribution for the random effects term. It is known that the parameter estimations under the mixed model have different interpretation than the marginal model because the marginal model integration over the random effect do not keep the logistic form. Using bridge distribution, matched the logistic shape of the conditional and marginal binary response models. The first contribution to use bridge distribution for the random

intercept logistic regression model is proposed by Wang and Louis (2003). We will start by the univariate longitudinal data structure. Let Y_{ij} be the binary response that measured at time $j,\ j = 1, 2, 3, ...J$ form independent observations $i,\ i = 1, 2, 3...N$. For each individual has a $C \times 1$ vector of covariate X_{ij}. Suppose the marginal distribution of the responses is Bernoulli with mean $E(Y_{ij}) = P(Y_{ij} = 1|X_{ij}, \beta_p) = \mu_{ij}$ through a logit link function, then responses model is:

$$logit(\mu_{ij}) = \beta_p^T X_{ij} \tag{2.4}$$

where $\beta_p = (\beta_0, \beta_1, \beta_2, ..., \beta_C)^T$ are the regression parameters from the marginal model. Wang and Louis (2003) used a bridge distribution for the random effect in the following mixed effects logistic model:

$$logit(\mu_{ij}|b_i, X_{ij}) = b_i + \phi\ \beta_s^T X_{ij} \tag{2.5}$$

where ϕ is the a cluster heterogeneity parameter. The relationship between the effects from the marginal regression model β_p and the mixed regression model β_s is related by ϕ such that:

$$\beta_p = \beta_s * \phi$$

Here will give a brief description of the method of Wang and Louis (2003). They introduced a CDF of bridge distribution $G(b)$ for the random effect to gain its advantage of keeping the marginal shape same as the conditional shape such as:

$$\int H(b + \beta_s^T X) dG(b) = H(r + \phi\beta_s^T X) \tag{2.6}$$

where H is a CDF of bridge distribution and ϕ is rescaling parameter between 0 and 1. The parameters β_s and r are unknown parameters and r is 0 when H is a CDF of symmetric distribution. The parameter β_s is regression effect and X is the covariates. By differentiate both sides of equation (2.6) respect to $\beta_s^T X$ and taking Fourier transformation F, then after organizing and using the Fourier Inversion theorem, they got the following equation:

$$g_\phi(x) = \frac{1}{2\pi} \int e^{i(\frac{r}{\phi - x})\xi} \frac{Fh(\frac{\xi}{\phi})}{Fh(\xi)} d\xi \tag{2.7}$$

If the function $H(.) = logit$ link function then $H(\beta_s^T X) = \dfrac{e^{\beta_s^T X}}{1 + e^{\beta_s^T X}}$. By plugging in Fourier transformation, see Wang and Louis (2003), then they got the pdf of bridge distribution:

$$g_\phi(x) = \frac{1}{2\pi} \frac{sin(\phi\pi)}{cosh(\phi x) + cos(\phi\pi)} \qquad (0 < \phi < 1, -\infty < x < \infty) \tag{2.8}$$

where $cosh(x) = \frac{e^x + e^{-x}}{2}$. The bridge distribution is symmetric and has slightly heavier tail than the normal distribution and lighter than logistic distribution with mean 0 and variance $\sigma_b^2 = \pi^2(\frac{1}{\phi^2} - 1/3)$. In addition, $\beta_p = \beta_s(1 - \rho_Y)$ where $\rho_Y = corr(Y_{ij}, Y_{ij'})$ is the intracluster correlation in the binary response. ϕ measure the heterogeneity across the clusters between $[0, 1]$. The CDF of the bridge distribution is :

$$G_\phi(x) = 1 - \frac{1}{\pi\phi}\,[\frac{\pi}{2} - arctan\frac{e^{\phi x} + cos(\pi\phi)}{sin(\pi\phi)}] \tag{2.9}$$

and the inverse of the cumulative density function is:

$$G_\phi^{-1}(x) = \frac{1}{\phi}\, log\frac{sin(\pi\phi x)}{sin(\pi\phi(1-x))} \tag{2.10}$$

Using the transformation $\tilde{b} = \Phi^{-1}(G(x))$, where Φ is CDF of standard Gaussian distention, then $\tilde{b} \sim N(0, 1)$. That leads to using the Gaussian-Hermite quadrature method to evaluate the integral over the bridge random effects and estimate MLE parameters. In addition, Parzen et al. (2011) have improved this model of bridge distribution for the random effect in the logistic model. Their contribution has two primary advantages. First, they constructed a model for distinct and correlated random bridge intercepts b_{ij} at each time point. The response Y_{ij} given bridge random intercepts follows the Bernoulli distribution $P(Y_{ij} = 1|b_{ij}, X_{ij}, \beta_s)$ instead of $P(Y_{ij} = 1|b_i, X_{ij}, \beta_s)$. Their method leads to better association modeling for within each subject correlation. Then, they use Copula to model the multivariate bridge random variables. Second, Parzen et al. (2011) recommend using Pearson correlation in terms of Kendall's τ to present the association between the Z's random variables due its advantage of invariance of the monotone transformation.

We exploit the advantage of the flexibility in the association structure in Parzen et al. (2011)'s method between the bridge random effects. It is a beneficial method to generate multivariate longitudinal data, controlling the within subject correlation over the outcomes and occasions using marginal model. We will generate it from mixed model using multivariate bridge random effects. First, Let Y_{ijk} be the binary response that measured at time j, $j = 1, 2, 3$ for observation i, $i = 1, 2, 3...N$ and for outcome $k = 1, 2$:

$$logit(E(Y_{ijk}|X_{ij}, b_{ijk})) = logit(P(Y_{ijk} = 1|b_{ijk}, X_{ij})) = \beta_{0k} + \beta_{1k}X_{ij} + b_{ijk} \tag{2.11}$$

where b_{ijk} is for distinct and correlated random bridge intercepts for each outcome $k = 1, 2$ at each occasion or time $j = 1, 2, 3$. Given the vector of the random effect b_{ijk}, the Y_{ijk} for subject

i is assumed to be independent Bernoulli random variables, $Y_{ijk}|b_{ijk} \sim Ber(P(Y_{ijk} = 1))$. The marginal model will be:

$$logit(E(Y_{ijk}|X_{ij})) = 1/\phi_{jk}\ (\beta_{0k} + \ \beta_{1k}X_{ij}) \tag{2.12}$$

where the parameter $0 < \phi_{jk} < 1$ is assumed to be toward zero to ensure the maximum heterogeneity of the random effect (clusters) for the response at time j for outcome k. For some reasons will be explained in the next section, we referred to ϕ_{jk} as ϕ which means all the bridge parameter have the same value. The contribution of subject i to the likelihood function is given by:

$$L_i = \int_{b_i} [\ \prod_{k=1}^{2} \prod_{j=1}^{3} P(Y_{ijk} = y_{ijk}|b_i, X_{ij})\ ] f_b(b_i) db_i, \tag{2.13}$$

where $f_b(b_i)$ is the joint density of $(b_{i11}, b_{i21}, b_{i31}, b_{i12}, b_{i22}, b_{i32})$. To simplify the notations, we will refer to the joint bridge random intercepts as $(b_{i1}, b_{i2}, b_{i3}, b_{i4}, b_{i5}, b_{i6})$. The likelihood function will be $\prod_{i=1}^{N} L_i$. The multivariate density of bridge random variables can be modeled using Copula model, a multivariate joint cumulative distribution function used to joint univariate marginal distribution when the inverse cumulative of each variable is uniform distribution on the interval [0,1], Sklar (1959). Here we use the Gaussian Copula to joint bridge random variables.

If $F_1(b_1), F_2(b_2), F_3(b_3), F_4(b_4), F_5(b_5), F_6(b_6)$ are the cumulative distribution functions for the random effect variables $(b_{i1}, b_{i2}, b_{i3}, b_{i4}, b_{i5}, b_{i6})$, then there exist a function C such that the joint CDF is:

$$C(u_1, u_2, u_3, u_3, u_4, u_5, u_6) = P(U_1 \leq u_1, U_2 \leq u_1, U_3 \leq u_3, U_4 \leq u_4, U_5 \leq u_5, U_6 \leq u_6)$$

where $U_1, U_2, ...U_6$ variables are $F_1(b_1), F_2(b_2)...., F_6(b_6)$ has uniformly distributed CDF's and C is the density of Gaussian Copula is given by:

$$C(u_1, u_2, u_3, u_3, u_4, u_5, u_6) = \Phi_{Z_1,Z_2,...Z_6,\Sigma}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3), \Phi^{-1}(u_4), \Phi^{-1}(u_5), \Phi^{-1}(u_6)) \tag{2.14}$$

where $\Phi_{Z_1,Z_2,...Z_6,\Sigma}$ is the CDF of a multivariate normal distribution with mean zero vector and variance covariane matrix is Σ. Then, the bridge variable can be obtained by $b_r = G^{-1}(\Phi(Z_r))$, where $r = 1, 2, ...6$ and $\Phi(.)$ is the CDF of univariate standard normal and $G^{-1}(.)$ is inverse cumulative distribution of marginal bridge distribution. To specify the correlation matrix Σ, we

need to specify the Pearson correlation $\rho_{ish} = Corr(Z_{is}, Z_{ih})$ for each pair of Z's random variables. Parzen et al. (2011) recommend using Pearson correlation in terms of Kendall's τ to present the association between the the Z's random variables due its advantage of invariance of the monotone transformation, as discussed in Hougaard (2000).

$$\rho_{ish} = sin(\pi\tau_{ish}/2)$$

Then, inducing the correlation in Copula random variables using Kendall's τ, will be produced in bridge random variables because the bridge random variable are monotone transformation of Z's random variables. The maximum likelihood estimates of the parameter can be obtained by maximizing the likelihood function using Copula method. Because the method does not have a closed form, maximum likelihood estimates can be implemented using numerical approximations.

2.4.2 Natural constraints

Most of the simulation methods for the binary data have some constraints related to response means or the correlation structure. Using bridge distribution for the random effect is also has some constraints. To explore the limitations, we assumed just two bridge random effects to generate correlated binary data for the GEE model:

$$logit(P(Y_{ik} = 1|b_i, \beta)) = \beta_{0k} + \beta_{1k}\, X_k + b_{ik} \tag{2.15}$$

where the parameters $\beta_{01} = 1,\ \beta_{02} = 1,\ \beta_{11} = 1,\ \beta_{12} = 1$, and $b_i = (b_{i1}, b_{i2})$ are distinct and correlated random bridge intercepts and $X = 1, 2$. Under the bridge distributional assumption, the rescalling parameter ϕ is the connection between the regression parameters in the marginal and the conditional logistic model such that:

$$\beta_p = \beta_s * \phi$$

where ϕ is the parameter that measures the heterogeneity between the clusters. Also, ϕ is related to the variance of bridge random effect, $\sigma_b^2 = \pi^2(\frac{1}{\phi^2} - 1/3)$. As we see in the following figure, the variability of bridge random effect convergences to zero when ϕ is larger for both the theatrical and empirical relationship. In the context of generating artificial binary responses, we are looking to assume the best ϕ value that leads to better estimations.

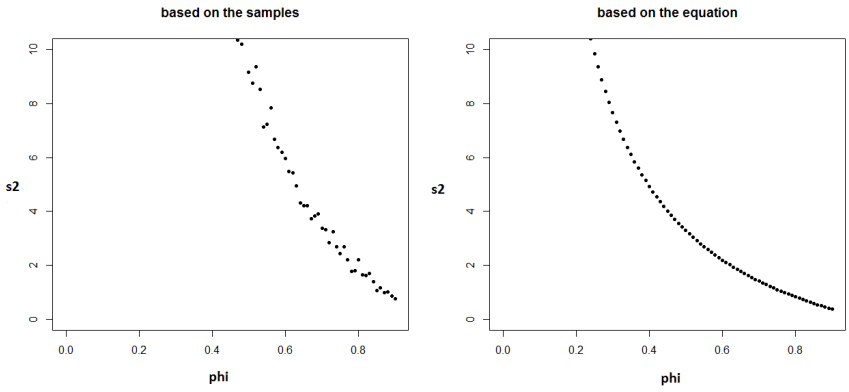

Figure 2.2: The relationship between the bridge parameter and its variance

In order to reach good estimation, it is important to investigate the connection between the correlation of the bridge random effects and the correlation of the binary responses. The efficiency of this method is based on the good induction of the desired correlation from bridge random effects into the binary responses $corr(Y_{i1}, Y_{i2}) \approx corr(b_{i1}, b_{i2})$. In figure 2.3, we explore the relationship between the Kendall's tau of the bridge random effect $corr(b_{i1}, b_{i2})$ and the Kendall's tau for the binary responses $corr(Y_{i1}, Y_{i2})$ for the range of correlation [-0.9,0.9] and for sample size=1000. The relationship is implemented for different values of ϕ, $\phi = 0.05, 0.30, 0.60, 0.80$. The best relationship is considered close to 45° degree line between the associations of bridge random effect and the binary responses. Thus, we recommend assuming $\phi = 0.05$ to induce the desired correlation from bridge random effect into the binary responses.

Secondly, the restriction on the estimation parameter β_s is imposed by ϕ since $\beta_p = \beta_s * \phi$. Wang and Louis (2003) said that the marginal parameter shrink toward zero when the heterogeneity is larger. Consequently, it is better to assume smaller values of β when ϕ is smaller. The last constraint is related to using Copula method of estimation. The R correlation matrix should be positive definite to generate Copula random variables.

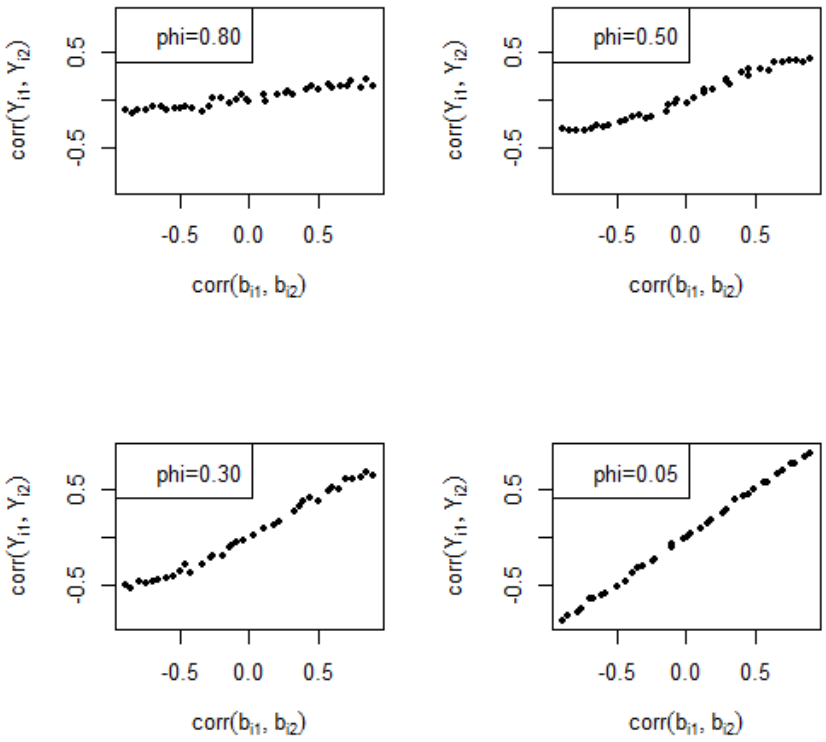

Figure 2.3: The relationship between the associations of (Y_{i1}, Y_{i2}) and the associations of (b_{i1}, b_{i2}) for different values of bridge parameter

2.5 Application to Simulation Design

We conducted a simulation study for the five correlation scenarios in order to explore the properties of using the proposed method. Specifying bridge distribution for the random effects is to generate multivariate longitudinal binary data. One of the goals of designing the simulation study was to determine the efficient sample size for specified model that leads to statistically- significant result in the treatment effect between the outcomes. A larger sample size certainly leads to more accurate parameter estimations, but would raise the research budget. Further, in clinical trials, it would require more human subjects who would be exposed to new treatments that may be harmful. In this section, we conducted a simulation study to estimate the efficient sample size in clinical trails needed to detect statistically significant results for the treatment using the proposed method.

Let $X_i = 0, 1$ is the treatment covariate and $t_j = 1, 2, 3$ is time covariate for three occasions. Let Y_{ijk} be the binary response that measured at time j, $j = 1, 2, 3$ for observation i, $i = 1, 2, 3...N$ and for outcome $k = 1, 2$. Then, the true logistic model be

$$logit(E(Y_{ijk}|X_i, b_i)) = logit(P_{ijk}) = \beta_{0k} + \beta_{1k}X_i + \beta_{2k}t_j + b_{ijk} \tag{2.16}$$

where $b_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4}, b_{i5}, b_{i6})$ are distinct but correlated random bridge intercept for each outcome at each occasion or time. Given the vector of the random effect b_i, the Y_{ijk} for subject i are assumed to be independent Bernoulli random variables, $Y_{ijk}|b_i \sim Ber(P(Y_{ijk} = 1))$. The marginal model will be:

$$logit(E(Y_{ijk}|X_i)) = 1/\phi\ (\beta_{0k} + \beta_{1k}X_i + \beta_{2k}t_j) \tag{2.17}$$

where the parameters $\phi = 0.05,\ \beta_{01} = 4,\ \beta_{02} = 2,\ \beta_{11} = 1,\ \beta_{12} = -3,\ \beta_{21} = 1,\ \beta_{22} = -5$. For the random effect model, we conducted the simulation study based on the GEE model of Shelton et al. (2004) to separate the estimated effects for each outcomes using Kronecker product. The simulation for the five correlation scenarios to explore the properties of the model when the correlation is induced over the outcomes and the occasions. Additional goal of this study was to estimate the effect sample size to reject the null hypothesis of the treatment group for the two outcomes, $H_0 : \beta_{11} = \beta_{12} = 0$.

2.6 The Results

We ran a simulation method for N=200 samples. The clinical trial is balance for each arm. For 200 samples, we computed the correlation mean for each correlation parameter and the results are in figure 2.4. The correlation means are calculated for the five scenarios. Starting from scenario 1, it can be seen as a reference scenario since we assume there is no correlation over occasions or outcomes. It seems it has good convergence close to zero of all its parameters. The correlation scenarios is 2 and 3 are supposed to be induced over the outcomes' parameters α and scenario 4 and 5 over the occasions' parameters υ. From figure 2.4, the estimated correlation matrix in scenarios 4 and 5 has a good results. Also, it is clear that the parameter estimation of the correlation matrix for scenarios 2 and 3 have increased bias over the time, meaning that the correlation between the

outcomes in time 1 is better than time 3. Adding the time dependent covariate in the longitudinal response model may effects the correlation for the binary responses.

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array} \\ \begin{pmatrix} 1 & \upsilon & \upsilon & \alpha & \tau & \tau \\ - & 1 & \upsilon & \tau & \alpha & \tau \\ - & - & 1 & \tau & \tau & \alpha \\ - & - & - & 1 & \upsilon & \upsilon \\ - & - & - & - & 1 & \upsilon \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.000 & 0.000 & 0.000 & -.002 & -.005 \\ - & 1 & -.003 & 0.001 & 0.006 & -.001 \\ - & - & 1 & -.001 & -.006 & -.005 \\ - & - & - & 1 & 0.000 & 0.002 \\ - & - & - & - & 1 & 0.002 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 1

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.003 & 0.000 & 0.571 & -.001 & 0.006 \\ - & 1 & 0.001 & 0.004 & 0.535 & 0.001 \\ - & - & 1 & 0.002 & -.009 & 0.497 \\ - & - & - & 1 & 0.003 & 0.008 \\ - & - & - & - & 1 & -.001 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 2

$$\begin{array}{c} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array} \\ \begin{pmatrix} 1 & 0.000 & -.002 & 0.770 & -.004 & -.002 \\ - & 1 & 0.000 & -.001 & 0.668 & 0.000 \\ - & - & 1 & 0.003 & 0.000 & 0.574 \\ - & - & - & 1 & -.001 & 0.003 \\ - & - & - & - & 1 & 0.001 \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

scenario 3

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.593 & 0.596 & 0.000 & -.000 & -.002 \\ - & 1 & 0.590 & 0.000 & 0.001 & -.001 \\ - & - & 1 & 0.002 & 0.001 & -.001 \\ - & - & - & 1 & 0.580 & 0.568 \\ - & - & - & - & 1 & 0.583 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 4

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.881 & 0.875 & -.000 & 0.000 & 0.000 \\ - & 1 & 0.881 & 0.000 & 0.000 & -.000 \\ - & - & 1 & 0.000 & -.000 & -.001 \\ - & - & - & 1 & 0.849 & 0.769 \\ - & - & - & - & 1 & 0.846 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 5

Figure 2.4: The estimated correlation structures using the proposed method

The second output is for the parameter estimations. Assuming the log odds the responses changes curvilinerly with the time and X, we got the means of estimated regression coefficients over 200 samples in table 3.3. Starting from scenario 1, we found the estimated effect of intercept means for the outcomes 1 and 2 respectively are 0.206, 0.089 over 200 samples and each sample size is n=400. The parameter estimations $\beta_{01}, \beta_{11}, \beta_{21}$ are the log odds of $P(Y_{i1} = 1)$ for intercepts, treatment and time covariates respectively and $\beta_{02}, \beta_{12}, \beta_{22}$ are for the second outcome. Generally for all the scenarios, the parameter estimations are approximately close to the true values unless in scenario 4 and 5. The standard deviation std in scenario 4 and 5 for treatment effect is large comparing with the other scenarios. That leads to conclude the strong correlation in the time factor may affects the bias of the regression coefficients especially in scenarios 4 and 5.

Table 2.2: The covariate parameter estimations

	senario 1		senario2		senario3		senario4		senario5	
$True\ \beta$	mean	std	mean	std	mean	std	mean	std	mean	std
$\beta_{01} = 0.2$	0.206	0.168	0.205	0.166	0.210	0.162	0.213	0.159	0.215	0.148
$\beta_{02} = 0.1$	0.089	0.178	0.120	0.166	0.114	0.156	0.1068	0.083	0.079	0.162
$\beta_{11} = 0.05$	0.059	0.112	0.046	0.117	0.063	0.111	0.051	0.171	0.035	0.221
$\beta_{12} = -0.15$	-0.152	0.117	-0.163	0.130	-0.138	0.114	-0.130	0.173	-0.121	0.214
$\beta_{21} = 0.05$	0.046	0.077	0.048	0.073	0.043	0.070	0.046	0.044	0.051	0.027
$\beta_{22} = -0.25$	-0.245	0.074	-0.259	0.072	-0.258	0.063	-0.249	0.045	-0.249	0.038

One of the goals of this case study was to estimate the best sample size for to detect significant treatment effect in outcome 1 or outcome 2. We applied the proposed method in the clinical trial model for range of sample sizes $n = (400, 800, 1200, 1600, 2000, 2400, 2800, 5000, 8000)$ and counted how many times the null hypothesis $H_0 : \beta_{11} = \beta_{12} = 0$ is rejected for each sample size versus at a least one of the parameter estimation is not zero. To get the study power, we estimated the power as a function of the sample size in order to estimate the best sample size leads to get 0.80 power value. In graph 2.5, we present the effect sample size for each scenario. In scenario 1, the best sample size for two arms is n=2200 that means approximately 1100 for each study arm. It is clear the effect sizes for scenarios 2 and 3, which expressed the correlation between the outcomes in each occasion, are lower than other scenarios. This happens may because this model is designed to separate the effects of the parameter for each outcome, then the correlation over occasions required higher sample size. The highest required sample sizes are approximately n=8850, 4880 for scenarios 4 and 5, respectively, when the correlation is induced over the occasions for each outcome.

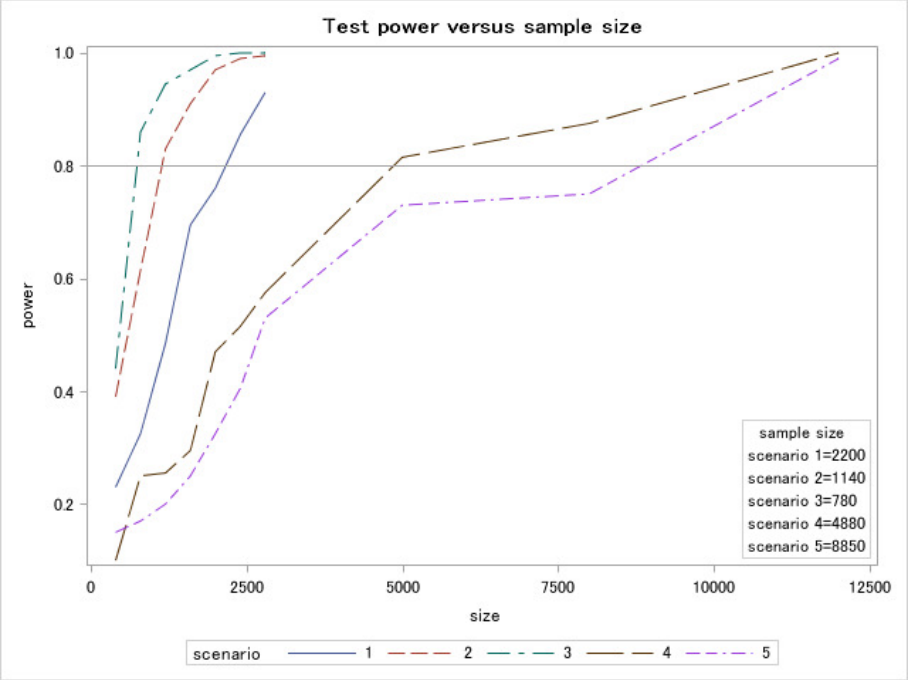

Figure 2.5: The sample size estimations over the correlation scenarios

In conclusion, the multivariate longitudinal data potentially has complicated correlation. The

correlation among the responses comes from the repeated measurements and the outcomes that are measured from the same observation. The estimations of each correlation pattern in multivariate structure for the five scenarios is computed. Scenarios 4 and 5 present the induced correlation over the occasions while scenarios 2 and 3 present the induced correlations between the outcomes. We saw in scenarios 2 and 3 the bias increases gradually over the time and maybe this due to the existence of the dependent coviarate in the model. Also, the parameter estimations over the five scenarios did not changed dramatically when we changed the source and the strength of the correlation over the scenarios unless in the treatment effects. It is clear the strong correlation produce more bias estimates. In fact, The effect sample size for the study model is also effected by the scenario. Based on our model, clinical trials require higher sample sizes for high correlations over the occasions, as we saw in scenarios 4 and 5.

2.7 Conclusion

Researchers have discussed variety methods to address problems related to generating correlated binary data for marginal models. In this paper, we describe a simple computed method using a linear mixed model via bridge distribution for the random effect. Using bridge distribution, it has the advantage to keep the same logistic shape for the marginal and conditional models. This method could reach good convergence for a desired R correlation matrix. It could be a good future work to study a comparison between the proposed method and Emrich and Piedmonte (1991) and Qaqish (2003) methods. Choosing the appropriate bridge parameter and parameter estimation of the marginal model would effect the results convergence of using bridge distribution. In conclusion, generate the binary responses for the marginal model using bridge distribution for the random effect could be good approach.

CHAPTER 3

A COMPARISON OF THREE MODELS IN MULTIVARIATE BINARY LONGITUDINAL DATA ANALYSIS

3.1 Summary

Multivariate longitudinal data analysis plays an important role in many biomedical and social problems. In this article, we present three methods for analyzing multiple and correlated binary outcomes; each one can be beneficial for determined aims. We review method one and method two and we proposed method three. The three methods estimate the marginal means using the GEE approach for multivariate binary longitudinal data. The first method addresses the question of estimating one group of covariate parameters for many binary outcomes while accounting for their multivariate structure. The second method addresses the question of estimating the covariate parameters for each binary outcome separately. The third method is an estimation of the covariate parameters for each combination of outcomes. Our goal is to investigate the difference among the parameter estimations of the three methods. In the simulation element, we present many scenarios related to different correlation structures. In the application element, we present a follow up study (Florida Dental Care Study) that measured three binary outcomes and five covariates in four intervals. That particular study is a useful explanation of the variation between outcomes since the outcomes were highly correlated.

3.2 Introduction

Many of the recent medical experiments and social research are characterized by multiple outcomes. In this research, we focus on the multivariate binary outcomes in longitudinal data. This situation occurs when there are many vectors of binary responses which are obtained on many occasions or visits and many covariates. There are multiple methods for analyzing multivariate binary longitudinal data, which differ depending on the research aims.

1) Reducing many outcomes into one summary outcome which requires a unique set of regression coefficients.

2) Analyzing each outcome separately with account for the correlations, which requires a set of regression coefficients for each outcome.

3) Analyzing the outcomes jointly and accounting for the correlations, which requires a set of regression coefficients for each combination case of them.

Lets start by refer "outcome" to the longitudinal readings within the same dependent variable and "response" for any longitudinal reading of dependent variable without specification. If the outcomes are uncorrelated, then we can analyze them as univariate longitudinal data. This independence assumption rarely happens in real applications because if the outcomes are taken from the same observation, then the correlation is expected to be observed among the outcomes. The best statistical analysis of multiple outcomes should account for the correlation among the outcomes and the occasions since it is longitudinal data.

Our investigating method of the longitudinal data analysis is the marginal models using GEE approach. The contribution of this study is a proposed method that is beneficial for the third aim. Studying the evolution of many combinations of multivariate binary outcomes over time is the main interest in the third method. We will convert multiple binary outcomes into one nominal multinomial outcome. Then, we will use a method of analyzing multinomial longitudinal data. All the methods in this study are marginal models using generalized estimation equation approach (GEE).

Liang and Zeger (1986) constructed the GEE approach of longitudinal data analysis for discrete and continuous outcomes. Then, Carey et al. (1993) applied the alternating logistic regression in multivariate binary model which is dependent on modeling the association between responses by the odds ratio. Rochon (1996) proposed a method to model two longitudinal outcomes, which can be mixture of binary and continuous types. Especially for the binary case, O'Brien and Fitzmaurice (2004) present two methods of analyzing multivariate binary longitudinal data. Their study is helpful for the first aim. Shelton et al. (2004) also build a SAS macro using GEE approach and it is helpful for the second aim. There are many statistical methods which are proposed for different types of outcomes. Choosing the correct method depends on the outcome type and the research aim. In this study, we focus in the binary outcomes in longitudinal data.

In section 1, we introduce an overview of the GEE marginal models for univariate longitudinal data analysis. Section 2 gives details of the GEE approach for multivariate cases for binary longitudinal data analysis. It is an explanation of the first and the second methods. Section 3 focuses on the third method, in which we utilize the nominal multinomial binary longitudinal data analysis in our proposed method. The simulation part is in section 5 and the application in section 6. The last section is the conclusion and future studies.

3.3 Marginal Models Using GEE Approach For Univariate Longitudinal Data Analysis

The generalized estimating equation (GEE) extends generalized linear modeling to longitudinal data. It leads to analyzing the marginal models of the correlated responses. The main target of marginal model inference is "population average model". It provides a unified method for analyzing different types of longitudinal responses without making assumptions about the distribution of the responses. The use of GEE approach has been proposed by Liang and Zeger (1986). Let $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, ..., Y_{iT_i})^T$ be a vector of T_i repeated measurements of i^{th} subject, $i = 1, 2, ...N$. There is $X_i = \left(X_{i1}^T, X_{i2}^T, X_{i3}^T, ..., X_{iT_i}^T\right)$ matrix of covariates, and $X_{i1}^T = (X_{i10}, X_{i11}, X_{i12}, ..., X_{i1C})$ are the covariate readings in time 1 and for time 2 is $X_{i2}^T = (X_{i20}, X_{i21}, X_{i22}, ..., X_{i2C})$, ... The corresponding occasions or visit numbers vector is $t_i = (t_{i1}, t_{i2}, t_{i3}, ..., T_i)^T$ at which the measurements are made. The GEE approach of marginal models separately models the mean response and within subject association among the repeated measurements. In marginal models, the covariance structure is regarded as nuisance since the goal is to make inferences about the mean response. Suppose the marginal expectation of the responses,

$$E(Y_{it}) = \mu_{it} \tag{3.1}$$

depends on the covariate X_i through a link function with:

$$g(\mu_{it}) = X_{it}\,\beta \tag{3.2}$$

where $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_C)^T$ are the regression parameters. The variance of each Y_i depends on the mean according to

$$var(Y_{it}) = \phi\, v(\mu_{it}) \tag{3.3}$$

where ϕ is a dispersion parameter and v(.) is known as a variance parameter. The covariance matrix of Y_i is given by

$$V_i = \phi \, A_i^{1/2} \, R(\gamma) \, A_i^{1/2} \tag{3.4}$$

where A_i is a diagonal matrix with $var(Y_{it}) = \phi \, v(\mu_{it})$ along the diagonal and $A^{1/2}$ is diagonal matrix with $\sqrt{\phi \, v(\mu_i)}$ along the diagonal. $R(\gamma)$ is the correlation matrix that is a function of γ where γ represents a vector of within subject association parameters. In GEE, R is the "working" correlation matrix and V_i is known as the "working" covariance matrix. The term "working" is used to express the V_i is approximation of the true covariance of the response. It is a function of ϕ, β via $v(\mu_i)$ and the within subject association γ. V_i can average dependence among the repeated responses over the subjects . A GEE estimator of β vector is obtained by solving the estimating equations.

$$\sum_{i=1}^{N} D_i{}' V_i^{-1} (y_i - \mu_i) = 0 \tag{3.5}$$

$$\mathbf{D_i} = \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \beta_0} & \frac{\partial \mu_{i1}}{\partial \beta_1} & \cdots & \frac{\partial \mu_{i1}}{\partial \beta_C} \\ \frac{\partial \mu_{i2}}{\partial \beta_0} & \frac{\partial \mu_{i2}}{\partial \beta_1} & \cdots & \frac{\partial \mu_{i2}}{\partial \beta_C} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \mu_{in_i}}{\partial \beta_0} & \frac{\partial \mu_{in_i}}{\partial \beta_1} & \cdots & \frac{\partial \mu_{in_i}}{\partial \beta_C} \end{pmatrix} \tag{3.6}$$

The solution requires an iterative algorithm; the two-stages of iterative method depend on β and γ:

1) $\widehat{\beta}$ is obtained as the solution of (3.5) by GEE. Given the current estimates of γ, ϕ , then V_i is estimated.

2) After getting the current estimates of $\widehat{\beta}$,then $\widehat{\gamma}$ and $\widehat{\phi}$ are obtained based on standardized residuals:

$$e_{it} = \frac{(Y_{it} - \widehat{\mu}_{it})}{\sqrt{v(\widehat{\mu}_{it})}} \tag{3.7}$$

In the binary response, as we will explain later, $\phi = 1$. The $\widehat{\gamma}$ can be estimated depends on the model of within subject association. For example, if we assume γ to be unstructured $\gamma_{tt'} = Corr(Y_{it}, Y_{it'})$, then $\gamma_{tt'} = \dfrac{\sum_{i=1}^{N} e_{it} e_{it'}}{\widehat{\phi} \, N}$. The GEE method for marginal models iterate between step 1 and step 2 until the convergence is achieved. Initial values of β are computed from fitting a generalized linear model

assuming independence among the observations. The GEE method yields consistent estimates of $\widehat{\beta}$ even if we misspecified the structure of the covariance matrix.

In large samples, $\widehat{\beta}$ has a multivariate normal distribution of mean β and

$$cov(\widehat{\beta}) = B^{-1}MB^{-1} \tag{3.8}$$

where $B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i$, $M = \sum_{i=1}^{N} D_i' V_i^{-1} cov(Y_i) V_i^{-1} D_i$ and $cov(Y_i) = \sum_{i=1}^{N} (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)'$. This is known as the empirical or "sandwich" estimator . Finally, the GEE approach to estimate the marginal model is regarded as semi-parametric since we estimate the equations without full specification of the distribution of the responses.

Our interest is in binary responses, and we will explain the application of GEE method on marginal models for binary responses. Let Y_i denote binary responses which are limited to two values (0 denoting "failure") and (1 denoting "success"). Then,

$$\mu_{it} = E(Y_{it}) = Pr(Y_{it} = 1) \tag{3.9}$$

Y_{it} has a Bernoulli distribution with probability of "success" μ_{it}. The logit link function relates the covariates to $E(Y_i) = \mu_i$, then $g(\mu_{it}) = logit(\mu_{it}) = log\dfrac{\mu_{it}}{1-\mu_{it}} = X_{it}\beta$ and $var(Y_{it}) = \mu_{it}(1-\mu_{it})$. Note that $(\phi = 1)$ in the marginal model for binary responses. The within subject association "γ" among binary responses in the case of assuming "unstructured" pairwise log odds ratio is:

$$Log(OR(Y_{it}, Y_{it'})) = \gamma_{tt'} \tag{3.10}$$

where $OR(Y_t, Y_{t'}) = \dfrac{Pr(Y_t = 1, Y_{t'} = 1)Pr(Y_t = 0, Y_{t'} = 0)}{Pr(Y_t = 1, Y_{t'} = 0)Pr(Y_t = 0, Y_{t'} = 1)}$.

3.4 Marginal Models Using GEE Approach For Multivariate Longitudinal Data Analysis

The multivariate longitudinal data model is a general approach of the univariate longitudinal model when we have more than one outcome. Each individual i has a vector of responses for different outcomes, $k = 1, 2, ..K$, which are measured at different times or occasions, $t = 1, 2 \ldots T_i$, and has cluster size $n_i = T_i K$. To simplify the notations, we will refer to T_i as T which is the number of occasions or visit numbers over all the observations :

$$Y_i = [Y_{i11}, Y_{i21}, ...Y_{iT1}, Y_{i12}, Y_{i22}, Y_{i32}, ..., Y_{iT2}, ..., Y_{i1K}, Y_{i2K}, ..., Y_{iTK}]^T$$

Let us model K vectors of outcomes measured corresponding to vector of times. Then, the structure is in the following figure:

$$\begin{matrix} ID \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ \vdots \\ N \\ N \\ \vdots \\ N \end{pmatrix} \end{matrix} \quad \begin{matrix} Y_{it1} & Y_{it2} & \cdots & Y_{itK} \\ \begin{pmatrix} y_{111} & y_{112} & \cdots & y_{11K} \\ y_{121} & y_{122} & \cdots & y_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1T1} & y_{1T2} & \cdots & y_{1TK} \\ y_{211} & y_{212} & \cdots & y_{21K} \\ y_{221} & y_{222} & \cdots & y_{22K} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_{N11} & y_{N12} & \cdots & y_{N1K} \\ y_{N21} & y_{N22} & \cdots & y_{N2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{NT1} & y_{NT2} & \cdots & y_{NTK} \end{pmatrix} \end{matrix} \quad \begin{matrix} time \\ \begin{pmatrix} 1 \\ 2 \\ \vdots \\ T \\ 1 \\ 2 \\ \vdots \\ \vdots \\ 1 \\ 2 \\ \vdots \\ T \end{pmatrix} \end{matrix}$$

Figure 3.1: Multivariate longitudinal data

For each outcome k, Rochon (1996) assumed the balance design for generalized linear model (GLM) form:

$$g_k(\mu_i^{(k)}) = X_i^{(k)} \ \beta^{(k)} \tag{3.11}$$

where $\mu_{ik} = E(Y_{ik})$, k=1,2. The $X_{ik}, k = 1, 2$ are matrices of covariates measured from the i^{th} individual and $g_k(.)$ are known link functions not necessarily identical.

It is straightforward to extend Rochon's work to accommodate multiple vectors of outcomes. Then,

$$var(Y_{it}^{(k)}) = \phi_k \ v(\mu_{it}^{(k)}) \tag{3.12}$$

where ϕ_k is dispersion parameter vector and $v(.^{(k)})^s$ are known variance functions. The covariance matrix of $Y_i^{(k)}$ is given by,

$$V_i^{(k)} = \phi_k \ (A_i^{(k)})^{1/2} R_{(kk)} \ (A_i^{(k)})^{1/2} \tag{3.13}$$

Letting $\phi_k = 1$ for binary outcomes and $A_i^{(k)}$ is a diagonal matrix with entities

$$v_k(\mu_{i1}^{(k)}), v_k(\mu_{i2}^{(k)}), ..., v_k(\mu_{iT}^{(k)})$$

The R_{kk} is the "working" correlation matrix among the responses of k^{th} outcomes. Rochon (1996) who used GLM to model each repeated outcome that is followed by "SUR", seemingly unrelated regression, combining ($k = 2$) outcome estimations. It is a generalization of the liner regression model of many regression equations in which each one has its dependent variable. Letting $Y_i = [{Y_i^{(1)}}^T, {Y_i^{(2)}}^T, ..., {Y_i^{(K)}}^T]^T$, then the SUR model is written as

$$g(\mu_i) = X_i\beta \tag{3.14}$$

Rochon (1996) assume $k = 2$ then the components are $g_1(.)$ and $g_2(.)$ and it can be extended to K components $g_1(.), g_2(.), ..., g_K(.)$. letting

$$X_i = \begin{bmatrix} X_i^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & X_i^{(2)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & X_i^{(K)} \end{bmatrix}, \ \mu_i = \begin{bmatrix} \mu_i^{(1)} \\ \mu_{i2}^{(2)} \\ \vdots \\ \mu_i^{(K)} \end{bmatrix}, \ \beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \\ \vdots \\ \beta^{(K)} \end{bmatrix}$$

$$V_i = \phi^{1/2} A_i^{1/2} R \, \phi^{1/2} A_i^{1/2} \tag{3.15}$$

where $A_i = \begin{bmatrix} A_i^{(1)} & 0 & 0 & \cdots & 0 \\ 0 & A_i^{(2)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_i^{(K)} \end{bmatrix}$, $\phi = \begin{bmatrix} \phi_1 I_n & 0 & 0 & \cdots & 0 \\ 0 & \phi_2 I_n & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \phi_K I_n \end{bmatrix}^{1/2}$ and

$$R = R(\gamma) = \begin{bmatrix} R_{11} & R_{12} & . & . & R_{1K} \\ R_{12}^T & R_{22} & . & . & R_{2K} \\ . & . & . & . & . \\ R_{1K}^T & R_{2K}^T & . & . & R_{KK} \end{bmatrix}$$

R_{11} is the correlation matrix of the repeated measurements within the first outcome. V_i is the working correlation matrix of K repeated measurements that has $q \times 1$ vector of γ parameters. The GEE estimator of β is obtained by solving

$$\sum D_i^T V_i^{-1} \, (Y_i - \mu_i) = 0 \tag{3.16}$$

$$\mathbf{D_i} = \begin{pmatrix} \frac{\partial \mu_i^{(1)}}{\partial \beta^T} & 0 & \dots & 0 \\ 0 & \frac{\partial \mu_i^{(2)}}{\partial \beta^T} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \frac{\partial \mu_i^{(K)}}{\partial \beta^T} \end{pmatrix} \tag{3.17}$$

Again as we write in the univariate case, there is no closed form of the solution of D_i and Rochon (1996) used procedure of Crowder (1985) for correlated binary data. The GEE estimations $\hat{\beta}$ are consistent for β under the assumption of multivariate normal distribution with mean β and variance Σ, the sandwich variance estimator is:

$$\hat{\Sigma} = B^{-1}MB^{-1},$$

where $B = \sum_{i=1}^{N} D_i' V_i^{-1} D_i$, $M = \sum_{i=1}^{N} D_i' V_i^{-1} cov(Y_i) V_i^{-1} D_i$ and $cov(Y_i) = \sum_{i=1}^{N} (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)'$. Rochon suggests double iterative methods between $\hat{\beta}$ and $\hat{\gamma}$ until the convergence.

Bandyopadhyay et al. (2011) implemented the joint analysis of Rochon (1996) but for four outcomes. The specification of the odds ratio matrix R, is the most important point on the GEE analysis of the marginal models. We need to estimate $\frac{(KT)(KT-1)}{2} = \binom{KT}{2}$ parameters. It is a large correlation matrix and it will rapidly increase when we have more parameters of R (correlation matrix). Since our interests are in binary multivariate longitudinal outcomes, then it is more natural to use the odds ratio to describe the association between the binary responses. O'Brien and Fitzmaurice (2004) applied two methods to reduce the parameter estimations of the within subject association matrix R. The two methods are Kronecker product and regression model for pairwise odds ratio. The two methods are equivalent to each other and have approximately the same results. We will focus on the regression model for pairwise odds ratio to be method 1 in the comparison part. Also, Shelton et al. (2004) built a SAS macro to analyze multivariate binary longitudinal outcomes using Kronecker product. They utilized a different approach than O'Brien and Fitzmaurice (2004) did. In this paper, we will consider the regression model for pairwise odds ratio of O'Brien and Fitzmaurice (2004) to be the statistical tool of the first aim and Kronecker product of Shelton et al. (2004) for the second aim. The next sections include additional explanations of these methods.

3.4.1 Regression modeling for odds ratio

Let Y_{ikt} be a binary response vector on subject i from outcome k and at time t, $(i = 1, 2 \ldots N)$, $(k = 1, 2, \ldots K)$ and $(t = 1, 2 \ldots T)$. For simplicity, assuming the vector of all responses:

$$Y_i = [Y_{i11}, Y_{i21}, ...Y_{iT1}, Y_{i12}, Y_{i22}, Y_{i32}, ..., Y_{iT2}, ..., Y_{i1K}, Y_{i2K}, ..., Y_{iTK}]^T$$

The GEE model is $g(\mu_i) = X_i\beta$ where $E(Y_i) = \mu_i$ and X_i is the covariate matrix and assuming the vector of responses for subject i is related to the covariate via logit link function.

The primary advantage of using GEE in multivariate binary data is working in two-way marginal relationships between the pairs of the responses for any dimension of outcomes and occasions without high order associations. The penalty of this advantage is the length of the cluster size which is $T \times K$. It will cause a larger correlation matrix R since the dimension of R matrix is $TK \times TK$. R is a diagonal and symmetric matrix, thus we need to estimate $\binom{TK}{2}$. For example, if we have 5 outcomes obtained at 5 occasions, then we need to estimate 300 parameters. We need to obtain a parsimonious model of working correlation R to be less than $\binom{TK}{2}$.

As we mentioned above, we can use the odds ratio instead of the correlation since the responses are binary. Then, the matrix R consists of pairwise log odds ratios. O'Brien and Fitzmaurice (2004) fit a regression model for marginal pairwise odds ratio to estimate less parameters than $\binom{TK}{2}$. Let γ be a vector of size $\binom{TK}{2}$ of all non-redundant pairwise odds ratio in R. The goal is reducing the length of the vector γ. Consider the matrix of the odds ratio R consist of three odds ratio types:

1- Let $\upsilon_{tk,tk'}$ be the inter-outcome odds ratio which compares the outcome k with the outcome $k^{'}$at time t:

$$\upsilon_{tk,tk'} = \frac{Pr(Y_{tk} = 1, Y_{tk'} = 1)Pr(Y_{tk} = 0, Y_{tk'} = 0)}{Pr(Y_{tk} = 1, Y_{tk'} = 0)Pr(Y_{tk} = 0, Y_{tk'} = 1)} \tag{3.18}$$

2- Let $\alpha_{tk,t'k}$ be the intra-outcome odds ratio which compares outcome k at time t with the same outcome at time $t^{'}$:

$$\alpha_{tk,t'k} = \frac{Pr(Y_{tk} = 1, Y_{t'k} = 1)Pr(Y_{tk} = 0, Y_{t'k} = 0)}{Pr(Y_{tk} = 1, Y_{t'k} = 0)Pr(Y_{tk} = 0, Y_{t'k} = 1)} \tag{3.19}$$

3- Let $\tau_{tk,t'k'}$ be the cross odds ratio which compares the outcome k at time t with outcome $k^{'}$ at time $t^{'}$:

$$\tau_{tk,t'k'} = \frac{Pr(Y_{tk} = 1, Y_{t'k'} = 1)Pr(Y_{tk} = 0, Y_{t'k'} = 0)}{Pr(Y_{tk} = 1, Y_{t'k'} = 0)Pr(Y_{tk} = 0, Y_{t'k'} = 1)} \tag{3.20}$$

The general regression model that relates the inter-outcome (υ), the intra-outcome (α) and the cross odds ratio (τ) to matrix Z via log link function, Carey et al. (1993), is:

$$log(OR) = Z\gamma \, , \tag{3.21}$$

where γ is a parsimonious vector of the odds ratio parameters that relate the $\binom{KT}{2}$ pairwise log odds ratio to Z. Here Z is a fixed indicator matrix that specifies the outcomes, times and interactions

between them which are under consideration. Using the regression model of (3.21) helps us to model the three types of association in matrix Z. For example, consider observations taken from two outcomes at two occasions and $OR(Y_{tk}, Y_{t'k'})$ represents the marginal odds ratio comparing the responses k at time t with response $k^{'}$ at time $t^{'}$, then the log odds ratio regression model and working correlation matrix R are:

$$log(OR(Y_{tk}, Y_{t'k'})) = \gamma_1 I_{(t=t',k\neq k')} + \gamma_2 I_{(t\neq t',k=k')} + \gamma_3 I_{(t\neq t',k\neq k')}$$

$$R = log(OR) = \begin{array}{c} \\ Y_{t_1k_1} \\ Y_{t_1k_2} \\ Y_{t_2k_1} \\ Y_{t_2k_2} \end{array} \begin{array}{c} \begin{array}{cccc} Y_{t_1k_1} & Y_{t_1k_2} & Y_{t_2k_1} & Y_{t_2k_2} \end{array} \\ \begin{pmatrix} 0 & \gamma_1 & \gamma_2 & \gamma_3 \\ - & 0 & \gamma_3 & \gamma_2 \\ - & - & 0 & \gamma_1 \\ - & - & - & 0 \end{pmatrix} \end{array}$$

Where $I(.)$ is an indicator function that equals 1 if the condition $(.)$ is satisfied. This model specifies that $log(\phi) = \gamma_1,\ log(\alpha) = \gamma_2,\ log(\tau) = \gamma_3$. Assuming the association are exchangeable within each type, inter-outcome υ, the intra-outcome α and cross odds ratio τ. Finally, the analysis of multivariate binary longitudinal model using regression model of odds ratio depends on $g(\mu_i) = X_i\beta$ and $log(OR) = Z\gamma$ to obtain a marginal model will produce a unique set of regression coefficients for all outcomes (first aim). In the following section, Kronecker Product approach will be helpful for the second aim that produces an estimation of a set of regression coefficient for each outcome separately.

3.4.2 Kronecker product

Shelton et al. (2004) built a SAS macro to estimate separate sets of regression coefficients for each binary outcome using a Kronecker product approach in longitudinal data. This method depends on the GEE estimation of the marginal models. The strength of this macro is creating a covariate design to allow for separate regression coefficients for each outcome.

Let Y_{ikt} be a binary response on subject i from outcome k and at time t, $(i = 1, 2..N)$, $(k = 1, 2, ..K)$ and $(t = 1, 2...T)$. For simplicity, assuming:

$$Y_i = [Y_{i11}, Y_{i21}, ...Y_{iT1}, Y_{i12}, Y_{i22}, Y_{i32}, ..., Y_{iT2}, ..., Y_{i1K}, Y_{i2K}, ..., Y_{iTK}]^T$$

The GEE model is $logit(\mu_i) = X_i\beta$ where $E(Y_i) = \mu_i$ and X_i is the covariate matrix and assuming the vector of responses for subject i is related to the covariate via logit link function.

Using Kronecker product of X matrix and K dimensional identity matrix be used to generate regression coefficients for each outcome. For example, suppose there are two outcomes and two covariates which are measured at four times. The outcome vector of all the binary outcomes is $KT \times 1$, and X_i covariate matrix $KT \times C$ (which C is number of covariate), then the block diagonal matrix that is resulted from $X \otimes I_K$ is TK by CK. The portion corresponding to the first observation is:

$$\begin{matrix} ID \\ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \end{matrix} \quad \begin{matrix} Y \\ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \end{matrix} \quad \begin{matrix} time \\ \begin{pmatrix} 1 \\ 2 \\ 1 \\ 2 \end{pmatrix} \end{matrix} \quad \begin{matrix} X_1 & X_2 \\ \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 1 & 2 \\ 3 & 4 \end{pmatrix} \end{matrix} \quad \begin{matrix} X & \otimes & I_K \\ \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 3 & 4 \end{pmatrix} \end{matrix}$$

Figure 3.2: Kronecker product example

The "working" correlation matrix contains the parameter estimations cross the times and the outcomes:

$$R(\gamma) = \begin{matrix} \\ Y_{t_1k_1} \\ Y_{t_2k_1} \\ Y_{t_1k_2} \\ Y_{t_2k_2} \end{matrix} \begin{matrix} \begin{matrix} Y_{t_1k_1} & Y_{t_2k_1} & Y_{t_1k_2} & Y_{t_2k_2} \end{matrix} \\ \begin{pmatrix} 1 & \gamma_1 & \gamma_2 & \gamma_3 \\ - & 1 & \gamma_3 & \gamma_2 \\ - & - & 1 & \gamma_1 \\ - & - & - & 1 \end{pmatrix} \end{matrix}$$

The Kronecker product method will generate a group of parameters estimations for each outcome. This method is beneficial in order to to fit more than one longitudinal outcome using GEE approach and accounts for the correlation among the outcomes. The result is separated estimation coefficients for each outcome, which is helpful for the second aim.

3.5 The Proposed Approach For Analyzing Binary Multivariate Longitudinal Data

Let Y_{itk} denote the value of k^{th} binary outcome measured at occasion or visit number t for individual i, where $i = 1, 2...N$, $k = 1, 2 \ldots K$ and $t = 1, 2, ..T$. Let X_{itc} denote the value of the

covariate c that is measured at occasion t for individual i where $c = 1, 2, \ldots C$ (C is the number of covariate variables). Each observation has a unique identification number 'ID'. The raw data are in the following figure:

$$\begin{array}{c} ID \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ \vdots \\ N \\ N \\ \vdots \\ N \end{pmatrix} \end{array} \quad \begin{array}{c} Y_{it1} \quad Y_{it2} \quad \cdots \quad Y_{itK} \\ \begin{pmatrix} y_{111} & y_{112} & \cdots & y_{11K} \\ y_{121} & y_{122} & \cdots & y_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1T1} & y_{1T2} & \cdots & y_{1TK} \\ y_{211} & y_{212} & \cdots & y_{21K} \\ y_{221} & y_{222} & \cdots & y_{22K} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_{N11} & y_{N12} & \cdots & y_{N1K} \\ y_{N21} & y_{N22} & \cdots & y_{N2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{NT1} & y_{NT2} & \cdots & y_{NTK} \end{pmatrix} \end{array} \quad \begin{array}{c} time \\ \begin{pmatrix} 1 \\ 2 \\ \vdots \\ T \\ 1 \\ 2 \\ \vdots \\ \vdots \\ 1 \\ 2 \\ \vdots \\ T \end{pmatrix} \end{array} \quad \begin{array}{c} X_{it1} \quad X_{it2} \quad \cdots \quad X_{itC} \\ \begin{pmatrix} x_{111} & x_{112} & \cdots & x_{11C} \\ x_{121} & x_{122} & \cdots & x_{12C} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1T1} & x_{1T2} & \cdots & x_{1TC} \\ x_{211} & x_{212} & \cdots & x_{21C} \\ x_{221} & x_{222} & \cdots & x_{22C} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N11} & x_{N12} & \cdots & x_{N1C} \\ x_{N21} & x_{N22} & \cdots & x_{N2C} \\ \vdots & \vdots & \ddots & \vdots \\ x_{NT1} & x_{NT2} & \cdots & x_{NTC} \end{pmatrix} \end{array}$$

Figure 3.3: Multivariate longitudinal data structure

Let $Y_1 \sim Bernolli(\pi_1)$, $Y_2 \sim Bernolli(\pi_2) \cdots Y_K \sim Bernolli(\pi_K)$. Assuming the outcomes are obtained in the same time for same C covariates. We will collapse K outcomes into one outcome Z. Since Multinomial trial process is a simple generation of Bernolli trial processes, then the new variable Z has multinomial distribution with $J = 2^K$ possible outcomes, $Z_1, Z_2, ..Z_J$. Suppose each possible outcome can occur with probability $p_1, p_2, ..., p_J$, then the probability of Z_1 occurs m_1 times, Z_2 occurs m_2 times....., Z_J occurs m_J times is following the probability mass function:

$$f(m_1, m_2, ..., m_J) = \binom{N}{m_1 m_2 .... m_J} p_1^{m_1} p_2^{m_2} ..... p_J^{m_J} \tag{3.22}$$

where $\sum_{j=1}^{J} p_j = 1$, $\sum_{j=1}^{J} m_j = N$. Then, $Z = (Z_1, Z_2, ..., Z_J)^T$ has a multinomial distribution with $p = (p_1, p_2, ..., p_J)$ parameters.

3.5.1 Collapsing binary outcomes method

To collapse many binary vectors into one multinomial vector, we suggest to do transformation from a binary coding system to a decimal coding system. For example, if we have three binary

outcomes, then the new variable Z has 2^3 possible outcomes which are 0,1,2...7. The new variable Z has multinomial distribution with $p = (p_0, p_1, ..., p_7)$ parameters, $p_0 = \prod_{k=1}^{K} (1 - \pi_k)$. Let $y_1 = 1\,, y_2 = 2\,, y_3 = 1$ then $z = 1 \times 2^0 + 0 \times 2^1 + 1 \times 2^2 = 5$

For K binary outcomes, the new variable Z has 2^K nominal levels, $J = 2^K$. An example of the overall design of data after producing the new outcome Z is in the following figure:

$$
\begin{array}{c}
ID \\
\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 2 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ N \\ N \\ \vdots \\ N \end{pmatrix}
\end{array}
\quad
\begin{array}{ccc}
Y_{it1} & Y_{it2} & Y_{itk} \\
\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}
\end{array}
\quad
\begin{array}{c}
Z_{it} \\
\begin{pmatrix} 3 \\ 4 \\ \vdots \\ 1 \\ 7 \\ \vdots \\ \vdots \\ \vdots \\ 6 \\ 3 \\ \vdots \\ 5 \end{pmatrix}
\end{array}
\quad
\begin{array}{c}
time \\
\begin{pmatrix} 1 \\ 2 \\ \vdots \\ 1 \\ 2 \\ \vdots \\ \vdots \\ \vdots \\ 1 \\ 2 \\ \vdots \\ T \end{pmatrix}
\end{array}
\quad
\begin{array}{cccc}
X_{it1} & X_{it2} & \cdots & X_{itP} \\
\begin{pmatrix} x_{111} & x_{112} & \cdots & x_{11C} \\ x_{121} & x_{122} & \cdots & x_{12C} \\ \vdots & \vdots & \ddots & \vdots \\ x_{211} & x_{212} & \cdots & x_{21C} \\ x_{221} & x_{222} & \cdots & x_{22C} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N11} & x_{N12} & \cdots & x_{N1C} \\ x_{N21} & x_{N22} & \cdots & x_{N2C} \\ \vdots & \vdots & \ddots & \vdots \\ x_{NT1} & x_{NT2} & \cdots & x_{NTC} \end{pmatrix}
\end{array}
$$

Figure 3.4: Collapsing binary outcomes

We will analyze the new outcome Z instead of K binary vectors. Since the data is longitudinal, we consider a correlation among the new variable Z_{it}. We have to use a method for analyzing correlated nominal multinomial responses.

For correlated multinomial responses, there are many methods using random effects model and marginal models. For random effect models, Hedeker (2003) adopts a method in Bock's model. For marginal models, the GEE approach avoids specification of the distribution of the multinomial outcome by adopting a "working" correlation matrix. However, Touloumis et al. (2013) propose a GEE approach using a local odds ratio parameterization. In the multinomial responses, we choose one of the possible outcomes to be a reference level and estimate the parameter in term of $J - 1$ remaining levels compared to the reference level.

3.5.2 Nominal multinomial longitudinal response

Let $Z_{it} \in 0, 1, 2, .., J - 1$ be the nominal multinomial response for individual i, $i \in 1, 2, ..N$ at occasion t , $t \in 1, 2, ..T$. Fahrmeir et al. (1994) assume a multinomial generalized liner model for

marginal expected mean vector for individual i at occasion t:

$$g(E(Z_{it})) = X_{it}\,\beta \tag{3.23}$$

Touloumis et al. (2013) employ local odds ratio in GEE approach. They identify γ as marginalized local odds ratio after summarizing the occasions in contingency tables for all nominal responses levels at each time pair. They model the associations γ in the marginalized tables using association models (see Goodman (1985)) that helps to reduce dimension of γ. The contingency table of each pair of responses is $(J-1)\times(J-1)$ table in which $j\times j^{'}$ cell is the probability of responses outcomes at row j and column $j^{'}$. Touloumis et al. (2013) define the local odds ratio at cutpoint $(j, j^{'})$ by $\gamma_{jj'}$ using a special case of homogeneous Goodman row and column effects (RC) given by:

$$log\,\gamma_{jj'} = \phi(\mu_j - \mu_{j+1})(\mu_{j'} - \mu_{j'+1}) \qquad for\;\; j, j^{'} = 1, 2, \ldots (J-1) \tag{3.24}$$

Then, $log\,\gamma_{jj'}$ decomposes the local odds ratio into two parts: the parameter ϕ that expresses the strength of the association and the scores μ_j for J response categories. We will have $J-2$ non redundant scores that are estimated as parameters. Then, the association models can capture the correlation patterns between correlated multinomial responses.

The GEE method in (3.25) for nominal multinomial responses has a link vector 'g' which is the baseline-category logit model. The GEE estimator $\hat{\beta}$ is the solution of the equation:

$$U(\beta, \gamma) = \frac{1}{N}\sum_{i=1}^{N} D_i V_i^{-1}(Z_i - E(Z_{it})) = 0 \tag{3.25}$$

where $D_i = \dfrac{\partial E(Z_{it})}{\partial \beta_i}$ and $V_i = V_i(\beta, \gamma)$ is a $T_i(J-1)\times T_i(J-1)$ "weight" matrix that is the working covariance matrix. $\hat{B}_G$ is the solution of $U(\beta, \gamma) = 0$ defined by Liang and Zeger (1986), where $\hat{\gamma}$ is a $\sqrt{N}$ consistent estimator of γ given β under mild regularly conditions. Liang and Zeger (1986) proved that $\hat{\beta}_G$ is consistent and $\sqrt{N}(\hat{\beta}_G - \beta)$ asymptotically follows multivariate normal distribution with mean zero and covariance matrix :

$$\Sigma_G = \lim_{N\to 0} N\,\Sigma_0^{-1}\,\Sigma_1\,\Sigma_0^{-1}, \tag{3.26}$$

where $\Sigma_0 = \sum_i D_i^{'} V_i^{-1} D_i \quad , \Sigma_1 = \sum_i D_i^{'} V_i^{-1}\, Cov(Z_i)\, V_i^{-1}\, D_i$, and $Cov(Z_i)$ is the $T_i(J-1)\times T_i(J-1)$ true covariance matrix for subject i. $Cov(Z_i)$ consist of :

$$Cov(Z_{itj}, Z_{it'j}|x_i) = \begin{cases} E(Z_{itj})(1 - E(Z_{itj})) & if\; t = t^{'} and\, j = j^{'} \\ -E(Z_{itj})E(Z_{itj'}) & if\; t = t^{'} and\, j \neq j^{'} \\ E(Z_{itj}Z_{it'j'})\; E(Z_{itj})E(Z_{it'j'}) & if\; O.W \end{cases}$$

where $E(Z_{itj}Z_{it'j'}) = P_r(Z_{itj} = Z_{it'j'} = 1|x_i) = P_r(Z_{it} = j, Z_{it'} = j^{'}|x_i)$.

The covariance Σ_G is robust and consistence can be estimated by ignoring the limit and use $(\hat{\beta_G}, \hat{\gamma})$ instead of (β_G, γ) and $(Y_i - E(\hat{Z_i}))(Y_i - E(\hat{Z_i}))^{'}$ instead of $Cov(Z_i)$. Touloumis et al. (2013) specify local odds ratio under generalized RC model satisfy

$$log\,\gamma_{tjt'j'} = \phi_{tt'}(\mu_{tj}^{tt^{'}} - \mu_{t(j+1)}^{tt^{'}})(\mu_{t'j}^{tt^{'}} - \mu_{t'(j+1)}^{tt^{'}}) \tag{3.27}$$

and they recommended two types of association for nominal multinomial correlated responses.

1) The time exchangeable structure, $log\,\gamma_{tjt'j'} = \phi(\mu_j - \mu_{j+1})(\mu_{j'} - \mu_{j'+1})$. It assumes a fixed odds ratio over all category outpoints $(j, j^{'})$ and there is no any time dependency between the response levels.

2) The RC structure, $log\,\gamma_{tjt'j'} = \phi_{tt'}(\mu_j^{tt^{'}} - \mu_{(j+1)}^{tt^{'}})(\mu_{j^{'}}^{tt^{'}} - \mu_{(j^{'}+1)}^{tt^{'}})$. This structure assumes more parameters for the odds ratio and it has fixed score parameters for rows and columns for any given pair $(t, t^{'})$ in (3.27). It allows for a time dependency between the correlated responses. Finally, using the local odds ratios in the GEE approach that is proposed by Touloumis et al. (2013) can be an appropriate method to analyze the nominal multinomial variable Z_{it} that results from collapsing many correlated binary outcomes.

After estimating the parameters β_G and γ of the nominal multinomial responses Z_{it}, we have to move to the "decoding" step. The decoding step is to transform each category in Z_{itj} response to its original definition. For example, the parameter estimation $\hat{\beta}_1$ for category $z = 3$ is for the original code "011". Then, $\hat{\beta}_1$ expresses the estimated parameter of the logit probabilities of accruing outcome 1 and outcome 2 simultaneously. In conclusion, using the GEE approach for nominal multinomial correlated responses could be helpful to analyze the joint distribution of many binary longitudinal outcomes.

The main difference between the GEE approach in Section 3 and this proposed approach is working in the the joint distribution of many outcomes. In the traditional method, we create a longer vector of responses for each individual, which are measured at different occasions and different outcomes. Then, we run the GEE approach for the marginal model with an unstructured covariance matrix. Specifying unstructured covariance guarantees that the correlation parameters or the odds ratios are estimated different for each occasion and outcome. The difficulty of this method is the total number of unknown parameters is extremely large $\frac{(TJ)(TJ-1)}{2}$.

O'Brien and Fitzmaurice (2004) present two methods for a parsimonious within subject association structure as we explained in section 3. Instead of creating a longer response vector of each individual and taking care of the covariance among the outcomes and occasions, the proposed method combines many outcomes into one outcome for each individual at each occasion and analyzes the data longitudinally using GEE approach for nominal multinomial correlated responses. This new method will cost more estimated parameters, but it has deeper potential for the joint distribution of the binary outcomes. Finally, accounting for all the combinations of binary outcomes and analyzing the longitudinal trends of these combinations over the time could be helpful in many biostatistics problems.

3.6 Illustration With Interpretation

In this section, we will explain three models to analyze multivariate binary longitudinal data. First, we assume there are two binary responses Y_{1t} and Y_{2t} that are obtained over three occasions, $t = 1, 2, 3$. There is an simple illustration of the three models are in the following subsections and we used them in the simulation part.

3.6.1 Model 1 [Regression modeling for odds ratio]

To illustrate the strategy: let there be Y_i, X_{i1} , X_{i2} *and* X_{i3} vectors with length $(n \times K \times T) =$ 3000:

$Y_i = [Y_{i11}, Y_{i21}, Y_{i12}, Y_{i22}, Y_{i13}, Y_{i23}]^T$

$X_i = [X_{i1}, X_{i1}, X_{i2}, X_{i2}, X_{i3}, X_{i3}]^T$

Assuming the covariates are the same for different outcomes at the same time, a logistic regression model is used to relate $\mu_{itk} = E(Y_{itk}) = Pr(Y_{itk} = 1)$ to X_{i1}, X_{i2} *and* X_{i3}, where k=1,2, is given by

$$logit(\mu_{itk}) = \beta_0 + \beta_1 X_i$$

.

The working covariance matrix of Y_{itk} is given by

$$V_i = A_i^{1/2} \, R(\gamma) \, A_i^{1/2} \tag{3.28}$$

where A_i is a diagonal matrix with $var(Y_{itk}) = \phi\, v(\mu_{itk})$ along the diagonal and $A^{1/2}$ is diagonal matrix with $\sqrt{\phi\, v(\mu_{itk})}$ along the diagonal. $R(\gamma)$ is the working correlation matrix that is a function of γ, where γ represents a vector of within subject association parameters.

$$R = \begin{array}{c} \\ Y_{t_1k_1} \\ Y_{t_1k_2} \\ Y_{t_2k_1} \\ Y_{t_2k_2} \\ Y_{t_3k_1} \\ Y_{t_3k_2} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{t_1k_1} & Y_{t_1k_2} & Y_{t_2k_1} & Y_{t_2k_2} & Y_{t_3k_1} & Y_{t_3k_2} \end{array} \\ \begin{pmatrix} 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 \\ - & 0 & \gamma_6 & \gamma_7 & \gamma_8 & \gamma_9 \\ - & - & 0 & \gamma_{10} & \gamma_{11} & \gamma_{12} \\ - & - & - & 0 & \gamma_{13} & \gamma_{14} \\ - & - & - & - & 0 & \gamma_{15} \\ - & - & - & - & - & 0 \end{pmatrix} \end{array}$$

The main idea in model 1 is modeling the working correlation matrix to capture the different types of the correlation in the multivariate longitudinal structure. The GEE estimator of β_0, β_1 and β_3 is obtained by solving the equation (3.5). Then, The iterative procedure is used to estimate β and γ in model 1. The basic assumption in GEE model is the independence between the observations.

3.6.2 Model 2 [Kronecker product]

Let $Y_{i1} = [Y_{i11}, Y_{i12}, Y_{i13}]^T$, $Y_{i2} = [Y_{i21}, Y_{i22}, Y_{i23}]^T$, $X_i = [X_{i1}, X_{i2}, X_{i3}]^T$. We used a Kronecker product method in subsection 3.2 and ran the following logistic regression model for the same X_{i1} *and* X_{i2} covariates:

$$logit(\mu_i) = \beta_{01} + \beta_{11} X_i + \beta_{02} + \beta_{12} X_i$$

where $\mu_i = E(Y_i) = Pr(Y_i = 1)$, (β_{01}, β_{11}) are the parameters of outcome Y_1 and (β_{02}, β_{12}) are the parameter estimations of outcome Y_2. Using the GEE approach in model 2 has the same steps that in model 1 but for different covariate matrix. The working correlation matrix using the SAS macro that are in Shelton et al. (2004) has the following structure:

$$R = \begin{array}{c} \\ Y_{t_1k_1} \\ Y_{t_2k_1} \\ Y_{t_3k_1} \\ Y_{t_1k_2} \\ Y_{t_2k_2} \\ Y_{t_3k_2} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{t_1k_1} & Y_{t_2k_1} & Y_{t_3k_1} & Y_{t_1k_2} & Y_{t_2k_2} & Y_{t_3k_2} \end{array} \\ \begin{pmatrix} 0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 \\ - & 0 & \gamma_6 & \gamma_7 & \gamma_8 & \gamma_9 \\ - & - & 0 & \gamma_{10} & \gamma_{11} & \gamma_{12} \\ - & - & - & 0 & \gamma_{13} & \gamma_{14} \\ - & - & - & - & 0 & \gamma_{15} \\ - & - & - & - & - & 0 \end{pmatrix} \end{array}$$

3.6.3 Model 3 [The proposed method]

In this model, we have to collapse the two outcomes into one outcome for each occasion. Let $Y_{i1} = [Y_{i11}, Y_{i12}, Y_{i13}]^T$, $Y_{i2} = [Y_{i21}, Y_{i22}, Y_{i23}]^T$, $X_i = [X_{i1}, X_{i2}, X_{i3}]^T$.

The new variable $Z_{it} = 2^0 \times Y_{i1} + 2^1 \times Y_{i2} = 2Y_{i2} + Y_{i1}$. The new variable $Z \in 0, 1, 2, 3$.

$$Z_{it} = \begin{cases} 0 & \text{when} \quad (Y_{i1} = 0) \ \& \ (Y_{i2} = 0) \\ 1 & \text{when} \quad (Y_{i1} = 0) \ \& \ (Y_{i2} = 1) \\ 2 & \text{when} \quad (Y_{i1} = 1) \ \& \ (Y_{i2} = 0) \\ 3 & \text{when} \quad (Y_{i1} = 1) \ \& \ (Y_{i2} = 1) \end{cases}$$

The third model is using Touloumis et al. (2013) method, assuming no correlation between the levels (j=1,2,3) of the multinomial responses at the same time, is:

$$log \frac{Pr(Z_{it} = j|X_{it})}{Pr(Z_{it} = 0|X_{it})} = \beta_{01} + \beta_{11}X_i + \beta_{02} + \beta_{12}X_i + \beta_{03} + \beta_{13}X_i$$

The baseline-category logit link function is used in the third model. The GEE estimator of $\hat{\beta}$ is a solution of (3.25). The local odds ratio that are in the $Cov(Z_i)$ under RC model is

$$log\, \gamma_{tjt'j'} = \phi_{tt'}(\mu_{tj}^{tt'} - \mu_{t(j+1)}^{tt'})(\mu_{t'j}^{tt'} - \mu_{t'(j+1)}^{tt'})$$

.

using the nominal multinoamial GEE model will generate a group of parameter estimations for 1,2 and 3 levels of the variable Z_{it}.

3.7 Simulation

We ran a simulation study to compare the parameter estimations for three methods. Assuming the distribution of the covariate X follows the normal as following:

$$X \ at \ time1 \sim Normal(1, 0.02)$$

$$X \ at \ time2 \sim Normal(2, 0.05)$$

$$X \ at \ time3 \sim Normal(3, 0.09)$$

The first method follows the first model in subsection 3.6.1 (modeling the odds ratio). The second model is in subsection 3.6.2 (Kronecker product) and the third model used is for the proposed

method discussed in 3.6.3. The three models that are used in the simulation procedures have been already explained in the last illustration part.

We did the simulation for GEE marginal regression models based on specified correlation structures. The method is a generate of correlated binary data using generalized liner mixed model using bridge distribution for the random effects term. The multivariate binary longitudinal data structure consists of six vectors of responses, Y_1 and Y_2 at time 1, 2 and 3. There are 15 correlation parameters in matrix R, which are needed to enable generate six vectors of binary outcomes. The three components building up the matrix R are the inter-outcome (υ), the intra-outcome (α) and the cross association (τ).

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ Y_{13} \\ Y_{23} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{21} & Y_{12} & Y_{22} & Y_{13} & Y_{23} \end{array} \\ \begin{pmatrix} 0 & \upsilon & \alpha & \tau & \alpha & \tau \\ - & 0 & \tau & \alpha & \tau & \alpha \\ - & - & 0 & \upsilon & \alpha & \tau \\ - & - & - & 0 & \tau & \alpha \\ - & - & - & - & 0 & \upsilon \\ - & - & - & - & - & 0 \end{pmatrix} \end{array}$$

where Y_{kt} for $k = 1, 2$ and (outcomes), $t = 1, 2, 3$ (occasions). The $\upsilon^{'}s$ represents the correlation between the outcomes in the same occasion. $\alpha^{'}s$ represents the correlation between the same outcomes at different occasions. $\tau^{'}s$ represents the correlation for different outcomes that measured at different occasions. We conducted five scenarios under the assumption of exchangeability for each correlation type υ, ϕ *and*, τ and $\tau = 0$ for all scenarios. The following table shows the values for each correlation structure and scenario in R matrix:

Table 3.1: The correlation scenarios of simulation study

	$\upsilon = 0.00$	$\upsilon = 0.50$	$\upsilon = 0.90$
$\alpha = 0.00$	scenario1	scenario2	scenario3
$\alpha = 0.50$	scenario4	-	-
$\alpha = 0.90$	scenario5	-	-

It is known that the parameter estimations under the mixed model have different interpretation than the marginal model for the binary responses because the marginal model integration over the random effect do not keep the logistic form. Using bridge distribution, will match the the logistic

shape of the conditional and marginal the binary response model. Wang and Louis (2003) introduce a bridge distribution for the random effect such as the marginal shape would remains the same as the conditional shape. First, Let Y_{ijk} be the binary response that measured at time j, $j = 1, 2, 3$ for observation i, $i = 1, 2, 3...N$ and for outcome $k = 1, 2$. We generated the correlated binary data using bridge method where the marginal model based on model 1 is:

$$logit(E(Y_{ijk}|X_{ij})) = logit(P_{ijk}) = 0.35 - 0.15X_{ij} \tag{3.29}$$

The Y_{ijk} for subject i are assumed to be independent Bernoulli random variables, $Y_{ijk} \sim Ber(P_{ijk})$. We did the same previous simulation steps but based on model 2 for vector of intercepts (0.2,0.1) and beta (0.05,-0.15) for outcome 1 and 2 respectively.

The goal of the simulation part was to investigate the differences between the parameter estimations of the three models under the six scenarios by control the correlation over the outcomes and the occasions. Then, we determined the simulation results and included them in table 3.2 and table 3.3 based on model 1 and table 3.4 and table 3.5 based on model 2. We aggregated the parameters for each covariate to investigate how the parameters changed over the three models.

Starting from the results table 3.2 in scenario 1, we find the mean=0.3491 from model 1, while the means for the second model are 0.3533 for the logit of $Pr(Y_{i1} = 1)$ and 0.3451 for the logit of $Pr(Y_{i2} = 1)$. The third model has three parameter estimations, β_{01} for the logit of the probability of occurring the second outcome $Pr(Y_{i1} = 0, Y_{i2} = 1)$ which is 0.3438 and β_{02} for the first outcome is 0.3504. We have to mention that the coding 1 expresses the logit of $Pr(Y_{i1} = 0, Y_{i2} = 1)$, then it is for the second outcome and code 2 for outcome 1. The parameter $\beta_{03} = 0.6986$ is estimation of the logit of $Pr(Y_{i1} = 1, Y_{i2} = 1)$, the joint case. In scenarios 2 and 3, the correlation increased among the outcomes ($\upsilon = 0.50, 0.90$) respectively. The results of the estimated joint parameters β_{03} have less deviations from the null and the parameters β_{01} and β_{02} of model 3 have more variations from the null. In scenario 4 and 5, the correlation among the occasions is increased ($\alpha = 0.50, 0.90$). Comparing with scenario 1 when $\alpha = 0.00$, the estimated joint parameters (β_{13}) approximately did not change over scenario 4 and 5. The results in table 3.3 of the covariate parameter estimation when the true value is -0.15. It is agreed with the intercepts' results. We conclude the simulation results based on model 1 does not have big variation over the three models.

Table 3.2: The intercept estimations based on model 1

	senario 1		senario2		senario3		senario4		senario5	
β_0	mean	std	mean	std	mean	std	mean	std	mean	std
$\beta_{0_{model1}}$	0.3491	0.0902	0.3500	0.1277	0.3457	0.1238	0.3507	0.0889	0.3400	0.0700
$\beta_{01_{model2}}$	0.3533	0.1388	0.3558	0.1544	0.3468	0.1289	0.3596	0.1161	0.3468	0.1044
$\beta_{02_{model2}}$	0.3451	0.1316	0.3448	0.1395	0.3456	0.1281	0.3443	0.1273	0.3407	0.0978
$\beta_{01_{model3}}$	0.3438	0.1992	-0.8651	0.2316	-2.6756	0.4452	0.3472	0.1692	0.3552	0.2017
$\beta_{02_{model3}}$	0.3504	0.2033	-0.8427	0.2437	-2.6605	0.4228	0.3571	0.1674	0.3562	0.2350
$\beta_{03_{model3}}$	0.6986	0.1845	0.4688	0.1718	0.3663	0.1322	0.7054	0.1798	0.6980	0.2590

Table 3.3: The covariate parameter estimations based on model 1

	senario 1		senario2		senario3		senario4		senario5	
β_1	mean	std	mean	std	mean	std	mean	std	mean	std
$\beta_{1_{model1}}$	-0.1483	0.0421	-0.1506	0.0580	-0.1500	0.0580	-0.1511	0.0324	-0.1446	0.0170
$\beta_{11_{model2}}$	-0.1503	0.0633	-0.1537	0.0689	-0.1512	0.0607	-0.1550	0.0439	-0.1465	0.0250
$\beta_{12_{model2}}$	-0.1464	0.0620	-0.1477	0.0642	-0.1493	0.0599	-0.1484	0.0439	-0.1464	0.0252
$\beta_{11_{model3}}$	-0.1469	0.0951	-0.0972	0.1120	-0.0651	0.1988	-0.1499	0.0685	-0.1537	0.0723
$\beta_{12_{model3}}$	-0.1497	0.0887	-0.1093	0.1098	-0.0820	0.1957	-0.1545	0.0670	-0.1544	0.0937
$\beta_{13_{model3}}$	-0.2967	0.0859	-0.2017	0.0778	-0.1590	0.0623	-0.3042	0.0671	-0.2971	0.1079

Generating the correlated binary data based on model 2 clarify many points. First, in table 3.4 the results of model 1 are not on the average of model 2's results for all the scenarios. Second, the results of model 3 for outcomes 1 and 2 (β_{02} and β_{01}) respectively are approximately agreed with model 2 results in scenario 1 and 4. Increasing the correlation over the outcomes in scenario 2 and 3 reveals more variation from the null hypothesis for the parameters (β_{02} and β_{01}) of model 3 and less variations in the joint parameter β_{03}. In table 3.5 for the covariate parameter estimation when the true values are -0.15 and 0.05 respectively it shows the same analytic points in table 3.4.

Table 3.4: The intercept estimations based on model 2

	senario 1		senario2		senario3		senario4		senario5	
β_0	mean	std	mean	std	mean	std	mean	std	mean	std
$\beta_{0_{model1}}$	0.6069	0.1049	0.6679	0.0892	0.6800	0.0759	0.4064	0.0981	0.2023	0.0950
$\beta_{01_{model2}}$	0.2087	0.1542	0.2021	0.1426	0.2137	0.1374	0.2092	0.1019	0.1962	0.0994
$\beta_{02_{model2}}$	0.1067	0.1292	0.1020	0.1323	0.1142	0.1376	0.1060	0.1259	0.1011	0.0965
$\beta_{01_{model3}}$	0.0809	0.2095	-1.0849	0.2964	-2.1438	0.5897	0.2104	0.1455	0.2055	0.2273
$\beta_{02_{model3}}$	0.1872	0.2040	-0.8813	0.1975	-1.7419	0.2453	0.1071	0.1933	0.1665	1.0218
$\beta_{03_{model3}}$	0.3135	0.2032	0.2060	0.1590	0.1864	0.1458	0.3168	0.1609	0.3566	0.7598

Table 3.5: The covariate parameter estimations based on model 2

	senario 1		senario2		senario3		senario4		senario5	
β_1	mean	std	mean	std	mean	std	mean	std	mean	std
$\beta_{1_{model1}}$	-0.2782	0.0504	-0.3085	0.0372	-0.3123	0.0287	-0.1784	0.0405	-0.0779	0.0264
$\beta_{11_{model2}}$	0.0468	0.0719	0.0481	0.0661	0.0430	0.0629	0.0452	0.0418	0.0508	0.0227
$\beta_{12_{model2}}$	-0.1549	0.0604	-0.1523	0.0640	-0.1584	0.0617	-0.1532	0.0462	-0.1499	0.0251
$\beta_{11_{model3}}$	-0.1447	0.1015	-0.2532	0.1483	-0.7718	0.3610	-0.1527	0.0797	-0.1825	0.4539
$\beta_{12_{model3}}$	0.0554	0.0951	0.1479	0.0839	0.3069	0.1057	0.0454	0.0626	0.0427	0.1268
$\beta_{13_{model3}}$	-0.1071	0.0956	-0.0708	0.0751	-0.0656	0.0667	-0.1086	0.0614	-0.1338	0.4606

In conclusion, the case of multivariate longitudinal data potentially has a complicated correlation. The correlation among the responses come from the repeated measurements and the outcomes that are measured from the same observation. The parameter estimations over the three models changed from each of the three models due to the strength of the correlation. We could understand it is not meaningful to use model 1 unless the means of the longitudinal outcomes are indeed equal and this is likely rare. Also, the results of model 3 vary from those of model 2 when the correlation over the outcomes increases. Starting from model 1 to model 3, most of the parameter estimations are likely to vary from the null hypothesis. Choosing the appropriate model depends on the application, and the question of the research.

3.8 Application

The Florida Dental Care Study (FDCS) is a longitudinal study of oral health and dental service utilization. It is conducted in four centers in North Florida in the United States. The sampling

methodology is in Gilbert et al. (1997). The sample size is 873 subjects who had baseline interview exams and four clinical examination interviews at 6, 12, 18 and 24 months past baseline visit. By the end of 24 months, 87.5% of 873 remained in the study. 3.3% refused to participate, 1.1% unable to participate due to medical issues, 3.3% were deceased and 4% were dropout. The issue of bias was discussed in Gilbert et al. (1998) by comparing the characteristics of the patients who remained in the study at 24 months, with those who did not for any reason. Also, there are three binary outcomes that were measured over four time intervals (0-6, 6-12, 12-18, 18-24 months). In this application, a subset of five covariates were chosen in table 3.6. The three binary outcomes measured in this study are 'problem oriented visit', 'dental cleaning' and 'dental check up'. At the end of each interval, each subject was asked whether he had visited a dentist within the past 6 months 'problem oriented visit' and whether this dental visit was for a 'dental cleaning' or 'check up'. The three binary outcomes are coded as (0='No',1='Yes').

Table 3.6: Covariates variables

Covariate	Definition
IRA	(1) The subject goes to a dentist regularly or occasionally whether or not has a problem, (0) The subject did go to a dental check-up once a year or more often in the previous 5 years.
Gender	(1) Female, (0) Male.
Cavit	(1) The subject reported having cavities (tooth decay) in the previous 6 months, (0) if not.
Loose	(l) The subject had a loose tooth, (0)had not
Able	(1) The subject able to pay unexpected US$ 500 dental bill, but with difficulty, (0) not able to pay

First, we consider the following three models:

Model1: $logit(\mu_i) = \beta_0 + \beta_1 X_{ira} + \beta_2 X_{gender} + \beta_3 X_{cavit} + \beta_4 X_{loose} + \beta_5 X_{able}$

Model2: $logit(\mu_i) = \sum_{k=1}^{3} (\beta_{k0} + \beta_{k1} X_{ira} + \beta_{k2} X_{gender} + \beta_{k3} X_{cavit} + \beta_{k4} X_{loose} + \beta_{k5} X_{able})$

Model3: $logit(\mu_i) = \sum_{j=1}^{7} (\beta_{j0} + \beta_{j1} X_{ira} + \beta_{j2} X_{gender} + \beta_{j3} X_{cavit} + \beta_{j4} X_{loose} + \beta_{j5} X_{able})$

where

$$j = \begin{cases} 1 & \text{when} \quad (Y_{i1} = 0)\ \&\ (Y_{i2} = 0)\ \&\ (Y_{i3} = 1) \\ 2 & \text{when} \quad (Y_{i1} = 0)\ \&\ (Y_{i2} = 1)\ \&\ (Y_{i3} = 0) \\ 3 & \text{when} \quad (Y_{i1} = 0)\ \&\ (Y_{i2} = 1)\ \&\ (Y_{i3} = 1) \\ 4 & \text{when} \quad (Y_{i1} = 1)\ \&\ (Y_{i2} = 0)\ \&\ (Y_{i3} = 0) \\ 5 & \text{when} \quad (Y_{i1} = 1)\ \&\ (Y_{i2} = 0)\ \&\ (Y_{i3} = 1) \\ 6 & \text{when} \quad (Y_{i1} = 1)\ \&\ (Y_{i2} = 1)\ \&\ (Y_{i3} = 0) \\ 7 & \text{when} \quad (Y_{i1} = 1)\ \&\ (Y_{i2} = 1)\ \&\ (Y_{i3} = 1) \end{cases}$$

The reference category is 0 when $(Y_{i1} = 0)\,\&\,(Y_{i2} = 0)\,\&\,(Y_{i3} = 0)$. Before running the three models, we measured the association between the binary responses using the odds ratio instead of the correlation. The odds ratio and the 95% confidence intervals are in table 3.7. The association parameters are larger than one and significant for all of the odds ratios because the 95% confidence intervals exclude the value one. Then, we expect a large differences between the three models. The strongest association is between 'cleaning' and 'check up' outcomes.

Table 3.7: Estimated odds ratio

Odds Ratio	Estimate	$95\%CI$
$OR(Y_1, Y_2)$	1.976815	(0.5122,0.8508)
$OR(Y_1, Y_3)$	2.080624	(0.5508,0.9145)
$OR(Y_2, Y_3)$	42.14729	(3.4971,3.9851)

We ran the three models with no constraints in the log odds ratio, then we got 66 unique estimated odds ratio ($\#\alpha = 12, \#\upsilon = 18, \#\tau = 36$) in models 1 and 2. The results of the three models are in tables 3.8, 3.9 and 3.10. The covariates Cavit and Able are significant in model 1 while Gender and IRA are close and Loose is not. In model 2, the covariates are estimated for each outcome separately. The Gender covariate still not significant for all. However, The Loose parameter estimation is changed completely from model 1 to model 2 to be significant for all. This change due to separate the effect of the covariate for each outcome, when they are highly correlated. In model 3, we find Gender parameter estimation still not significant for all. Loose parameter estimation becomes significant only for outcome $Y_1, Y_{1\&2}, Y_{1\&3}$ *and* $Y_{1,2\&3}$ which means the high association between the 'problem visit' outcome and the others hides the effect of Loose covariate in the outcome in model 1. To get a closer look at the parameters' differences in the three models, we aggregate the parameter estimations of all models for each covariate in figure 3.5.

Table 3.8: Model 1 results

Parm	**Estimate**	**std**	**Z**	**ProbZ**
Intercept	-1.1302	0.0651	-17.36	$< .0001$
IRA	0.2379	0.1273	1.87	0.0616
Gender	-0.1649	0.0901	-1.83	0.0673
Cavit	0.6235	0.0795	7.84	$< .0001$
Loose	0.0818	0.081	1.01	0.3126
Able	-0.2715	0.086	-3.16	0.0016

Table 3.9: Model 2 results

Parm	**Estimate**	**std**	**Z**	**ProbZ**
Intercept1	-1.306	0.0827	-15.79	$< .0001$
IRA_y1	0.1447	0.1768	0.82	0.4129
Gender_y1	-0.1904	0.116	-1.64	0.1006
Cavit_y1	1.0137	0.1059	9.58	$< .0001$
Loose_y1	0.5367	0.1191	4.51	$< .0001$
Able_y1	-0.1008	0.1077	-0.94	0.3491
Intercept2	-1.1788	0.0979	-12.04	$< .0001$
IRA_y2	0.4656	0.1836	2.54	0.0112
Gender_y2	-0.1652	0.1323	-1.25	0.2118
Cavit_y2	0.3919	0.1113	3.52	0.0004
Loose_y2	-0.4316	0.1245	-3.47	0.0005
Able_y2	-0.4329	0.1257	-3.44	0.0006
Intercept3	-0.8009	0.0945	-8.48	$< .0001$
IRA_y3	0.328	0.1898	1.73	0.0839
Gender_y3	-0.1609	0.1277	-1.26	0.2075
Cavit_y3	0.2894	0.1065	2.72	0.0066
Loose_y3	-0.4101	0.117	-3.51	0.0005
Able_y3	-0.5959	0.1245	-4.79	$< .0001$

Table 3.10: Model 3 results

Parm	**Estimate**	**std**	**Z**	**ProbZ**
Intercept 1	2.31644	0.15412	15.0297	$< 2e^{-16}$
IRA_1	-0.44468	0.28851	-1.5413	0.12325
Gender_1	0.25399	0.20448	1.2421	0.2142
Cavit_1	-1.59776	0.16413	-9.7347	$< 2e^{-16}$
Loose_1	0.30088	0.22217	1.3543	0.17565
Able_1	0.56503	0.19391	2.9139	0.00357
Intercept 2	0.2618	0.20957	1.2493	0.21157
IRA_2	-0.31291	0.41567	-0.7528	0.45157
Gender_2	0.32744	0.26896	1.2174	0.22343
Cavit_2	-2.22835	0.32981	-6.7565	$< 2e^{-16}$
Loose_2	0.40086	0.27408	1.4626	0.14359
Able_2	0.08664	0.2673	0.3241	0.74585
Intercept 3	-1.64295	0.3346	-4.9101	$< 2e^{-16}$
IRA_3	0.34044	0.53157	0.6404	0.52188
Gender_3	0.09152	0.38695	0.2365	0.81304
Cavit_3	-1.67004	0.46435	-3.5965	0.00032
Loose_3	0.4977	0.41352	1.2036	0.22876
Able_3	0.89498	0.36628	2.4434	0.01455
Intercept 4	1.08689	0.14938	7.2758	$< 2e^{-16}$
IRA_4	0.02381	0.3449	0.069	0.94496
Gender_4	0.06488	0.21685	0.2992	0.76478
Cavit_4	-2.02666	0.21162	-9.577	$< 2e^{-16}$
Loose_4	-0.51225	0.25708	-1.9926	0.0463
Able_4	-0.07702	0.20244	-0.3805	0.7036
Intercept 5	0.7074	0.17148	4.1254	$< 4e^{-5}$
IRA_5	-0.22746	0.32465	-0.7006	0.48353
Gender_5	0.09493	0.2191	0.4333	0.66481
Cavit_5	-0.92957	0.19173	-4.8484	$< 2e^{-16}$
Loose_5	0.98945	0.22902	4.3204	$< 2e^{-05}$
Able_5	0.54333	0.20724	2.6218	0.00875
Intercept 6	-0.26746	0.214	-1.2498	0.21136
IRA_6	-0.19855	0.43232	-0.4593	0.64604
Gender_6	-0.20593	0.29128	-0.707	0.47957
Cavit_6	-0.77173	0.26226	-2.9426	0.00325
Loose_6	0.68872	0.29773	2.3133	0.02071
Able_6	-0.06312	0.26756	-0.2359	0.81351
Intercept 7	-1.41215	0.25849	-5.463	$< 2e^{-16}$
IRA_7	0.10559	0.46596	0.2266	0.82072
Gender_7	0.1786	0.32314	0.5527	0.58046
Cavit_7	-0.83001	0.33215	-2.4989	0.01246
Loose_7	0.94458	0.35612	2.6524	0.00799
Able_7	0.53696	0.30734	1.7471	0.08062

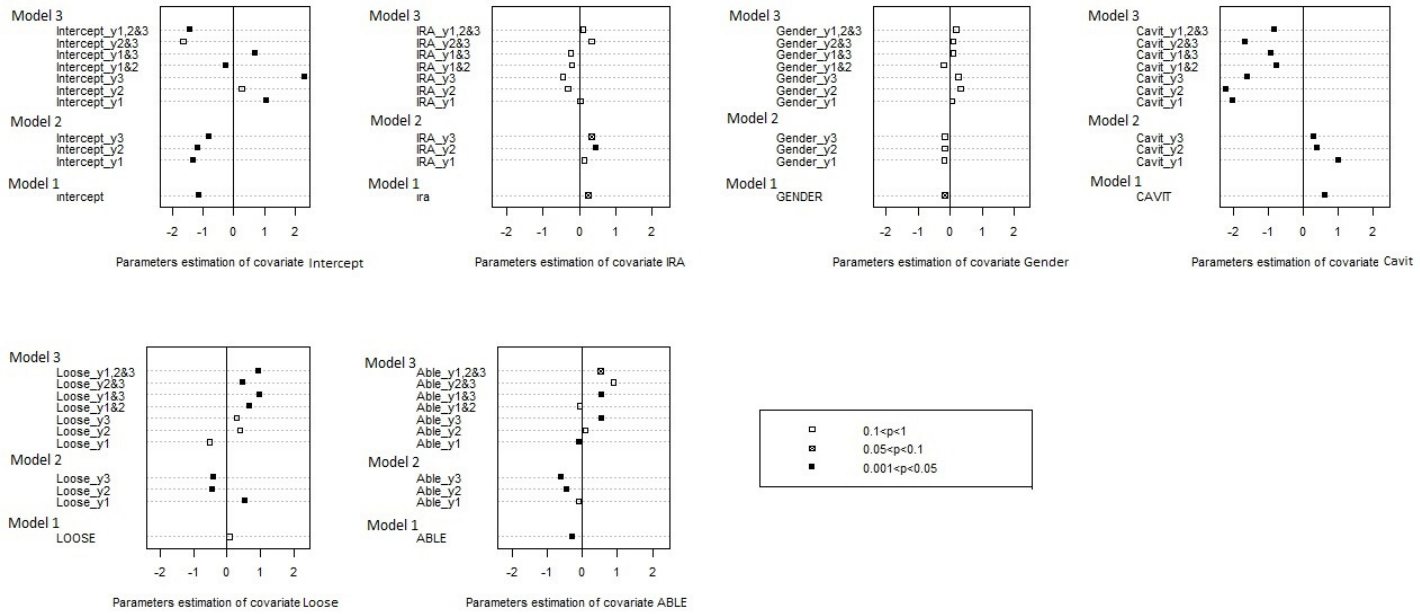

Figure 3.5: The parameter estimations over the three models

The dotplot in figure 3.5 indicates many points. First, the variation of the parameter estimation between model 1, model 2 and model 3 are relatively. The variation in IRA and Gender parameter estimations over the three models is limited due to insignificant results. IRA parameter estimation is close to being significant in model 1, then model 2 explains the source of the effect which is the second outcome. The third model clarifies the effect after breaking up the responses for each joint case and finding that there is no significant estimations. In a contrary case, the parameter estimation of Able covariate starts significant in model 1 and 2 expect for the first outcome, then model 3 illustrates some significant parameters related to the significant association between outcomes 1 and 3. The Loose variable begins as not significant in model 1 but it is significant in model 2 for all outcomes while model 3 clarifies the sources of effect which is the associations between the three outcomes. The case of the intercept is clearly starting significant in model 1 and model 2, but model 3 explains the cause of the significance which are outcomes 1 and 3. The parameter of Cavit covariate shows how the third model changes the sign of the parameter to negative after using the coding method.

In summary, The FSCS study of three outcomes 'problem visit', 'cleaning' and 'check up' and four intervals is a good example of highly correlated responses that are used to investigate the differences of the parameter estimation of three models. Over this example, we see that the effect of each cavriate on the responses varies over the three models. The most significant covariates over the three models are Cavity, Loose tooth, and the ability to pay an unexpected US$ 500 dental bill, but with difficulty.

3.9 Discussion and Future work

In this article, we review two existing methods for analyzing binary multivariate longitudinal data and proposed a third method. In most longitudinal experiments, there are more than one outcome are obtained from the subjects over many occasions. Usually, the researcher is interested in investigating the effect of many independent variables on theses outcomes. We present three different methods to be beneficial in this case. The aim of this article is to give a more general overview of analyzing binary multivariate longitudinal data in case of expecting a correlation among the outcomes other than the correlation among the repeated measurements.

If the researcher is interested in estimating one group the estimated effects of covariates for all outcomes and accounting of the multivariate structure, then the first method is the best. Also, if the researcher is eager to separate the effect of the covariate on the outcomes, then the second method is helpful. However, the joint analysis of all of the outcomes in the third method is beneficial to investigate the relationship between the correlated outcomes and how the covariate are effecting on them. The third method consists of encoding and decoding stages. The encoding stage is a step of defining each case of relationships between the outcomes into simple number after converting many outcomes in one outcomes . It is an idea of store of all of the possible combinations of the outcomes into one outcomes. We converted three binary outcomes into one multinomial outcome. The decoding stage is a step of returning each code to its original definition.

Then, as a result, the joint analysis in method 3 is limited to a small number of outcomes because the number of parameters of estimation is rapidly increasing when the number of outcomes is increased. It would be beneficial for future work to improve the third method in order to reduce the dimension of the parameters estimations by adding a filtering stage between the encoding and decoding steps. We presented a simulation part for many scenarios and application. The FCSD example is applied to the three methods and we saw how were the parameter estimations of the three methods are changed over many covariates. Each method explains more details starting from model 1, and moving to model 2 and 3; each method answers a different question. In conclusion, the ultimate chosen model depends on the research goal.

CHAPTER 4

MISSING DATA ANALYSIS FOR BINARY MULTIVARIATE LONGITUDINAL DATA THROUGH A SIMULATION STUDY

4.1 Summary

Longitudinal data play an important role in many biomedical and social problems. Multivariate longitudinal data is a generalization of univariate longitudinal data where there are many dependent outcomes obtained at many occasions. Missing data frequently occur in longitudinal studies for many reasons such as subjects' moving or medical issues. Missingness could affect the bias and precision of parameter estimations. For multivariate longitudinal data, the effect of missingness has not been investigated carefully – e.g., the induced bias may be associated with the degree of correlation over the outcomes or occasions. In this paper, we fill this gap with a dedicated simulation to study the missingness effects on the parameter estimations for multivariate longitudinal data. We investigated missing data analysis for binary multivariate longitudinal data via GEE models, controlling the correlation over the occasions and outcomes. Generalized estimating equations (GEE) are commonly used to analyze longitudinal data due to its advantage of avoiding full specification of the response distribution. The simulation study is conducted to evaluate the effects of the correlation in the parameter estimations for missing data mechanisms, that are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Also, several analysis methods for handling missing data are used to reduce the bias of the parameter estimations when there is missingness in the responses, namely Completed cases (CC), Mean substitution (MS), Last Observation Carried Forward (LOCF) and Regression Imputation (RI). The inverse propensity weighed (IPW) GEE is used for the MAR mechanisms to investigate its performance with dropout missingness. Based on evaluation measurements such as root mean square error (RMSE) and coverage probability (CP), we found our results in agreement with Little and Rubin (1987) who indicated MCAR is the appropriate mechanism for the longitudinal data. We found that severe correlation over the occasions may affect the parameter estimations for both

complete and incomplete data. Also, we found the MS handling method, after extension to accommodate the multivariate structure, largely has good results even when one outcome is MAR and the second is MCAR. The inverse propensity weighed GEE shows some good results to treat the dropout in MAR mechanism especially when the correlation is induced over the outcomes.

4.2 Introduction

The nature of longitudinal data is repeated measurements over many occasions. Missingness frequently occurs during a longitudinal study because of circumstances such as a subject moving, medical illness or administrative reasons. In this paper, our focus is multivariate longitudinal data, for which there are more than one outcome measured at many occasions. The missingness in multivariate longitudinal data can happen such that all outcomes are missing at a particular time or partially such that only a subset of outcomes is not observed at that time. Missingness is considered a big issue in statistical analysis since its effects are to reduce the statistical power and increase bias. The statistical analysis should pay attention to the problem of missing data and use some statistical missing data handling methods to reduce the bias and reach better estimates.

Missing data mechanisms were classified by Little and Rubin (1987) into three types: 1) missing completely at random (MCAR), when the missingness is independent of the observed and unobserved responses; 2) missing at random (MAR), when the missingness may depend on observed responses, but is independent of the unobserved responses; 3) missing not at random (MNAR), when missingness depends on both observed and unobserved responses. MNAR is often called "informative missingness." In real life, the missingness mechanism is unknown and it is not a trivial issue to assume that missingness is MCAR or MAR. It is important to assume the appropriate mechanism based on the nature of the study to avoid biased estimates. In this paper, we generated artificial data through a simulation study to perform statistical analysis of incomplete binary multivariate longitudinal data in order to answer many questions. First, are the estimated correlation parameters different for the completed data and incomplete data, controlling the correlation over the outcomes and occasions? Second, are the estimated regression parameters valid for the three cases of missingness in all outcomes, in just partial outcomes, or mixed missing mechanisms for the outcomes? The third question is about the best method for handling missing data among CC, MS, LOCF and RI to treat the incomplete binary multivariate longitudinal data.

Many researchers have conducted simulation studies to investigate the effects of missing data handling methods for univariate longitudinal data. For example, Myers (2000) conducted a simulation for longitudinal data to compare the complete case method and multiple imputation methods in clinical trials and found some limitations of using multiple imputation. Also, Touloumi et al. (2001) designed a simulation study to investigate the impact of missing data due to drop out in longitudinal studies for six methods such as GEE, weighted and unweighted ordinary least square. They found the MCAR assumption is doing well for all their methods while MAR and MNAR generate biased estimates. Newman (2003) compared six missing data techniques based on simulation study. The comparison was for the three mechanisms of the missingness (MCAR, MAR, and MNAR) and for three levels of missingness 25%, 50% and 75%. Their results support a multiple imputation approach. Also, Hening (2009) performed a comparison for five imputation methods for the missing data (mean substitution, median substitution, zero values, hot deck and MI) with 20% missing rates for the first year students retention data in Ohio university. His dissertation indicates that mean imputation and median imputation yield good performance in precision. Ali et al. (2011) investigated four imputation methods (complete case analysis, mean substitution, and multiple imputation with and without inclusion of the outcome in the imputation model) under MCAR, MAR and MNAR based on simulation study to figure out the best approach. They found, based on their model, that the estimates for multiple imputation were least biased and most accurate.

The purpose of this paper is an investigation of the effect of missing data and the performance of missing data handling methods, especially for binary multivariate longitudinal data through a simulation study. Multivariate binary longitudinal data are generated for specified correlation structures and for different missing data mechanisms. The organization of this paper is as follows: section 3 contains the simulation design. It contains the model details, correlation scenarios, missing data mechanisms, handling missing data methods and simulation evaluation measurements. In section 4, the simulation results are produced. Finally, discussion and concluding remarks are in section 5.

4.3 Simulation Design

The multivariate longitudinal data structure is an extension of the univariate longitudinal structure to more than one outcome. Each individual i has a vector of responses for different outcomes,

$k = 1, 2, \ldots, K$. Also, each individual is measured at different occasions, $j = 1, 2 \ldots, J_i$, and has cluster size $n_i = J_i K$. Let us model K vectors of outcomes measured corresponding to a vector of times. Then, the structure is in the following figure:

$$\begin{array}{c} ID \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ N \\ N \\ \vdots \\ N \end{pmatrix} \end{array} \quad \begin{array}{c} \begin{array}{cccc} Y_{ij1} & Y_{ij2} & \cdots & Y_{ijK} \end{array} \\ \begin{pmatrix} y_{111} & y_{112} & \cdots & y_{11K} \\ y_{121} & y_{122} & \cdots & y_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1J1} & y_{1J2} & \cdots & y_{1JK} \\ y_{211} & y_{212} & \cdots & y_{21K} \\ y_{221} & y_{222} & \cdots & y_{22K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N11} & y_{N12} & \cdots & y_{N1K} \\ y_{N21} & y_{N22} & \cdots & y_{N2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{NJ1} & y_{NJ2} & \cdots & y_{NJK} \end{pmatrix} \end{array} \quad \begin{array}{c} time \\ \begin{pmatrix} 1 \\ 2 \\ \vdots \\ J \\ 1 \\ 2 \\ \vdots \\ 1 \\ 2 \\ \vdots \\ J \end{pmatrix} \end{array}$$

Figure 4.1: Multivariate longitudinal responses structure

To simplify these notations, we will refer to J_i as J, which is the number of occasions for all individuals. The vector of completed responses for subject i is:

$$Y_i = [Y_{i11}, Y_{i21}, \ldots, Y_{iJ1}, Y_{i12}, Y_{i22}, Y_{i32}, \ldots, Y_{iJ2}, ..., Y_{i1K}, Y_{i2K}, \ldots, Y_{iJK}]^T.$$

We conducted a simulation study in order to explore the changes of the parameter estimations for different missing data mechanisms when the data structure is multivariate longitudinal binary data. One of the goals to design the simulation study was to control the missing data pattern and the correlation for multivariate structure through the regression model.

4.3.1 Simulation model

We will use generalized estimating equations (GEE), an extension of generalized linear modeling to longitudinal data, to specify the simulation model and for the analysis. Thus, we specify a marginal model for the correlated responses. Estimation via GEE yields consistent regression parameter estimations despite the lack of full likelihood specification when the data are complete. This consistency also holds when data are MCAR, as shown in Little and Rubin (1987). We fit the GEE model of Shelton et al. (2004) to estimate the effects for each outcomes separately

using a Kronecker product approach to account for the correlation structure (see section 3.4.2). Let $X_i = 0, 1$ be the treatment assignment covariate. The time covariate is $t_j = 1, 2, 3$ for three occasions. Let Y_{ijk} be the binary response that is measured at time j, $j = 1, 2, 3$ for observation i, $i = 1, 2, 3, \ldots, N$ and for outcome $k = 1, 2$. Then, we assume the logistic model

$$logit(E(Y_{ijk}|X_i)) = \beta_{0k} + \beta_{1k}X_i + \beta_{2k}t_j \tag{4.1}$$

with the true parameter values $\beta_{01} = 0.2$, $\beta_{02} = 0.1$, $\beta_{11} = 0.05$, $\beta_{12} = -0.15$, $\beta_{21} = 0.05$, and $\beta_{22} = -0.25$. Given the vector of treatment covariate X_i, the Y_{ijk} for subject i is assumed to follow the Bernoulli distribution, $Y_{ijk}|X_i \sim Ber(E(Y_{ijk}|X_i))$. We did the simulation for the marginal regression model based on specified correlation structures and for different missing data mechanisms. We applied the method from chapter two that used the bridge distribution for the random effect in the mixed model to generate the correlated binary data. We generated $N = 250$ samples of correlated binary data. We conducted the simulation study for eight missing data mechanisms and for five scenarios (correlation structures) to explore the properties of the model when the correlation is induced over the outcomes and the occasions. In the following subsections, there are more explanations about correlation scenarios, the missing data mechanisms, handling missing data methods and evaluation measurements.

4.3.2 Correlation scenarios

Multivariate longitudinal data has a complicated correlation structure in which the correlation has two factors, over the occasions and over the outcomes. This simulation study explores the effects of increasing correlation over each of these two factors. We start by explaining the nature of the correlation in multivariate longitudinal data. The correlation matrix $R(\gamma)$ is a function of $\gamma_{jk,j'k'}$, where $\{\gamma_{jk,j'k'}\}$ represents the collection of within subject correlation parameters of size $\binom{JK}{2}$ of all non-redundant pairwise correlation parameters. For our setup with $J = 3$ and $K = 2$, there are 15 correlation parameters given by:

$$\gamma_{jk,j'k'} = \frac{P(Y_{jk} = 1, Y_{j'k'} = 1) - P(Y_{jk} = 1)P(Y_{j'k'} = 1)}{\sqrt{P(Y_{jk} = 1)P(Y_{j'k'} = 1)(1 - P(Y_{jk} = 1))(1 - P(Y_{j'k'} = 1))}}$$

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{21} & Y_{31} & Y_{12} & Y_{22} & Y_{32} \end{array} \\ \begin{pmatrix} 1 & \gamma_{11,21} & \gamma_{11,31} & \gamma_{11,12} & \gamma_{11,22} & \gamma_{11,32} \\ - & 1 & \gamma_{21,31} & \gamma_{21,12} & \gamma_{21,22} & \gamma_{21,32} \\ - & - & 1 & \gamma_{31,12} & \gamma_{31,22} & \gamma_{31,32} \\ - & - & - & 1 & \gamma_{12,22} & \gamma_{12,32} \\ - & - & - & - & 1 & \gamma_{22,32} \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

To reduce size of the unknown correlation parameter vector γ, we assume there are three components that build up the correlation structure R in the multivariate longitudinal data. The inter-outcome (υ), the intra-outcome (α); and the cross association (τ). The αs represent the correlation between the outcomes in the same occasion. The υs represent the correlation within outcomes at different occasions, and τs represent the correlation between different outcomes measured at different occasions. Thus, R can be written as follows:

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array} \\ \begin{pmatrix} 1 & \upsilon_1 & \upsilon_2 & \alpha_1 & \tau_1 & \tau_2 \\ - & 1 & \upsilon_3 & \tau_3 & \alpha_2 & \tau_4 \\ - & - & 1 & \tau_5 & \tau_6 & \alpha_3 \\ - & - & - & 1 & \upsilon_4 & \upsilon_5 \\ - & - & - & - & 1 & \upsilon_6 \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

We conducted five scenarios under the assumption of exchangeability for each correlation type υ, α and τ and $\tau = 0$ for all scenarios. The following table shows the values for each correlation parameter in each scenario for the correlation matrix R :

Table 4.1: The scenarios of correlation design

	$\alpha = 0.00$	$\alpha = 0.60$	$\alpha = 0.90$
$\upsilon = 0.00$	scenario 1	scenario 2	scenario 3
$\upsilon = 0.60$	scenario 4	-	-
$\upsilon = 0.90$	scenario 5	-	-

4.3.3 Missing data mechanisms

It is important to investigate the nature of the missing data to get valid inference. The missing data mechanism can be defined as the probability distribution of the missing indicator variable $R_i = (0, 1)$ that takes value 1 when the response is missing, and 0 if not. The vector of complete responses for subject i is:

$$Y_i = [Y_{i11}, Y_{i21}, \ldots, Y_{iJ1}, Y_{i12}, Y_{i22}, Y_{i32}, \ldots, Y_{iJ2}, ..., Y_{i1K}, Y_{i2K}, \ldots, Y_{iJK}]^T$$

Let R_i be the vector of response missingness indicators,

$$R_i = [R_{i11}, R_{i21}, \ldots, R_{iJ1}, R_{i12}, R_{i22}, R_{i32}, \ldots, R_{iJ2}, ..., R_{i1K}, R_{i2K}, \ldots, R_{iJK}]^T,$$

with $R_{ijk} = 1$ when Y_{ijk} is not observed, and $R_{ijk} = 0$ when Y_{ijk} is observed. In this paper, we do not consider missingness in the covariates. Given R_i, the complete set of responses Y_i can be divided into two groups: Y_i^O and Y_i^M. Y_i^O denotes the vector of observed responses and Y_i^M the vector of missing responses. Setting up the three mechanisms is done by the approaches in Hedeker and Gibbons (2006). Assuming the responses at the first time are fully observed, then the three missing mechanisms are as follows:

1- Missing completely at random MCAR:

The missing data pattern is considered to be MCAR if the probability that the responses are missing is independent of both Y_i^O and Y_i^M,

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

The missing data can be missing arbitrary or non-arbitrary. Here for MCAR mechanism, we will set up arbitrary missingness. In the simulation design, we assume the missingness is 25% at time 2 and 25% at time 3. Because we generate drop out missingness for the MAR and MNAR mechanisms, we checked that the drop out patterns in our MCAR data satisfied MCAR. Let $D_i = 0$ for subject i who has data at all time points $t = 1, 2, 3$ (no drop out), $D_i = 1$ for the dropout case at time 1 and $D_i = 2$ when the dropout occurs at time 2. Also, let $last$ denote the last observed value of the response for subject i. Then, a logistic regression model for each outcome is done separately as follows:

$$log\Big[\frac{P(D_i = j|D_i \geq j)}{1 - P(D_i = j|D_i \geq j)}\Big] = \beta_{0j} + \beta_1\, last_i + \beta_2 X_i + \beta_3 last_i\, t_j.$$

The MCAR mechanism implies that β_1 and β_3 are zero. Hence, a test for MCAR is as test of the null hypothesis $H_0 : \beta_1 = \beta_3 = 0$, which is not rejected when $\hat{\beta}_1$ and $\hat{\beta}_3$ are both not statistically different from zero. If the simulated data set led to rejection of H_0, we discarded those data and generated a new data set until H_0 was not rejected. Designing incomplete data to be compatible with the MCAR assumption could be done by different patterns, but ensuring the drop out pattern is MCAR should be sufficient for our purposes.

2- Missing at random MAR:

The missing at random mechanism implies the missingness probability is independent of Y_i^M, i.e.,

$$P(R_i|Y_i) = P(R|Y_i^O).$$

Here the design is much easier than for MCAR. We just design the missingness to be related the observed data and independent of unobserved data. In this simulation design, we set up the drop out after the first time point. If the logit of the response is lower than a specified cutpoint, then the subject drops out at the next time point for all the subsequent times. Thus we could ensure missingness depends on observed data. Then, we can assume the missingness is MAR in this simulation design.

3- Missing not at random MNAR:

The missing not at random mechanism holds when the probability the responses are missing is not independent of either Y_i^M or Y_i^O. Then we set the missingness related to the observed and unobserved data. After the first time point, if the logit of the response is lower than a specified cutpoint, then the subject drops out at that time point for all the subsequent times. Here the cause of the missingness comes from observed and unobserved data, satisfying MNAR.

To address the multivariate structure, we further classified mechanisms as "both missingness" when missingness occurs in both outcomes, and "partial missingness" when only one outcome is incomplete. The term "mixed missingness" when the two outcomes have different missing mechanisms. Defining different missing patterns is in order to study the changes in the regression effects estimations and the correlation parameters. In the following table, there are eight missing data mechanisms.

Table 4.2: The mechanisms of incomplete-dataset

Type	**Outcome1**	**Outcome2**	**Code**
Both	MCAR	MCAR	MCAR_MCAR
Both	MAR	MAR	MAR_MAR
Both	MNAR	MNAR	MNAR_MNAR
Partial	Completed	MCAR	COMP_MCAR
Partial	Completed	MAR	COMP_MAR
Partial	Completed	MNAR	COMP_MNAR
Mixed	MCAR	MAR	MCAR_MAR
Mixed	MCAR	MNAR	MCAR_MNAR

4.3.4 Handling missing data

There are many traditional methods of handling missing data in longitudinal studies. We will conduct four methods as follows:

1- Completed Case Analysis

This method is defined by deleting all subjects who have missing data. It includes only the completed cases. The primary condition to use this method for valid inference is the MCAR assumption. It will yield unbiased estimates of the mean responses when the assumption of the missing data is MCAR. However, the completed case analysis is not appealing when the size of the completed subjects is small relative to the whole number of subjects because it will reduce the statistical power and increase the standard error of the estimates (Allison, 1999). Also, if the data is not MCAR, then the parameter estimates could be biased.

Imputation

The imputation approach is widely used in practice. The basic idea is filling the missing values with imputed values. The imputed values are chosen by many methods. We will discuss three methods: Mean Substitution (MS), Last Observation Carried Forward (LOCF) and Regression Imputation (RI). The advantage of using imputation methods is the possibility to use the standard statistical analysis methods for the completed data after the imputation.

2- Mean Substitution (MS)

In this method, we easily use single imputation for each missing data by the mean of available observed values. This method may causes bias in the parameter estimations and under estimate the variability (Cook et al., 2004). Thus, it may artificially reduce the variability of the parameter

estimations. In this paper, we accommodate the multivariate structure by imputing the mean of each outcome, for each occasion, and for the same treatment. That means when we have two outcomes, three occasions and treatment binary variable, we will impute the missing values by 12 means instead of single mean. Each missing value is imputed by the mean of its outcome, its occasion, and treatment.

3- Last Observation Carried Forward (LOCF)

The LOCF is frequently used in clinical trials. It is a single imputation method that is widely used for longitudinal data, despite that it leads to bias expect for MCAR mechanism. In this method, each missing value is replaced by the last observed value for that subject. It is easy to impute LOCF method, but it is unrealistic to assume the observations following the dropouts remain unchanged unless the dropout is due to cure. Shao and Zhong (2003) conducted studies and proved that LOCF method could have biased results and reduce the precision of the parameter estimations.

4- Regression Imputation (RI)

The idea behind the regression imputation is simple. It is based on assuming the distribution of the missing responses is close to the distribution of the observed responses. Here, we impute the missing responses using the prediction from the regression equation derived from the observed responses. Then, we remodel the completed responses after the imputation. Using the prediction method to impute the missing responses seems not appealing when the cause of the drop out is related to the efficiency of the treatment. For example, when the sick subjects tends to drop out more than the healthy subjects, this leads to different estimates of the observed and missing responses. In MAR and MNAR assumptions, the there is a statistical difference between responses means of the observed and missing responses.

5- Inverse Probability Weight method (IPW)

The GEE method can be adapted to treat the dropout missingness for MAR mechanism. One applicable method is the inverse probability weighted (IPW) GEE approach of Preisser et al. (2002) and Robins and Rotnitzky (1995). It provides a method to handle the dropout missingness in the lack of good reasons to assume the MCAR mechanism. The idea is to weight the observed data to account for the probability of unobserved data. The weights are obtained by estimating the probability of dropout as a function of observed responses prior to dropouts and covariates. The

probability of not dropping out is called the propensity score. Here in the multivariate longitudinal data, the appropriate way is to estimate the propensity score weights for each outcome k. Let π_{ijk} denote the conditional probability of subject i being observed at time j for outcome k given the readings history of that subject at prior times.

$$\pi_{ijk} = P(R_{ijk} = 0 \,|R_{i1k} = R_{i2k} = \ldots = R_{i,j-1,k}, X_i, Y_{i1k}, Y_{i2k}, \ldots, Y_{i,j-1,k})$$

The appropriate weight for Y_{ijk} is the inverse of the unconditional probability of Y_{ijk} being observed, this probability being the cumulative product of π_{ijk} for $j = 1, 2, .., J$:

$$w_{ijk} = \frac{1}{\pi_{i1k}\pi_{i2k}...\pi_{iJk}}.$$

In our simulation context, the model for missingness estimates the probability of not dropping out at given time point, given the previous response, treatment values and their interaction. The logistic regression model for each outcome k is:

$$logit(\pi_{ijk}) = \beta_{0k} + \beta_{1k}\, t_{ij} + \beta_{3k} X_i + \beta_{4k} Y_{i,j-1,k} \; + \beta_{5k}\, X_i\, Y_{i,j-1,k}.$$

Thus score propensity is estimated for each subject at each time. Then, we analyze the data using weighted GEE model of Shelton et al. (2004) using the"working independence" correlation matrix.

4.3.5 Simulation evaluation

We used five criteria to measure the performance of each scenario and mechanism for the correlated binary data. These are the average, standard deviation (std), the bias to express the difference between the true and estimated parameters, the root of mean square error (RMSE), and the coverage probability (CP) associated with the usual 95% confidence interval. Using the RMSE is good because it is the square root of MSE that measures the precision of the estimates. The coverage probability is the proportion of the nominal 95% confidence intervals from simulated datasets that contain the true parameter values.

4.4 The Results

The general strategy of this study is to first generate the artificial multivariate longitudinal data for two outcomes and three occasions and for five correlation scenarios based on the model in

4.3.1. Then, we performed the eight missing mechanisms described in 4.3.3. First, we will present the estimates of incomplete data. The correlation parameters in matrix R are estimated using sample correlation values for each mechanism and for each correlation scenario after generating the correlated binary data. In scenario 1, where all correlation parameters are assumed to be zero, the estimated parameters are close to zero. In scenarios 2 and 3, we induced the correlation between the outcomes in times 1, 2 and 3 respectively, α_1, α_2, and α_3. Table 4.3 reports the estimated means of 224 samples of the outcomes parameters. In scenario 2 and scenario 3, the true correlation parameters are 0.6 and 0.9, respectively. The estimated correlation parameters in the completed data exhibit increasing bias over time, and this may be due to existence of the time covariate. Comparing the estimated correlation parameters for the complete and incomplete data for different mechanisms, we note many points. We found α_1 did not change for all the missing mechanisms because the baseline is full without missingness. The MNAR assumption has a clear effect on α_2 and α_3 over the both, partial and mixed mechanisms.

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{c} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array} \\ \begin{pmatrix} 1 & \upsilon_1 & \upsilon_2 & \alpha_1 & \tau_1 & \tau_2 \\ - & 1 & \upsilon_3 & \tau_3 & \alpha_2 & \tau_4 \\ - & - & 1 & \tau_5 & \tau_6 & \alpha_3 \\ - & - & - & 1 & \upsilon_4 & \upsilon_5 \\ - & - & - & - & 1 & \upsilon_6 \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

Table 4.3: The within outcomes correlation parameter estimations in scenario 2 and 3.

	Scenario 2			**Senario 3**		
	α_1	α_2	α_3	α_1	α_2	α_3
Complete	0.572	0.538	0.496	0.772	0.668	0.573
MCAR_MCAR	0.572	0.539	0.495	0.772	0.668	0.572
MAR_MAR	0.572	0.539	0.496	0.772	0.668	0.576
MNAR_MNAR	0.572	0.372	0.197	0.772	0.512	0.124
COMP_MCAR	0.572	0.539	0.495	0.772	0.668	0.572
COMP_MAR	0.572	0.538	0.495	0.772	0.668	0.576
COMP_MNAR	0.572	0.404	0.254	0.772	0.512	0.124
MCAR_MAR	0.572	0.539	0.495	0.772	0.668	0.576
MCAR_MNAR	0.572	0.405	0.253	0.772	0.512	0.142

In table 4.4, the estimated correlation parameters for time factor υ_1, υ_2 and υ_3 represent the correlation between the responses at (time1, time2), (time1, time3) and (time2, time3) within the first outcome and υ_4, υ_5 and υ_6 within the second outcome. In scenarios 4 and 5, we induced the correlation over the occasions factor. We found the MCAR_MCAR and COMP_MCAR yield results very close to parameters in the completed data but the MCAR_MAR has been affected in scenario 5, the scenario that has strongest correlation over the occasions. Additionally, the results of the remaining MAR and MNAR for both, partial and mixed missingness are biased and have been affected by the correlation apart from the baseline parameters α_1 and α_4. We found the MCAR assumption for both outcomes is the appropriate assumption for the GEE model. Now we can figure out its advantage to keep the correlation parameter close to the correlation parameters in the complete data. Also, we may find the missingness pattern of MAR and MNAR mechanisms could affect the parameter estimations due to the bias in the correlation parameters before we fit the model.

Table 4.4: The within occasions correlation parameter estimations in scenario 4 and 5.

	Senario 4						**Senario 5**					
	within outcome 1			**within outcome 2**			**within outcome 1**			**within outcome 2**		
	υ_1	υ_2	υ_3	υ_4	υ_5	υ_6	υ_1	υ_2	υ_3	υ_4	υ_5	υ_6
Complete	0.592	0.589	0.587	0.584	0.567	0.581	0.883	0.876	0.878	0.852	0.773	0.849
MCAR_MCAR	0.590	0.589	0.589	0.583	0.568	0.583	0.883	0.877	0.881	0.852	0.772	0.849
MAR_MAR	0.497	0.371	0.263	0.496	0.371	0.275	0.844	0.642	0.621	0.813	0.442	0.626
MNAR_MNAR	0.445	0.324	0.303	0.423	0.274	0.291	0.811	0.746	0.746	0.743	0.447	0.645
COMP_MCAR	0.590	0.589	0.589	0.584	0.567	0.581	0.883	0.877	0.881	0.852	0.773	0.849
COMP_MAR	0.590	0.589	0.589	0.496	0.371	0.275	0.883	0.877	0.881	0.813	0.442	0.626
COMP_MNAR	0.590	0.589	0.589	0.423	0.274	0.291	0.883	0.877	0.881	0.743	0.447	0.645
MCAR_MAR	0.592	0.589	0.587	0.496	0.371	0.275	0.883	0.876	0.878	0.813	0.442	0.626
MCAR_MNAR	0.592	0.589	0.587	0.423	0.274	0.291	0.883	0.876	0.878	0.743	0.447	0.645

In model 4.1, there are three regression coefficients associated with the intercept, X and time covariates for each outcome after we fit the GEE model assuming the unstructured within subject correlation. The parameter estimations of the X covariate are β_{11}, β_{12} for scenarios 1, 2 and 4 in tables 4.5, 4.6, and 4.7, respectively. The remaining results for the time and intercept effects are in appendix A. Generally from all the scenarios, we found the results of MCAR_MCAR and COMP_MCAR mechanisms are close to the regression coefficients of the completed data without

missngness. The MAR_MAR and MNAR_MNAR have shown some bias results. In scenarios 2 and 4, the bias clearly affects the MAR more than MNAR in the parameters of the complete outcome. For the partial mechanisms, we found in scenario 1 that the estimated effects appear unbiased for the for the first outcome, which is complete. We conclude that assuming MCAR, which is implicit for GEE models, is affected by MAR and MNAR when they are mixed, especially with the strong correlation over the occasions.

Table 4.5: The estimates of X covariate in scenario 1 for the two outcomes.

	True $\beta_{11} = 0.05$					True $\beta_{12} = -0.15$				
type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.055	0.112	0.005	0.112	94.20	-0.148	0.103	0.002	0.103	95.98
MCAR_MCAR	0.059	0.115	0.009	0.116	96.88	-0.145	0.115	0.005	0.114	95.54
MAR_MAR	0.055	0.121	0.005	0.121	92.92	-0.153	0.121	-0.003	0.121	94.81
MNAR_MNAR	0.049	0.130	-0.001	0.130	94.20	-0.130	0.125	0.020	0.126	94.20
COMP_MCAR	0.056	0.112	0.006	0.112	94.20	-0.145	0.115	0.005	0.114	95.98
COMP_MAR	0.055	0.111	0.005	0.111	93.84	-0.150	0.121	0.000	0.121	94.79
COMP_MNAR	0.055	0.112	0.005	0.112	93.75	-0.130	0.125	0.020	0.126	94.20
MCAR_MAR	0.059	0.115	0.009	0.115	96.74	-0.152	0.122	-0.002	0.122	95.35
MCAR_MNAR	0.058	0.116	0.008	0.116	96.43	-0.130	0.125	0.020	0.126	94.20

Table 4.6: The estimates of X covariate in scenario 2 for the two outcomes.

	True $\beta_{11} = 0.05$					True $\beta_{12} = -0.15$				
type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.047	0.107	-0.003	0.107	95.24	-0.150	0.105	0.000	0.105	94.37
MCAR_MCAR	0.047	0.116	-0.003	0.116	95.24	-0.152	0.119	-0.002	0.119	93.51
MAR_MAR	0.048	0.160	-0.002	0.159	96.54	-0.149	0.132	0.001	0.132	93.94
MNAR_MNAR	0.047	0.120	-0.003	0.120	96.97	-0.139	0.120	0.011	0.121	94.81
COMP_MCAR	0.047	0.106	-0.003	0.106	95.24	-0.151	0.114	-0.001	0.114	93.94
COMP_MAR	0.057	0.155	0.007	0.154	95.24	-0.143	0.138	0.007	0.138	92.64
COMP_MNAR	0.048	0.107	-0.002	0.106	95.24	-0.141	0.120	0.009	0.120	94.37
MCAR_MAR	0.057	0.143	0.007	0.142	94.78	-0.145	0.129	0.005	0.129	93.04
MCAR_MNAR	0.051	0.116	0.001	0.115	94.81	-0.137	0.121	0.013	0.121	93.51

Table 4.7: The estimates of X covariate in scenario 4 for the two outcomes.

	True $\beta_{11} = 0.05$					True $\beta_{12} = -0.15$				
type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.054	0.163	0.004	0.163	93.10	-0.180	0.163	-0.030	0.166	91.81
MCAR_MCAR	0.057	0.167	0.007	0.167	93.53	-0.179	0.166	-0.029	0.168	92.24
MAR_MAR	0.134	0.629	0.084	0.630	95.83	-0.114	0.683	0.036	0.680	95.83
MNAR_MNAR	0.045	0.172	-0.005	0.172	93.45	-0.167	0.174	-0.017	0.175	92.14
COMP_MCAR	0.054	0.163	0.004	0.163	93.97	-0.180	0.166	-0.030	0.168	91.81
COMP_MAR	0.052	0.334	0.002	0.332	90.91	-0.146	0.499	0.004	0.496	96.10
COMP_MNAR	0.053	0.164	0.003	0.164	93.48	-0.163	0.162	-0.013	0.162	92.61
MCAR_MAR	0.041	0.418	-0.009	0.416	94.05	-0.180	0.444	-0.030	0.442	94.05
MCAR_MNAR	0.055	0.167	0.005	0.167	93.94	-0.160	0.166	-0.010	0.166	92.64

Now, we present the analysis after handling the missing data mechanisms for the incomplete data. We handled the missing data using four different methods: CC, MS, LOCF and RI as they are described in section 4.3.4. In table 4.8, the results of the X covariate parameters estimations after handling the missingness just for the mechanism MCAR_MCAR while the rest of the intercepts and time covariate are in appendix B. Since the appropriate mechanism for the GEE models is the assumption MCAR, we will analyze the handling methods for the mechanisms COMP_MCAR, MCAR_MCAR, COMP_MAR and MCAR_MAR. In table 4.8, the results of the estimated X effects on the log odds of response $P(Y_{ijk} = 1)$ for the MCAR_MCAR mechanism. We present four evaluation measurements. We found using the MS and LOCF methods have good means, CP and also less bias in scenario 1 for the estimated parameters of the treatment effects. In scenario 2 where the correlation is induced between the outcomes, we found the methods LOCF and MS have better means close to the true values, good CP and less bias. For the results of scenario 4 where the correlation is induced over the occasions, we found the parameter estimation of the completed data are already biased and we don't find good method to reduce it over the four methods.

In table 4.9, the results of handling the mechanism COMP_MCAR for estimated X effects on the log odds of response $P(Y_{ijk} = 1)$. This mechanism is a mixture when the first outcome is complete and the second is MCAR. We found the best method in all the scenarios is MS with good means, less standard deviation and bias for the estimated parameters of the two outcomes unless in the parameter β_{12} in scenario 4. This estimated parameter is already bias in the completed data and MS method have shown a slight reduction in the bias criteria. We found the estimates of the

completed outcome were not affected by the MCAR mechanism in the second outcome regardless of the correlation induced between the outcomes or occasions.

In table 4.10 where the mechanism is mixed between the MCAR for the first outcomes and MAR for the second outcome. Here it is a good point to testify the stability of the estimated parameters of the first outcome when the parameters of the second outcome has MAR set up pattern for its missing data. We found the MS method has been good for all the scenarios for the two outcomes with good means and less std. This generalization has exception in the second parameter β_{12} in scenario 1 and β_{11} in scenario 4 when the there a slight difference between the incomplete data and MS results. Here we indicated the mean substitution imputation reduce the bias of the MAR missing mechanism in the second outcomes and doing good results to impute the first outcome when its mechanism is MCAR. Finally, it seems the MS method has good means and less standard deviation and less bias in most the results of handling the missing data.

In table 4.11, we have mixed missing mechanisms for the two outcomes. While the first outcome is complete, the second outcome has MAR assumption for its missing data. Here the goal is testify the precision of the parameters of first outcomes if the they will be affected by the MAR assumption, especially when the two outcomes are correlated. Here the the additional method IPW is added since it is appropriate to reduce the bias in the MAR assumption. Generally, we found the RI method has good means, CP and less std for all the scenarios while the CC has the worst results. Based on the bias results we found, RI and MS and IPW reduced the bias from the completed data in scenarios 2 and 4 where correlation is increased. Generally, we can conclude from model parameters the MS and IPW have some good results to treat the missingness in the mechanism COMP_MAR.

Table 4.12 shows results where the mechanism is MAR for both outcomes. Based on the available methods we conducted in this study to handle the missing data, we don't find clear good results to treat the bias in MAR assumption for both outcomes. Also, we found some good results to reduce the bias using IPW method in scenario 1 and scenario 4 where we increase the correlation over the occasion. In table 4.13, the both outcomes has the MNAR assumption. Here The IPW method fails to reduce the bias at for the scenarios. Also, there is no method that shows good results to treat the missngness. The MAR and MNAR are considered to be problematic mechanisms for the parameter estimations in incomplete data.

Table 4.8: The estimates of X covariate for different imputation methods of MCAR_MCAR.

	True $\beta_{11} = 0.05$				True $\beta_{12} = -0.15$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0529	0.0469	0.0549	**Complete**	-0.1479	-0.1474	-0.1778
CC	0.0635	0.0396	0.0562	**CC**	-0.1480	-0.1581	-0.1835
LOCF	0.0567	0.0506	0.0572	**LOCF**	-0.1448	-0.1502	-0.1801
RI	0.0260	0.0109	0.0191	**RI**	-0.1554	-0.1695	-0.1872
MS	0.0571	0.0431	0.0567	**MS**	-0.1463	-0.1582	-0.1810
Incomplete	0.0589	0.0468	0.0566	**Incomplete**	-0.1443	-0.1518	-0.1791
		std				**std**	
Complete	0.1116	0.1056	0.1595	**Complete**	0.1020	0.1062	0.1617
CC	0.1413	0.1482	0.2152	**CC**	0.1396	0.1428	0.2255
LOCF	0.1233	0.1260	0.1672	**LOCF**	0.1238	0.1280	0.1654
RI	0.1044	0.1182	0.1500	**RI**	0.0986	0.1096	0.1444
MS	0.1203	0.1196	0.1715	**MS**	0.1226	0.1234	0.1776
Incomplete	0.1154	0.1161	0.1663	**Incomplete**	0.1145	0.1188	0.1642
		bias				**bias**	
Complete	0.0032	-0.0031	0.0062	**Complete**	0.0018	0.0026	-0.0263
CC	0.0136	-0.0104	0.0072	**CC**	0.0024	-0.0081	-0.0313
LOCF	0.0073	0.0006	0.0085	**LOCF**	0.0047	-0.0002	-0.0286
RI	-0.0237	-0.0391	-0.0293	**RI**	-0.0052	-0.0195	-0.0359
MS	0.0072	-0.0069	0.0077	**MS**	0.0038	-0.0082	-0.0289
Incomplete	0.0092	-0.0032	0.0079	**Incomplete**	0.0055	-0.0018	-0.0275
		CP				**CP**	
Complete	94.00	95.60	93.60	**Complete**	96.40	94.40	92.00
CC	95.11	92.21	95.26	**CC**	95.11	93.51	91.38
LOCF	94.22	94.81	93.97	**LOCF**	95.11	93.51	92.67
RI	94.67	93.07	93.53	**RI**	96.00	94.81	91.81
MS	92.89	91.77	88.36	**MS**	92.89	90.91	87.93
Incomplete	96.89	95.24	93.53	**Incomplete**	95.56	93.51	92.24

Table 4.9: The estimates of X covariate for different imputation methods of COMP_MCAR.

	True $\beta_{11} = 0.05$				True $\beta_{12} = -0.15$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0529	0.0469	0.0549	**Complete**	-0.1479	-0.1474	-0.1778
CC	0.0635	0.0396	0.0562	**CC**	-0.1480	-0.1581	-0.1835
LOCF	0.0554	0.0465	0.0538	**LOCF**	-0.1444	-0.1514	-0.1802
RI	0.0550	0.0460	0.0536	**RI**	-0.1624	-0.1632	-0.1936
MS	0.0553	0.0471	0.0530	**MS**	-0.1496	-0.1497	-0.1794
Incomplete	0.0554	0.0465	0.0538	**Incomplete**	-0.1444	-0.1514	-0.1802
		std				**std**	
Complete	0.1116	0.1056	0.1595	**Complete**	0.1020	0.1062	0.1617
CC	0.1413	0.1482	0.2152	**CC**	0.1396	0.1428	0.2255
LOCF	0.1120	0.1063	0.1623	**LOCF**	0.1145	0.1145	0.1646
RI	0.1118	0.1063	0.1634	**RI**	0.0998	0.1032	0.1434
MS	0.1124	0.1075	0.1626	**MS**	0.1213	0.1263	0.1765
Incomplete	0.1120	0.1063	0.1623	**Incomplete**	0.1145	0.1145	0.1646
		bias				**bias**	
Complete	0.0032	-0.0031	0.0062	**Complete**	0.0018	0.0026	-0.0263
CC	0.0136	-0.0104	0.0072	**CC**	0.0024	-0.0081	-0.0313
LOCF	0.0057	-0.0035	0.0051	**LOCF**	0.0053	-0.0014	-0.0286
RI	0.0053	-0.0040	0.0049	**RI**	-0.0130	-0.0132	-0.0421
MS	0.0056	-0.0029	0.0043	**MS**	0.0004	0.0003	-0.0278
Incomplete	0.0057	-0.0035	0.0051	**Incomplete**	0.0053	-0.0014	-0.0286
		CP				**CP**	
Complete	94.00	95.60	93.60	**Complete**	96.40	94.40	92.00
CC	95.11	92.21	95.26	**CC**	95.11	93.51	91.38
LOCF	94.22	95.24	93.97	**LOCF**	96.00	93.94	91.81
RI	94.22	94.81	93.97	**RI**	96.89	93.07	92.24
MS	94.22	94.37	93.97	**MS**	93.78	89.61	87.07
Incomplete	94.22	95.24	93.97	**Incomplete**	96.00	93.94	91.81

Table 4.10: The estimates of X covariate for different imputation methods of MCAR_MAR.

	True $\beta_{11} = 0.05$				True $\beta_{12} = -0.15$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0529	0.0469	0.0549	**Complete**	-0.1479	-0.1474	-0.1778
CC	0.0733	0.1110	0.0250	**CC**	-0.1025	-0.1188	-0.0963
LOCF	0.0586	0.0495	0.0562	**LOCF**	-0.1701	-0.1699	-0.1817
RI	0.0227	0.0152	0.0087	**RI**	-0.1536	-0.1608	-0.1777
MS	0.0534	0.0480	0.0598	**MS**	-0.1433	-0.1500	-0.1421
Incomplete	0.0590	0.0571	0.0413	**Incomplete**	-0.1519	-0.1453	-0.1799
		std				**std**	
Complete	0.1116	0.1056	0.1595	**Complete**	0.1020	0.1062	0.1617
CC	0.2222	0.2773	0.3348	**CC**	0.2040	0.2185	0.2938
LOCF	0.1141	0.1159	0.1659	**LOCF**	0.1360	0.1342	0.1644
RI	0.1017	0.1091	0.1441	**RI**	0.0894	0.0954	0.1223
MS	0.1284	0.1264	0.1757	**MS**	0.1520	0.1391	0.1740
Incomplete	0.1150	0.1426	0.4179	**Incomplete**	0.1219	0.1291	0.4438
		bias				**bias**	
Complete	0.0032	-0.0031	0.0062	**Complete**	0.0018	0.0026	-0.0263
CC	0.0224	0.0610	-0.0232	**CC**	0.0478	0.0312	0.0546
LOCF	0.0089	-0.0006	0.0075	**LOCF**	-0.0206	-0.0199	-0.0299
RI	-0.0269	-0.0348	-0.0399	**RI**	-0.0041	-0.0108	-0.0265
MS	0.0038	-0.0020	0.0108	**MS**	0.0066	0.0000	0.0092
Incomplete	0.0093	0.0071	-0.0087	**Incomplete**	-0.0024	0.0047	-0.0299
		CP				**CP**	
Complete	94.00	95.60	93.60	**Complete**	96.40	94.40	92.00
CC	93.78	93.51	92.24	**CC**	95.11	94.37	95.26
LOCF	96.44	95.24	93.97	**LOCF**	95.11	94.37	93.10
RI	94.22	92.21	93.97	**RI**	95.56	92.64	93.53
MS	91.11	89.61	88.36	**MS**	84.00	88.31	83.19
Incomplete	96.76	94.78	94.05	**Incomplete**	95.37	93.04	94.05

Table 4.11: The estimates of X covariate for different imputation methods of COMP_MAR.

	True $\beta_{11} = 0.05$				True $\beta_{12} = -0.15$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0529	0.0469	0.0549	**Complete**	-0.1479	-0.1474	-0.1778
CC	0.0536	0.1200	0.0353	**CC**	-0.1037	-0.1099	-0.0950
LOCF	0.0552	0.0482	0.0537	**LOCF**	-0.1706	-0.1691	-0.1828
RI	0.0550	0.0456	0.0537	**RI**	-0.1534	-0.1534	-0.1725
MS	0.0551	0.0458	0.0532	**MS**	-0.1475	-0.1531	-0.1374
IPW	0.0526	0.0447	0.0532	**IPW**	-0.1530	-0.1521	-0.1585
Incomplete	0.0551	0.0575	0.0525	**Incomplete**	-0.1496	-0.1425	-0.1465
		std				**std**	
Complete	0.1116	0.1056	0.1595	**Complete**	0.1020	0.1062	0.1617
CC	0.1675	0.1922	0.2599	**CC**	0.1576	0.1507	0.2181
LOCF	0.1114	0.1062	0.1624	**LOCF**	0.1358	0.1343	0.1644
RI	0.1111	0.1035	0.1637	**RI**	0.0964	0.0900	0.1144
MS	0.1118	0.1116	0.1634	**MS**	0.1495	0.1397	0.1773
IPW	0.1718	0.1043	0.1181	**IPW**	0.3062	0.1792	0.2085
Incomplete	0.1112	0.1545	0.3340	**Incomplete**	0.1209	0.1383	0.4995
		bias				**bias**	
Complete	0.0032	-0.0031	0.0062	**Complete**	0.0018	0.0026	-0.0263
CC	0.0036	0.0700	-0.0130	**CC**	0.0457	0.0401	0.0568
LOCF	0.0055	-0.0018	0.0050	**LOCF**	-0.0212	-0.0191	-0.0310
RI	0.0053	-0.0044	0.0051	**RI**	-0.0040	-0.0034	-0.0210
MS	0.0054	-0.0042	0.0044	**MS**	0.0025	-0.0031	0.0138
IPW	-0.0053	0.0032	0.0026	**IPW**	-0.0021	-0.0085	-0.0030
Incomplete	0.0055	0.0075	0.0025	**Incomplete**	-0.0001	0.0075	0.0035
		CP				**CP**	
Complete	94.00	95.60	93.60	**Complete**	96.40	94.40	92.00
CC	93.78	94.37	92.24	**CC**	94.22	95.67	94.40
LOCF	94.22	95.67	93.10	**LOCF**	94.67	94.37	93.10
RI	94.22	95.67	93.10	**RI**	96.44	95.24	96.12
MS	93.78	93.51	93.53	**MS**	84.44	86.58	84.91
IPW	93.98	95.18	95.52	**IPW**	93.57	93.57	93.72
Incomplete	93.87	95.24	90.91	**Incomplete**	94.81	92.64	96.10

Table 4.12: The estimates of X covariate for different imputation methods of MAR_MAR.

	True $\beta_{11} = 0.05$				True $\beta_{12} = -0.15$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0532	0.0469	0.0562	**Complete**	-0.1482	-0.1474	-0.1763
CC	0.0715	0.0883	0.0228	**CC**	-0.0867	-0.1372	-0.1012
LOCF	0.0587	0.0454	0.0537	**LOCF**	-0.1716	-0.1702	-0.1812
RI	0.0162	0.0129	0.0029	**RI**	-0.1552	-0.1625	-0.1759
MS	0.0604	0.0468	0.0365	**MS**	-0.1492	-0.1454	-0.1315
IPW	0.0524	0.0447	0.1252	**IPW**	-0.1538	-0.1507	-0.1222
Incomplete	0.0552	0.0480	0.1340	**Incomplete**	-0.1528	-0.1486	-0.1136
		std				**std**	
Complete	0.1116	0.1056	0.1595	**Complete**	0.1020	0.1062	0.1617
CC	0.2345	0.2092	0.3507	**CC**	0.2105	0.1668	0.2699
LOCF	0.1353	0.1268	0.1650	**LOCF**	0.1362	0.1366	0.1643
RI	0.1128	0.1087	0.1452	**RI**	0.0943	0.0933	0.1208
MS	0.1318	0.1287	0.1578	**MS**	0.1360	0.1419	0.1745
IPW	0.4272	0.1151	0.1509	**IPW**	0.5577	0.1790	0.2062
Incomplete	0.1207	0.1597	0.6291	**Incomplete**	0.1210	0.1323	0.6835
		bias				**bias**	
Complete	0.0032	-0.0031	0.0062	**Complete**	0.0018	0.0026	-0.0263
CC	0.0215	0.0383	-0.0272	**CC**	0.0633	0.0128	0.0489
LOCF	0.0087	-0.0046	0.0037	**LOCF**	-0.0215	-0.0202	-0.0312
RI	-0.0338	-0.0371	-0.0471	**RI**	-0.0052	-0.0125	-0.0258
MS	0.0104	-0.0032	-0.0135	**MS**	0.0008	0.0046	0.0185
IPW	-0.0053	0.0752	0.0024	**IPW**	-0.0007	0.0278	-0.0038
Incomplete	0.0052	-0.0020	0.0840	**Incomplete**	-0.0028	0.0014	0.0364
		CP				**CP**	
Complete	93.98	95.58	93.98	**Complete**	96.39	94.38	92.37
CC	95.54	95.67	92.17	**CC**	94.64	95.67	93.48
LOCF	92.86	96.54	94.81	**LOCF**	94.20	93.94	93.07
RI	91.96	90.91	93.51	**RI**	95.54	95.24	95.24
MS	88.84	87.45	91.77	**MS**	88.39	86.58	82.25
IPW	92.77	94.78	94.94	**IPW**	93.98	94.38	96.20
Incomplete	92.92	96.54	95.83	**Incomplete**	94.81	93.94	95.83

Table 4.13: The estimates of X covariate for different imputation methods of MNAR_MNAR.

	True $\beta_{11} = 0.05$				True $\beta_{12} = -0.15$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0532	0.0469	0.0562	**Complete**	-0.1482	-0.1474	-0.1763
CC	0.0754	0.1091	0.0196	**CC**	-0.1041	-0.1171	-0.0939
LOCF	0.0584	0.0478	0.0566	**LOCF**	-0.1515	-0.1407	-0.1748
RI	0.0118	0.0087	0.0019	**RI**	-0.1762	-0.1791	-0.1834
MS	0.0454	0.0498	0.0392	**MS**	-0.1139	-0.1178	-0.1230
IPW	0.0452	0.0457	0.0391	**IPW**	-0.1160	-0.1210	-0.1370
Incomplete	0.0485	0.0470	0.0461	**Incomplete**	-0.1301	-0.1391	-0.1657
		std				**std**	
Complete	0.1116	0.1056	0.1595	**Complete**	0.1020	0.1062	0.1617
CC	0.2696	0.2447	0.3645	**CC**	0.2390	0.1701	0.3182
LOCF	0.1403	0.1310	0.1708	**LOCF**	0.1459	0.1386	0.1600
RI	0.1176	0.1207	0.1531	**RI**	0.0983	0.1062	0.1314
MS	0.1444	0.1337	0.1771	**MS**	0.1536	0.1397	0.1868
IPW	0.1661	0.1259	0.1241	**IPW**	0.1983	0.1409	0.1378
Incomplete	0.1299	0.1202	0.1711	**Incomplete**	0.1247	0.1204	0.1726
		bias				**bias**	
Complete	0.0032	-0.0031	0.0062	**Complete**	0.0018	0.0026	-0.0263
CC	0.0254	0.0591	-0.0304	**CC**	0.0459	0.0329	0.0561
LOCF	0.0084	-0.0022	0.0066	**LOCF**	-0.0015	0.0093	-0.0247
RI	-0.0382	-0.0413	-0.0481	**RI**	-0.0262	-0.0291	-0.0334
MS	-0.0046	-0.0002	-0.0108	**MS**	0.0361	0.0322	0.0270
IPW	-0.0043	-0.0109	-0.0048	**IPW**	0.0290	0.0130	0.0341
Incomplete	-0.0015	-0.0030	-0.0039	**Incomplete**	0.0199	0.0109	-0.0157
		CP				**CP**	
Complete	93.98	95.58	93.98	**Complete**	96.39	94.38	92.37
CC	91.52	93.51	89.18	**CC**	93.30	96.97	92.21
LOCF	95.54	96.97	95.24	**LOCF**	92.41	92.64	93.94
RI	91.07	90.48	92.64	**RI**	96.43	94.81	93.07
MS	88.84	87.88	91.34	**MS**	82.14	84.42	80.95
IPW	92.77	95.58	95.18	**IPW**	94.78	96.79	91.57
Incomplete	94.20	96.97	93.86	**Incomplete**	94.20	94.81	92.54

4.5 Conclusion

We presented the analysis of incomplete multivariate binary longitudinal data. The correlation parameters for the complete and incomplete data are estimated. We found the missingness affects the estimated correlation among the occasions more than between the outcomes. Also, we found the estimates of MAR and MNAR of incomplete data are affected by the induced correlation over the occasions and the outcomes. This agrees with results of Little and Rubin (1987) about the bias of estimates of MAR or MNAR mechanisms for GEE models. After imputing the incomplete data with four missing data handling methods, we conclude many points. The mean substitution based on the multivariate structure mostly has good estimates in the mechanism COMP_MCAR, MCAR_MAR and COMP_MAR. It has been a good remark to find the MS imputation reduced the effects of MAR assumption in the mixed mechanisms and generated mostly less biased results. In the mechanism MCAR_MCAR, the LOCF method has good results. Also, using the weighted GEE for the full and mixed MAR assumptions shows some good results. Generally, we recommend using the mean substitution based on the multivariate structure to impute the missing data. It could be a good future work to use the multiple imputation to handle the missingness in the multivariate structure based on Shelton et al. (2004)'s model.

CHAPTER 5

CONCLUSION

The three contributions of this dissertation are about three different topics in the analysis of binary multivariate longitudinal data. As we pointed out, the main problem in multivariate longitudinal data is the complicated correlation of within subject measurements. Accounting for this correlation will reduce the bias of the parameter estimations and increase the power of the model. At the last station of this dissertation, we conclude some points. First, from chapter two we found the method of using the bridge distribution for the random effect in the liner mixed model could be a good method, if we run the model based on some constraints. It is a method to generate correlated binary data based on specified R correlation matrix for the marginal model. Using the bridge distribution has the advantage of keeping the same logistic shape for the marginal and conditional model.

In the second contribution we revised two exciting methods and proposed the third method to analyze the binary multivariate longitudinal data. The first method estimates one group of the covariate effects and the second method separated the effects of the covariate for each outcome. The third method estimates the effects for each outcome and for the joint cases between them. All the three methods account for the multivariate structure and the complicated correlation. We found the second and third methods are good methods to analyze the multivariate longitudinal data for the binary responses while the third one has more deeper analysis and more parameters to estimate.

In the third contribution, we analyzed incomplete multivariate longitudinal data. We presented a simulation study using the method of generating the correlated binary data from the first contribution. We studied the effects of the correlation on the parameter estimations while controlling the correlation over the outcomes and occasions. Also, we control the type of the missing mechanisms MCAR, MAR or MNAR for both, partial or mixed outcomes. We found ourselves in agreement with Rubin's result that the estimations of MAR and MNAR are affected by the missingness in GEE models. After analyzing the incomplete data, we handled the missingness using the four meth-

ods of Completed cases (CC), Mean substitution (MS), Last Observation Carried Forward (LOCF) and Regression Imputation (RI). We found the MS method based on the multivariate structure for each outcome and occasion has mostly good results. Also, we found when the first outcome has missing data that is MCAR or is complete without missingness and the second outcome has MAR missingness, then the imputation using MS method produces mostly good results. Imputing the missing data is a wide research topic and it could be good future work to use a Bayesian method and multiple imputation to impute the missing data.

Finally, analyzing binary multivariate data has a substantial role in recent research. Many of the recent longitudinal studies have binary measurements and correlation is induced among the responses. A good method accounting for the correlation in the longitudinal data leads to good parameter estimations. This dissertation contributes a straightforward investigation of the changes in covariate effect estimates while accounting for within subject correlation in the binary multivariate longitudinal data.

APPENDIX A

THE PARAMETER ESTIMATIONS FOR THE TWO OUTCOMES

Table A.1: The estimates of X covariate in scenario 1 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.055	0.112	0.005	0.112	94.20	-0.148	0.103	0.002	0.103	95.98
MCAR_MCAR	0.059	0.115	0.009	0.116	96.88	-0.145	0.115	0.005	0.114	95.54
MAR_MAR	0.055	0.121	0.005	0.121	92.92	-0.153	0.121	-0.003	0.121	94.81
MNAR_MNAR	0.049	0.130	-0.001	0.130	94.20	-0.130	0.125	0.020	0.126	94.20
COMP_MCAR	0.056	0.112	0.006	0.112	94.20	-0.145	0.115	0.005	0.114	95.98
COMP_MAR	0.055	0.111	0.005	0.111	93.84	-0.150	0.121	0.000	0.121	94.79
COMP_MNAR	0.055	0.112	0.005	0.112	93.75	-0.130	0.125	0.020	0.126	94.20
MCAR_MAR	0.059	0.115	0.009	0.115	96.74	-0.152	0.122	-0.002	0.122	95.35
MCAR_MNAR	0.058	0.116	0.008	0.116	96.43	-0.130	0.125	0.020	0.126	94.20

Table A.2: The estimates of X covariate in scenario 2 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.047	0.107	-0.003	0.107	95.24	-0.150	0.105	0.000	0.105	94.37
MCAR_MCAR	0.047	0.116	-0.003	0.116	95.24	-0.152	0.119	-0.002	0.119	93.51
MAR_MAR	0.048	0.160	-0.002	0.159	96.54	-0.149	0.132	0.001	0.132	93.94
MNAR_MNAR	0.047	0.120	-0.003	0.120	96.97	-0.139	0.120	0.011	0.121	94.81
COMP_MCAR	0.047	0.106	-0.003	0.106	95.24	-0.151	0.114	-0.001	0.114	93.94
COMP_MAR	0.057	0.155	0.007	0.154	95.24	-0.143	0.138	0.007	0.138	92.64
COMP_MNAR	0.048	0.107	-0.002	0.106	95.24	-0.141	0.120	0.009	0.120	94.37
MCAR_MAR	0.057	0.143	0.007	0.142	94.78	-0.145	0.129	0.005	0.129	93.04
MCAR_MNAR	0.051	0.116	0.001	0.115	94.81	-0.137	0.121	0.013	0.121	93.51

Table A.3: The estimates of X covariate in scenario 3 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.060	0.105	0.010	0.105	95.69	-0.144	0.103	0.006	0.103	95.69
MCAR_MCAR	0.060	0.115	0.010	0.115	95.69	-0.140	0.112	0.010	0.112	94.83
MAR_MAR	0.058	0.182	0.008	0.182	96.12	-0.143	0.139	0.007	0.139	93.97
MNAR_MNAR	0.053	0.115	0.003	0.115	96.98	-0.150	0.118	0.000	0.118	94.40
COMP_MCAR	0.060	0.105	0.010	0.105	95.69	-0.140	0.108	0.010	0.108	93.97
COMP_MAR	0.063	0.200	0.013	0.200	94.40	-0.140	0.135	0.010	0.135	93.97
COMP_MNAR	0.047	0.136	-0.003	0.136	94.83	-0.144	0.117	0.006	0.117	95.69
MCAR_MAR	0.053	0.182	0.003	0.182	94.37	-0.138	0.142	0.012	0.142	94.37
MCAR_MNAR	0.061	0.112	0.011	0.113	93.97	-0.139	0.115	0.011	0.115	94.40

Table A.4: The estimates of X covariate in scenario 4 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.054	0.163	0.004	0.163	93.10	-0.180	0.163	-0.030	0.166	91.81
MCAR_MCAR	0.057	0.167	0.007	0.167	93.53	-0.179	0.166	-0.029	0.168	92.24
MAR_MAR	0.134	0.629	0.084	0.630	95.83	-0.114	0.683	0.036	0.680	95.83
MNAR_MNAR	0.045	0.172	-0.005	0.172	93.45	-0.167	0.174	-0.017	0.175	92.14
COMP_MCAR	0.054	0.163	0.004	0.163	93.97	-0.180	0.166	-0.030	0.168	91.81
COMP_MAR	0.052	0.334	0.002	0.332	90.91	-0.146	0.499	0.004	0.496	96.10
COMP_MNAR	0.053	0.164	0.003	0.164	93.48	-0.163	0.162	-0.013	0.162	92.61
MCAR_MAR	0.041	0.418	-0.009	0.416	94.05	-0.180	0.444	-0.030	0.442	94.05
MCAR_MNAR	0.055	0.167	0.005	0.167	93.94	-0.160	0.166	-0.010	0.166	92.64

Table A.5: The estimates of X covariate in scenario 5 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.062	0.169	0.012	0.169	96.52	-0.136	0.159	0.014	0.159	98.26
MCAR_MCAR	0.063	0.171	0.013	0.171	96.09	-0.137	0.162	0.013	0.162	98.26
MAR_MAR	0.101	1.508	0.051	1.504	81.33	-0.060	0.622	0.090	0.626	84.67
MNAR_MNAR	0.076	1.249	0.026	1.245	94.29	-0.144	0.563	0.006	0.561	94.29
COMP_MCAR	0.063	0.168	0.013	0.168	96.52	-0.137	0.161	0.013	0.161	97.39
COMP_MAR	0.475	1.402	0.425	1.365	85.71	-0.716	1.082	-0.566	1.151	85.71
COMP_MNAR	0.105	0.952	0.055	0.909	100.00	-0.007	1.099	0.143	1.058	100.00
MCAR_MAR	-0.242	0.807	-0.292	0.823	100.00	0.131	0.658	0.281	0.687	100.00
MCAR_MNAR	-0.197	0.750	-0.247	0.768	100.00	-0.283	0.542	-0.133	0.542	100.00

Table A.6: The estimates of time covariate in scenario 1 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.045	0.061	-0.005	0.061	95.54	-0.247	0.067	0.003	0.067	93.30
MCAR_MCAR	0.047	0.065	-0.003	0.065	96.43	-0.247	0.072	0.003	0.072	94.20
MAR_MAR	0.040	0.081	-0.010	0.081	91.04	-0.266	0.160	-0.016	0.161	71.70
MNAR_MNAR	0.584	0.080	0.534	0.540	0.00	0.483	0.091	0.733	0.739	0.00
COMP_MCAR	0.045	0.061	-0.005	0.061	95.54	-0.247	0.072	0.003	0.1	94.20
COMP_MAR	0.045	0.062	-0.005	0.062	94.79	-0.266	0.160	-0.016	0.160	72.99
COMP_MNAR	0.045	0.061	-0.005	0.061	94.64	0.483	0.091	0.733	0.739	0.00
MCAR_MAR	0.046	0.066	-0.004	0.066	95.81	-0.274	0.175	-0.024	0.177	71.16
MCAR_MNAR	0.047	0.065	-0.003	0.065	95.98	0.483	0.092	0.733	0.739	0.00

Table A.7: The estimates of time covariate in scenario 2 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.052	0.060	0.002	0.060	95.67	-0.248	0.063	0.002	0.063	95.24
MCAR_MCAR	0.053	0.064	0.003	0.064	96.97	-0.248	0.067	0.002	0.067	96.97
MAR_MAR	0.059	0.103	0.009	0.103	90.91	-0.293	0.242	-0.043	0.245	68.40
MNAR_MNAR	0.593	0.081	0.543	0.549	0.00	0.417	0.093	0.667	0.673	0.00
COMP_MCAR	0.052	0.061	0.002	0.060	95.67	-0.248	0.066	0.002	0.066	97.40
COMP_MAR	0.053	0.061	0.003	0.061	96.10	-0.283	0.225	-0.033	0.227	68.83
COMP_MNAR	0.057	0.062	0.007	0.062	95.24	0.211	0.102	0.461	0.472	0.00
MCAR_MAR	0.054	0.064	0.004	0.064	97.39	-0.281	0.218	-0.031	0.219	70.43
MCAR_MNAR	0.066	0.065	0.016	0.067	94.81	0.298	0.098	0.548	0.557	0.00

Table A.8: The estimates of time covariate in scenario 3 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.055	0.067	0.005	0.067	93.53	-0.248	0.064	0.002	0.064	94.83
MCAR_MCAR	0.056	0.070	0.006	0.070	95.26	-0.246	0.069	0.004	0.069	96.12
MAR_MAR	0.062	0.094	0.012	0.094	90.52	-0.295	0.248	-0.045	0.251	71.55
MNAR_MNAR	0.635	0.082	0.585	0.590	0.00	0.384	0.082	0.634	0.639	0.00
COMP_MCAR	0.054	0.067	0.004	0.067	93.10	-0.247	0.069	0.003	0.069	94.40
COMP_MAR	0.057	0.068	0.007	0.068	93.10	-0.300	0.263	-0.050	0.267	68.97
COMP_MNAR	0.061	0.069	0.011	0.070	93.10	-0.055	0.167	0.195	0.257	31.90
MCAR_MAR	0.058	0.070	0.008	0.071	95.24	-0.277	0.234	-0.027	0.235	68.40
MCAR_MNAR	0.084	0.072	0.034	0.080	91.81	0.203	0.110	0.453	0.466	1.29

Table A.9: The estimates of time covariate in scenario 4 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.045	0.042	-0.005	0.042	94.83	-0.245	0.042	0.005	0.042	95.26
MCAR_MCAR	0.045	0.047	-0.005	0.047	92.67	-0.246	0.043	0.004	0.043	96.98
MAR_MAR	0.316	0.106	0.266	0.286	37.50	-0.430	0.195	-0.180	0.264	47.22
MNAR_MNAR	0.484	0.078	0.434	0.441	0.44	0.173	0.103	0.423	0.435	0.87
COMP_MCAR	0.045	0.042	-0.005	0.042	94.40	-0.247	0.043	0.003	0.043	96.98
COMP_MAR	0.042	0.037	-0.008	0.037	97.40	-0.348	0.186	-0.098	0.209	64.94
COMP_MNAR	0.045	0.042	-0.005	0.042	94.35	0.202	0.096	0.452	0.462	0.00
MCAR_MAR	0.050	0.127	0.000	0.127	91.67	-0.332	0.169	-0.082	0.187	76.19
MCAR_MNAR	0.045	0.047	-0.005	0.047	93.94	0.201	0.101	0.451	0.462	0.43

Table A.10: The estimates of time covariate in scenario 5 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.051	0.021	0.001	0.021	96.52	-0.246	0.031	0.004	0.031	94.78
MCAR_MCAR	0.050	0.025	0.000	0.025	94.78	-0.244	0.035	0.006	0.035	92.61
MAR_MAR	0.314	0.089	0.264	0.279	15.33	-0.412	0.093	-0.162	0.186	62.00
MNAR_MNAR	0.253	0.097	0.203	0.225	23.57	-0.358	0.081	-0.108	0.135	83.57
COMP_MCAR	0.051	0.021	0.001	0.021	96.96	-0.244	0.035	0.006	0.035	93.48
COMP_MAR	0.044	0.021	-0.006	0.020	100.00	-0.332	0.107	-0.082	0.129	85.71
COMP_MNAR	0.049	0.021	-0.001	0.020	100.00	-0.361	0.106	-0.111	0.150	63.64
MCAR_MAR	0.015	0.081	-0.035	0.085	90.91	-0.367	0.080	-0.117	0.140	81.82
MCAR_MNAR	0.036	0.073	-0.014	0.072	100.00	-0.329	0.092	-0.079	0.120	94.12

Table A.11: The estimates of the intercept in scenario 1 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.208	0.149	0.008	0.149	94.20	0.089	0.153	-0.011	0.153	94.64
MCAR_MCAR	0.204	0.151	0.004	0.150	93.75	0.088	0.163	-0.012	0.163	92.41
MAR_MAR	0.215	0.162	0.015	0.163	91.51	0.118	0.245	0.018	0.246	83.49
MNAR_MNAR	-0.349	0.170	-0.549	0.575	9.82	-0.678	0.177	-0.778	0.798	0.45
COMP_MCAR	0.208	0.148	0.008	0.148	94.20	0.088	0.163	-0.012	0.163	92.41
COMP_MAR	0.209	0.148	0.009	0.148	93.36	0.116	0.242	0.016	0.242	84.36
COMP_MNAR	0.210	0.149	0.010	0.149	94.20	-0.678	0.178	-0.778	0.798	0.45
MCAR_MAR	0.205	0.151	0.005	0.150	92.56	0.129	0.266	0.029	0.267	83.26
MCAR_MNAR	0.205	0.151	0.005	0.150	94.20	-0.678	0.178	-0.778	0.798	0.45

Table A.12: The estimates of the intercept in scenario 2 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.197	0.140	-0.003	0.140	95.24	0.092	0.135	-0.008	0.135	96.10
MCAR_MCAR	0.197	0.139	-0.003	0.138	97.84	0.093	0.141	-0.007	0.141	96.10
MAR_MAR	0.195	0.182	-0.005	0.181	94.37	0.156	0.362	0.056	0.366	78.36
MNAR_MNAR	-0.379	0.156	-0.579	0.599	5.63	-0.628	0.170	-0.728	0.748	0.43
COMP_MCAR	0.197	0.140	-0.003	0.139	95.24	0.093	0.139	-0.007	0.139	96.54
COMP_MAR	0.196	0.147	-0.004	0.146	95.67	0.139	0.325	0.039	0.326	80.09
COMP_MNAR	0.185	0.141	-0.015	0.141	94.81	-0.401	0.178	-0.501	0.531	14.72
MCAR_MAR	0.194	0.150	-0.006	0.150	97.39	0.137	0.318	0.037	0.319	79.57
MCAR_MNAR	0.174	0.141	-0.026	0.143	96.54	-0.493	0.177	-0.593	0.618	6.06

Table A.13: The estimates of the intercept in scenario 3 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.180	0.150	-0.020	0.151	93.53	0.088	0.146	-0.012	0.146	96.12
MCAR_MCAR	0.179	0.152	-0.021	0.153	93.97	0.085	0.154	-0.015	0.154	96.55
MAR_MAR	0.184	0.168	-0.016	0.168	93.53	0.158	0.384	0.058	0.387	81.90
MNAR_MNAR	-0.451	0.163	-0.651	0.671	1.29	-0.611	0.158	-0.711	0.728	0.00
COMP_MCAR	0.181	0.150	-0.019	0.151	93.10	0.086	0.155	-0.014	0.155	95.69
COMP_MAR	0.191	0.168	-0.009	0.168	93.53	0.165	0.406	0.065	0.410	79.31
COMP_MNAR	0.175	0.164	-0.025	0.165	93.10	-0.115	0.241	-0.215	0.323	65.95
MCAR_MAR	0.191	0.172	-0.009	0.172	94.37	0.130	0.354	0.030	0.355	81.82
MCAR_MNAR	0.134	0.155	-0.066	0.168	91.81	-0.401	0.180	-0.501	0.532	15.09

Table A.14: The estimates of the intercept in scenario 4 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.212	0.132	0.012	0.133	96.55	0.102	0.131	0.002	0.131	96.55
MCAR_MCAR	0.212	0.137	0.012	0.137	95.69	0.103	0.134	0.003	0.134	96.55
MAR_MAR	-0.114	0.388	-0.314	0.497	68.06	0.374	0.387	0.274	0.472	77.78
MNAR_MNAR	-0.225	0.156	-0.425	0.452	19.21	-0.311	0.156	-0.411	0.439	20.09
COMP_MCAR	0.213	0.133	0.013	0.133	96.55	0.104	0.134	0.004	0.134	96.55
COMP_MAR	0.215	0.218	0.015	0.217	90.91	0.316	0.320	0.216	0.385	83.12
COMP_MNAR	0.213	0.132	0.013	0.132	96.52	-0.344	0.144	-0.444	0.467	14.78
MCAR_MAR	0.220	0.343	0.020	0.341	95.24	0.310	0.321	0.210	0.382	83.33
MCAR_MNAR	0.213	0.136	0.013	0.136	94.81	-0.344	0.145	-0.444	0.467	15.58

Table A.15: The estimates of the intercept in scenario 5 for the two outcomes respectively

type	**mean**	**std**	**bias**	**RMSE**	**CP**	**mean**	**std**	**bias**	**RMSE**	**CP**
Complete	0.191	0.125	-0.009	0.125	94.35	0.083	0.132	-0.017	0.133	95.65
MCAR_MCAR	0.191	0.131	-0.009	0.131	94.35	0.080	0.135	-0.020	0.136	95.22
MAR_MAR	-0.174	0.756	-0.374	0.841	69.33	0.209	0.315	0.109	0.332	78.00
MNAR_MNAR	-0.091	0.708	-0.291	0.763	93.57	0.231	0.360	0.131	0.382	91.43
COMP_MCAR	0.190	0.125	-0.010	0.126	94.35	0.081	0.135	-0.019	0.136	95.22
COMP_MAR	0.054	0.656	-0.146	0.625	85.71	0.467	0.629	0.367	0.689	71.43
COMP_MNAR	0.151	0.505	-0.049	0.484	90.91	0.039	0.790	-0.061	0.756	90.91
MCAR_MAR	0.464	0.433	0.264	0.490	100.00	0.032	0.443	-0.068	0.428	90.91
MCAR_MNAR	0.336	0.438	0.136	0.446	94.12	0.201	0.355	0.101	0.359	100.00

APPENDIX B

THE PARAMETER ESTIMATIONS AFTER HANDLING INCOMPLETE DATA

Table B.1: The estimates of time covariate for handling missingness in MCAR_MCAR

	True $\beta_{21} = 0.05$				True $\beta_{22} = -0.25$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0467	0.0490	0.0464	**Complete**	-0.2527	-0.2514	-0.2469
CC	0.0509	0.0455	0.0469	**CC**	-0.2507	-0.2495	-0.2440
LOCF	0.0389	0.0402	0.0362	**LOCF**	-0.2003	-0.2030	-0.1943
RI	0.0358	0.0380	0.0309	**RI**	-0.1257	-0.1267	-0.1158
MS	0.0466	0.0508	0.0450	**MS**	-0.2463	-0.2496	-0.2458
Incomplete	0.0470	0.0529	0.0453	**Incomplete**	-0.2478	-0.2477	-0.2463
		std				**std**	
Complete	0.0622	0.0609	0.0415	**Complete**	0.0677	0.0631	0.0415
CC	0.0862	0.0816	0.0579	**CC**	0.0880	0.0835	0.0574
LOCF	0.0572	0.0552	0.0410	**LOCF**	0.0650	0.0626	0.0367
RI	0.0694	0.0694	0.0568	**RI**	0.0426	0.0375	0.0534
MS	0.0693	0.0670	0.0534	**MS**	0.0762	0.0730	0.0530
Incomplete	0.0648	0.0638	0.0469	**Incomplete**	0.0718	0.0674	0.0433
		bias				**bias**	
Complete	-0.0030	-0.0008	-0.0038	**Complete**	-0.0024	-0.0010	0.0029
CC	0.0015	-0.0045	-0.0035	**CC**	-0.0006	0.0005	0.0061
LOCF	-0.0109	-0.0098	-0.0140	**LOCF**	0.0498	0.0470	0.0556
RI	-0.0139	-0.0120	-0.0195	**RI**	0.1245	0.1234	0.1339
MS	-0.0031	0.0008	-0.0054	**MS**	0.0040	0.0004	0.0041
Incomplete	-0.0027	0.0029	-0.0050	**Incomplete**	0.0026	0.0023	0.0036
		CP				**CP**	
Complete	95.60	95.20	94.80	**Complete**	92.80	95.20	95.20
CC	94.22	96.54	91.38	**CC**	94.22	97.40	95.26
LOCF	95.11	96.97	90.95	**LOCF**	86.67	87.45	71.98
RI	90.67	91.77	86.64	**RI**	53.33	50.22	27.16
MS	92.89	92.64	92.67	**MS**	91.11	92.64	93.10
Incomplete	96.44	96.97	92.67	**Incomplete**	94.22	96.97	96.98

Table B.2: The estimates of time covariate for handling missingness in COMP_MCAR

	True $\beta_{21} = 0.05$				True $\beta_{22} = -0.25$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0467	0.0490	0.0464	**Complete**	-0.2527	-0.2514	-0.2469
CC	0.0509	0.0455	0.0469	**CC**	-0.2507	-0.2495	-0.2440
LOCF	0.0449	0.0518	0.0452	**LOCF**	-0.2476	-0.2481	-0.2467
RI	0.0449	0.0507	0.0453	**RI**	-0.1287	-0.1228	-0.1176
MS	0.0448	0.0517	0.0450	**MS**	-0.2435	-0.2501	-0.2446
Incomplete	0.0449	0.0518	0.0452	**Incomplete**	-0.2476	-0.2481	-0.2467
		std				**std**	
Complete	0.0622	0.0609	0.0415	**Complete**	0.0677	0.0631	0.0415
CC	0.0862	0.0816	0.0579	**CC**	0.0880	0.0835	0.0574
LOCF	0.0612	0.0606	0.0422	**LOCF**	0.0720	0.0662	0.0433
RI	0.0611	0.0607	0.0420	**RI**	0.0434	0.0371	0.0533
MS	0.0612	0.0609	0.0420	**MS**	0.0777	0.0741	0.0532
Incomplete	0.0612	0.0606	0.0422	**Incomplete**	0.0720	0.0662	0.0433
		bias				**bias**	
Complete	-0.0030	-0.0008	-0.0038	**Complete**	-0.0024	-0.0010	0.0029
CC	0.0015	-0.0045	-0.0035	**CC**	-0.0006	0.0005	0.0061
LOCF	-0.0048	0.0018	-0.0050	**LOCF**	0.0028	0.0019	0.0032
RI	-0.0048	0.0007	-0.0049	**RI**	0.1214	0.1272	0.1322
MS	-0.0049	0.0017	-0.0052	**MS**	0.0067	-0.0001	0.0054
Incomplete	-0.0048	0.0018	-0.0050	**Incomplete**	0.0028	0.0019	0.0032
		CP				**CP**	
Complete	95.60	95.20	94.80	**Complete**	92.80	95.20	95.20
CC	94.22	96.54	91.38	**CC**	94.22	97.40	95.26
LOCF	95.56	95.67	94.40	**LOCF**	94.22	97.40	96.98
RI	96.00	96.54	94.83	**RI**	52.89	45.45	27.16
MS	95.56	96.10	94.40	**MS**	89.78	92.64	92.24
Incomplete	95.56	95.67	94.40	**Incomplete**	94.22	97.40	96.98

Table B.3: The estimates of time covariate for handling missingness in COMP_MAR

	True $\beta_{21} = 0.05$				True $\beta_{22} = -0.25$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0467	0.0490	0.0464	**Complete**	-0.2527	-0.2514	-0.2469
CC	0.0365	-0.2609	0.0443	**CC**	-0.4483	-0.4754	-0.4012
LOCF	0.0452	0.0544	0.0452	**LOCF**	-0.8588	-0.8384	-0.3508
RI	0.0447	0.0679	0.0452	**RI**	0.0280	0.0250	0.2795
MS	0.0448	0.0511	0.0452	**MS**	-0.2533	-0.2568	0.3348
IPW	0.0462	0.0537	0.0450	**IPW**	-0.2756	-0.2761	-0.2917
Incomplete	0.0444	0.0534	0.0421	**Incomplete**	-0.2687	-0.2827	-0.3476
		std				**std**	
Complete	0.0622	0.0609	0.0415	**Complete**	0.0677	0.0631	0.0415
CC	0.1176	0.1583	0.0565	**CC**	0.1266	0.1276	0.0931
LOCF	0.0618	0.0609	0.0420	**LOCF**	0.0705	0.0693	0.0419
RI	0.0622	0.0587	0.0421	**RI**	0.1079	0.0792	0.0942
MS	0.0612	0.0622	0.0425	**MS**	0.1023	0.0935	0.0659
IPW	0.0430	0.0643	0.0718	**IPW**	0.0951	0.1287	0.1326
Incomplete	0.0620	0.0614	0.0369	**Incomplete**	0.1601	0.2251	0.1863
		bias				**bias**	
Complete	-0.0030	-0.0008	-0.0038	**Complete**	-0.0024	-0.0010	0.0029
CC	-0.0128	-0.3109	-0.0063	**CC**	-0.1982	-0.2254	-0.1516
LOCF	-0.0046	0.0044	-0.0050	**LOCF**	-0.6082	-0.5884	-0.1010
RI	-0.0050	0.0179	-0.0049	**RI**	0.2781	0.2750	0.5291
MS	-0.0050	0.0011	-0.0050	**MS**	-0.0031	-0.0068	0.5843
IPW	0.0037	-0.0050	-0.0038	**IPW**	-0.0261	-0.0417	-0.0256
Incomplete	-0.0053	0.0034	-0.0079	**Incomplete**	-0.0159	-0.0327	-0.0976
		CP				**CP**	
Complete	95.60	95.20	94.80	**Complete**	92.80	95.20	95.20
CC	89.78	35.93	94.83	**CC**	54.67	47.62	65.52
LOCF	94.22	94.81	94.40	**LOCF**	0.00	0.00	35.78
RI	94.67	94.81	94.83	**RI**	9.33	3.90	0.00
MS	95.11	95.67	94.40	**MS**	80.00	82.25	0.00
IPW	94.38	97.19	93.72	**IPW**	85.54	88.35	76.23
Incomplete	94.81	96.10	97.40	**Incomplete**	72.64	68.83	64.94

Table B.4: The estimates of time covariate for handling missingness in MAR_MAR

	True $\beta_{21} = 0.05$				True $\beta_{22} = -0.25$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0470	0.0492	0.0462	**Complete**	-0.2524	-0.2510	-0.2471
CC	-0.1429	-0.3548	0.0556	**CC**	-0.4554	-0.5350	-0.4070
LOCF	-0.5123	-0.5066	-0.1094	**LOCF**	-0.8568	-0.8791	-0.3512
RI	0.0234	0.0473	0.3400	**RI**	0.0267	0.0279	0.2789
MS	0.0414	0.0481	0.5449	**MS**	-0.2531	-0.2532	0.3345
IPW	0.0412	0.0473	0.3129	**IPW**	-0.2746	-0.2733	-0.4236
Incomplete	0.0403	0.0594	0.3160	**Incomplete**	-0.2663	-0.2925	-0.4300
		std				**std**	
Complete	0.0622	0.0609	0.0415	**Complete**	0.0677	0.0631	0.0415
CC	0.1841	0.1793	0.1252	**CC**	0.1660	0.1437	0.1191
LOCF	0.0605	0.0537	0.0403	**LOCF**	0.0716	0.0726	0.0424
RI	0.0748	0.0702	0.0690	**RI**	0.1067	0.0811	0.0927
MS	0.0804	0.0796	0.0670	**MS**	0.0978	0.0933	0.0646
IPW	0.1384	0.0739	0.0989	**IPW**	0.1055	0.1285	0.1335
Incomplete	0.0805	0.1030	0.1056	**Incomplete**	0.1601	0.2419	0.1946
		bias				**bias**	
Complete	-0.0030	-0.0008	-0.0038	**Complete**	-0.0024	-0.0010	0.0029
CC	-0.1929	-0.4048	0.0056	**CC**	-0.2054	-0.2850	-0.1570
LOCF	-0.5623	-0.5566	-0.1594	**LOCF**	-0.6068	-0.6291	-0.1012
RI	-0.0266	-0.0027	0.2900	**RI**	0.2767	0.2779	0.5289
MS	-0.0086	-0.0019	0.4949	**MS**	-0.0031	-0.0032	0.5845
IPW	-0.0027	0.2629	-0.0088	**IPW**	-0.0233	-0.1736	-0.0246
Incomplete	-0.0097	0.0094	0.2660	**Incomplete**	-0.0163	-0.0425	-0.1800
		CP				**CP**	
Complete	95.58	95.18	94.78	**Complete**	92.77	95.18	95.18
CC	74.11	27.71	97.39	**CC**	66.52	36.36	76.96
LOCF	0.00	0.00	1.30	**LOCF**	0.00	0.00	35.50
RI	87.95	92.21	0.00	**RI**	8.04	3.46	0.00
MS	85.71	87.45	0.00	**MS**	81.70	82.25	0.00
IPW	93.57	90.36	36.71	**IPW**	85.94	89.16	47.44
Incomplete	91.04	90.91	37.50	**Incomplete**	71.70	68.40	47.22

Table B.5: The estimates of time covariate for handling missingness in MNAR_MNAR

	True $\beta_{21} = 0.05$				True $\beta_{22} = -0.25$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0470	0.0492	0.0462	**Complete**	-0.2524	-0.2510	-0.2471
CC	0.5821	1.1156	0.2172	**CC**	0.5204	0.5533	-0.1916
LOCF	0.4047	0.3960	0.1166	**LOCF**	0.2288	0.2287	-0.0654
RI	0.3346	0.3337	0.4229	**RI**	0.3130	0.3069	0.4012
MS	0.5890	0.6039	0.6927	**MS**	0.5231	0.5192	0.6159
IPW	0.5951	0.5919	0.4208	**IPW**	0.5139	0.4531	0.2032
Incomplete	0.5840	0.5928	0.4839	**Incomplete**	0.4834	0.4166	0.1724
		std				**std**	
Complete	0.0622	0.0609	0.0415	**Complete**	0.0677	0.0631	0.0415
CC	0.1643	0.1629	0.1228	**CC**	0.1453	0.1262	0.1094
LOCF	0.0619	0.0586	0.0423	**LOCF**	0.0532	0.0535	0.0362
RI	0.0796	0.0742	0.0697	**RI**	0.1254	0.1009	0.1127
MS	0.0853	0.0901	0.0888	**MS**	0.1052	0.1022	0.0802
IPW	0.0672	0.0823	0.0813	**IPW**	0.0834	0.0928	0.0898
Incomplete	0.0795	0.0813	0.0783	**Incomplete**	0.0910	0.0928	0.1029
		bias				**bias**	
Complete	-0.0030	-0.0008	-0.0038	**Complete**	-0.0024	-0.0010	0.0029
CC	0.5321	1.0656	0.1672	**CC**	0.7704	0.8033	0.0584
LOCF	0.3547	0.3460	0.0666	**LOCF**	0.4788	0.4787	0.1846
RI	0.2846	0.2837	0.3729	**RI**	0.5630	0.5569	0.6512
MS	0.5390	0.5539	0.6427	**MS**	0.7731	0.7692	0.8659
IPW	0.5419	0.3708	0.5451	**IPW**	0.7031	0.4532	0.7639
Incomplete	0.5340	0.5428	0.4339	**Incomplete**	0.7334	0.6666	0.4224
		CP				**CP**	
Complete	95.58	95.18	94.78	**Complete**	92.77	95.18	95.18
CC	9.38	0.00	72.73	**CC**	0.00	0.00	93.07
LOCF	0.00	0.00	60.17	**LOCF**	0.00	0.00	0.43
RI	3.13	1.30	0.00	**RI**	0.00	0.00	0.00
MS	0.00	0.00	0.00	**MS**	0.00	0.00	0.00
IPW	0.00	0.00	0.00	**IPW**	0.00	0.00	1.20
Incomplete	0.00	0.00	0.44	**Incomplete**	0.00	0.00	0.88

Table B.6: The estimates of time covariate for handling missingness in MCAR_MAR

	True $\beta_{21} = 0.05$				True $\beta_{22} = -0.25$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.0467	0.0490	0.0464	**Complete**	-0.2527	-0.2514	-0.2469
CC	0.0466	-0.2675	0.0458	**CC**	-0.4519	-0.4731	-0.4001
LOCF	0.0467	0.0474	0.0452	**LOCF**	-0.8589	-0.8427	-0.3506
RI	0.0338	0.0571	0.0289	**RI**	0.0294	0.0282	0.2794
MS	0.0438	0.0544	0.0446	**MS**	-0.2534	-0.2586	0.3357
Incomplete	0.0457	0.0538	0.0497	**Incomplete**	-0.2768	-0.2808	-0.3321
		std				**std**	
Complete	0.0622	0.0609	0.0415	**Complete**	0.0677	0.0631	0.0415
CC	0.1623	0.2122	0.0856	**CC**	0.1560	0.1648	0.1335
LOCF	0.0664	0.0634	0.0469	**LOCF**	0.0706	0.0701	0.0419
RI	0.0669	0.0652	0.0564	**RI**	0.1078	0.0851	0.0917
MS	0.0687	0.0724	0.0562	**MS**	0.0945	0.0982	0.0629
Incomplete	0.0657	0.0637	0.1275	**Incomplete**	0.1754	0.2176	0.1689
		bias				**bias**	
Complete	-0.0030	-0.0008	-0.0038	**Complete**	-0.0024	-0.0010	0.0029
CC	-0.0021	-0.3175	-0.0049	**CC**	-0.2024	-0.2231	-0.1503
LOCF	-0.0030	-0.0026	-0.0051	**LOCF**	-0.6083	-0.5927	-0.1008
RI	-0.0162	0.0071	-0.0214	**RI**	0.2792	0.2782	0.5291
MS	-0.0058	0.0044	-0.0059	**MS**	-0.0031	-0.0086	0.5852
Incomplete	-0.0039	0.0038	-0.0003	**Incomplete**	-0.0243	-0.0308	-0.0821
		CP				**CP**	
Complete	95.60	95.20	94.80	**Complete**	92.80	95.20	95.20
CC	90.67	58.87	93.53	**CC**	70.22	65.37	77.59
LOCF	96.00	96.97	92.67	**LOCF**	0.00	0.00	37.93
RI	91.56	93.94	88.36	**RI**	7.11	6.06	0.00
MS	94.22	91.77	88.36	**MS**	81.78	80.52	0.00
Incomplete	95.83	97.39	91.67	**Incomplete**	70.83	70.43	76.19

Table B.7: The estimates of intercept part for handling missingness in MCAR_MCAR

	True $\beta_{01} = 0.2$				True $\beta_{02} = 0.1$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.2068	0.2020	0.2085	**Complete**	0.0987	0.0969	0.1050
CC	0.1923	0.2175	0.2103	**CC**	0.1030	0.1072	0.1076
LOCF	0.2177	0.2144	0.2255	**LOCF**	0.0101	0.0205	0.0202
RI	0.2253	0.2288	0.2400	**RI**	-0.0202	-0.0099	-0.0144
MS	0.2051	0.2030	0.2128	**MS**	0.0861	0.1011	0.1034
Incomplete	0.2041	0.1967	0.2117	**Incomplete**	0.0886	0.0934	0.1030
		std				**std**	
Complete	0.1492	0.1403	0.1330	**Complete**	0.1542	0.1364	0.1316
CC	0.2039	0.1835	0.1787	**CC**	0.2075	0.1920	0.1843
LOCF	0.1376	0.1295	0.1344	**LOCF**	0.1512	0.1356	0.1291
RI	0.1477	0.1452	0.1324	**RI**	0.1039	0.0979	0.0938
MS	0.1575	0.1428	0.1444	**MS**	0.1704	0.1484	0.1429
Incomplete	0.1507	0.1387	0.1373	**Incomplete**	0.1634	0.1410	0.1344
		bias				**bias**	
Complete	0.0066	0.0015	0.0086	**Complete**	-0.0013	-0.0040	0.0048
CC	-0.0089	0.0175	0.0104	**CC**	0.0031	0.0072	0.0065
LOCF	0.0173	0.0144	0.0254	**LOCF**	-0.0895	-0.0795	-0.0801
RI	0.0252	0.0288	0.0403	**RI**	-0.1206	-0.1099	-0.1144
MS	0.0048	0.0030	0.0134	**MS**	-0.0142	0.0011	0.0028
Incomplete	0.0039	-0.0033	0.0120	**Incomplete**	-0.0116	-0.0066	0.0028
		CP				**CP**	
Complete	94.00	95.60	96.00	**Complete**	94.00	96.00	96.40
CC	93.78	95.67	95.69	**CC**	93.78	96.97	95.69
LOCF	96.89	96.54	96.12	**LOCF**	89.33	93.51	93.97
RI	91.56	94.37	93.53	**RI**	95.56	95.24	96.12
MS	91.56	94.81	94.40	**MS**	92.00	93.94	94.40
Incomplete	93.78	97.84	95.69	**Incomplete**	92.44	96.10	96.55

Table B.8: The estimates of intercept part for handling missingness in COMP_MCAR

	True $\beta_{01} = 0.2$				True $\beta_{02} = 0.1$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.2068	0.2020	0.2085	**Complete**	0.0987	0.0969	0.1050
CC	0.1923	0.2175	0.2103	**CC**	0.1030	0.1072	0.1076
LOCF	0.2086	0.1971	0.2130	**LOCF**	0.0884	0.0931	0.1041
RI	0.2088	0.2007	0.2128	**RI**	-0.0122	-0.0189	-0.0084
MS	0.2089	0.1972	0.2135	**MS**	0.0834	0.0936	0.1035
Incomplete	0.2086	0.1971	0.2130	**Incomplete**	0.0884	0.0931	0.1041
		std				**std**	
Complete	0.1492	0.1403	0.1330	**Complete**	0.1542	0.1364	0.1316
CC	0.2039	0.1835	0.1787	**CC**	0.2075	0.1920	0.1843
LOCF	0.1483	0.1397	0.1329	**LOCF**	0.1633	0.1387	0.1341
RI	0.1478	0.1392	0.1333	**RI**	0.1030	0.0942	0.0943
MS	0.1484	0.1400	0.1322	**MS**	0.1729	0.1477	0.1441
Incomplete	0.1483	0.1397	0.1329	**Incomplete**	0.1633	0.1387	0.1341
		bias				**bias**	
Complete	0.0066	0.0015	0.0086	**Complete**	-0.0013	-0.0040	0.0048
CC	-0.0089	0.0175	0.0104	**CC**	0.0031	0.0072	0.0065
LOCF	0.0085	-0.0029	0.0131	**LOCF**	-0.0118	-0.0069	0.0038
RI	0.0086	0.0007	0.0128	**RI**	-0.1118	-0.1189	-0.1086
MS	0.0087	-0.0028	0.0135	**MS**	-0.0167	-0.0065	0.0030
Incomplete	0.0085	-0.0029	0.0131	**Incomplete**	-0.0118	-0.0069	0.0038
		CP				**CP**	
Complete	94.00	95.60	96.00	**Complete**	94.00	96.00	96.40
CC	93.78	95.67	95.69	**CC**	93.78	96.97	95.69
LOCF	94.22	95.24	96.55	**LOCF**	92.44	96.54	96.55
RI	94.22	95.24	96.12	**RI**	97.78	94.81	96.12
MS	93.78	95.24	96.55	**MS**	90.22	92.64	94.83
Incomplete	94.22	95.24	96.55	**Incomplete**	92.44	96.54	96.55

Table B.9: The estimates of intercept part for handling missingness in COMP_MAR

	True $\beta_{01} = 0.2$				True $\beta_{02} = 0.1$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.2068	0.2020	0.2085	**Complete**	0.0987	0.0969	0.1050
CC	0.2242	1.5925	0.2306	**CC**	1.2676	1.3847	1.9707
LOCF	0.2081	0.1879	0.2127	**LOCF**	0.8124	0.7777	0.2402
RI	0.2094	0.1696	0.2124	**RI**	-0.2372	-0.2320	-0.4394
MS	0.2090	0.1997	0.2129	**MS**	0.0962	0.1044	-0.5284
IPW	0.2085	0.2003	0.2119	**IPW**	0.1380	0.1370	0.3149
Incomplete	0.2095	0.1958	0.2149	**Incomplete**	0.1195	0.1390	0.3161
		std				**std**	
Complete	0.1492	0.1403	0.1330	**Complete**	0.1542	0.1364	0.1316
CC	0.2562	0.3360	0.2116	**CC**	0.2843	0.3001	0.2463
LOCF	0.1485	0.1396	0.1328	**LOCF**	0.1550	0.1499	0.1317
RI	0.1509	0.1390	0.1333	**RI**	0.0943	0.0796	0.0833
MS	0.1492	0.1444	0.1348	**MS**	0.1914	0.1843	0.1538
IPW	0.1486	0.1479	0.1570	**IPW**	0.2203	0.2258	0.2372
Incomplete	0.1482	0.1467	0.2178	**Incomplete**	0.2422	0.3247	0.3204
		bias				**bias**	
Complete	0.0066	0.0015	0.0086	**Complete**	-0.0013	-0.0040	0.0048
CC	0.0221	1.3925	0.0316	**CC**	1.1677	1.2847	1.8704
LOCF	0.0079	-0.0121	0.0128	**LOCF**	0.7122	0.6777	0.1400
RI	0.0091	-0.0304	0.0124	**RI**	-0.3369	-0.3320	-0.5393
MS	0.0088	-0.0003	0.0129	**MS**	-0.0039	0.0044	-0.6279
IPW	0.0003	0.0119	0.0085	**IPW**	0.0370	0.2149	0.0380
Incomplete	0.0092	-0.0042	0.0149	**Incomplete**	0.0161	0.0390	0.2161
		CP				**CP**	
Complete	94.00	95.60	96.00	**Complete**	94.00	96.00	96.40
CC	92.44	1.73	96.12	**CC**	0.89	0.00	0.00
LOCF	94.22	94.81	96.12	**LOCF**	0.44	0.00	84.05
RI	94.22	94.81	96.12	**RI**	28.44	28.14	0.00
MS	93.33	94.81	96.12	**MS**	90.22	87.88	2.16
IPW	92.77	95.98	93.72	**IPW**	87.55	89.56	78.03
Incomplete	93.40	95.67	90.91	**Incomplete**	83.96	80.09	83.12

Table B.10: The estimates of intercept part for handling missingness in MAR_MAR

	True $\beta_{01} = 0.2$				True $\beta_{02} = 0.1$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.2066	0.2015	0.2086	**Complete**	0.0987	0.0960	0.1048
CC	1.1399	1.9927	1.5072	**CC**	1.2715	1.6252	1.9826
LOCF	0.8730	0.8577	0.4144	**LOCF**	0.8102	0.8447	0.2403
RI	0.2526	0.2100	-0.0537	**RI**	-0.2355	-0.2290	-0.4364
MS	0.2107	0.2032	-0.2963	**MS**	0.0972	0.0944	-0.5303
IPW	0.2150	0.2080	-0.1097	**IPW**	0.1366	0.1302	0.3695
Incomplete	0.2151	0.1947	-0.1138	**Incomplete**	0.1176	0.1562	0.3738
		std				**std**	
Complete	0.1492	0.1403	0.1330	**Complete**	0.1542	0.1364	0.1316
CC	0.4050	0.3875	0.3575	**CC**	0.3890	0.3309	0.3299
LOCF	0.1471	0.1345	0.1292	**LOCF**	0.1579	0.1538	0.1325
RI	0.1595	0.1496	0.1372	**RI**	0.0935	0.0833	0.0868
MS	0.1693	0.1599	0.1526	**MS**	0.1852	0.1801	0.1524
IPW	0.3140	0.1576	0.1875	**IPW**	0.3285	0.2256	0.2379
Incomplete	0.1624	0.1816	0.3876	**Incomplete**	0.2455	0.3624	0.3869
		bias				**bias**	
Complete	0.0066	0.0015	0.0086	**Complete**	-0.0013	-0.0040	0.0048
CC	0.9399	1.7927	1.3072	**CC**	1.1715	1.5252	1.8826
LOCF	0.6730	0.6577	0.2144	**LOCF**	0.7102	0.7447	0.1403
RI	0.0526	0.0100	-0.2537	**RI**	-0.3355	-0.3290	-0.5364
MS	0.0107	0.0032	-0.4963	**MS**	-0.0028	-0.0056	-0.6303
IPW	0.0080	-0.3097	0.0150	**IPW**	0.0302	0.2695	0.0366
Incomplete	0.0151	-0.0053	-0.3138	**Incomplete**	0.0176	0.0562	0.2738
		CP				**CP**	
Complete	93.98	95.58	95.98	**Complete**	93.98	95.98	96.39
CC	27.68	0.00	2.17	**CC**	6.25	0.00	0.00
LOCF	0.00	0.00	64.50	**LOCF**	0.45	0.00	84.85
RI	91.07	93.07	59.31	**RI**	29.02	33.77	0.00
MS	91.52	90.48	11.26	**MS**	88.39	88.31	1.30
IPW	93.17	95.18	65.82	**IPW**	87.95	89.96	78.48
Incomplete	91.51	94.37	68.06	**Incomplete**	83.49	78.36	77.78

Table B.11: The estimates of intercept part for handling missingness in MNAR_MNAR

	True $\beta_{01} = 0.2$				True $\beta_{02} = 0.1$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.2066	0.2015	0.2086	**Complete**	0.0987	0.0960	0.1048
CC	-0.3559	-0.9178	0.9871	**CC**	-0.7545	-0.7572	1.5935
LOCF	-0.1449	-0.1320	0.1399	**LOCF**	-0.3694	-0.3706	-0.0975
RI	-0.0373	-0.0359	-0.1077	**RI**	-0.4370	-0.4239	-0.5002
MS	-0.3579	-0.3922	-0.4214	**MS**	-0.7498	-0.7416	-0.7772
IPW	-0.3659	-0.3792	-0.1846	**IPW**	-0.7324	-0.6915	-0.3757
Incomplete	-0.3492	-0.3790	-0.2244	**Incomplete**	-0.6782	-0.6285	-0.3105
		std				**std**	
Complete	0.1492	0.1403	0.1330	**Complete**	0.1542	0.1364	0.1316
CC	0.3516	0.3006	0.3292	**CC**	0.3220	0.2575	0.3098
LOCF	0.1472	0.1391	0.1337	**LOCF**	0.1438	0.1317	0.1297
RI	0.1658	0.1493	0.1422	**RI**	0.0999	0.0911	0.0913
MS	0.1799	0.1669	0.1677	**MS**	0.1986	0.1811	0.1634
IPW	0.1574	0.1665	0.1648	**IPW**	0.1809	0.1811	0.1783
Incomplete	0.1699	0.1555	0.1559	**Incomplete**	0.1773	0.1698	0.1563
		bias				**bias**	
Complete	0.0066	0.0015	0.0086	**Complete**	-0.0013	-0.0040	0.0048
CC	-0.5559	-1.1178	0.7871	**CC**	-0.8545	-0.8572	1.4935
LOCF	-0.3449	-0.3320	-0.0601	**LOCF**	-0.4694	-0.4706	-0.1975
RI	-0.2373	-0.2359	-0.3077	**RI**	-0.5370	-0.5239	-0.6002
MS	-0.5579	-0.5922	-0.6214	**MS**	-0.8498	-0.8416	-0.8772
IPW	-0.5792	-0.3846	-0.5659	**IPW**	-0.7915	-0.4757	-0.8324
Incomplete	-0.5492	-0.5790	-0.4244	**Incomplete**	-0.7782	-0.7285	-0.4105
		CP				**CP**	
Complete	93.98	95.58	95.98	**Complete**	93.98	95.98	96.39
CC	62.95	2.60	29.87	**CC**	24.55	9.52	0.00
LOCF	32.59	36.80	93.07	**LOCF**	7.59	4.33	68.83
RI	64.73	67.97	44.16	**RI**	0.45	0.00	0.00
MS	9.82	4.33	3.46	**MS**	0.45	0.00	0.00
IPW	9.24	5.22	32.93	**IPW**	0.00	0.80	24.50
Incomplete	9.82	5.63	19.30	**Incomplete**	0.45	0.43	20.18

Table B.12: The estimates of intercept part for handling missingness in MCAR_MAR

	True $\beta_{01} = 0.2$				True $\beta_{02} = 0.1$		
	scen 1	**scen 2**	**scen 4**		**scen 1**	**scen 2**	**scen 4**
		mean				**mean**	
Complete	0.2068	0.2020	0.2085	**Complete**	0.0987	0.0969	0.1050
CC	0.1851	1.6227	0.2278	**CC**	1.2800	1.3951	1.9867
LOCF	0.2043	0.2040	0.2120	**LOCF**	0.8123	0.7850	0.2395
RI	0.2298	0.1878	0.2480	**RI**	-0.2392	-0.2321	-0.4358
MS	0.2127	0.1954	0.2106	**MS**	0.0945	0.1054	-0.5277
Incomplete	0.2058	0.1940	0.2198	**Incomplete**	0.1319	0.1366	0.3099
		std				**std**	
Complete	0.1492	0.1403	0.1330	**Complete**	0.1542	0.1364	0.1316
CC	0.3693	0.4678	0.2826	**CC**	0.3592	0.3936	0.3528
LOCF	0.1519	0.1398	0.1374	**LOCF**	0.1554	0.1507	0.1323
RI	0.1495	0.1383	0.1354	**RI**	0.0956	0.0891	0.0822
MS	0.1588	0.1537	0.1505	**MS**	0.1794	0.1905	0.1500
Incomplete	0.1507	0.1498	0.3429	**Incomplete**	0.2660	0.3179	0.3210
		bias				**bias**	
Complete	0.0066	0.0015	0.0086	**Complete**	-0.0013	-0.0040	0.0048
CC	-0.0182	1.4227	0.0286	**CC**	1.1808	1.2951	1.8859
LOCF	0.0041	0.0040	0.0122	**LOCF**	0.7121	0.6850	0.1393
RI	0.0297	-0.0122	0.0482	**RI**	-0.3385	-0.3321	-0.5356
MS	0.0122	-0.0046	0.0111	**MS**	-0.0055	0.0054	-0.6273
Incomplete	0.0055	-0.0060	0.0198	**Incomplete**	0.0289	0.0366	0.2099
		CP				**CP**	
Complete	94.00	95.60	96.00	**Complete**	94.00	96.00	96.40
CC	90.67	7.36	94.40	**CC**	4.89	4.76	0.00
LOCF	93.78	96.54	94.83	**LOCF**	0.44	0.00	84.91
RI	90.67	96.54	93.10	**RI**	29.78	29.00	0.00
MS	92.00	92.64	93.10	**MS**	88.44	87.45	1.72
Incomplete	92.59	97.39	95.24	**Incomplete**	82.87	79.57	83.33

REFERENCES

Ali, A., Dawson, S., Blows, F., Provenzano, E., Ellis, I., Baglietto, L., Huntsman, D., Caldas, C., and Pharoah, P. (2011). Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. *British journal of cancer*, 104(4):693–699.

Allison, P. D. (1999). *Multiple regression: A primer*. Pine Forge Press.

Bandyopadhyay, S., Ganguli, B., and Chatterjee, A. (2011). A review of multivariate longitudinal data analysis. *Statistical methods in medical research*, 20(4):299–330.

Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.

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(3):820–828.

Crowder, M. (1985). Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 229–237.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304.

Fahrmeir, L., Tutz, G., Hennevogl, W., and Salem, E. (1994). *Multivariate statistical modelling based on generalized linear models*, volume 2. Springer New York.

Gilbert, G. H., Duncan, R. P., Kulley, A. M., Coward, R. T., and Heft, M. W. (1997). Evaluation of bias and logistics in a survey of adults at increased risk for oral health decrements. *Journal of public health dentistry*, 57(1):48–58.

Gilbert, G. H., Duncan, R. P., and Vogel, W. B. (1998). Determinants of dental care use in dentate adults: six-monthly use during a 24-month period in the florida dental care study. *Social Science & Medicine*, 47(6):727–737.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, pages 10–69.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in medicine*, 22(9):1433–1446.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*, volume 451. John Wiley & Sons.

Hening, D. A. (2009). *Missing Data Imputation Method Comparison in Ohio University Student Retention Database*. PhD thesis, Ohio University.

Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer Science & Business Media.

Kang, S.-H. and Jung, S.-H. (2001). Generating correlated binary variables with complete specification of the joint distribution. *Biometrical Journal*, 43(3):263–269.

Lee, A. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician*, 47(3):209–215.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Lunn, A. D. and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika*, 85(2):487–490.

Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug Information Journal*, 34(2):525–533.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, 6(3):328–362.

O'Brien, L. M. and Fitzmaurice, G. M. (2004). Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):177–193.

Parzen, M., Ghosh, S., Lipsitz, S., Sinha, D., Fitzmaurice, G. M., Mallick, B. K., and Ibrahim, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *The annals of applied statistics*, 5(1):449.

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(20):3035–3054.

Preisser Jr, J. S. and Qaqish, B. F. (2012). A comparison of methods for generating correlated binary variates with specified marginal means and correlations.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455–463.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Rochon, J. (1996). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics*, pages 740–750.

Shao, J. and Zhong, B. (2003). Last observation carry-forward and last observation analysis. *Statistics in medicine*, 22(15):2429–2441.

Shelton, B. J., Gilbert, G. H., Liu, B., and Fisher, M. (2004). A sas macro for the analysis of multivariate longitudinal binary outcomes. *Computer Methods and Programs in Biomedicine*, 76(2):163–175.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Touloumi, G., Babiker, A., Pocock, S., and Darbyshire, J. (2001). Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study. *Statistics in medicine*, 20(24):3715–3728.

Touloumis, A., Agresti, A., and Kateri, M. (2013). Gee for multinomial responses using a local odds ratios parameterization. *Biometrics*, 69(3):633–640.

Towers, C. V., Rumney, P. J., and Ghamsary, M. G. (2010). Longitudinal study of cd4+ cell counts in hiv-negative pregnant patients. *The Journal of Maternal-Fetal & Neonatal Medicine*, 23(10):1091–1096.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research*, 23(1):42–59.

Wang, Z. and Louis, T. A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, 90(4):765–775.

BIOGRAPHICAL SKETCH

Hissah Alzahrani was born in 1983. After finishing high school, Hissah completed a Bachelor of CS/Statistics degree, and a master of Statistics at King Abdul Aziz University. Following a decade in the work force in this discipline, Hissah traveled to FSU in United States to pursue graduate work.