Analyzing Incomplete Longitudinal Binary Data using Approximate Likelihood Methods

by

Seyed Abdolmotalleb Izad-Shenas

A Thesis submitted to the
Faculty of Graduate and Post Doctoral Affairs
in partial fulfilment of the requirements for the degree of
Master of Science in Mathematics
with Concentration in Statistics

Ottawa-Carleton Institute of
Mathematics and Statistics

Department of Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada
November 2020

Abstract

In longitudinal studies, an outcome variable and a set of covariates are observed repeatedly on the same subject over time. Within-subject correlation results from repeated observations on the same subject, and in order to obtain valid estimates, any method that is used to analyze the longitudinal data, should consider this correlation.

Also, missingness in the outcome variable is a common problem in longitudinal data. It complicates the analysis, especially when the missingness is nonignorable, i.e., when the missing mechanism depends on the unobserved values of the outcome variable. In this case, usual approaches to missing data including imputation-based techniques fail to obtain valid inferences. We need to define a joint distribution to account for the missing data model and the longitudinal response model, at the same time.

In this thesis, we investigate two approximate likelihood methods, namely, bivariate pseudo-likelihood (BPL) proposed by Sinha et al. [1], and independent pseudolikelihood (IPL) proposed by Troxel et al. [2], along with the exact likelihood method, to analyze longitudinal data with nonignorable, nonmonotone missingness in the outcome variable. We use the marginal binary models, and we present results of a simulation study and results from an application of these three methods to the RAND HRS longitudinal data.

Results from our simulation study shows that with nonignorable missingness, the BPL

estimator is considerably more efficient than the IPL estimator and it yields smaller MSEs and smaller relative biases, especially when the within-subject correlation is high. Investigating standard errors of estimators in our applications on real world data shows similar findings, especially when the sample size is small. Overall, compared to Sinha et al. [1], our findings show new evidence for better performance of the BPL method over the IPL method when the length of response vector is relatively short.

I dedicate this thesis to

Fahimeh and Hannah

my beloved wife and daughter

for giving me a glimpse of infinity,

just by their mere existence.

Acknowledgments

I would like to express my deepest appreciation to Professor Sanjoy K. Sinha for supervising me as his M.Sc. student, and providing me with encouragement and patience throughout the duration of this thesis. With his ongoing support and step-by-step guidance, I was able to conduct necessary research, and complete this work. Without his invaluable contribution, knowledge and experience, I had no chance to write this thesis or conduct accompanying research. I have been honored for having the opportunity to be his M.Sc. student.

I would also like to extend my sincere thanks to Nicole Gaertner, the Graduate Administrator of the school, and other people in the school of Mathematics and Statistics at Carleton University, for their amazing helps and valuable suggestions throughout my studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Longitudinal studies are different from cross-sectional designs in that, for each individual subject, the outcome is *repeatedly* measured over a period of time. Instead of having a fixed, one-time record of the data, as is the case with cross-sectional designs, in longitudinal studies one can introduce *time* element into the analysis, and gain the capacity to analyse additional parameters, such as *cohort*, and *age* effects [3], which are not available otherwise. Longitudinal designs are used widely, and a good example can be found in Multicenter AIDS Cohort Study (MACS), in which 369 infected AIDS patients are followed over time, and their disease progression is monitored by repeated measurements of CD4+ cell counts in their blood [3]. The complete MACS dataset has 2376 CD4+ cell count measurements, where the outcome was measured more than 6 *times*, on average, for each subject, and by incorporating the time component into their study, researchers introduce the within-subject correlation structure. This is not a trivial correlation structure, and if we ignore it during our data analysis, and try to rather infer by using classical statistical methods, we may get biased, inefficient estimates of regression parameters.

There is a special case where the outcome variable in a longitudinal design is *binary*. There is remarkable amount of literature on methods for the analysis of longitudinal

binary data; however, this field of data analysis and estimation is complicated in two ways. The first problem arises when one tries to estimate regression parameters in the presence of nuisance parameters. When we analyze longitudinal binary data, our main focus is on modelling the effect of a covariate vector on *marginal* expectations. Although we need to account for *within-subject* correlation parameters, they are considered *nuisance* parameters. Also, as the length of the response vector increases, i.e., as the number of time points increases, the nuisance parameters *proliferate* rapidly, which makes classical estimation methods, e.g., maximum likelihood estimation, ineffective. Therefore, we need alternative estimation methods which treat nuisance correlation parameters differently, and focus on effectively estimating main parameters of interest.

On the other hand, the second problem arises because marginal modelling of longitudinal binary data is often complicated by the missingness in the response vector, especially when the missingness at a specific time point depends on the unobserved value of outcome at that time point, i.e., when the missing data mechanism in *nonignorable*. We focus on a subclass of missing pattern, in which a subject may return to study after an initial missingness, i.e., *nonmonotone* missingness. Nonignorable missingness is common in different research settings. For example, in clinical trials, when the dependent variable is defined as a dichotomous outcome, e.g., response/lack of response to treatment by a new medication, and a patient who feels sicker at a designated time point is more likely to show up for a clinic visit at that time point, as compared to a patient who is feeling less sick.

Estimation methods used commonly in marginal modelling of longitudinal binary data suffer from the presence of nonignorable nonmonotone missingness. Generalized estimating equations and relevant methods are sensitive to misspecification in models. Maximum likelihood-based estimators in the multivariate Bahadur model [4] are only of practical use when the length of the response vector is ≤ 3. Approximate likelihood

methods have been developed to overcome the difficulties in modelling longitudinal binary data with nonignorable missingness in the outcome variable. In this thesis, we focus on analyzing longitudinal binary data with nonmonotone and nonignorable missingness in the response vector. Our main objective is to compare the performance of two pseudolikelihood methods, namely, *Bivariate Pseudolikelihood*, proposed by Sinha et al. [1], and *Independent Pseudolikelihood*, proposed by Troxel et al. [2] along with the maximum likelihood estimation.

We organize this thesis into six chapters. In Chapter 2, we review generalized linear models and their estimation methods. An extension of these models to the longitudinal binary case is given in Chapter 3, where a review of relevant literature on estimating marginal models with complete binary data is also given. Chapter 4 presents a literature review on missing data problems in the context of longitudinal designs, in general, and also discusses two approximate likelihood methods for analyzing incomplete longitudinal binary data with nonignorable missingness, namely bivariate and independent pseudolikelihood. In Chapter 5, the results of a simulation study are presented, and the performance of the approximate and exact maximum likelihood estimators in longitudinal binary models is assessed. In Chapter 6, an application of these three methods using the RAND HRS longitudinal data is presented. Our conclusions are stated in Chapter 7.

Chapter 2

Generalized Linear Models

In this chapter, we review the generalized linear model (GLM). We discuss components of a GLM, likelihood functions for GLMs, and their estimation methods. We rely on [5], [6], and [7] for the review. We also discuss the GLM for a special case when the outcome variable is binary, and present the logit model as an example. This chapter will serve to build the ground for next chapters on longitudinal binary modelling.

2.1 Introduction

A common problem arises when working with longitudinal data sets in health sciences, medical, biological, and related areas, where the dependent variable is defined as a dichotomous response, i.e., as a binary outcome. For example, in clinical trials it is very common to look for relative efficacy of an intervention among test groups based on some cut-off levels, and conclude whether the subject falls into the success or failure group.

The classical linear regression method fails in two circumstances: when the outcome variable is restricted to binary or count data; and when the variance of the outcome depends on its mean. In order to overcome the requirement of a continuous outcome

variable, earlier efforts used log-normal transformations in order to apply the probit model [8]; however, the advent of the GLM [9] has been a turning point in developing a comprehensive way to deal with non-linearity issues.

Let $Y = (y_1, ..., y_n)'$ be a vector of n realized observations from a random variable Y, which are independently distributed, with mean vector μ. Also, let X be a $n \times p$ matrix of covariates, and μ and β be $n \times 1$ and $p \times 1$ vectors, respectively. For ordinary linear models, depending on preference in indexing covariates and parameters, or assuming a matrix notation, a specification of vector μ in terms of unknown parameters $\beta = (\beta_1, ..., \beta_p)'$ is denoted either as

$$\mu = \sum_{j=1}^{p} X_j \beta_j,$$

or as

$$\mu = X\beta,$$

where βs are parameters to be estimated. Indexing an observation by i, the systematic component of the ordinary linear model is given by

$$E(Y_i) = \mu_i = \sum_{j=1}^{p} X_{ij} \beta_j, \qquad i = 1, ..., n,$$

where x_{ij} is the jth covariate for observation i. For the systematic component, it is assumed that all covariates are observed, and they are measured effectively and without error. For the random component, an assumption of independence and constant variance of errors is necessary. The components of the random variable Y are independent, normally distributed, with the constant variance of σ^2 and mean

response

$$E(Y) = \mu,$$

where $\mu = X\beta$. Now, we review components of a GLM.

2.2 Components of GLM

The GLM is a generalization of the ordinary linear model.

1. **Systematic Part** or **Linear Predictor** (η): The linear predictor η is a linear function of regressors, and depending on the preference in indexing covariates and parameters, or assuming a matrix notation, is denoted either by

$$\eta = \sum_{j=1}^{p} X_j \beta_j,$$

or by

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi},$$

for $i = 1, ..., n$.

2. **Random Part** or **Variance Function** (Y): The random component denotes the conditional distribution of the response variable Y_i, given model covariates; along with a *variance function*, which is denoted as

$$var(y_i) \equiv \phi v(\mu_i),$$

where the dispersion parameter ϕ is a constant. For some specific functions $a(.)$, $b(.)$, and $c(.)$, vector Y is assumed to consist of independent measurements from

the exponential family of distributions, that is,

$$Y_i \sim indep.\ f_{Y_i}(y_i),$$

where

$$f_{Y_i}(y_i) = \exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} - c(y_i, \phi)\right\}. \tag{2.1}$$

The parameters θ and ϕ are called *natural* and *dispersion* parameters, respectively.

3. **Link Function** $g(.)$: This function provides a link between the systematic and random components, and is a smooth, invertible function which transforms $\mu_i = E(Y_i)$ to the systematic component, such that

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}, \tag{2.2}$$

or, equivalently,

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_1 x_{1i} + ... + \beta_p x_{pi}).$$

Some of commonly used link functions along with their inverses are listed in Table 2.1.

Table 2.1: Common Link Functions.

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Logit	$\log \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$

2.3 Likelihood Function for GLM

By indexing elements of Y, and using (2.1), a general form of the log-likelihood function for β under the GLMs is given by

$$\ell(\beta) = \sum_{i=1}^{n} \ell_i = \sum_{i=1}^{n} \log f(y_i, \theta_i, \phi),$$

where ℓ_i depends on the model parameters β. Using this general form, the log-likelihood for a single observation y_i is obtained as

$$\ell(\theta_i, \phi; y_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi),$$

and the log-likelihood for n independent observations is given by

$$\ell(\theta, \phi; y) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}. \tag{2.3}$$

Under regularity conditions

$$E\left[\frac{\partial \log f_{Y_i}(y_i)}{\partial \theta_i}\right] = 0, \quad \text{and} \tag{2.4}$$

$$var\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \theta_i}\right) = -E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \theta_i^2}\right]. \tag{2.5}$$

Then using (2.1) and (2.4), we get

$$E\left[\left\{y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}\right\}/\phi\right] = 0,$$

or

$$E(y_i) = \mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}. \tag{2.6}$$

Also, using (2.1), (2.5), and (2.6), we obtain

$$var\left(\left\{y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}\right\}/\phi\right) = -E\left[-\frac{1}{\phi}\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}\right],$$

or

$$var\left(\frac{y_i - \mu_i}{\phi}\right) = \frac{1}{\phi}\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$$

or

$$var(y_i) = \phi\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2},$$

$$\equiv \phi v(\mu_i), \tag{2.7}$$

where the *variance function* $v(\mu_i)$ provides the variance of y_i and is dependent on the mean of y_i. Two other useful results that follow from the previous results include

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{v(\mu_i)}, \tag{2.8}$$

and

$$\frac{\partial \mu_i}{\partial \beta} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} x_i. \tag{2.9}$$

2.4 Likelihood Estimation for GLM

Recall the GLM given by equation (2.2). This model provides the expected values of n observations in terms of a limited number of regression parameters β. To obtain their estimating equations, we differentiate the log-likelihood function (2.3) with respect to each regression parameter. We use the chain rule of differentiation given by

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}.\frac{\partial \theta_i}{\partial \mu_i}.\frac{\partial \mu_i}{\partial \eta_i}.\frac{\partial \eta_i}{\partial \beta_j}, \qquad \text{for all } j = 0, 1, 2, \cdots, p.$$

Then summing over n independent observations, we obtain

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= \frac{1}{\phi}\sum_{1}^{n}\left[y_i\frac{\partial \theta_i}{\partial \beta} - \frac{\partial b(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta}\right] \\
&= \frac{1}{\phi}\sum_{1}^{n}(y_i - \mu_i)\frac{\partial \theta_i}{\partial \beta} \qquad \text{(by using (2.6))} \\
&= \frac{1}{\phi}\sum_{1}^{n}(y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta}.
\end{aligned}$$

Now, letting $w_i = [v(\mu_i)g_\mu^2(\mu_i)]^{-1}$ and using (2.8) and (2.9), we obtain

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= \frac{1}{\phi}\sum_{1}^{n}\frac{(y_i - \mu_i)}{v(\mu_i)g_\mu(\mu_i)}x_i \\
&= \frac{1}{\phi}\sum_{1}^{n}(y_i - \mu_i)w_i g_\mu(\mu_i)x_i.
\end{aligned}$$

Let W be a diagonal matrix with diagonal elements equal to w_i, and let Δ be a diagonal matrix with diagonal elements equal to $g_\mu(\mu_i)$. The **maximum likelihood equations for GLM** are given by

$$\mathbf{X'W\Delta y} = \mathbf{X'W\Delta}\mu,$$

where W, Δ and μ involve the unknown β. Then the score functions for GLM in matrix form is obtained as

$$\frac{1}{\phi}\mathbf{X'W\Delta}(y-\mu) = 0. \tag{2.10}$$

The second derivative of the log-likelihood function given in (2.3) is obtained as

$$\frac{\partial^2 \ell}{\partial\beta\partial\beta'} = -\frac{1}{\phi}\mathbf{X'W\Delta}\frac{\partial\mu}{\partial\beta'} + \frac{1}{\phi}\mathbf{X'}\frac{\partial\mathbf{W\Delta}}{\partial\beta'}(y-\mu),$$

which denotes **the Hessian** matrix. The expected value of the negative of the Hessian, $E(H)$, is obtained as

$$\begin{aligned} -E\left[\frac{\partial^2 \ell}{\partial\beta\partial\beta'}\right] &= \frac{1}{\phi}\mathbf{X'W\Delta}\frac{\partial\mu}{\partial\beta'} + 0 \\ &= \frac{1}{\phi}\mathbf{X'W\Delta\Delta^{-1}X}, \end{aligned}$$

or

$$E(H) = \frac{1}{\phi}\mathbf{X'WX},$$

and given the fact that

$$-E\left[\frac{\partial^2 \ell}{\partial\beta\partial\phi}\right] = 0,$$

the **asymptotic variance** of the ML estimator $\hat{\beta}$ is given by

$$var_{\infty}(\hat{\beta}) = \phi(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.$$

There is no analytic solution to the likelihood equations in the forms presented in (2.10). However, these likelihood equations can be solved using numerical methods. For example, an iterative equation based on **Fisher scoring** technique is given by

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} + I(\beta^{(m)})^{-1}\frac{\partial \ell}{\partial \beta}\bigg|_{\beta=\beta^{(m)}} \\ &= \beta^{(m)} + \{E(H)\}.\left(\frac{\partial \ell}{\partial \beta}\right) \\ &= \beta^{(m)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(y-\mu),\end{aligned}$$

where superscript (m) denotes the mth iteration of the algorithm.

2.5 Quasi-likelihood Estimation for GLM

In absence of *a priori* knowledge of the distribution of Y, we derive likelihood-like functions with fewer assumptions. Starting from (2.4) and (2.5), we define their quasi-likelihood equivalents by differentiating with respect to μ_i instead of θ_i. Therefore, the quasi-likelihood analogue of (2.4) is obtained as

$$E\left[\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right] = 0, \tag{2.11}$$

where $Y = (y_1, \cdots, y_n)'$ is a random variable with $E(Y_i) = (\mu_i)$, and $var(Y_i) \propto v(\mu_i)$. The quasi-likelihood analogue of (2.5), denoted by V, is obtained as

$$
\begin{aligned}
V &= var\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) \\
&= var\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\right) \qquad \text{(using the chain rule)} \\
&= \left[var\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \theta_i}\right)\right]\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2,
\end{aligned}
$$

and using (2.5), we obtain

$$
\begin{aligned}
V &= \left(-E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \theta_i^2}\right]\right)\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2 \\
&= \frac{1}{\phi}\left[\frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}\right]\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2.
\end{aligned}
$$

Noting the definition and derivation of $v(\mu_i)$ in (2.7), we have

$$
V = [v(\mu_i)/\phi]\left(\frac{\partial \theta_i}{\partial \mu_i}\right)^2.
$$

Now, using (2.8) we get

$$
V = [v(\mu_i)/\phi]\frac{1}{v(\mu_i)^2},
$$

or, equivalently,

$$
var\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) = \frac{1}{\phi v(\mu_i)}. \tag{2.12}
$$

Thus

$$
var\left(\frac{\partial \log f_{Y_i}(y_i)}{\partial \mu_i}\right) = -E\left[\frac{\partial^2 \log f_{Y_i}(y_i)}{\partial \mu_i^2}\right] = \frac{1}{\phi v(\mu_i)}. \tag{2.13}
$$

We define a quantity q_i such that it has the properties shown in (2.11) and (2.13), when it is replaced by $\partial \log f_{Y_i}(y_i)/\partial(\mu_i)$. This quantity is given by

$$q_i = \frac{y_i - \mu_i}{\phi v(\mu_i)},$$

and it is easy to show that

$$E(q_i) = 0, \tag{2.14}$$

and

$$-E(\partial q_i/\partial \mu_i) = \frac{1}{\phi v(\mu_i)}, \tag{2.15}$$

and q_i is the first derivative of a quantity Q_i, given by

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi v(t)} dt,$$

with respect to μ_i. We simply denote q_i as the log quasi-likelihood contribution of y_i. Then maximum quasi-likelihood (**MQL**) estimator of β is the solution to the MQL equations given by

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{n} Q_i = 0,$$

which gives

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi v(\mu_i)} \frac{\partial \mu_i}{\partial \beta} = 0,$$

and using (2.9) we obtain

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi v(\mu_i) g_\mu(\mu_i)} x_i' \;=\; 0,$$

or, equivalently, in matrix notation

$$\frac{1}{\phi} \mathbf{X}' \mathbf{W} \boldsymbol{\Delta} (y - \mu) \;=\; 0. \tag{2.16}$$

By comparing (2.16) with (2.10), one can see that both the MQL and the likelihood estimates for the GLM are exactly the same. The reason is that when we construct Q_i, we only use information about the variance-mean relationship. It is the same information we use when we build traditional likelihood functions. The Δ matrix for (2.16) and (2.10) is given by

$$\Delta = \begin{bmatrix} \frac{\partial \theta_1}{\partial \eta_1} & 0 & \cdots & 0 \\ 0 & \frac{\partial \theta_2}{\partial \eta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \frac{\partial \theta_n}{\partial \eta_n} \end{bmatrix}.$$

Despite giving same results, the MQL method has certain advantages over the ML method. The MQL does not need certain assumptions for the distribution of Y. Also, it assumes $var(Y_i) \propto v(\mu_i)$ up to a certain constant, ϕ, so that the MQL method is useful to accommodate overdispersion in the outcome variable Y.

2.6 GLM for Binary Outcome Variable

An important implication of GLM is its application when the outcome variable is not continuous. More specifically, when the dependent variable Y takes a binary form, e.g., in clinical trials when deciding on patient outcomes, in credit decisions and approvals, or email processing and filtering of spam emails. Binary GLM models are specified by their systematic component, given by

$$\eta = g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

and their random component specified by the conditional distribution

$$Y|X \sim Bernoulli(p).$$

In this case, the exponential family distribution discussed in (2.1) is a Bernoulli distribution. The link function $g(.)$ maps $y \in [0, 1]$ to the entire real line, $(-\infty, +\infty)$. Three most common link functions associated with the binary GLM are

$$\eta_i = \log\left(\frac{p_i}{1 - p_i}\right),$$

for the logit model,

$$\eta_i = \Phi^{-1}(p_i)$$

for the probit model, where Φ^{-1} is the inverse of standard normal distribution function, and

$$\eta_i = \log(-\log(1-p_i)),$$

for the complementary log-log model.

2.6.1 The Logit model: ML estimation

Let $Y = (y_1, ..., y_n)'$ represent a vector of n Bernoulli trials, and let p_i denote the probability of success for the ith trial. Also, let X denote a matrix of covariates where x_i' denotes its ith row, and let β denote a $p \times 1$ vector of regression coefficients. Let

$$p_i = \ Pr(y_i = 1|x_i) = p_i(x_i'\beta)$$

be the probability of success for the ith subject. Then

$$1 - p_i = \ Pr(y_i = 0|x_i) = 1 - p_i(x_i'\beta)$$

represents the probability of failure. We assume

$$E(y_i|x_i) = p_i(x_i'\beta) = \ \frac{1}{1 + e^{(-x_i'\beta)}},$$

and

$$y_i|x_i \sim \ Bernoulli\left(\frac{1}{1 + e^{(-x_i'\beta)}}\right),$$

for the systematic and random components of the logit model, respectively. For the logit model, we can write

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = x_i'\beta, \tag{2.17}$$

or, equivalently, in terms of odds,

$$\frac{p_i}{1-p_i} = \exp(x_i'\beta).$$

The logit model can also be written in th form

$$p_i = E(y_i|x_i,\beta) = Pr(y_i = 1|x_i,\beta) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}. \tag{2.18}$$

To estimate the regression coefficients β, we can use the maximum likelihood method. Recall the Bernoulli density

$$f_{Y_i}(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

The likelihood of β for a single observation, $\{y_i, x_i\}$, is given by

$$L(\beta; y_i, x_i) = p_i^{y_i}(1-p_i)^{1-y_i},$$

and, by independence of all trials, the likelihood for n observations is obtained as

$$L(\beta; y, X) = \prod_{i=1}^{n}[p_i(x_i'\beta)]^{y_i}[1-p_i(x_i'\beta)]^{1-y_i}.$$

The log-likelihood is given by

$$\ell(\beta; y, X) = \sum_{i=1}^{n} y_i \log p_i(x_i'\beta) + (1 - y_i) \log[1 - p_i(x_i'\beta)]. \tag{2.19}$$

Taking the derivative of the log-likelihood with respect to β, we have (after some algebra)

$$\begin{aligned}\frac{\partial \ell(\beta)}{\partial \beta} &= \sum_{i=1}^{n} \frac{\partial \ell_i(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \ell_i(\beta)}{\partial p_i} . \frac{\partial p_i}{\partial \eta_i} . \frac{\partial \eta_i}{\partial \beta} \\ &= \sum_{i=1}^{n} \{y_i - p_i(x_i'\beta)\} x_i.\end{aligned}$$

The score function is defined by

$$S(\beta; y, X) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \{y_i - p_i(x_i'\beta)\} x_i.$$

The Hessian is obtained as

$$H(\beta; y, X) = -\sum_{i=1}^{n} p_i(1 - p_i) x_i x_i'.$$

Using the Newton-Raphson method with Fisher scoring, the ML estimate of β may be obtained from the iterative equations

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} + I(\beta^{(m)})^{-1} \frac{\partial \ell}{\partial \beta}\bigg|_{\beta=\beta^{(m)}} \\ &= \beta^{(m)} + \{\mathbf{X}'\mathbf{W}\mathbf{X}\}^{-1}\{\mathbf{X}'(y - p)\},\end{aligned}$$

where $p = (p_1, p_2, \cdots, p_n)'$, and $p_i = E(y_i|x_i, \beta)$. Also, W is a diagonal matrix in form

$$W = \begin{bmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & p_n(1-p_n) \end{bmatrix}.$$

Chapter 3

Longitudinal Models

Longitudinal studies are a heterogeneous group of studies, mainly observational in nature, which are used in different domains of science, such as medicine, biology, and social sciences, in order to evaluate the relationship between covariates and outcomes, over a period of time. These studies can be prospective, retrospective, or can simply include a set of repeated cross sectional studies. Advantages of longitudinal studies are many, which include the ability to follow the change in the response over time. However, statistical analysis of longitudinal data are often complex, as it needs to deal with additional within-subject correlations inherent to these data, and also, to deal with the problem of incomplete data, which are very common in longitudinal settings.

The variability in longitudinal data originates from two sources. The first type of variation comes from correlated measurements within an individual subject over time. The second type of variation lies in measurements across different subjects. When the outcome variable is continuous, the modelling of these variations is straightforward, where assuming an approximate Gaussian distribution for the outcome variable, we can simply use traditional mean and covariance parametrizations [10] to model the data set. On the other hand, in case of a non-continuous outcome variable, e.g., longitudinal binary data, the modelling is complicated. We review available literature

on modelling complete longitudinal binary data in some details here.

3.1 Marginal Models for Longitudinal data

Marginal models are a group of regression models that characterize the *marginal expectation* of the outcome variable Y as a function of covariates. In contrast to random effects and transitional models, marginal models perform the regression of the outcome variable on available covariates, separately from modelling within-individual associations of the outcome variable. The main focus of marginal models is on inference about population averages.

Assume a longitudinal design on N independent individuals in which an outcome variable and a set of covariates are measured over a predefined period of T. Let Y_{ij} denote the outcome (or, equivalently, response) variable, either continuous or discrete, for the ith individual on the jth occasion, where $i = 1, \cdots, N$, and $j = 1, \cdots, n_i$. Then $Y_i = (y_{i1}, ..., y_{n_i})'$ denotes the vector of n_i response measurements for the ith individual. In **unbalanced** longitudinal data, different subjects may not be measured on a common set of time points, and to account for that, we assume that each Y_{ij} is observed at time t_{ij}. Also, let $x_{ij} = (x_{1ij}, \cdots, x_{pij})'$ denote a vector of associated covariates with each instance of Y_{ij}. This vector may include time-varying (such as age at each time point, time since baseline, and income status at each time point) or time-invariant (such as gender, identity, and race) variables. Then any marginal model assumes three main features, as discussed below.

1. **Conditional expectation** of each response Y_{ij} depends on covariates through a known link function $g(.)$, and given $E(Y_{ij}|x_{ij}) = \mu_{ij}$, we obtain

$$g(\mu_{ij}) = \eta_{ij} = x'_{ij}\beta.$$

2. **Conditional variance** of each response Y_{ij} given x_{ij}, depends on the mean, as

$$var(Y_{ij}) = \phi v(\mu_{ij}),$$

where v is a known variance function, and ϕ is a scalar parameter to be estimated.

3. **Within-subject association** of repeated responses, or correlation between Y_{ij} and Y_{ik} is a function of marginal means and an additional vector of association parameters, α, such that

$$\rho(Y_{ij}, Y_{ik}) = h(\mu_{ij}, \mu_{ik}, \alpha).$$

The ML estimation for marginal models requires full specifications of all three features mentioned above. The first two features are simple extensions to the components of a GLM discussed earlier (see Section 2.2). However, specifying within-subject association is a real challenge when the outcome variable is discrete. For this reason, as we will see later, the likelihood estimation of marginal models is not straightforward, and we need **approximate** and **non-likelihood** methods.

3.1.1 Marginal Models for Longitudinal Binary Data

Assume a longitudinal study of N independent subjects observed at T time points. Then each subject from $i = 1, \cdots, N$ has a $T \times 1$ vector of binary responses denoted by $Y_i = (Y_{i1}, \cdots, Y_{iT})'$, such that

$$Y_{it} = \begin{cases} 1, & \text{if subject } i \text{ is a success at time } t \\ 0, & \text{otherwise.} \end{cases}$$

Also, let $x_{it} = (x_{1it}, \cdots, x_{pit})'$ be a $P \times 1$ vector of covariates for subject i at time

point t. Then the marginal model for the Y_{it} is given by

$$f(y_{it}|x_{it}) = \exp[y_{it}.\theta_{it} - \log(1 + \exp(\theta_{it}))], \tag{3.1}$$

where θ_{it} is the link function for the logit, and is given here as

$$\begin{aligned}\theta_{it} = logit(\mu_{it}) &= \log \frac{\mu_{it}}{1 - \mu_{it}} \\ &= \log \frac{Pr(Y_{it} = 1)}{Pr(Y_{it} = 0)} = x_{it}^{'}\beta.\end{aligned}$$

Then the conditional expectation of each response denoted by $E(Y_{it})$, conditional variance of each response denoted by $var(Y_{it})$, and within-subject association of repeated responses denoted by $\rho(Y_{it}, Y_{it'})$ are obtained as

$$\begin{aligned}E(Y_{it}) &= Pr(Y_{it} = 1|x_{it}, \beta) \\ &= \mu_{it} = \mu_{it}(\beta), \\ var(Y_{it}) &= \mu_{it}(1 - \mu_{it}),\end{aligned}$$

and

$$\rho(Y_{it}, Y_{it'}) = \alpha, \tag{3.2}$$

respectively.

3.1.1.1 Association of Binary Outcomes

Consider equation (3.2). We are interested in measuring the association between two categorical, or more specifically, between two binary random variables [11]. Let X_i, $i = 1, 2, ..., T$ and Y_k, $k = 1, ..., T'$ to be two categorical random variables. Then the

total mass in two-dimensional distribution of $\{x_i, y_k\}$ is given by

$$p_{ik} = Pr(X_i = x_i, Y_i = y_k),$$

where $\sum_{i,k} p_{ik} = 1$. Now, p_{ik} makes a contingency matrix with T rows and T' columns, where sums of rows and sums of columns are denoted by "$p_{i.}$" and "$p_{.k}$", respectively. In fact, $p_{i.}$ and $p_{.k}$ denote the marginal distributions of X_i and Y_k, respectively, and are obtained as

$$p_{i.} = Pr(X = x_i) = \sum_k p_{ik},$$

and

$$p_{.k} = Pr(Y = y_k) = \sum_i p_{ik}.$$

The mean square contingency of X_i and Y_k distribution is obtained as

$$r_\phi = \sum_{i,k} \frac{p_{ik} - p_{i.}p_{.k}}{\sqrt{p_{i.}p_{.k}}},$$

where the coefficient r_ϕ is a measure of the correlation between two categorical random variables X_i and Y_k, with the similar interpretation to the Pearson correlation ρ.

In the special case of $T = T' = 2$, we obtain binary random variables. For any two binary random variables X_i and Y_k, the contingency table denotes joint distribution of X_i and Y_k, and is given by

X,Y	$Y_k = 1$	$Y_k = 0$	Total
$X_i = 1$	p_{11}	p_{10}	$p_{1.}$
$X_i = 0$	p_{01}	p_{00}	$p_{0.}$
Total	$p_{.1}$	$p_{.0}$	$p_{..}$

Then the coefficient r_ϕ is obtained as

$$r_\phi = \frac{(p_{11}p_{00} - p_{10}p_{01})}{\sqrt{p_{1.}p_{0.}p_{.1}p_{.0}}}. \tag{3.3}$$

For any binary random variable, say Y_k, probabilities of success and failure are denoted by the marginal probabilities

$$p_{.1} = E(Y_k = 1) = \mu_{Y_k},$$

and

$$p_{.0} = 1 - E(Y_k = 1) = 1 - \mu_{Y_k},$$

respectively. Therefore, we equivalently obtain

$$r_\phi = \frac{(p_{11}p_{00} - p_{10}p_{01})}{\sqrt{\mu_{X_i}(1 - \mu_{X_i})\mu_{Y_k}(1 - \mu_{Y_k})}}. \tag{3.4}$$

Now, to describe the odds ratio of any two binary random variables X_i and Y_k, we define the conditional distributions

$$Pr(Y_k|X_i = 0),$$

and

$$Pr(Y_k|X_i = 1),$$

which are shown in the contingency table below, as derived from the joint distribution of X_i and Y_k:

X,Y	$Y_k = 1$	$Y_k = 0$
$X_i = 1$	$a = p_{11}/(p_{11} + p_{10})$	$b = p_{10}/(p_{11} + p_{10})$
$X_i = 0$	$c = p_{01}/(p_{01} + p_{00})$	$d = p_{00}/(p_{01} + p_{00})$

Then the odds ratio is defined by

$$OR = \frac{Pr(Y_k = 1|X_i = 1)Pr(Y_k = 0|X_i = 1)}{Pr(Y_k = 1|X_i = 0)Pr(Y_k = 0|X_i = 0)} = \frac{a/b}{c/d},$$

or, equivalently,

$$OR = \frac{p_{11}p_{00}}{p_{10}p_{01}}. \tag{3.5}$$

An advantage of using odds ratios over correlations is the fact that correlation between any two binary outcomes constrains the joint probability distribution of $Pr(X_i = 1, Y_i = 1)$ by their expectations.

3.2 Estimation of Longitudinal Binary Models

In this section, we review methods for estimating marginal models with binary outcomes. We discuss likelihood-based methods, pseudolikelihoods, and non-likelihood methods including generalized estimating equations. We rely on [3], [12], and [13] for the review.

3.2.1 Likelihood-based Methods

To build a ground for the next chapters of the thesis, we introduce likelihood-based approaches to analyzing longitudinal binary models. We specifically review the maximum likelihood and the Bahadur model for multivariate binary data.

3.2.1.1 The Maximum Likelihood Estimation

The maximum likelihood equations are obtained by the naive assumption of independence of the elements of the response vector Y_i, which is an equivalent representation for Independence Estimating Equations of Liang and Zeger [14]. Relying on results obtained in Section 2.4, we start from equation (3.1). We obtain the joint distribution of binary responses for subject i as

$$f(y_i|X_i) = \exp\left[\sum_{t=1}^{T} y_{it}.\theta_{it} - \sum_{t=1}^{T} \log(1+\exp(\theta_{it}))\right].$$

Then the contribution of each subject i to the log-likelihood equations is obtained by

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta} &= \left(\frac{\partial \mu_i}{\partial \beta}\right)' \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \ell_i}{\partial \theta_i} \\ &= \left(\frac{\partial \mu_i}{\partial \beta}\right)' \frac{y_i - \mu_i}{var(Y_i)}.\end{aligned}$$

Using the matrix form

$$\left(\frac{\partial \mu_i}{\partial \beta}\right)' = X_i'\Delta_i,$$

where Δ_i is a $T \times T$ diagonal matrix with the diagonal elements $var(Y_{i1}), \cdots, var(Y_{iT})$, the maximum likelihood estimator $\hat{\beta}$ is obtained as a solution to

$$\sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \beta} = \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta} \right)' \frac{y_i - \mu_i}{var(Y_i)}$$
$$= \sum_{i=1}^{N} X_i^{'}(y_i - \mu_i) = 0.$$

The ML method gives estimates for β, which are consistent and asymptotically normal, especially when the autocorrelation between the responses is small. However, if the correlation is large, then $\hat{\beta}$ may be inefficient. Also, obtaining joint-likelihood under the naive assumption of independence ignores correlations among binary responses, and gives inconsistent estimates of the asymptotic variance of $\hat{\beta}$. As a result, the standard errors of time-varying covariates are usually overestimated, and those of time-invariant covariates may be underestimated [12].

3.2.1.2 The Bahadur Model

Assume T binary responses from a subject i in the form of $Y_i = (Y_{i1}, \cdots, Y_{iT})'$. The joint distribution of this vector follows a multinomial distribution with a probability vector of 2^T dimension, which is fully parametrized with $2^T - 1$ parameters. Bahadur [4], described the joint distribution of such Y_i vector in terms of its marginal means as

$$f(y_{i1}, \cdots, y_{iT}) = Pr(Y_{i1} = y_{i1}, \cdots, Y_{iT} = y_{iT}) = f^*(y_i)c^*(y_i),$$

where

$$f^*(y_i) = \prod_{j=1}^{T} \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{(1-y_{ij})},$$

and

$$c^*(y_i) = 1 + \sum_{j<k} \rho_{ijk} z_{ij} z_{ik} + \sum_{j<k<l} \rho_{ijkl} z_{ij} z_{ik} z_{il} + \cdots + \rho_{i12\cdots T}\, z_{i1} z_{i2} \cdots z_{iT},$$

with

$$\begin{aligned} z_{ij} &= \frac{y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1-\mu_{ij})}} \\ \rho_{ijk} &= E(z_{ij} z_{ik}) \\ \rho_{ijkl} &= E(z_{ij} z_{ik} z_{il}) \\ &\vdots \\ \rho_{i12\cdots T} &= E(z_{i1} z_{i2} \cdots z_{iT}). \end{aligned}$$

Here the joint distribution of Y_i is the product of an independence model denoted by $f^*(y_i)$, and a correction factor denoted by $c^*(y_i)$, with the latter viewed as a model for overdispersion [15]. The Bahadur representation of a trivariate response variable with $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})' = (1,1,1)'$ may be obtained as

$$\begin{aligned} Pr(Y_{i1} = 1, Y_{i2} = 1, Y_{i3} = 1) = {} & \mu_{i1}\mu_{i23} + \mu_{i2}\mu_{i13} + \mu_{i3}\mu_{i12} - 2\mu_{i1}\mu_{i2}\mu_{i3} \\ & + \rho_{123}\sqrt{\mu_{i1}(1-\mu_{i1})\mu_{i2}(1-\mu_{i2})\mu_{i3}(1-\mu_{i3})}, \end{aligned}$$

where

$$\begin{aligned} \mu_{ijk} &= Pr(Y_{ij} = 1, Y_{ik} = 1) \\ &= \mu_{ij}\mu_{ik} + \rho_{ijk}\sqrt{\mu_{ij}(1-\mu_{ij})\mu_{ik}(1-\mu_{ik})}. \end{aligned}$$

Similar to other marginal models, the Bahadur model is reproducible for any

subset of the response vector. This model can fully specify the joint distribution of the responses in closed form, given $2^T - T - 1$ marginal correlations denoted as $\rho_i = (\rho_{i12}, \rho_{i13}, \cdots, \rho_{i12\cdots T})'$ are specified. Clearly, the likelihood of a whole sample of size N is obtained as

$$L = \prod_{i=1}^{N} f(y_{i1}, \cdots, y_{iT}),$$

but in doing so, the Bahadur model needs parametrization of higher order associations, in terms of correlation parameters. As showed in equation (3.4), for discrete random variables, the correlations are constrained by the expectation of the response variable, Y_i. In the case of the binary response, the correlation is more restricted such that

$$max(0, \mu_{ij} + \mu_{ik} - 1) < Pr(Y_{ij} = 1, Y_{ik} = 1) < min(\mu_{ij}, \mu_{ik}),$$

therefore, the Bahadur model requires many inequality constraints on the model parameters, and except in cases where the number of time points is relatively small, maximization of likelihood equations is very difficult. Consequently, the Bahadur model has not been widely adopted in modelling longitudinal binary data.

3.2.2 The Pseudo-likelihood Method

Although the likelihood-based approaches are very popular estimation approaches, their performance are negatively affected by the presence of nuisance parameters. The pseudo maximum likelihood has been proposed by Gong and Samaniego [16], which can be used for estimation in the presence of such parameters.

Let $X_1, \cdots, X_n$ denote a random sample from a member of two parameter family of distributions of $\mathbb{R}$ such that

$$\mathscr{F} = \{F_{\theta,\gamma}\},$$

and also a member of the probability model given by

$$P = \{P_{\theta,\gamma} : \theta \in \Theta \subset \mathbb{R}^j, \gamma \in \Gamma \subset \mathbb{R}^k\}.$$

Also, let $\theta = (\theta_1, \cdots, \theta_j)'$ and $\gamma = (\gamma_1, \cdots, \gamma_k)'$ denote the vectors of structural and nuisance parameters, respectively. Let θ_0 and γ_0 denote the true (and unknown) values of the structural and nuisance parameters, respectively. Now, let P_{θ_0,γ_0} denote the true probability distribution that gives the observed data. Then the scientific interest is to find an estimate of θ_0, namely $\hat{\theta}_i$, $i = 1, \cdots, j$, which is a subset of Θ. In a classical likelihood approach, the likelihood function is obtained in a general form of

$$L(\theta, \gamma) = L(\theta, \gamma | X_1, \cdots, X_n),$$

and the log-likelihood function takes the form

$$\ell(\theta, \gamma) = \sum_{i=1}^{n} \log P_{\theta,\gamma}(X_i),$$

which gives score equation in the form

$$S(\hat{\theta}, \hat{\gamma}) = \left[\frac{\partial}{\partial\theta} l(\theta, \gamma)\right]_{\theta=\hat{\theta},\gamma=\hat{\gamma}} = 0.$$

In the presence of nuisance parameters γ, no solution to this equation exists in a closed form. Under regularity conditions, one can use some methods other than the likelihood to find a consistent estimator $\hat{\gamma} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_n)'$. Now, $\hat{\gamma}$ is used to replace the nuisance parameter γ in the model and represents the true (and unknown) parameter γ_0. Then the pseudo likelihood function is obtained as

$$L_{PL}(\theta, \hat{\gamma}) = L(\theta, \hat{\gamma}|X_1, \cdots, X_n),$$

and the pseudo log-likelihood function is obtained as

$$\ell_{PL}(\theta, \hat{\gamma}) = \sum_{i=1}^{n} \log P_{\theta,\hat{\gamma}}(X_i).$$

The pseudo score equation is given by

$$S_{PL} = \frac{\partial}{\partial \theta} \log L(\theta, \hat{\gamma}) = 0,$$

where $\hat{\gamma}$ is a "$\sqrt{n}$-consistent" estimate of γ. Gong and Samaniego [16] showed that a consistent and asymptotically normal estimator of θ, $\hat{\theta}_{PL}$, is obtained by solving the pseudo score equations, S_{PL}, which , under some regularity conditions, is consistent if $\hat{\gamma}$ is consistent, and its efficiency is dependent upon the efficiency of $\hat{\gamma}$.

Zhao and Prentice [17] used a pseudo ML-based method in the regression analysis of binary data using a quadratic exponential family. We continue with the same notation introduced at the start of Section 3.2. The model is parametrized in terms of marginal means and pairwise correlations. Assuming higher order associations equal to zero, they used log-linear specification for the joint distribution of Y_i, given by

$$f(y_i, \Theta_i, \Pi_i) = \exp\{\Theta_i^{'} y_i + \Pi_i^{'} w_i - A(\Theta_i, \Pi_i)\}, \tag{3.6}$$

where Θ_i, and Π_i are canonical parameters, such that, $\Theta_i = (\theta_{i1}, \cdots, \theta_{iT})^{'}$, and $\Pi_i = (\pi_{i12}, \cdots, \pi_{iT-1,T}, \cdots, \pi_{i12\cdots T})^{'}$, respectively, and $W_i = (Y_{i1}Y_{i2}, \cdots, Y_{iT-1}Y_{iT})^{'}$ is a $T(T-1)/2$ vector of Y_i's two-way cross products, and $A(\Theta_i, \Pi_i)$ is a normalizing constant. Now, let

$$S_{ist} = (Y_{is} - \mu_{is})(Y_{it} - \mu_{it}),$$

and

$$\sigma_{ist} = E(S_{ist}).$$

Then by using (3.6) and by using some link function which transforms (Θ_i, Π_i) to (μ_i, σ_i), one by one, we model the mean and covariance of the outcome variable as functions of β and α, respectively. The likelihood score equations for β and α are derived as

$$\sum_{i=1}^{N} D_i' V_i^{-1} f_i = 0, \tag{3.7}$$

or, equivalently,

$$\sum_{i=1}^{N} \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \sigma_i}{\partial \beta} & \frac{\partial \sigma_i}{\partial \alpha} \end{bmatrix}' \begin{bmatrix} var(Y_i) & cov(Y_i, S_i) \\ cov(S_i, Y_i) & var(S_i) \end{bmatrix}^{-1} \begin{bmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{bmatrix} = 0.$$

Note that equation (3.7) gives the pseudo maximum likelihood estimators $(\hat{\beta}, \hat{\alpha})$, which are equivalent to the ML estimates when true associations among Y_i at three- or higher-order are zero. The consistency of $\hat{\beta}$ and $\hat{\alpha}$ is adversely affected by model misspecification for the mean and pairwise correlations. Continuing from (3.7), the Fisher information matrix is given by

$$I = \sum_{i=1}^{N} D_i' V_i^{-1} D_i,$$

or, equivalently, by

$$I = \sum_{i=1}^{N} \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \sigma_i}{\partial \beta} & \frac{\partial \sigma_i}{\partial \alpha} \end{bmatrix}' \begin{bmatrix} var(Y_i) & cov(Y_i, S_i) \\ cov(S_i, Y_i) & var(S_i) \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \sigma_i}{\partial \beta} & \frac{\partial \sigma_i}{\partial \alpha} \end{bmatrix}.$$

According to Zhao and Prentice [17], given that $\hat{\beta}$ and $\hat{\alpha}$ are pseudo maximum likelihood estimates satisfying equation (3.7), under regularity conditions, the quantity $\{(\hat{\beta} - \beta), (\hat{\alpha} - \alpha)\}$ is asymptotically normal with mean zero. A consistent estimate of the covariance matrix for $(\hat{\beta}, \hat{\alpha})$ is given by

$$I^{-1}\Big(\sum_{i=1}^{N} D_i' V_i^{-1} f_i f_i' V_i^{-1} D_i\Big) I^{-1}.$$

3.2.3 Non-likelihood Methods

In this section, we briefly discuss non-likelihood methods for fitting marginal binary models, which include generalized estimating equations (GEEs).

3.2.3.1 Generalized estimating equations(GEEs)

The GEE [14] is a semiparametric method that uses the quasi-likelihood estimation to extend the GLM to longitudinal data. Instead of full specification of the joint distribution of Y_i for all subjects, where $i = 1, ..., N$, the GEE specifies the marginal distribution of Y_{it} by relying on a working correlation matrix for repeated observations of each subject at times $t = 1, ..., T$.

We rely on the same notation as used in Section 3.1.1. Let A_i denote an $N \times N$ diagonal matrix, where $V(\mu_{ij})$ is its jth diagonal element, given by

$$V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij}).$$

Let $R_i(\alpha)$ denote an $N \times N$ diagonal matrix of working correlations among N repeated observations for subject i. Assuming α as a vector of nuisance parameters, the GEE estimator of β is obtained by solving

$$U_\beta(\beta, \alpha) = \sum_{i=1}^{N} D_i'[V(\hat{\alpha})]^{-1}(y_i - \mu_i) = 0, \tag{3.8}$$

where $\hat{\alpha}$ is a consistent estimator of α,

$$D_i = \partial\mu_i/\partial\beta,$$

and $v(\alpha)$ is a working variance–covariance matrix for Y_i, given by

$$V(\alpha) = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}.$$

We can numerically estimate the regression parameters by using the iterative Fisher's scoring algorithm given by

$$\beta^{(m+1)} = \beta^{(m)} + \left\{ \sum_{i=1}^{N} D_i' V_i(\hat{\alpha})^{-1} D_i \right\}^{-1} \left\{ \sum_{i=1}^{N} D_i' V_i(\hat{\alpha})^{-1} (y_i - \mu_i) \right\} \Bigg|_{\beta=\beta^{(m)}}.$$

Let us define

$$B_\beta = \sum_{i=1}^{N} D_i' V_i(\hat{\alpha})^{-1} D_i,$$

and

$$H_\beta = \sum_{i=1}^{N} D_i' V_i(\hat{\alpha})^{-1} (y_i - \mu_i) V_i'(\hat{\alpha}).$$

Then under some regularity conditions, it can be shown that (3.8) provides consistent estimators, where the sandwich variance of the resulting estimators $\hat{\beta}$ is obtained as

$$Var(\beta) = \lim_{N \to +\infty} N\, B_{\beta}^{-1}\, H_{\beta}\, B_{\beta}^{-1}.$$

As discussed earlier in this chapter, especially in Section 3.1.1.1, within-subject correlated responses are a main feature of all longitudinal data. Therefore, we may use the GEE for longitudinal binary data analysis, as it avoids the specification of the joint densities. However, as we will see in Chapter 4, the GEE and related methods have stringent assumptions on missing data mechanisms, and they often fail when there are nonignorable missing data, which is the main focus of the current thesis.

Chapter 4

Methods for Analyzing Longitudinal Data with Nonignorable Missingness

In longitudinal studies, repeated observations are taken either from human subjects, in case of clinical trials, psychological, and medical studies, or from non-human subjects, in case of biological, veterinary, and agricultural studies. These studies are more acutely associated with missing data problem, compared to cross-sectional, because multiple measurements are taken from same subjects over time. The missingness or non-response simply occurs when an intended observation is not obtained on a certain measurement occasion. In general, missingness reduces the precision of estimations, which simply results from loss of information. However, in longitudinal studies, an important implication of missingness is biased estimates and misleading inferences. In order to avoid it, we should make specific assumptions about the reasons for missingness, and in course of data analysis, we should incorporate missing data mechanisms in our model. In the current chapter, we discuss analysis of incomplete longitudinal data by reviewing taxonomy of missing data, and introducing models from literature which are used to analyze longitudinal binary data with nonignorable missingness.

4.1 Taxonomy of Missing Data

In this section, we review general taxonomy of missing data, including missing data patterns, mechanisms, and methods [18] [19] [13]. Assume a longitudinal design on N subjects, each with $t = 1, ..., T$ intended measurements. Let $Y_i = (Y_{i1}, ..., Y_{iT})^{'}$ denote the response vector for subject i, $i = 1, ..., N$, with complete set of responses. Let each Y_i response vector be associated with a $T \times P$ matrix of covariates X_i which are completely observed.

4.1.1 Missing Data Patterns

Monotone missingness or dropout happens if once a subject misses a measurement occasion, that subject is never observed again, i.e., for each response vector Y_i, a dropout occurs whenever Y_{it} is missing, then for all $t' > t$, $Y_{it'}$ are also missing. In case of monotone missingness, it is easy to evaluate likelihood functions because we can factor them in terms of conditional densities. On the other hand, **non-monotone** missingness happens if after an initial occurrence of a missing value, the subject is observed at least once, i.e., for each vector Y_i, a nonmonotone missingness occurs whenever Y_{it} is missing, for some $t' > t$, $Y_{it'}$ is observed. In case of nonmonotone missingness, it is not easy to evaluate likelihood functions because a simple factorization is almost impossible.

4.1.2 Missing Data Mechanisms

Continuing with notation used at the beginning of this section, we define a set of indicator random variables for members of Y_i vector and we denote them as

$R_i = (R_{i1}, ..., R_{iT})'$, such that

$$R_{it} = \begin{cases} 1, & \text{if } Y_{it} \in Y_i^o \\ 0, & \text{if } Y_{it} \in Y_i^m, \end{cases}$$

whence Y_i is partitioned into two subsets of observed data, denoted by Y_i^o, and missing data, denoted by Y_i^m, such that $Y_i = \{Y_i^o, Y_i^m\}$. We assume that the distribution of R_i depends on a non-response parameter γ, response vector Y_i, and covariates matrix X_i. We denote it as $f(r_i|y_i, X_i, \gamma)$. On the other hand, we assume that the distribution of Y_i depends on some set of regression parameters β, and within-subject association parameters α. We denote it as $f(y_i|X_i, \beta)$, or, equivalently,

$$f(y_i|X_i, \alpha, \beta) = f(y_i^o, y_i^m|X_i, \alpha, \beta).$$

Then the joint distribution of Y_i and R_i is given by

$$f(y_i, r_i|X_i, \alpha, \beta, \gamma) = f(y_i|X_i, \alpha, \beta) f(r_i|y_i, X_i, \gamma). \tag{4.1}$$

Using the conditional distribution of R_i, given Y_i (or, equivalently, given Y_i^o and Y_i^m) and X_i, three types of the missing data mechanisms are described below.

1. Missing completely at random (**MCAR**) occurs when the distribution of R_i is independent of both Y_i^o and Y_i^m, and might be, independent of X_i. In this case,

   $$Pr(R_i|Y_i^o, Y_i^m, X_i) = Pr(R_i),$$

   or, equivalently,

   $$f(r_i|y_i, X_i, \gamma) = f(r_i|\gamma).$$

Under MCAR, the observed data are a random sample of the whole data, and all of its moments and its joint distribution are the same as the complete data. Hence a simple complete case analysis gives unbiased and valid estimates. [20] An example of MCAR mechanism occurs during rotating panel study design, when sample units of equal sizes are brought in and out of a survey at a preset pattern.

2. Missing at random (**MAR**) occurs when the distribution of R_i is only independent of Y_i^m. In this case, we have

$$Pr(R_i|Y_i^o, Y_i^m, X_i) = Pr(R_i|Y_i^o, X_i),$$

or, equivalently,

$$f(r_i|y_i, X_i, \gamma) = f(r_i|y_i^o, X_i, \gamma). \tag{4.2}$$

Under MAR, given the observed data, the probability of missingness only depends on the observed, but not missing data. For example, if in a longitudinal design, a subject is taken out of study once the outcome variable falls below a certain level, the missingness is conditional on observed outcome. Under MAR missing data mechanism, the joint distribution of Y_i^o and missing data mechanism R_i is given by

$$\begin{aligned} f(y_i^o, r_i|X_i, \alpha, \beta, \gamma) &= f(r_i|y_i^o, X_i, \gamma). \int f(y_i^o, y_i^m|X_i, \alpha, \beta) dY_i^m \\ &= f(r_i|y_i^o, X_i, \gamma). f(y_i^o|X_i, \alpha, \beta). \end{aligned} \tag{4.3}$$

Equation (4.3) shows the contribution of subject i to the likelihood function. If both missing data model parameters γ and response model parameters, $\{\beta, \alpha\}$, are variation independent, then the MAR mechanism is said to be **ignorable**. On the other hand, if these parameters are variation dependent, then the MAR mechanism

is said to be **nonignorable**.

3. Not missing at random (**NMAR**) occurs when the distribution of R_i depends on Y_i^o and at least some (if not all) components of Y_i^m. This mechanism is always considered **nonignorable**, and its factorization as shown in equation (4.3) is not possible.

The NMAR mechanism occurs very commonly in longitudinal studies. Ibrahim et al. mention that in clinical trials (e.g., AIDS and cancer) patient's participations can be affected by the side effects of the treatment, and therefore, the outcome can impact missingness [19]. In their research on modelling missing data in AIDS clinical trials, Sinha et al. [1] argue that AIDS patients with more profound CD4 lymphopenia are more likely to show up at the clinic appointments, as they feel more ill, and therefore, they are more likely to come to every doctor's visit. Therefore, if the CD4 lymphocyte count itself is the outcome variable of interest, its missingness will depend on the unobserved values of the outcome variable.

4.1.3 Ignorablility of Missingness

Whether a missingness mechanism is "ignorable" or "nonignorable" is determined largely on the basis of inference. In frequentist inference, missingness is only ignorable when its mechanism is MCAR. On the other hand, in likelihood and Bayesian inferences, generally, missingness under MCAR and MAR mechanisms is considered ignorable, but the NMAR mechanism is considered nonignorable.

Little and Rubin [18], and Diggle et. al. [3] provide an account of ignorability of MAR mechanism under likelihood inference. Let $f(y_i^o, y_i^m, r_i)$ denote the joint PDF of (Y_i^o, Y_i^m, R_i). By standard factorization, we obtain

$$f(y_i^o, y_i^m, r_i) = f(y_i^o, y_i^m) f(r_i | y_i^o, y_i^m). \tag{4.4}$$

The joint PDF of observed random variables denoted by (Y_i^o, R_i) is given by

$$f(y_i^o, r_i) = \int f(y_i^o, y_i^m) f(r_i | y_i^o, y_i^m) dy_i^m. \tag{4.5}$$

Assuming a MAR mechanism, $f(r_i|y_i^o, y_i^m)$ only depends on y_i^o and does not depend on y_i^m, and (4.5) is equivalent to

$$\begin{aligned} f(y_i^o, r_i) &= f(r_i|y_i^o) \int f(y_i^o, y_i^m) dy_i^m \\ &= f(r_i|y_i^o) f(y_i^o). \end{aligned} \tag{4.6}$$

Then we obtain the log-likelihood function as

$$\log L = \log f(r_i|y_i^o) + \log f(y_i^o), \tag{4.7}$$

and it is noted that the log-likelihood function in (4.7) is maximized by maximizing two RHS terms, separately. The first term, $logf(r_i|y_i^o)$, provides no information about the distribution of Y_i^o and is simply ignored. Therefore, we conclude that under likelihood inference, the MAR missingness is an ignorable mechanism.

4.1.4 Missing Data Methods

Little and Rubin [21] list 'completely recorded units', 'weighting-based', 'imputation-based', and 'model-based' approaches for handling missing data. These are briefly discussed here.

1. Completely recorded units are used when the amount of missing data is small and we can discard instances where some variables are not recorded and analyse the complete subset of data. This approach is not efficient on small data sets

or on subpopulations and may give to seriously biased inferences. In *complete case analysis* or listwise deletion, we run analysis on cases where all variables are recorded. It is a simple method and gives comparable univariate statistics, but, loss of precision always occurs in some degrees and if the missing data mechanism is not MCAR, it gives biased estimates. On the other hand, *available case analysis* or pairwise deletion, uses all available data for univariate descriptive statistics. In analysing bivariate or multivariate statistics, this approach usually gives inferior results compared to CCA, especially when the variables are highly correlated.

2. Weighting-based method: Inference on a complete sample is based on design weights of sample units, which are expressed by the Horwitz-Thompson (HT) estimator of the population mean, given by

$$\left(\sum_{i=1}^{n} \pi_i^{-1} y_i\right)\left(\sum_{i=1}^{n} \pi_i^{-1}\right),$$

where y_i is the value of the random variable Y for unit i in the population. When we are inferring on a sample with missingness, we need to integrate a weighting element into the HT estimator, defined as the probability of response for unit i, denoted as $\hat{p}_i$. By using this extra weighting element, we adjust our estimator for the missingness as if it was part of the design. Now, the HT estimator of the population mean is obtained as

$$\sum_{i=1}^{n} (\pi_i \hat{p}_i)^{-1} y_i / \sum_{i=1}^{n} (\pi_i \hat{p}_i)^{-1},$$

where the sums denote the oversampled responding unit, handled by inclusion of $\hat{p}_i$.

3. Imputation-based methods entail filling the missing values using a procedure, such

as mean, regression, or hot deck imputation. Then the analysis is performed on resultant complete set of data.

4. Model-based methods are the most comprehensive approaches developed for handling of missing data, and one needs to define a model for the observed data before running the analysis. Model-based approaches are flexible as they give estimates of variance, and they provide possibility of evaluating underlying assumptions.

4.2 Models for Nonignorable Missing Data

When the missing data are nonignorable, we use joint distributions to account for missingness. Instead of $Pr(Y_i)$, we need to model $Pr(Y_i, R_i)$ which can't be factorized into two independent parts. In order to decompose it, we can use either selection or pattern-mixture models. In **selection models**, we model the longitudinal response vector, before modelling the conditional probability of missingness, and we obtain

$$Pr(Y_i, R_i) = Pr(Y_i)Pr(R_i|Y_i),$$

or, equivalently,

$$f(y_i, r_i|X_i, \alpha, \beta, \gamma) = f_{Y_i}(y_i|X_i, \alpha, \beta) f_{R_i|Y_i}(r_i|x_i, Y_i, \gamma).$$

In **pattern-mixture models**, we model the longitudinal response vector given missing patterns, before modelling missingness, and we obtain

$$Pr(Y_i, R_i) = Pr(Y_i|R_i)Pr(R_i),$$

or, equivalently,

$$f(y_i, r_i|X_i, \alpha, \beta, \gamma) = f_{Y_i|R_i}(y_i|X_i, R_i, \alpha, \beta) f_{R_i}(r_i|x_i, \gamma).$$

In modelling nonignorable missing data, the GEE and relevant approaches are generally inferior to the ML-based methods, as they are sensitive to model misspecification, and may yield biased results. On the other hand, the presence of nuisance parameters in the analysis of longitudinal binary data complicates the use of classical ML estimation methods. When the number of time points in a longitudinal design is greater than 3, the number of nuisance parameters (and as a result, the number of ML estimating equations) grow very quickly, which makes it impossible to reach to an analytical solution. Instead, when one tries to maximize the ML functions numerically, the convergence becomes too slow, impractical, or even impossible. Overall, using the ML estimation for longitudinal binary data with nonignorable missingness is only advised when the length of the response vector Y_i, is less than or equal to three [1].

In the remainder of this chapter, we review two **approximate likelihood** approaches from selection methods family, which are used to analyze longitudinal binary data with nonignorable missingness. These methods are *Independent Pseudolikelihood* (**IPL**) proposed by Troxel et al. [2], and *Bivariate Pseudolikelihood* (**BPL**) proposed by Sinha et al. [1]. Also, we briefly review the classical ML estimation method for the analysis of longitudinal binary data with nonignorable missingness. Before reviewing these models, we introduce necessary notation. We continue with the same notation as used in this chapter, mainly in Section 4.1.2. Assume

$$Y_{it} \sim Bernoulli(p_{it}), \tag{4.8}$$

where p_{it} is the probability of success for Y_{it}, $t = 1, ..., T$, and, using equation (2.18) it is obtained as

$$p_{it} = E(Y_{it}|x_{it}, \beta) = P(Y_{it} = 1|x_{it}, \beta) = \frac{\exp(x'_{it}\beta)}{1 + \exp(x'_{it}\beta)}, \tag{4.9}$$

knowing that our interest is in the regression parameters β and the within-subject correlations are considered nuisance parameters. Using equation (3.4), the marginal correlation between any two responses Y_{it} and $Y_{it'}$ is obtained as

$$\rho(Y_{it}, Y_{it'}) = \frac{Pr(Y_{it} = 1, Y_{it'} = 1) - p_{it}p_{it'}}{\sqrt{p_{it}(1 - p_{it})p_{it'}(1 - p_{it'})}},$$

which gives the multivariate Bahadur distribution (see Section 3.2.1.2). On the other hand, using (3.5), marginal odds ratio between any two responses Y_{it} and $Y_{it'}$ is obtained as

$$OR(Y_{it}, Y_{it'}) = \frac{Pr(Y_{it} = 1, Y_{it'} = 1)Pr(Y_{it} = 0, Y_{it'} = 0)}{Pr(Y_{it} = 1, Y_{it'} = 0)Pr(Y_{it} = 0, Y_{it'} = 1)},$$

which gives the multivariate Plackett distribution. Now, we assume that all subjects are observed at baseline time $t = 1$. Then the R_i vector will have $T - 1$ binary random variables, such that $R_i = (R_{i2}, ..., R_{iT})'$. We assume the marginal distribution of R_{it} is given by

$$R_{it} \sim Bernoulli(\pi_{it}),$$

where π_{it} is the probability of success (being observed) for R_{it} when $t = 2, ..., T$, and using equation (4.2), it is obtained as

$$\pi_{it} = Pr(R_{it}|y_{it}, x_{it}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 y_{it} + x'_{it}\gamma_2)}{1 + \exp(\gamma_0 + \gamma_1 y_{it} + x'_{it}\gamma_2)}, \tag{4.10}$$

which defines a nonignorable missing data mechanism if the parameter γ_1 is not zero. It is equivalent to saying that the probability of missingness depends on an outcome value, which is not observed.

4.2.1 Independent Pseudolikelihood (IPL)

Troxel et al. [2] propose an approximate likelihood method using the naive assumption of independence among repeated measures of the response vector Y_i. Partitioning data into observed and missing subsets, they get

$$f(y_{i1}, r_{i1}, ..., y_{iT}, r_{iT}) = f(y_i^o, y_i^m, r_i).$$

The marginal distribution of (Y_{it}, R_{it}) at time t is denoted as $f(y_{it}, r_{it}|x_{it}, \beta, \gamma)$, and is obtained as

$$f(y_{it}, r_{it}|x_{it}, \beta, \gamma) = f(y_{it}|x_{it}, \beta) f(r_{it}|y_{it}, x_{it}, \gamma), \tag{4.11}$$

where $f(y_{it}|x_{it}, \beta)$ is a Bernoulli density with probability of success equal to p_{it} obtained in (4.9), and $f(r_{it}|y_{it}, x_{it}, \gamma)$ is also a Bernoulli density with probability of success equal to π_{it} obtained in (4.10). They set pairwise correlation coefficients ρ_{st} for two responses equal to zero. Following Gong and Samaniego [16], it means Troxel et al. [2] set a consistent estimator, $\hat{\alpha}$, for association parameter α and they set its value equal to zero. Now, the pseudolikelihood is obtained as

$$
\begin{aligned}
L_{IPL}(\beta,\gamma) &= L((\beta,\gamma|\hat{\alpha}=0) \\
&= \prod_{i}^{N} f(y_i^o, y_i^m, r_i) \\
&= \prod_{i}^{N} \left\{ \sum_{Y_i^m} f(y_i^o, r_i) \right\},
\end{aligned}
$$

and forcing independence among observations over time, the pseudolikelihood is given by

$$
\begin{aligned}
L_{IPL}(\beta,\gamma) &= \prod_{i=1}^{N}\prod_{t=1}^{T} \{f(y_{it}|x_{it},\beta) f(r_{it}|y_{it},x_{it},\gamma)\}^{rit} \\
&\quad \times \left\{ \sum_{y_{it}=0}^{1} f(y_{it}|x_{it},\beta) f(r_{it}|y_{it},x_{it},\gamma) \right\}^{(1-rit)} \\
&= \prod_{i=1}^{N}\prod_{t=1}^{T} \{f(y_{it}|x_{it},\beta)\pi_{it}\}^{rit} \\
&\quad \times \left\{ \sum_{y_{it}=0}^{1} f(y_{it}|x_{it},\beta)(1-\pi_{it}) \right\}^{(1-rit)}. \qquad (4.12)
\end{aligned}
$$

The IPL estimator $\hat{\nu} = (\hat{\beta}, \hat{\gamma})$ is obtained by solving the IPL pseudoscore equation given by

$$
S(\nu) = \frac{\partial \log L_{IPL}(\nu)}{\partial \nu} = 0, \qquad (4.13)
$$

and under the independence condition, it can be shown that at true values of $\nu = (\beta, \gamma)$, we have

$$
E[S(\nu)] = 0, \qquad (4.14)
$$

and it can be shown that the IPL estimator $\hat{\nu}$ is asymptotically normal, consistent estimator of β and γ [2], and its asymptotic sandwich variance is obtained by

$$var(\hat{\nu}) \approx \left\{ \frac{\partial S(\nu)}{\partial \nu} \right\}^{-1} \left(\sum_{i=1}^{N} S_i(\nu) S_i^T(\nu) \right) \left\{ \frac{\partial S(\nu)}{\partial \nu} \right\}^{-1}. \tag{4.15}$$

4.2.2 Bivariate Pseudolikelihood (BPL)

Sinha et al. [1] propose an approximate likelihood method assuming a vector of pairwise association parameters α which denotes pairwise correlations between universally observed Y_{i1} for each subject i at baseline time $t = 1$ and each Y_{it} where $t > 1$. The marginal distribution of (Y_{it}, R_{it}) at time t is denoted as $f(y_{it}, r_{it}|y_{i1}, x_{it}, \beta, \alpha, \gamma)$, and is obtained as

$$f(y_{it}, r_{it}|y_{i1}, x_{it}, \beta, \alpha, \gamma) = f(y_{it}|y_{i1}, x_{it}, \beta, \alpha) f(r_{it}|y_{i1}, y_{it}, x_{it}, \gamma). \tag{4.16}$$

The second factor on the right side is the missing data mechanism, $f(r_{it}|y_{i1}, y_{it}, x_{it}, \gamma)$, and is modelled using the logistic regression model

$$f(r_{it}|y_{i1}, y_{it}, x_{it}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{it} + \gamma_3 x_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{it} + \gamma_3 x_{it})}. \tag{4.17}$$

The first term in (4.16) is decomposed by specifying the marginal distribution of $Y_{i1}, i = 1, ..., N$, for all subjects, when $Y_{i1} \sim Bernoulli(p_{i1})$. It is denoted as $\prod_{i=1}^{N} f(y_{i1}|x_{it}, \beta)$ and is given by

$$\prod_{i=1}^{N} f(y_{i1}|x_{it}, \beta) = \prod_{i=1}^{N} {p_{i1}}^{y_{i1}} (1 - p_{i1})^{(1-y_{i1})}. \tag{4.18}$$

We also need to specify $f(y_{it}|y_{i1}, x_{it}, \beta, \alpha)$. Sinha et al. [1] adopt a bivariate Bahadur model to specify this conditional distribution from the joint bivariate density of (y_{i1}, y_{it})

such that the marginal logistic model for y_{it} holds. The joint density is denoted as $f(y_{i1}, y_{it}|x_{it}, \beta, \alpha)$, and is given by

$$f(y_{i1}, y_{it}|x_{it}, \beta, \alpha) = {p_{i1}}^{y_{i1}}(1-p_{i1})^{(1-y_{i1})}{p_{it}}^{y_{it}}(1-p_{it})^{(1-y_{it})} \times \left\{1 + \alpha_{1it}\frac{(y_{i1}-p_{i1})(y_{it}-p_{it})}{\sqrt{p_{i1}(1-p_{i1})p_{it}(1-p_{it})}}\right\}, \tag{4.19}$$

where $\alpha_{1it} = Corr(y_{i1}, y_{it}|x_{it})$. It is easy to show that for the bivariate density in (4.19), $Pr(y_{it}|y_{i1})$ follows a Bernoulli density with probability of success equal to

$$p_{it1} = Pr(y_{it} = 1|y_{i1}, x_{it}, \beta, \alpha),$$

which is obtained as

$$p_{it1} = p_{it}\left\{1 + \alpha_{1it}\frac{(1-p_{it})(y_{i1}-p_{i1})}{\sqrt{p_{it}(1-p_{it})p_{i1}(1-p_{i1})}}\right\}.$$

Then the first term on the right side of (4.16) is obtained as

$$f(y_{it}|y_{i1}, x_{it}, \beta, \alpha) = {p_{it1}}^{y_{it}}(1-p_{it1})^{(1-y_{it})}. \tag{4.20}$$

Now, using results (4.17) through (4.20) and assuming independence of the joint density of $f(y_{it}, r_{it}|y_{i1}, x_{it}, \beta, \alpha, \gamma)$ over all subjects $i = 1, ..., N$, the BPL function is obtained as

$$
\begin{aligned}
L_{BPL}(\beta,\alpha,\gamma) &= \prod_{i=1}^{N} L_{BPL,i}(\beta,\alpha,\gamma) \\
&= \prod_{i=1}^{N} f(y_{i1}|x_{it},\beta) \prod_{t=2}^{T} \{f(y_{it},r_{it}|y_{i1},x_{it},\beta,\alpha,\gamma)\}^{r_{it}} \\
&\quad \times \left\{ \sum_{y_{it}=0}^{1} f(y_{it},r_{it}|y_{i1},x_{it},\beta,\alpha,\gamma) \right\}^{(1-r_{it})}. \qquad (4.21)
\end{aligned}
$$

The BPL estimator $\hat{\nu} = (\hat{\beta}, \hat{\alpha}, \hat{\gamma})$ is obtained by solving the BPL pseudoscore equations given by

$$
S(\nu) = \sum_{i=1}^{N} S_i(\nu) = \sum_{i=1}^{N} \frac{\partial \log L_{BPL,i}(\nu)}{\partial \nu} = 0. \qquad (4.22)
$$

Given correctly specified terms in (4.16), it can be shown that at true values of $\nu = (\beta, \alpha, \gamma)$, we have

$$
E[S(\nu)] = 0, \qquad (4.23)
$$

even if the true independence for in $f(y_{it}, r_{it}|y_{i1}, x_{it}, \beta, \alpha, \gamma)$ for $t = 2, ..., T$ is not verified. It can be shown that the BPL estimator $\hat{\nu}$ is asymptotically normal, consistent estimators of β, α, and γ [1], and its asymptotic sandwich variance of is obtained by

$$
var(\hat{\nu}) \approx \left\{ \frac{\partial S(\nu)}{\partial \nu} \right\}^{-1} \left(\sum_{i=1}^{N} S_i(\nu) S_i^T(\nu) \right) \left\{ \frac{\partial S(\nu)}{\partial \nu} \right\}^{-1}. \qquad (4.24)
$$

4.2.3 Maximum Likelihood (ML)

Recall the result given in (4.1), which denotes the standard factorization of the joint distribution of the response vector Y_i, and missing data indicators R_i, given by

$$f(y_i, r_i|X_i, \alpha, \beta, \gamma) = f(y_i|X_i, \alpha, \beta) f(r_i|y_i, X_i, \gamma).$$

If all Y_i are observed, the likelihood of complete-data is obtained as

$$L_{complete}(\beta, \alpha, \gamma) = \prod_{i=1}^{N} f(y_i, r_i|X_i, \alpha, \beta, \gamma).$$

On the other hand, if we assume ignorable missingness in data, the likelihood of incomplete data is obtained as

$$L_{ignor}(\beta, \alpha, \gamma) = \prod_{i=1}^{N} \int f(y_i|X_i, \alpha, \beta) dY_i^m = \prod_{i=1}^{N} f(y_i^o|X_i, \alpha, \beta).$$

However, if we assume nonignorable missingness in the data, we make inference based on the observed data by integrating the missing data out of the joint density of $f(y_i, r_i)$. We obtain the likelihood function as

$$\begin{aligned} L_{ML}(\beta, \alpha, \gamma|y^o, r, X) &= \prod_{i=1}^{N} \int f(y_i, r_i|X_i, \beta, \alpha, \gamma) dY_i^m \\ &= \prod_{i=1}^{N} \left\{ \sum_{Y_i^m} f(y_i|\beta, \alpha) f(r_i|y_i, \gamma) \right\} \\ &= \prod_{i=1}^{N} \left\{ \sum_{Y_i^m} f(y_i^o, y_i^m|\beta, \alpha) f(r_i|y_i^o, y_i^m, \gamma) \right\}. \end{aligned} \tag{4.25}$$

For simplicity, we often assume

$$f(r_i|y_i,\gamma) = \prod_{t=1}^{T} f(r_{it}|y_{it},\gamma),$$

where

$$f(r_{it}|y_{it},\gamma) = {\pi_{it}}^{r_{it}}(1-\pi_{it})^{(1-r_{it})}.$$

The density $f(y_i|\beta,\alpha)$ is modelled using a multivariate Bahadur model [4] (as discussed in Section 3.2.1.2), given by

$$\begin{aligned} f(y_i|X_i,\beta,\alpha) &= f(y_{i1},y_{i2},...,y_{iT}|x_{it},\beta,\alpha) \\ &= \prod_{t=1}^{T} {p_{it}}^{y_{it}}(1-p_{it})^{(1-y_{it})} \\ &\quad \times \Bigg\{1 + \sum_{t_1<t_2} \alpha_{it_1t_2} z_{it_1} z_{it_2} + \sum_{t_1<t_2<t_3} \alpha_{it_1t_2t_3} z_{it_1} z_{it_2} z_{it_3} \\ &\quad + ... + \alpha_{i12...T}\, z_{i1} z_{i2} ... z_{iT} \Bigg\}, \end{aligned} \tag{4.26}$$

where p_{it} is given by equation (4.9) and

$$\begin{aligned} z_{it} &= \frac{y_{it}-p_{it}}{\sqrt{p_{it}(1-p_{it})}}, \\ \alpha_{it_1t_2} &= E(z_{it_1} z_{it_2}), \\ \alpha_{it_1t_2t_3} &= E(z_{it_1} z_{it_2} z_{it_3}), \\ &\vdots \\ \alpha_{12...T} &= E(z_{i1} z_{i2} ... z_{iT}). \end{aligned}$$

The ML estimator $\hat{\nu} = (\hat{\beta},\hat{\alpha},\hat{\gamma})$ is obtained by solving the ML score equations given

by

$$S(\nu) = \sum_{i=1}^{N} S_i(\nu) = \sum_{i=1}^{N} \frac{\partial \log L_{ML,i}(\nu)}{\partial \nu} = 0. \tag{4.27}$$

Given correctly specified terms in (4.25), it can be shown that at true values of $\nu = (\beta, \alpha, \gamma)$, we have

$$E[S(\nu)] = 0. \tag{4.28}$$

Also, under some regularity conditions, the ML estimator $\hat{\nu}$ is asymptotically normal, such that

$$(\hat{\nu} - \nu) \sim N(0, I_\nu^{-1}),$$

where I_ν is given by

$$I_\nu = \left\{ \frac{-\partial^2 \log L(\nu | r, y^o, X)}{\partial \nu \partial \nu'} \right\}.$$

Chapter 5

Simulation Study

We conduct a simulation study to compare three models discussed in Section 4.2 to analyze longitudinal binary data with nonignorable missingness. These models are independent pseudolikelihood (IPL) proposed by Troxel et al. [2], bivariate pseudolikelihood (BPL) proposed by Sinha et al. [1], and the classical ML. We assume each subject i from $i = 1, ..., N$, is measured repeatedly over time, and the number of time points are $T = 3$. We first discuss the settings of simulations, and we continue with simulation results.

5.1 Settings of simulations

We set the simulation study by introducing the 'response model', 'missingness model', and 'estimation methods' used in simulations.

5.1.1 Response model for simulations

Assume a dichotomous response vector Y_i for each subject i from $i = 1, ..., N$, where $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})'$ follows a trivariate Bahadur model [4] as explained in Section

3.2.1.2. The joint density of Y_i is given by

$$Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}|x_i, \beta, \alpha) = \prod_{t=1}^{3} {p_{it}}^{y_{it}}(1-p_{it})^{(1-y_{it})}$$
$$\times\{1 + \alpha_{12}z_{i1}z_{i2} + \alpha_{13}z_{i1}z_{i3} + \alpha_{23}z_{i2}z_{i3} + \alpha_{123}z_{i1}z_{i2}z_{i3}\}, \qquad (5.1)$$

where

$$z_{it} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}},$$
$$\alpha_{kl} = \frac{E[(y_{ik} - p_{ik})(y_{il} - p_{il})|x_i, \beta]}{\sqrt{p_{ik}(1-p_{ik})p_{il}(1-p_{il})}},$$
$$\alpha_{123} = \frac{E[(y_{i1} - p_{i1})(y_{i2} - p_{i2})(y_{i3} - p_{i3})|x_i, \beta]}{\sqrt{p_{i1}(1-p_{i1})p_{i2}(1-p_{i2})p_{i3}(1-p_{i3})}}.$$

Also, we adopt a single, time-invariant covariate from the uniform distribution, denoted as $X_i \sim U(0, 2)$, which is assumed to be fully observed for all subjects at baseline. Now, the marginal distribution of each binary outcome Y_{it} is assumed to be Bernoulli with probability of success denoted as p_{it}. Using equation (4.9), we obtain

$$p_{it} = E(Y_{it}|X_i, \beta) = P(Y_{it} = 1|X_i, \beta) = \frac{\exp(\beta_0 + x_{it}^{'}\beta_1 + \beta_\tau(t-1))}{1 + \exp(\beta_0 + x_{it}^{'}\beta_1 + \beta_\tau(t-1))}, \qquad (5.2)$$

and, as stated, the length of the response vector for all three estimation methods are set at $T = 3$. Then we consider the pairwise correlations among repeated responses, wherever applicable by the model under simulation, and we ignore all third- or higher-order moments. We choose exchangeable correlation model for true correlations, but we varied the correlation parameter such that $\alpha \in \{0.1, 0.25, 0.40\}$. This enables us to study the properties of models when the association parameter α changes from relatively low to relatively high correlations.

5.1.2 Missingness model for simulations

The missingness model for the current simulation study is considered non-monotone and nonignorable. Sinha et al. [1] obtain a model for missing data mechanism before defining ignorability. Using the logistic regression model obtained in (4.10) and assuming that all subjects are observed at baseline $t = 1$, the true nonignorable missing data mechanism is given by

$$f(r_i|y_i, \gamma) = \prod_{t=2}^{3} {\pi_{it}}^{rit}(1 - \pi_{it})^{(1-rit)},$$

where π_{it} is specifically modeled by the logit as

$$\text{logit}(\pi_{it}) = \text{logit}\{Pr(R_{it} = 1|y_{it}, \gamma)\} = \gamma_0 + \gamma_1 y_{it}. \tag{5.3}$$

When the missingness parameter γ_1 is such that $\gamma_1 \neq 0$, the missing data mechanism is nonignorable, and the probability of missingness, π_{it}, depends on an outcome which is not observed.

5.1.3 Estimation methods for simulations

For the current simulation study, our main interest is to estimate the regression parameters $\beta = \{\beta_0, \beta_1, \beta_\tau\}$ as shown in equation (5.2). We review the estimation for each of the three methods under simulation in more details.

5.1.3.1 Independent Pseudolikelihood (IPL)

For Troxel et al.'s IPL estimator [2], we assume independence among repeated responses in vector Y_i for subject i. Using (4.11) and (5.2), we have $Y_{it} \sim Bernoulli(p_{it})$, and

$$f(y_i|X_i,\beta) = \prod_{t=1}^{3} f(y_{it}|x_{it},\beta) = \prod_{t=1}^{3} {p_{it}}^{y_{it}}(1-p_{it})^{(1-y_{it})}.$$

From (4.11) and (5.3), we have $R_{it} \sim Bernoulli(\pi_{it})$, and assuming independence among responses for each subject i, we obtain

$$f(r_i|y_i,\gamma) = \prod_{t=1}^{3} f(r_{it}|y_{it},\gamma) = \prod_{t=1}^{3} {\pi_{it}}^{r_{it}}(1-\pi_{it})^{(1-r_{it})}.$$

Then using equation (4.12), the IPL function for current simulations is given by

$$L_{IPL}(\beta,\gamma) = \prod_{i=1}^{N}\prod_{t=1}^{3}\{f(y_{it}|x_{it},\beta)\pi_{it}\}^{rit} \times \left\{\sum_{y_{it}=0}^{1} f(y_{it}|x_{it},\beta)(1-\pi_{it})\right\}^{(1-rit)}.$$

The IPL estimator $\hat{\nu} = (\hat{\beta},\hat{\gamma})$ is obtained by solving the IPL pseudoscore equations

$$S(\nu) = \frac{\partial \log L_{IPL}(\nu)}{\partial \nu} = 0.$$

The estimates of $\nu = (\beta,\gamma)$ are calculated using the Optim package in R, relying on Byrd et. al. [22]. It is based on a quasi-Newton algorithm for optimization, and provides an option for box constraints on parameters. Consequently, the asymptotic variance of the IPL estimator $\hat{\nu} = (\hat{\beta},\hat{\gamma})$ is obtained using equation (4.15).

5.1.3.2 Bivariate Pseudolikelihood (BPL)

For Sinha et al.'s BPL estimator [1], we assume that Y_{i1} is observed for all subjects at the baseline time $t = 1$. As discussed in Section 4.2.2, in order to obtain the BPL

function, we use the bivariate Bahadur model for the joint distribution of (y_{i1}, y_{it}) . Then the conditional distribution of Y_{it} is obtained as

$$f(y_{it}|y_{i1}, x_{it}, \beta, \alpha) = {p_{it1}}^{y_{it}}(1 - p_{it1})^{(1-y_{it})},$$

where p_{it1} is given by

$$p_{it1} = p_{it}\left\{1 + \alpha_{1it}\frac{(1 - p_{it})(y_{i1} - p_{i1})}{\sqrt{p_{it}(1 - p_{it})p_{i1}(1 - p_{i1})}}\right\}.$$

Then using equations (4.16)-(4.18), the BPL function for current simulations is obtained as

$$L_{BPL}(\beta, \alpha, \gamma) = \prod_{i=1}^{N} f(y_{i1}|x_{it}, \beta) \prod_{t=2}^{3}\{f(y_{it}, r_{it}|y_{i1}, x_{it}, \beta, \alpha, \gamma)\}^{r_{it}} \times \left\{\sum_{y_{it}=0}^{1} f(y_{it}, r_{it}|y_{i1}, x_{it}, \beta, \alpha, \gamma)\right\}^{(1-r_{it})}.$$

The BPL estimator $\hat{\nu} = (\hat{\beta}, \hat{\alpha}, \hat{\gamma})$ is obtained by solving the BPL pseudoscore equations given as

$$S(\nu) = \frac{\partial \log L_{BPL}(\nu)}{\partial \nu} = 0.$$

The estimates of $\nu = (\beta, \alpha, \gamma)$ were calculated using the Optim package in R, relying on Byrd et. al. [22]. It is based on a quasi-Newton algorithm for optimization, and provides an option for box constraints on parameters, which is useful in defining the association parameter α in Sinha et al.'s BPL. We specify lower and upper bounds of $[-1, 1]$ for α, such that the parameter doesn't fall outside possible probability space. The asymptotic variance of the BPL estimator $\hat{\nu} = (\hat{\beta}, \hat{\alpha}, \hat{\gamma})$ is obtained from (4.24).

5.1.3.3 Maximum Likelihood (ML)

For the ML estimator, we assume that Y_{i1} is observed for all subjects at the baseline time $t = 1$. In order to obtain the ML function, we need to specify the density $f(y_i^o, y_i^m|\beta, \alpha)$ in (4.25). We use the multivariate Bahadur model for $f(y_i^o, y_i^m|\beta, \alpha)$, where the marginal density of y_{i1} is given by

$$f(y_{i1}|x_{it}, \beta) = {p_{i1}}^{y_{i1}}(1 - p_{i1})^{(1-y_{i1})},$$

for p_{i1} as given in equation (5.2). Now, we need to specify the conditional density of y_{i2}, and y_{i3}. Using the Bahadur density given in equation (4.26), the conditional density of y_{i2} given y_{i1} is obtained as

$$f(y_{i2}|y_{i1}, x_{it}, \beta, \alpha) = {p_{i2.1}}^{y_{i2}}\ (1 - p_{i2.1})^{(1-y_{i2})},$$

where

$$p_{i2.1} = p_{i2}\left\{1 + \alpha_{12}\frac{(y_{i1} - p_{i1})(1 - p_{i2})}{\sqrt{p_{i1}(1 - p_{i1})p_{i2}(1 - p_{i2})}}\right\},$$

and similarly, the conditional density of y_{i3} given y_{i1} and y_{i2} is obtained as

$$f(y_{i3}|y_{i1}, y_{i2}, x_{it}, \beta, \alpha) = {p_{i3.12}}^{y_{i3}}\ (1 - p_{i3.12})^{(1-y_{i3})},$$

where

$$p_{i3.12} = p_{i3}\left\{1 + \frac{\alpha_{13}z_{i1}z_{i3} + \alpha_{23}z_{i2}z_{i3} + \alpha_{123}z_{i1}z_{i2}z_{i3}}{1 + \alpha_{12}z_{i1}z_{i2}}\right\},$$

where

$$z_{i1} = (y_{i1} - p_{i1})/\sqrt{p_{i1}(1-p_{i1})}$$
$$z_{i2} = (y_{i2} - p_{i2})/\sqrt{p_{i2}(1-p_{i2})}$$
$$z_{i3} = (1 - p_{i3})/\sqrt{p_{i3}(1-p_{i3})}.$$

To specify the missingness mechanism in (4.25), we have

$$f(r_i|y_i^o, y_i^m) = \prod_{t=2}^{3} f(r_{it}|y_i^o, y_i^m, \gamma),$$

where

$$f(r_{it}|y_i^o, y_i^m, \gamma) = {\pi_{it}}^{r_{it}}(1-\pi_{it})^{(1-r_{it})},$$

with

$$\pi_{it} = \frac{\exp(\gamma_0 + \gamma_1 y_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{it})},$$

for $t = 2, 3$.

5.2 Results of Simulations

We ran two sets of simulations, each with 1000 replications. We name these simulation sets, set A, and set B, respectively. Each set used two sample sizes $N = \{240, 120\}$. We refer them as the large and small sample sizes, respectively. The objective of simulations for both sets is to estimate regression parameters $\beta = (\beta_0, \beta_1, \beta_\tau)$ as stated

in the response model (5.2). The vector of association parameters is defined as

$$\boldsymbol{\alpha} = (\{\alpha_{st}\}, \alpha_{123})',$$

where applicable, and using equation (5.3). The true missingness parameters for both sets are specified as

$$\boldsymbol{\gamma} = \{\gamma_0 = -0.2, \gamma_1 = 1\},$$

such that $\gamma_1 \neq 0$, so that the missingness mechanism is nonignorable. The true values of the regression parameters β for sets A and B are specified as

$$\boldsymbol{\beta_A} = (-0.2, 0.6, -0.2),$$

and

$$\boldsymbol{\beta_B} = (1, -0.5, -0.5),$$

respectively.

The results of simulations are presented in two sets of tables. The empirical biases, mean squared errors (MSE) and coverage probabilities are presented in Tables 5.1-5.4. On the other hand, Tables 5.5-5.8 provide percentage relative biases and relative efficiencies of the estimators.

5.2.1 Bias of Estimators

For a simulation study with S replicates of datasets, suppose we have the estimates $\hat{\theta}^{(s)}$ $(s = 1, ..., S)$, where $\hat{\theta}^{(s)}$ is the value of the estimator $\hat{\theta}$ for the sth replicate. Then the empirical bias of the estimator $\hat{\theta}$ of θ is obtained as

$$\text{Bias}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}^{(s)} - \theta.$$

The corresponding empirical percentage relative bias is obtained as

$$\%\text{Bias}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\theta}^{(s)} - \theta}{\theta}\right) \times 100.$$

The percentage relative bias, %Bias, provides a better insight into the performance of estimators compared to absolute values of the bias.

It is clear from the empirical results in Tables 5.1-5.8 that the two pseudo-likelihood approaches (BPL and IPL) produce roughly unbiased estimators of the regression coefficients under all simulation settings considered. The ML method can produce some bias in estimation, but the bias tends to decrease as the sample size increases. For example, as shown in Tables 5.5 and 5.6, for the correlation parameter $\rho = 0.25$, the empirical percentage relative biases of the ML estimates of (β_1, β_τ) are obtained as $(1.80\%, -4.25\%)$ at $N = 120$; the biases, however, reduce to $(0.75\%, -0.75\%)$ at the larger sample size $N = 240$. The BPL method generally provides smaller bias as compared to the IPL method. For example, as shown in Table 5.6, when estimating (β_1, β_τ) at $\rho = 0.40$ and $N = 240$, the BPL method provides empirical percentage relative biases of $(0.82\%, 0.25\%)$, whereas the IPL method provides larger percentage relative biases of $(1.67\%, -6.15\%)$. The bias of the IPL estimator of β_τ generally increases with the increased value of the correlation parameter ρ.

5.2.2 MSE and Efficiency of Estimators

Using the same notation as used earlier in Section 5.2.1, the empirical mean squared error of the estimator $\hat{\theta}$ of θ may be obtained as

$$\text{MSE}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\theta}^{(s)} - \theta\right)^2.$$

The empirical MSE of $\hat{\theta}$ can also be expressed in the form

$$\text{MSE}(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\theta}^{(s)} - \bar{\hat{\theta}}\right)^2 - \text{Bias}^2(\hat{\theta}),$$

where $\bar{\hat{\theta}} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}^{(s)}$, the average of $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(s)}$.

The efficiency of any two estimators $\hat{\theta_1}$ and $\hat{\theta_2}$, given they are unbiased estimators of the population parameter θ, is simply defined as

$$e(\hat{\theta_1}, \hat{\theta_2}) := \frac{E[(\hat{\theta_2} - \theta)^2]}{E[(\hat{\theta_1} - \theta)^2]} := \frac{var(\hat{\theta_2})}{var(\hat{\theta_1})}, \tag{5.4}$$

where $e(\hat{\theta_1}, \hat{\theta_2})$ is the efficiency of $(\hat{\theta_1})$ relative to $(\hat{\theta_2})$ [23]. As stated earlier, we assume that the contribution of the bias to the MSE is negligible [1]. Then the relative efficiency of the BPL estimator relative to the IPL estimator is obtained as

$$e(BPL, IPL) := \frac{var(\hat{\beta}_{IPL})}{var(\hat{\beta}_{BPL})} := \frac{MSE_{(IPL)}}{MSE_{(BPL)}}, \tag{5.5}$$

and similarly, one can obtain $e(BPL, ML)$ and $e(IPL, ML)$, which denote the efficiency of the BPL estimator relative to the ML estimator and the efficiency

of the IPL estimator relative to the ML estimator, respectively.

Here we investigate the efficiency gains of the BPL estimators over the IPL estimators. We assume that the bias of any estimator makes a negligible contribution to the MSE. It appears that the BPL estimators generally provide gains in efficiency over the IPL estimators for β_τ and for correlations greater than 0.10, and for both sample sizes ($N = 120$ and $N = 240$). For all settings in Tables 5.1-5.6, the BPL estimators provide MSEs that are generally smaller than those of the corresponding IPL estimators. For the time effect β_τ, the BPL estimator is generally much more efficient than the IPL estimator when the correlation coefficient is moderate to high. For example, when $N = 120$ and $\rho = 0.25$, for the time effect β_τ, the IPL estimator is only 67% as efficient as the BPL estimator. Also, when $N = 120$ and $\rho = 0.40$, for the time effect β_τ, the IPL estimator is only 42% as efficient as the BPL estimator. As expected, the two pseudo-likelihood methods provide estimators that are nearly as efficient as the ML estimators when the correlation is weak (e.g., $\rho = 0.10$).

5.2.3 Coverage Probabilities of the Estimators

The empirical coverage probability (CP) of an estimator $\hat{\theta}$ of θ may be obtained (for 95% confidence interval) as

$$\mathrm{CP}(\hat{\theta}) = \sum_{s=1}^{S} \frac{I(\theta \in \{\hat{\theta}^{(s)} \pm 1.96\,\mathrm{S.E.}(\hat{\theta}^{(s)})\})}{S} \times 100,$$

where I is an indicator variable and S.E. denotes the standard error of an estimator. For a 95% confidence interval, the true coverage probability is 95%. The empirical coverage probability of an estimator $\hat{\theta}$ simply gives a count of confidence intervals that contains the true population parameter θ, assuming we sample repeatedly from

the population.

Tables 5.1-5.4 present empirical coverage probabilities of 95% confidence intervals for the regression parameters. It appears that the IPL method generally provides poor coverage probabilities for the time effect β_τ when the correlation is moderate or high. For example, when $N = 120$ and $\rho = 0.25$, Table 5.1 shows that for the time effect β_τ, the IPL method provides an empirical coverage probability of 90.3%, whereas the BPL and ML methods provide coverage probabilities of 93.6% and 94.6%, respectively. Also, when $N = 120$ and $\rho = 0.40$, Table 5.1 shows that for the time effect β_τ, the IPL method provides an empirical coverage probability of 90.4%, whereas both BPL and ML methods provide a better coverage probability of 93.8%.

Table 5.1: Empirical biases, mean squared errors (MSEs) and coverage probabilities of ML, BPL, and IPL estimators for exchangeable correlation. **Set A**: $\boldsymbol{\beta} = (-0.2, 0.6, -0.2)$, **N=120**.

	Bias			MSE			Coverage Probability		
	ML	BPL	IPL	ML	BPL	IPL	ML	BPL	IPL
N=120									
$\rho = 0.10$									
β_0	-0.0154	-0.0153	-0.0174	0.0844	0.0841	0.0847	95.2	95.3	95.4
β_1	0.0192	0.0188	0.0219	0.0586	0.0583	0.0587	94.4	94.4	94.7
β_τ	0.0158	0.0147	0.0195	0.0719	0.0736	0.0785	89.8	89.9	88.8
$\rho = 0.25$									
β_0	-0.0020	-0.0037	-0.0060	0.1007	0.0996	0.1021	95.0	95.5	95.5
β_1	0.0108	0.0125	0.0162	0.0733	0.0734	0.0748	94.5	95.2	94.7
β_τ	0.0085	0.0116	0.0295	0.0424	0.0471	0.0696	94.6	93.6	90.3
$\rho = 0.40$									
β_0	-0.0046	-0.0043	-0.0065	0.1118	0.1138	0.1171	93.5	95.9	95.3
β_1	0.0129	0.0142	0.0180	0.0807	0.0856	0.0877	94.5	95.6	96.0
β_τ	0.0026	0.0037	0.0306	0.0243	0.0289	0.0691	93.8	93.8	90.4

Table 5.2: Empirical biases, mean squared errors (MSEs) and coverage probabilities of ML, BPL, and IPL estimators for exchangeable correlation. **Set A**: $\boldsymbol{\beta} = (-0.2, 0.6, -0.2)$, **N=240**.

	Bias			MSE			Coverage Probability		
	ML	BPL	IPL	ML	BPL	IPL	ML	BPL	IPL
N=240									
$\rho = 0.10$									
β_0	-0.0107	-0.0105	-0.0111	0.0457	0.0458	0.0457	95.1	95.2	95.0
β_1	0.0138	0.0139	0.0151	0.0292	0.0294	0.0295	94.8	94.9	94.5
β_τ	0.0004	0.0029	0.0051	0.0341	0.0367	0.0395	93.8	92.9	91.9
$\rho = 0.25$									
β_0	-0.0027	-0.0024	-0.0038	0.0490	0.0501	0.0502	94.6	94.9	94.4
β_1	0.0045	0.0045	0.0069	0.0343	0.0355	0.0359	94.6	94.7	94.6
β_τ	0.0015	0.0045	0.0137	0.0201	0.0219	0.0333	93.7	94.0	93.4
$\rho = 0.40$									
β_0	-0.0053	-0.0036	-0.0068	0.0524	0.0564	0.0569	94.2	94.7	94.0
β_1	0.0076	0.0049	0.0100	0.0386	0.0419	0.0426	94.0	94.6	94.2
β_τ	0.0001	-0.0005	0.0123	0.0127	0.0150	0.0298	95.1	95.0	93.9

Table 5.3: Empirical biases, mean squared errors (MSEs) and coverage probabilities of ML, BPL, and IPL estimators for exchangeable correlation. **Set B**: $\boldsymbol{\beta} = (1, -0.5, -0.5)$, **N=120**.

	Bias			MSE			Coverage Probability		
	ML	BPL	IPL	ML	BPL	IPL	ML	BPL	IPL
N=120									
$\rho = 0.10$									
β_0	0.0171	0.0182	0.0201	0.1067	0.1063	0.1063	94.6	94.5	94.6
β_1	-0.0071	-0.0072	-0.0091	0.0645	0.0642	0.0649	93.9	93.6	93.5
β_τ	0.0150	0.0157	0.0150	0.0674	0.0679	0.0694	91.8	91.1	91.6
$\rho = 0.25$									
β_0	0.0373	0.0362	0.0406	0.1024	0.1039	0.1089	96.0	95.5	95.8
β_1	-0.0245	-0.0232	-0.0246	0.0653	0.0669	0.0699	94.9	95.2	94.2
β_τ	-0.0037	0.0015	0.0169	0.0398	0.0452	0.0672	93.3	93.4	92.5
$\rho = 0.40$									
β_0	0.0256	0.0169	0.0277	0.1154	0.1254	0.1290	93.8	94.8	94.3
β_1	-0.0107	-0.0032	-0.0088	0.0749	0.0832	0.0842	93.7	94.8	94.6
β_τ	-0.0257	-0.0224	0.0077	0.0278	0.0310	0.0668	93.4	93.2	90.4

Table 5.4: Empirical biases, mean squared errors (MSEs) and coverage probabilities of ML, BPL, and IPL estimators for exchangeable correlation. **Set B**: $\boldsymbol{\beta} = (1, -0.5, -0.5)$, **N=240**.

	Bias			MSE			Coverage Probability		
	ML	BPL	IPL	ML	BPL	IPL	ML	BPL	IPL
N=240									
$\rho = 0.10$									
β_0	0.0114	0.0117	0.0126	0.0511	0.0515	0.0509	95.3	95.6	95.3
β_1	-0.0093	-0.0093	-0.0102	0.0284	0.0286	0.0284	94.8	94.8	94.6
β_τ	0.0005	0.0014	0.0010	0.0321	0.0328	0.0346	92.5	92.4	92.4
$\rho = 0.25$									
β_0	0.0069	0.0063	0.0115	0.0526	0.0531	0.0550	95.4	96.3	95.9
β_1	-0.0018	-0.0014	-0.0043	0.0321	0.0322	0.0327	95.1	95.2	95.2
β_τ	-0.0048	-0.0069	0.0088	0.0203	0.0223	0.0335	93.3	93.6	93.0
$\rho = 0.40$									
β_0	0.0148	0.0123	0.0153	0.0546	0.0589	0.0597	94.5	94.9	94.8
β_1	-0.0136	-0.0108	-0.0130	0.0340	0.0359	0.0363	94.6	95.6	96.0
β_τ	-0.0082	-0.0066	-0.0011	0.0130	0.0151	0.0300	93.7	94.2	92.8

Table 5.5: Empirical percentage relative biases(%Bias) and relative efficiencies of ML, BPL and IPL estimators. **Set A**: $\boldsymbol{\beta} = (-0.2, 0.6, -0.2)$, for small sample size.

	%Bias			Relative Efficiency		
	ML	BPL	IPL	$e(BPL, ML)$	$e(IPL, ML)$	$e(BPL, IPL)$
N=120						
$\rho = 0.10$						
β_0	7.70	7.65	8.70	1.004	0.996	1.007
β_1	3.20	3.13	3.65	1.005	0.998	1.007
β_τ	-7.90	-7.35	-9.75	0.977	0.916	1.067
$\rho = 0.25$						
β_0	1.00	1.85	3.00	1.011	0.986	1.025
β_1	1.80	2.08	2.70	0.999	0.980	1.019
β_τ	-4.25	-5.80	-14.75	0.900	0.609	1.478
$\rho = 0.40$						
β_0	2.30	2.15	3.25	0.982	0.955	1.029
β_1	2.15	2.37	3.00	0.943	0.920	1.025
β_τ	-1.30	-1.85	-15.30	0.841	0.352	2.391

Table 5.6: Empirical percentage relative biases(%Bias) and relative efficiencies of ML, BPL and IPL estimators. **Set A**: $\boldsymbol{\beta} = (-0.2, 0.6, -0.2)$, for large sample size.

	%Bias			Relative Efficiency		
	ML	BPL	IPL	$e(BPL, ML)$	$e(IPL, ML)$	$e(BPL, IPL)$
N=240						
$\rho = 0.10$						
β_0	5.35	5.25	5.55	0.998	1.000	0.998
β_1	2.30	2.32	2.52	0.993	0.990	1.003
β_τ	-0.20	-1.45	-2.55	0.929	0.863	1.076
$\rho = 0.25$						
β_0	1.35	1.20	1.90	0.978	0.976	1.002
β_1	0.75	0.75	1.15	0.966	0.955	1.011
β_τ	-0.75	-2.25	-6.85	0.918	0.604	1.521
$\rho = 0.40$						
β_0	2.65	1.80	3.40	0.929	0.921	1.009
β_1	1.27	0.82	1.67	0.921	0.906	1.017
β_τ	-0.05	0.25	-6.15	0.847	0.426	1.987

Table 5.7: Empirical percentage relative biases(%Bias) and relative efficiencies of ML, BPL and IPL estimators. **Set B**: $\boldsymbol{\beta} = (1, -0.5, -0.5)$, for small sample size.

	%Bias			Relative Efficiency		
	ML	BPL	IPL	$e(BPL, ML)$	$e(IPL, ML)$	$e(BPL, IPL)$
N=120						
$\rho = 0.10$						
β_0	1.71	1.82	2.01	1.004	1.004	1.000
β_1	1.42	1.44	1.82	1.005	0.994	1.011
β_τ	-3.00	-3.14	-3.00	0.993	0.971	1.022
$\rho = 0.25$						
β_0	3.73	3.62	4.06	0.986	0.940	1.048
β_1	4.90	4.64	4.92	0.976	0.934	1.045
β_τ	0.74	-0.30	-3.38	0.881	0.592	1.487
$\rho = 0.40$						
β_0	2.56	1.69	2.77	0.920	0.895	1.029
β_1	2.14	0.64	1.76	0.900	0.890	1.012
β_τ	5.14	4.48	-1.54	0.897	0.416	2.155

Table 5.8: Empirical percentage relative biases(%Bias) and relative efficiencies of ML, BPL and IPL estimators. **Set B**: $\boldsymbol{\beta} = (1, -0.5, -0.5)$, for large sample size.

	%Bias			Relative Efficiency		
	ML	BPL	IPL	$e(BPL, ML)$	$e(IPL, ML)$	$e(BPL, IPL)$
N=240						
$\rho = 0.10$						
β_0	1.14	1.17	1.26	0.992	1.004	0.988
β_1	1.86	1.86	2.04	0.993	1.000	0.993
β_τ	-0.10	-0.28	-0.20	0.979	0.928	1.055
$\rho = 0.25$						
β_0	0.69	0.63	1.15	0.991	0.956	1.036
β_1	0.36	0.28	0.86	0.997	0.982	1.016
β_τ	0.96	1.38	-1.76	0.910	0.606	1.502
$\rho = 0.40$						
β_0	1.48	1.23	1.53	0.927	0.915	1.014
β_1	2.72	2.16	2.60	0.947	0.937	1.011
β_τ	1.64	1.32	0.22	0.861	0.433	1.987

Chapter 6

Application: Rand HRS Data

We present an application of the three methods as discussed in Section 4.2 on a real-world longitudinal dataset in order to analyze longitudinal binary data with nonignorable missingness. These methods are independent pseudolikelihood (IPL) proposed by Troxel et al. [2], bivariate pseudolikelihood (BPL) proposed by Sinha et al. [1], and the classical ML estimator. In this chapter, we introduce the RAND HRS longitudinal data which we use for the application, and review its relevant descriptive statistics in some detail. Then we compare the three methods by discussing their estimates and corresponding standard errors.

6.1 RAND HRS Data

We use the RAND HRS Longitudinal File 2016 (v2) dataset [24], which is publicly available to the researchers after registering with their website. The Health and Retirement Study (HRS) is a national household panel survey, which studies the US population aged 50 or more and their spouses, and provides many different aspects related to population ageing in an extraordinarily rich and complex way. This survey has been conducted by the Institute for Social Research at the University of Michigan

since its introduction in 1992. It had an annual longitudinal design until 1996, and has run bi-annually afterwards. On the other hand, the RAND HRS Longitudinal File 2016 (v2) dataset has been created by the RAND Center for the Study of Aging, through cleaning and processing of variables derived from original HRS data files, and making them more accessible to the researchers [25].

In each annual or bi-annual wave, the HRS survey collects data on wide range of variables, including demographics, and indicators of health, financial, income, social security, pension, health insurance, family structure, and employment status. It also provides detailed information on respondent's usage of different health services, including medical, drug, and hospital expenditures. In this chapter, we narrow our focus on three most recent waves of the HRS survey, namely, waves 11 through 13, conducted on 2012, 2014, and 2016. We choose a vector of covariates, which includes binary indicators of seven major chronic disease states. We also choose a unique binary outcome variable, which is a transformation to number of respondent's medical visits during each wave, dichotomized at a specific cutoff.

6.1.1 Summary Statistics of RAND HRS data

In this section, we review the vector of outcome variables, denoted by Y_i, and the vector of covariates, denoted by x.

6.1.1.1 Response Vector

We define a vector of response variables $Y_i = (y_{i1}, y_{i2}, y_{i3})'$ as 'number of doctor visits' in 2-years period preceding each wave interview, corresponding to waves 11, 12, and 13, respectively. Boxplots of the responses over three waves are given in Figure 6.1. As seen in Figure 6.1, all responses in vector Y_i are highly skewed to the right, where for all y_{it}, $t = 1, 2, 3$, the median and mean of the data are around 5 and 9, respectively, while the data have a wide range of 900. Also, each of the three responses shows a

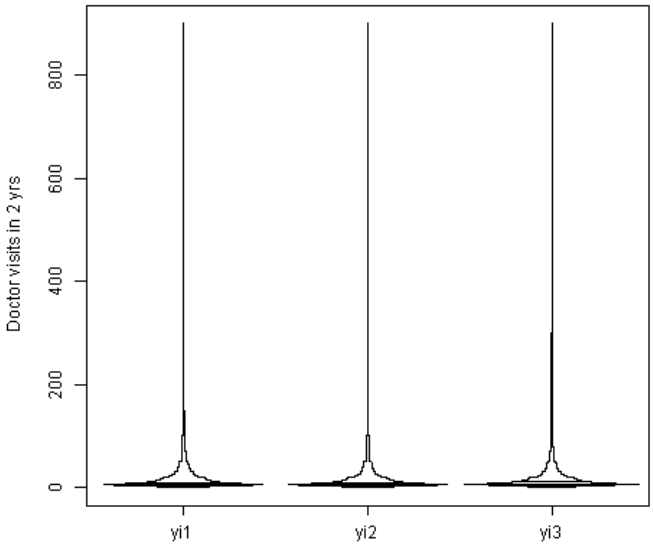

Figure 6.1: Boxplots of responses over three waves

high standard deviation ranging from 18.8 to 25.4. Using these descriptives, in order to build a model with binary responses, we transform each y_{it}, $t = 1, 2, 3$, to a binary variable, such that

$$y_{it} = \begin{cases} 1, & \text{if } y_{it} \geq 8 \\ 0, & \text{otherwise.} \end{cases}$$

Now, as stated in Section 4.2, we assume that at the baseline $t = 1$, the outcome variable y_{i1} and covariates X_i are observed completely, and we remove all observations where y_{i1} and/or one of covariates are missing. The resultant dataset has $N = 18321$ observations. We call it 'full data'. Also, we build a second dataset, by taking a random sample of size $N = 5000$ without replacement from full data. We call it 'partial data'. Table 6.1 summarizes frequency of values for the vector of responses, Y_i, for full and partial data, along with their missingness status. We assume that

all subjects are observed at baseline, therefore, no missing data are reported for y_{i1}. Based on the frequency of values, the partial data are a good representation of the full data and are suitable for our applications.

Table 6.1: Percentage of responses under different labels.

	Labels		
	Missing	0	1
Full data			
y_{i1}	0	62.6	37.4
y_{i2}	17.59	52.4	30
y_{i3}	28.71	41.7	29.6
Partial data			
y_{i1}	0	63.5	36.5
y_{i2}	18	52	29.8
y_{i3}	29.4	41	29.7

As shown in Table 6.1, the missingness for $t = 2$ and 3 are about 18% and 29%, respectively. Also, the missingness in the data is **nonmonotone**. Additionally, we analyze the effect of a covariate, e.g., the presence or absence of arthritis, on missingness. When subjects lack arthritis, probabilities of missingness in y_{i2} and y_{i3} are 16% and 25%, respectively. When subjects have arthritis, probabilities of missingness in y_{i2} and y_{i3} are 19%, and 32%, respectively. A Chi-square test shows that the proportions of missingness among subjects with arthritis vs. subjects without arthritis are different with a p-value $< 2.2e - 16$. It confirms that the missingness

is dependent on the health status of respondents at the time of interview. A sicker person is more prone to miss an interview session in non-compulsory surveys. In other words, the missingness is dependent on the unobserved outcome itself, and therefore, is considered **nonignorable**.

6.1.1.2 Covariates

We define a vector of binary covariates, $x = (x_1, ..., x_7)'$, as the presence or absence of the following medical conditions at baseline $t = 1$:

1. High blood pressure (x_1)
2. Diabetes (x_2)
3. Cancer (except skin cancer, x_3)
4. Chronic lung disease (except bronchial asthma, x_4)
5. Chronic heart disease (including coronary heart disease and heart failure, x_5)
6. Stroke (x_6)
7. Arthritis (x_7)

The average missingness among all covariates are less than 1.3%. As stated earlier, in order to satisfy our models requirements, we remove all observations where one or more covariates are missing. Table 6.2 summarizes frequency of values for the covariate vector x, for full and partial data, after removing observations with missingness. Based on the frequency of values, the partial data appear to be a good representation of the full data and are suitable for our applications. Also, after fitting a multiple logistic regression, we found that all covariates in the model are significantly associated with the binary outcome variable. Therefore, we include all covariates in our applications.

Table 6.2: Summary statistics of binary covariates in HRS data. Percentages of respondents are shown under each category.

	Covariates						
	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Full data							
0	39.9	76.6	85.8	90	76.7	92.8	43.6
1	60.1	23.4	14.2	10	23.3	7.2	56.4
Partial data							
0	40	77	85.7	90.8	76.4	93.1	43.8
1	60	23	14.3	9.2	23.6	6.9	56.2

6.2 Settings of Applications

We apply the three methods introduced in Section 4.2 on full and partial data, in order to fit the longitudinal binary model with nonignorable missingness. We are interested in estimating the regression parameters β, as shown in (5.2). We compare the approximate likelihood estimators with the exact ML estimators.

6.2.1 Response model for applications

We assume a vector of binary responses $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})'$, for each subject i ($i = 1, ..., N$), where Y_i follows a trivariate Bahadur model [4] as explained in Section 3.2.1.2, with joint probabilities as obtained in (5.1). We assume that all subjects are fully observed at the baseline and no missingness is present in Y_{i1}. Also, we adopt

a vector of seven time-invariant binary covariates $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7})'$, which are fully observed. Now, the marginal distribution of each binary outcome Y_{it} is Bernoulli with a probability of success p_{it}. Using (4.9), p_{it} is obtained as

$$p_{it} = E(Y_{it}|x_{it}, \beta) = P(Y_{it} = 1|x_{it}, \beta) = \frac{\exp(\beta_0 + x_i'\beta^* + \beta_\tau(t-1))}{1 + \exp(\beta_0 + x_i'\beta^* + \beta_\tau(t-1))},$$

where β^* is a vector of regression parameters $\beta^* = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)'$. Also, for all three estimation methods, the length of the response vector is set at $T = 3$. We consider only the pairwise correlations among repeated measures, wherever applicable by the model under study, and we ignore all third- or higher-order moments.

6.2.2 Missingness data model

Here we assume a nonignorable nonmonotone missingness mechanism. Also, we assume that all subjects are observed at baseline time $t = 1$. Then the vector of missingness indicators R_i has $T - 1 = 2$ binary random variables, such that, $R_i = (R_{i2}, R_{i3})'$. The marginal distribution of R_{it}, $t = 2, 3$ is assumed Bernoulli, where the BPL method defines the response probability π_{it} as

$$\pi_{it} = Pr(R_{it} = 1|y_{i1}, y_{it}, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{it})}.$$

For the ML and IPL methods, we ignore the effect of y_{i1} in the missing data model, so that

$$\pi_{it} = Pr(R_{it} = 1|y_{it}, \gamma) = \frac{\exp(\gamma_0 + \gamma_2 y_{it})}{1 + \exp(\gamma_0 + \gamma_2 y_{it})}.$$

6.3 Results

Table 6.3 presents the ML, BPL, and IPL estimates with their corresponding standard errors and z-values of the regression parameters of the mean response model and nuisance parameters of the missing data model for the original (full) HRS data. Table 6.4 repeats the results for a subset of the HRS data (partial data).

From Tables 6.3 and 6.4, we see that the estimates from the three methods are very similar. Standard errors of the estimates for the original (full) data as shown in Table 6.3 are also similar, as expected for a large dataset of $N = 18321$ individuals considered in the original data. For a subset of $N = 5000$ individuals as considered in Table 6.4, the BPL method, however, yields smaller estimated standard errors as compared to the IPL method. For example, from Table 6.4 we see that for the time effect, the estimated relative efficiency (ratio of estimated variances) is $(0.0194/0.0277) \times 100\% = 70\%$ for IPL versus BPL. This highlights the potential gains in efficiency from the use of the BPL method under a finite sample, in particular when the correlation among repeated measures is relatively strong.

From the parameter estimates of the missing data model, the response probability tends to decrease for individuals with increased number of doctor visits. From the estimates of the regression parameters of the mean response model, we also observe that an individual with a chronic health condition tends to make more doctor visits, as compared to an otherwise healthy individual.

Table 6.3: ML, BPL and IPL estimates of regression coefficients in mean response model and nuisance parameters in missing data model for original (full) HRS data.

	ML			BPL			IPL		
Variable	Estimate	SE	z-value	Estimate	SE	z-value	Estimate	SE	z-value
Intercept (β_0)	-1.4295	0.0239	-59.81	-1.4504	0.0243	-59.69	-1.4731	0.0243	-60.62
HBP (β_1)	0.3711	0.0234	15.86	0.3756	0.0246	15.27	0.3795	0.0244	15.55
Diabetes (β_2)	0.4312	0.0266	16.21	0.4530	0.0277	16.35	0.4589	0.0277	16.57
Cancer (β_3)	0.5266	0.0317	16.61	0.5280	0.0341	15.48	0.5446	0.0339	16.06
Lung (β_4)	0.4904	0.0378	12.97	0.5060	0.0402	12.59	0.5122	0.0401	12.77
Heart (β_5)	0.5109	0.0270	18.92	0.5282	0.0286	18.47	0.5360	0.0284	18.87
Stroke (β_6)	0.2951	0.0438	6.74	0.3050	0.0450	6.78	0.3043	0.0454	6.70
Arthritis (β_7)	0.5187	0.0226	22.95	0.5210	0.0236	22.08	0.5372	0.0235	22.86
Time (β_τ)	0.4513	0.0095	47.51	0.4615	0.0100	46.15	0.4661	0.0100	46.61
Correlation (ρ)	0.2565	0.0045	57.00	0.2456	0.0058	42.34			
Intercept (γ_0)	10.9363	7.9857	1.37	8.0037	0.3494	22.91	8.1265	0.3393	23.95
y_{i1} (γ_1)				0.4806	0.0336	14.30			
y_{it} (γ_2)	-10.6854	7.9860	-1.34	-7.9966	0.3492	-22.90	-7.8740	0.3395	-23.19

Table 6.4: ML, BPL and IPL estimates of regression coefficients in mean response model and nuisance parameters in missing data model for a subset of HRS data (partial data).

	ML			BPL			IPL		
Variable	Estimate	SE	z-value	Estimate	SE	z-value	Estimate	SE	z-value
Intercept (β_0)	-1.4598	0.0456	-32.01	-1.4915	0.0466	-32.01	-1.5186	0.0528	-28.76
HBP (β_1)	0.3520	0.0447	7.87	0.3749	0.0468	8.01	0.3758	0.0472	7.96
Diabetes β_2)	0.4454	0.0510	8.73	0.4679	0.0528	8.86	0.4765	0.0535	8.91
Cancer (β_3)	0.4765	0.0601	7.93	0.4724	0.0628	7.52	0.4977	0.0637	7.81
Lung (β_4)	0.3747	0.0736	5.09	0.3931	0.0718	5.47	0.4002	0.0730	5.48
Heart (β_5)	0.5948	0.0515	11.55	0.6029	0.0543	11.10	0.6086	0.0543	11.21
Stroke (β_6)	0.1475	0.0847	1.74	0.1659	0.0874	1.90	0.1539	0.0901	1.71
Arthritis (β_7)	0.5390	0.0435	12.39	0.5483	0.0455	12.05	0.5690	0.0495	11.49
Time (β_τ)	0.4879	0.0185	26.37	0.4998	0.0194	25.76	0.4945	0.0277	17.85
Correlation (ρ)	0.2498	0.0085	29.39	0.2417	0.0111	21.77			
Intercept (γ_0)	7.8043	3.9155	1.99	6.7698	2.9884	2.27	4.6536	2.0901	2.23
y_{i1} (γ_1)				0.5703	0.0640	8.91			
y_{it} (γ_2)	-7.5783	3.9186	-1.93	-6.8250	2.9959	-2.28	-4.4090	2.1295	-2.07

Chapter 7

Conclusions

In the current research, we focused on a common problem in longitudinal data analysis when there is nonignorable missingness in the outcome variable. We assumed that the length of the response vector is fixed at $T = 3$, $t = 1, ..., T$, and all subjects are observed completely at baseline, $t = 1$. Also the vector of covariates is considered fully observed. We conducted a simulation study and an application on real world data in order to compare the performance of two approximate likelihoods, i.e., the BPL and IPL estimators, and the classical ML estimator. We compared three methods using estimates of regression parameters, percentage relative biases of the estimates, MSE, relative efficiency of estimators, and coverage probabilities of the estimates.

In our simulation study, similar to Sinha et al. [1], we observed that the BPL estimator consistently gives smaller relative biases compared to the IPL estimator, and it performs better under moderate to strong within-subject correlations, or when the sample size is increased. Similarly [1], we showed that the BPL estimator is considerably more efficient than the IPL estimator, and yields smaller MSEs, especially when the within-subject correlation is high. In our applications on real world data, we found that at smaller sample size, the BPL estimator consistently yields smaller standard errors compared to the IPL estimator. However, when the sample size is

extremely large, two models yield very close relative efficiencies, and the IPL behaves similar to the BPL method at the asymptotes. Compared to Sinha et al. [1] where $T = 5$, the findings from current research shows new evidence for better performance of the BPL method over the IPL method, even when the response vector is relatively short at $T = 3$. On the other hand, we expect from Sinha et al. [1] that the ML estimator outperforms the approximate likelihoods when the length of the response vector is ≤ 3. Our results prove this expectation; however, we showed that the BPL estimator yields biases and standard errors competitive to the ML estimator.

List of References

[1] S. K. Sinha, A. B. Troxel, S. R. Lipsitz, D. Sinha, G. M. Fitzmaurice, G. Molenberghs, and J. G. Ibrahim, "A bivariate pseudolikelihood for incomplete longitudinal binary data with nonignorable nonmonotone missingness," *Biometrics*, vol. 67, no. 3, pp. 1119–1126, 2011.

[2] A. B. Troxel, S. R. Lipsitz, and D. P. Harrington, "Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data," *Biometrika*, vol. 85, no. 3, pp. 661–672, 1998.

[3] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, S. Zeger, *et al.*, *Analysis of longitudinal data*. Oxford University Press, 2002.

[4] R. Bahadur, "On classification based on responses to n dichotomous items," *Studies in item analysis and prediction*, vol. 6, pp. 169–176, 1961.

[5] P. McCullagh and J. Nelder, "Generalized linear models., 2nd edn.(chapman and hall: London)," *Standard book on generalized linear models*, 1989.

[6] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2001.

[7] T. Amemiya, *Advanced econometrics*. Harvard university press, 1985.

[8] C. I. Bliss, "The method of probits.," *Science*, 1934.

[9] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[10] J. H. Ware, "Linear models for the analysis of longitudinal studies," *The American Statistician*, vol. 39, no. 2, pp. 95–101, 1985.

[11] H. Cramér, *Mathematical methods of statistics*, vol. 43. Princeton university press, 1999.

[12] G. M. Fitzmaurice, N. M. Laird, and A. G. Rotnitzky, "Regression models for discrete longitudinal responses," *Statistical Science*, pp. 284–299, 1993.

[13] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, *Longitudinal data analysis*, vol. 000. Chapman Hall/CRC, 2009.

[14] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[15] L. Declerck, M. Aerts, and G. Molenberghs, "Behaviour of the likelihood ratio test statistic under a bahadur model for exchangeable binary data," *Journal of Statistical Computation and Simulation*, vol. 61, no. 1-2, pp. 15–38, 1998.

[16] G. Gong and F. J. Samaniego, "Pseudo maximum likelihood estimation: theory and applications," *The Annals of Statistics*, pp. 861–869, 1981.

[17] L. P. Zhao and R. L. Prentice, "Correlated binary regression using a quadratic exponential model," *Biometrika*, vol. 77, no. 3, pp. 642–648, 1990.

[18] J. Little Roderick and B. Rubin Donald, "Statistical analysis with missing data," *Hoboken, NJ: Wiley*, 1987.

[19] J. G. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: a review," *Test*, vol. 18, no. 1, pp. 1–43, 2009.

[20] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring, "Missing-data methods for generalized linear models: A comparative review," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 332–346, 2005.

[21] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2002.

[22] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[23] D. Wackerly, W. Mendenhall, and R. L. Scheaffer, *Mathematical statistics with applications*. Cengage Learning, 2008.

[24] Health and R. Study, "Rand hrs longitudinal file 2016 (v2) dataset," 2020. Santa Monica, CA: RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration.

[25] D. Bugliari, N. Campbell, C. Chan, O. Hayden, J. Hayes, M. Hurd, R. Main, J. Mallett, C. McCullough, E. Meijer, *et al.*, "Rand hrs longitudinal file 2016 (v2) documentation," *Santa Monica, CA: RAND Center for the Study of Aging*, 2020.

APPENDIX

R Code Used in Simulaton Study

Appendix A

```
# ------------------------------------

binary.miss.dat <- function(k=100, n=3, theta=c(1,-.5,-.5, .4, -.5,
1))
{
x <- round(runif(k, min=0, max=2), digits=2)
# x <- rep(c(0, 1, 2), c(k/3, k/3, k/3))
t0 <- c(1, 2, 3)
eta1 <- theta[1] + theta[2] * x + theta[3] * (t0[1]-1)
eta2 <- theta[1] + theta[2] * x + theta[3] * (t0[2]-1)
eta3 <- theta[1] + theta[2] * x + theta[3] * (t0[3]-1)

p1   <- exp(eta1)/(1+exp(eta1))
p2   <- exp(eta2)/(1+exp(eta2))
p3   <- exp(eta3)/(1+exp(eta3))

# print(p1)
# print(p2)
# print(p3)
```

```
y1 <- rbinom(k, 1, p1)
rho <- theta[4]
rho12 <- rho
rho13 <- rho
rho23 <- rho
# rho12 = rho ^ abs(1 - 2)
    # rho13 = rho ^ abs(1 - 3)
    # rho23 = rho ^ abs(2 - 3)
p2.1 <- p2 * (1 + rho12 * (y1-p1) * (1-p2)/ sqrt(p1*(1-p1) *
p2*(1-p2)))

y2 <- rbinom(k, 1, p2.1)

rho123 <- 0
z1 <- (y1-p1) / sqrt(p1*(1-p1))
z2 <- (y2-p2) / sqrt(p2*(1-p2))
z3 <- (1-p3)  / sqrt(p3*(1-p3))
p3.12 <- p3 * (1 + (rho13 * z1 * z3 + rho23 * z2 *z3 + rho123 * z1 *
z2 * z3) / (1 + rho12 * z1 * z2))

# print(p3.12)

y3 <- rbinom(k, 1, p3.12)

r1 <- rep(1, k) # r1 = 1 if y observed, and r1=0 if missing
zeta.2 <- theta[5] + theta[6] * y2
zeta.3 <- theta[5] + theta[6] * y3
```

```
pi.2 <- exp(zeta.2)/(1+exp(zeta.2))
pi.3 <- exp(zeta.3)/(1+exp(zeta.3))
r2 <- rbinom(k, 1, pi.2)
r3 <- rbinom(k, 1, pi.3)
dat <- cbind(y1, y2, y3, x, r1, r2, r3)
dat
}
# data0 <- binary.miss.dat(k=200)

# -------------- Exact likelihood --------------

exact.like <- function(dat=data0, k=100, n=3, initial=c(1,.5,-.5, .4,
-.2,1))
# it uses the function "optim", which is available in R
{
y <- dat[, c("y1", "y2", "y3")]
x <- dat[, "x"]
r <- dat[, c("r1", "r2", "r3")]
t0 <- c(1, 2, 3)
obj.fn <- function(theta)
{
log.dens <- 0

for (i in 1:k)
{
eta1 <- theta[1] + theta[2] * x[i] + theta[3] * (t0[1]-1)
     p1   <- exp(eta1)/(1+exp(eta1))
```

```
eta2 <- theta[1] + x[i] * theta[2] + theta[3] * (t0[2]-1)
     p2 <- exp(eta2)/(1+exp(eta2))
eta3 <- theta[1] + x[i] * theta[2] + theta[3] * (t0[3]-1)
     p3 <- exp(eta3)/(1+exp(eta3))

rho <- theta[4]
rho12 = rho
    rho13 = rho
    rho23 = rho
# rho12 = rho ^ abs(1 - 2)
    # rho13 = rho ^ abs(1 - 3)
    # rho23 = rho ^ abs(2 - 3)
rho123 <- 0

f.y1 <- dbinom(y[i,1], 1, p1)

if(all(r[i,2:3] == c(1,1))) yj.mat <- matrix(y[i, 2:3], 1, 2)
if(all(r[i,2:3] == c(0,1))) yj.mat <- rbind(c(0, y[i,3]), c(1,
y[i,3]))
if(all(r[i,2:3] == c(1,0))) yj.mat <- rbind(c(y[i,2], 0), c(y[i,2],
1))
if(all(r[i,2:3] == c(0,0))) yj.mat <- rbind(c(0,0), c(1,0), c(0,1),
c(1,1))

n.row <- nrow(yj.mat)
prod.dens.j <- 0
for(m in 1 : n.row)
```

```
{
y1 <- y[i, 1]

y2m <- yj.mat[m, 1]
y3m <- yj.mat[m, 2]

p2.1 <- p2 * (1 + rho12 * (y1-p1) * (1-p2)/ sqrt(p1*(1-p1) *
p2*(1-p2)))

p2.1 <- ifelse(p2.1 < 0, 0.00001, p2.1)
p2.1 <- ifelse(p2.1 > 1, 0.99999, p2.1)

fy2.y1 <- dbinom(y2m, 1, p2.1)

z1 <- (y1-p1)  / sqrt(p1*(1-p1))
z2 <- (y2m-p2) / sqrt(p2*(1-p2))
z3 <- (1-p3)   / sqrt(p3*(1-p3))

p3.12 <- p3 * (1 + (rho13 * z1 * z3 + rho23 * z2 *z3 + rho123 * z1 *
z2 * z3) / (1 + rho12 * z1 * z2))

p3.12 <- ifelse(p3.12 < 0, 0.00001, p3.12)
p3.12 <- ifelse(p3.12 > 1, 0.99999, p3.12)

fy3.y1y2 <- dbinom(y3m, 1, p3.12)

r2 <- r[i,2]
```

```
r3 <- r[i,3]
zeta.2 <- theta[5] + theta[6] * y2m
zeta.3 <- theta[5] + theta[6] * y3m
pi.2 <- exp(zeta.2)/(1+exp(zeta.2))
pi.3 <- exp(zeta.3)/(1+exp(zeta.3))
fr2 <- dbinom(r2, 1, pi.2)
fr3 <- dbinom(r3, 1, pi.3)
prod.dens.j0 <- fy2.y1 * fy3.y1y2 * fr2 * fr3
prod.dens.j <- prod.dens.j + prod.dens.j0
}

     log.dens0 <- log(f.y1 * prod.dens.j)
log.dens  <- log.dens + log.dens0
  }
-log.dens
}

like.fit <- optim(par=initial, fn=obj.fn, hessian=TRUE, method =
c("L-BFGS-B"))
cat(like.fit$message, "\n")
estimate <- round(like.fit$par, digits=4)
std.err <- round(sqrt(diag(solve(like.fit$hessian))), digits=4)
objective <- like.fit$value

list(estimate=estimate, std.err=std.err, objective=objective)
}
```

```
# -------------- Bivariate.Pseudo.Like --------------

pseudo.like <- function(dat=data0, k=100, n=3, initial=c(1,.5,-.5, .4,
-.2,1))
# it uses the function "optim", which is available in R
{
y <- dat[, c("y1", "y2", "y3")]
x <- dat[, "x"]
r <- dat[, c("r1", "r2", "r3")]
t0 <- c(1, 2, 3)
obj.fn <- function(theta)
{
log.dens <- 0

for (i in 1:k)
{
eta1 <- theta[1] + theta[2] * x[i] + theta[3] * (t0[1]-1)
     p1   <- exp(eta1)/(1+exp(eta1))
f.y1 <- dbinom(y[i,1], 1, p1)

prod.dens.j <- 1
       for (j in 2:n)
     {
r.ij <- r[i,j]

if(r.ij==1)
{
```

```
     etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

rho <- theta[4]
pj.1 <- pj * (1 + rho * (y[i,1]-p1) * (1-pj)/ sqrt(p1*(1-p1) *
pj*(1-pj)))

pj.1 <- ifelse(pj.1 < 0, 0.00001, pj.1)
pj.1 <- ifelse(pj.1 > 1, 0.99999, pj.1)



fyj.y1 <- dbinom(y[i,j], 1, pj.1)
zeta.j <- theta[5] + theta[6] * y[i,j]
pi.j <- exp(zeta.j)/(1+exp(zeta.j))
fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.j0 <- fyj.y1 * fr.yj

}
else
{
etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

rho <- theta[4]
pj.1 <- pj * (1 + rho * (y[i,1]-p1) * (1-pj)/ sqrt(p1*(1-p1) *
pj*(1-pj)))
```

```
pj.1 <- ifelse(pj.1 < 0, 0.00001, pj.1)
pj.1 <- ifelse(pj.1 > 1, 0.99999, pj.1)



yy <- 0

fyj.y1 <- dbinom(yy, 1, pj.1)
zeta.j <- theta[5] + theta[6] * yy
pi.j <- exp(zeta.j)/(1+exp(zeta.j))
fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.1 <- fyj.y1 * fr.yj

yy <- 1

fyj.y1 <- dbinom(yy, 1, pj.1)

zeta.j <- theta[5] + theta[6] * yy
pi.j <- exp(zeta.j)/(1+exp(zeta.j))

fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.2 <- fyj.y1 * fr.yj
prod.dens.j0 <- prod.dens.1 + prod.dens.2

}

prod.dens.j <- prod.dens.j * prod.dens.j0
```

```
     }

     log.dens0 <- log(f.y1 * prod.dens.j)
log.dens  <- log.dens + log.dens0
  }
-log.dens
}

pseudo.fit <- optim(par=initial, fn=obj.fn, hessian=TRUE, method =
c("L-BFGS-B"))
cat(pseudo.fit$message, "\n")
estimate <- round(pseudo.fit$par, digits=4)
std.err <- round(sqrt(diag(solve(pseudo.fit$hessian))), digits=4)

# -------------- Standard errors --------------

theta <- pseudo.fit$par
rho <- theta[4]

log.deriv <- 0

for (i in 1:k)
{
eta1 <- theta[1] + theta[2] * x[i] + theta[3] * (t0[1]-1)
     p1   <- exp(eta1)/(1+exp(eta1))
f.y1 <- dbinom(y[i,1], 1, p1)
dlog1.dbeta <- (y[i,1] - p1) * c(1, x[i], (t0[1]-1))
```

```
dlog1.drho  <- 0
dlog1.dgam  <- c(0,0)
dlog1.dtheta <- c(dlog1.dbeta, dlog1.drho, dlog1.dgam)
dlog2.dtheta <- 0
       for (j in 2:n)
     {

r.ij <- r[i,j]

if(r.ij==1)
{
     etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

pj.1 <- pj * (1 + rho * (y[i,1]-p1) * (1-pj)/ sqrt(p1*(1-p1) *
pj*(1-pj)))

zeta.j <- theta[5] + theta[6] * y[i,j]
pi.j   <- exp(zeta.j)/(1+exp(zeta.j))



xx.1 <- c(1, x[i], (t0[1]-1))
xx.j <- c(1, x[i], (t0[j]-1))

den <- sqrt(p1*(1-p1) * pj*(1-pj))
term1 <- - (1-pj) * p1 * (1-p1) * xx.1 / den
```

```
term2 <- - (y[i,1] - p1) * pj * (1-pj) * xx.j / den
term3 <- - 0.5 * (1-pj) * (y[i,1] - p1) * ((1-2*p1) * xx.1 + (1-2*pj)
* xx.j) / den

dpj.i.dbeta <- pj.1 * (1-pj) * xx.j + pj * rho * (term1 + term2 +
term3)



dlog2.dbeta <- (y[i,j] - pj.1)/(pj.1*(1-pj.1)) * dpj.i.dbeta
dlog2.drho  <- (y[i,j] - pj.1)/(pj.1*(1-pj.1)) *
pj * (y[i,1] - p1) * (1-pj)/ sqrt(p1*(1-p1) * pj*(1-pj))

dlog2.dgam <- (r[i,j] - pi.j) * c(1, y[i, j])
dlog2.dtheta0 <- c(dlog2.dbeta, dlog2.drho, dlog2.dgam)

}
else
{
etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

pj.1 <- pj * (1 + rho * (y[i,1]-p1) * (1-pj)/ sqrt(p1*(1-p1) *
pj*(1-pj)))

yy <- 0

zeta.j <- theta[5] + theta[6] * yy
```

```
pi.j   <- exp(zeta.j)/(1+exp(zeta.j))

fyj.y1 <- dbinom(yy, 1, pj.1)
fr.yj  <- dbinom(r[i,j], 1, pi.j)
prod.dens.1 <- fyj.y1 * fr.yj



xx.1 <- c(1, x[i], (t0[1]-1))
xx.j <- c(1, x[i], (t0[j]-1))

den <- sqrt(p1*(1-p1) * pj*(1-pj))
term1 <- - (1-pj) * p1 * (1-p1) * xx.1 / den
term2 <- - (y[i,1] - p1) * pj * (1-pj) * xx.j / den
term3 <- - 0.5 * (1-pj) * (y[i,1] - p1) * ((1-2*p1) * xx.1 + (1-2*pj)
* xx.j) / den

dpj.i.dbeta <- pj.1 * (1-pj) * xx.j + pj * rho * (term1 + term2 +
term3)



dlog2.dbeta <- (yy - pj.1)/(pj.1*(1-pj.1)) * dpj.i.dbeta
dlog2.drho  <- (yy - pj.1)/(pj.1*(1-pj.1)) *
pj * (y[i,1] - p1) * (1-pj)/ sqrt(p1*(1-p1) * pj*(1-pj))
dlog2.dgam <- (r[i,j] - pi.j) * c(1, yy)
dlog2.dtheta.1 <- prod.dens.1 * c(dlog2.dbeta, dlog2.drho, dlog2.dgam)
```

```
yy <- 1

zeta.j <- theta[5] + theta[6] * yy
pi.j   <- exp(zeta.j)/(1+exp(zeta.j))
fyj.y1 <- dbinom(yy, 1, pj.1)
fr.yj  <- dbinom(r[i,j], 1, pi.j)
prod.dens.2 <- fyj.y1 * fr.yj



xx.1 <- c(1, x[i], (t0[1]-1))
xx.j <- c(1, x[i], (t0[j]-1))

den <- sqrt(p1*(1-p1) * pj*(1-pj))
term1 <- - (1-pj) * p1 * (1-p1) * xx.1 / den
term2 <- - (y[i,1] - p1) * pj * (1-pj) * xx.j / den
term3 <- - 0.5 * (1-pj) * (y[i,1] - p1) * ((1-2*p1) * xx.1 + (1-2*pj)
* xx.j) / den

dpj.i.dbeta <- pj.1 * (1-pj) * xx.j + pj * rho * (term1 + term2 +
term3)



dlog2.dbeta <- (yy - pj.1)/(pj.1*(1-pj.1)) * dpj.i.dbeta
dlog2.drho  <- (yy - pj.1)/(pj.1*(1-pj.1)) *
pj * (y[i,1]-p1) * (1-pj)/ sqrt(p1*(1-p1) * pj*(1-pj))
dlog2.dgam <- (r[i,j] - pi.j) * c(1, yy)
dlog2.dtheta.2 <- prod.dens.2 * c(dlog2.dbeta, dlog2.drho, dlog2.dgam)
```

```
dlog2.dtheta0 <- (dlog2.dtheta.1 + dlog2.dtheta.2)/(prod.dens.1 +
prod.dens.2)

}
dlog2.dtheta <- dlog2.dtheta + dlog2.dtheta0

     }

     log.deriv0 <- dlog1.dtheta + dlog2.dtheta
log.deriv <- log.deriv + log.deriv0 %*% t(log.deriv0)

  }
sandwich.var <- solve(pseudo.fit$hessian) %*% log.deriv %*%
solve(pseudo.fit$hessian)
sand.std.err <- round(sqrt(diag(sandwich.var)), digits=4)

objective <- pseudo.fit$value

list(estimate=estimate, std.err=std.err, sand.std.err=sand.std.err,
objective=objective)
}

#-------------- Pesudo.Like.Troxel --------------
pseudo.like.troxel <- function(dat=data0, k=100, n=3,
initial=c(1,.5,-.5, -.2,1))
# it uses the function "optim", which is available in R
```

```
{
y <- dat[, c("y1", "y2", "y3")]
x <- dat[, "x"]
r <- dat[, c("r1", "r2", "r3")]
t0 <- c(1, 2, 3)
obj.fn <- function(theta)
{
log.dens <- 0

for (i in 1:k)
{
eta1 <- theta[1] + theta[2] * x[i] + theta[3] * (t0[1]-1)
     p1 <- exp(eta1)/(1+exp(eta1))
f.y1 <- dbinom(y[i,1], 1, p1)

prod.dens.j <- 1
       for (j in 2:n)
     {
r.ij <- r[i,j]

if(r.ij==1)
{
     etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

rho <- 0
pj.1 <- pj
```

```
fyj.y1 <- dbinom(y[i,j], 1, pj.1)
zeta.j <- theta[4] + theta[5] * y[i,j]
pi.j   <- exp(zeta.j)/(1+exp(zeta.j))
fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.j0 <- fyj.y1 * fr.yj

}
else
{
etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

rho <- 0
pj.1 <- pj
yy <- 0

fyj.y1 <- dbinom(yy, 1, pj.1)
zeta.j <- theta[4] + theta[5] * yy

pi.j <- exp(zeta.j)/(1+exp(zeta.j))
fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.1 <- fyj.y1 * fr.yj

yy <- 1

fyj.y1 <- dbinom(yy, 1, pj.1)
```

```
zeta.j <- theta[4] + theta[5] * yy
pi.j <- exp(zeta.j)/(1+exp(zeta.j))

fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.2 <- fyj.y1 * fr.yj
prod.dens.j0 <- prod.dens.1 + prod.dens.2
}

prod.dens.j <- prod.dens.j * prod.dens.j0

     }

     log.dens0 <- log(f.y1 * prod.dens.j)
log.dens  <- log.dens + log.dens0
  }
-log.dens
}

pseudo.fit <- optim(par=initial, fn=obj.fn,  hessian=TRUE, method =
c("L-BFGS-B"))
cat(pseudo.fit$message, "\n")
estimate <- round(pseudo.fit$par, digits=4)
std.err <- round(sqrt(diag(solve(pseudo.fit$hessian))), digits=4)
objective <- pseudo.fit$value

# -------------- Standard errors --------------
```

```
theta <- pseudo.fit$par
rho <- 0

log.deriv <- 0

for (i in 1:k)
{
eta1 <- theta[1] + theta[2] * x[i] + theta[3] * (t0[1]-1)
     p1   <- exp(eta1)/(1+exp(eta1))
f.y1 <- dbinom(y[i,1], 1, p1)
dlog1.dbeta <- (y[i,1] - p1) * c(1, x[i], (t0[1]-1))
dlog1.dgam  <- c(0,0)
dlog1.dtheta <- c(dlog1.dbeta, dlog1.dgam)
dlog2.dtheta <- 0
       for (j in 2:n)
     {

r.ij <- r[i,j]

if(r.ij==1)
{
     etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

pj.1 <- pj

zeta.j <- theta[4] + theta[5] * y[i,j]
```

```
pi.j <- exp(zeta.j)/(1+exp(zeta.j))

xx.j <- c(1, x[i], (t0[j]-1))
dpj.i.dbeta <- pj.1 * (1-pj) * xx.j
dlog2.dbeta <- (y[i,j] - pj.1)/(pj.1*(1-pj.1)) * dpj.i.dbeta

dlog2.dgam <- (r[i,j] - pi.j) * c(1, y[i, j])
dlog2.dtheta0 <- c(dlog2.dbeta, dlog2.dgam)

}
else
{
etaj <- theta[1] + x[i] * theta[2] + theta[3] * (t0[j]-1)
     pj <- exp(etaj)/(1+exp(etaj))

pj.1 <- pj

yy <- 0

zeta.j <- theta[4] + theta[5] * yy
pi.j <- exp(zeta.j)/(1+exp(zeta.j))

fyj.y1 <- dbinom(yy, 1, pj.1)
fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.1 <- fyj.y1 * fr.yj

xx.j <- c(1, x[i], (t0[j]-1))
```

```
dpj.i.dbeta <- pj.1 * (1-pj) * xx.j
dlog2.dbeta <- (yy - pj.1)/(pj.1*(1-pj.1)) * dpj.i.dbeta
dlog2.dgam <- (r[i,j] - pi.j) * c(1, yy)
dlog2.dtheta.1 <- prod.dens.1 * c(dlog2.dbeta, dlog2.dgam)



yy <- 1

zeta.j <- theta[4] + theta[5] * yy
pi.j <- exp(zeta.j)/(1+exp(zeta.j))
fyj.y1 <- dbinom(yy, 1, pj.1)
fr.yj <- dbinom(r[i,j], 1, pi.j)
prod.dens.2 <- fyj.y1 * fr.yj

xx.j <- c(1, x[i], (t0[j]-1))

dpj.i.dbeta <- pj.1 * (1-pj) * xx.j

dlog2.dbeta <- (yy - pj.1)/(pj.1*(1-pj.1)) * dpj.i.dbeta
dlog2.dgam <- (r[i,j] - pi.j) * c(1, yy)
dlog2.dtheta.2 <- prod.dens.2 * c(dlog2.dbeta, dlog2.dgam)

dlog2.dtheta0 <- (dlog2.dtheta.1 + dlog2.dtheta.2)/(prod.dens.1 +
prod.dens.2)

}
```

```
dlog2.dtheta <- dlog2.dtheta + dlog2.dtheta0

     }

     log.deriv0 <- dlog1.dtheta + dlog2.dtheta
log.deriv <- log.deriv + log.deriv0 %*% t(log.deriv0)

  }
sandwich.var <- solve(pseudo.fit$hessian) %*% log.deriv %*%
solve(pseudo.fit$hessian)
sand.std.err <- round(sqrt(diag(sandwich.var)), digits=4)

list(estimate=estimate, std.err=std.err, sand.std.err=sand.std.err,
objective=objective)
}

#-------------- Simulation --------------
set.seed(01)

simul.pseudo.like <- function(sim=2, k=100, n=3, theta=c(1,.5,-.5, .4,
-.2,1))
{
like.est    <- NULL
pseudo.est    <- NULL
pseudo.troxel.est <- NULL
for(s in 1:sim)
```

```
{
cat(s)

data0 <- binary.miss.dat(k=k, n=n, theta=theta)
like.fit          <- exact.like(dat=data0, k=k, n=n, initial=theta)
pseudo.fit        <- pseudo.like(dat=data0, k=k, n=n, initial=theta)
pseudo.troxel.fit <- pseudo.like.troxel(dat=data0, k=k, n=n,
initial=theta[-4])

like.est0          <- c(like.fit$estimate,          like.fit$std.err)
pseudo.est0        <- c(pseudo.fit$estimate,
pseudo.fit$sand.std.err)
pseudo.troxel.est0 <- c(pseudo.troxel.fit$estimate,
pseudo.troxel.fit$sand.std.err)

like.est          <- rbind(like.est, like.est0)
pseudo.est        <- rbind(pseudo.est, pseudo.est0)
pseudo.troxel.est <- rbind(pseudo.troxel.est, pseudo.troxel.est0)
cat("\n")
}

list(like.est=like.est, pseudo.est=pseudo.est,
pseudo.troxel.est=pseudo.troxel.est)
}

 # <DONE>  simul.pseudo4 <- simul.pseudo.like(sim=200, k=240, n=3,
theta=c(-0.2, 0.6, -0.2,  0.10,  -0.2, 1))
```

```
 # <DONE>  simul.pseudo5 <- simul.pseudo.like(sim=200, k=240, n=3,
theta=c(-0.2, 0.6, -0.2,  0.25,  -0.2, 1))
 # <DONE>  simul.pseudo6 <- simul.pseudo.like(sim=200, k=240, n=3,
theta=c(-0.2, 0.6, -0.2,  0.40,  -0.2, 1))

bias <- function(dat=simul.pseudo.result3, beta=c(0.5, 1, -0.4))
{
b1 <- apply(dat$like.est,          2, mean)[1:3] - beta
b2 <- apply(dat$pseudo.est,        2, mean)[1:3] - beta
b3 <- apply(dat$pseudo.troxel.est, 2, mean)[1:3] - beta

v1 <- apply(dat$like.est, 2, var)[1:3]
v2 <- apply(dat$pseudo.est, 2, var)[1:3]
v3 <- apply(dat$pseudo.troxel.est, 2, var)[1:3]
r1 <- v1 + b1^2
r2 <- v2 + b2^2
r3 <- v3 + b3^2
bias <- cbind(b1, b2, b3)
mse <- cbind(r1, r2, r3)

# coverage probabilities
n0 <- nrow(dat$like.est)

like.covpr0   <- abs(dat$like.est[, 1:3]          - matrix(rep(beta,
c(n0,n0,n0)), n0, 3)) <= 1.96 * dat$like.est[, 7:9]
new.covpr0    <- abs(dat$pseudo.est[, 1:3]        - matrix(rep(beta,
c(n0,n0,n0)), n0, 3)) <= 1.96 * dat$pseudo.est[, 7:9]
```

```
troxel.covpr0 <- abs(dat$pseudo.troxel.est[, 1:3] - matrix(rep(beta,
c(n0,n0,n0)), n0, 3)) <= 1.96 * dat$pseudo.troxel.est[, 6:8]

like.covpr   <- apply(like.covpr0,   2, sum)/n0 * 100
new.covpr    <- apply(new.covpr0,    2, sum)/n0 * 100
troxel.covpr <- apply(troxel.covpr0, 2, sum)/n0 * 100

covpr <- cbind(like.covpr, new.covpr, troxel.covpr)

result <- round(cbind(bias, mse, covpr), digits=4)
dimnames(result) <- list(c("beta0", "beta1", "beta2"), c("like.bias",
"new.bias", "troxel.bias",
  "like.mse", "new.mse", "troxel.mse", "like.covpr", "new.covpr",
"troxel.covpr"))
result
}
```