Journal of Statistical Research *ISSN 0256 - 422 X*
2010, Vol. 44, No. 1, pp. 135-146
Bangladesh

A LIKELIHOOD RATIO TEST FOR NONIGNORABLE MISSINGNESS IN INCOMPLETE BINARY LONGITUDINAL DATA

Sanjoy K. Sinha
School of Mathematics and Statistics, Carleton University
Ottawa, ON, Canada K1S 5B6
Email: sinha@math.carleton.ca

SUMMARY

Missing data are common in many clinical studies. When missingness is nonignorable, a full likelihood analysis of the data requires incorporating a missing data model into the observed data likelihood function. In this article, we study the bias of the ML estimator when the corresponding maximum likelihood is obtained under a misspecified missing data model. We further explore a likelihood ratio statistic for testing the missing data mechanism in binary longitudinal data. The empirical level and power of the test are investigated in small simulations. We also present an example using some real data obtained from a longitudinal study.

Keywords and phrases: Asymptotic bias; likelihood ratio; logistic regression; longitudinal data; missing data.

AMS Classification: MSC 2000: Primary 62F03; secondary 62F12.

1 Introduction

We encounter missing data problems in many experimental studies, including surveys and clinical trials. Little and Rubin (2002) discuss various missing data patterns or mechanisms, which concern the relationship between missingness and the values of the variables in the data. If missingness does not depend on the values of the data, missing or observed, then the data are called missing completely at random (MCAR). A less restrictive assumption is that missingness depends only on the observed values of the variables in the data, and not on the values that are missing. In this case, the missing data mechanism is called missing at random (MAR). When the data are MAR, the likelihood-based inference does not depend on the missing data mechanism (Rubin, 1976). If missingness depends on the values of the missing variables, then the missing data are called nonignorable. In the case of nonignorable missing data, it is necessary to model the distribution of the missing data mechanism. Little (1995) reviews methods for modelling the data and the missing data mechanism simultaneously,

and presents a number of examples to illustrate likelihood-based inferences via maximum likelihood or Bayesian methods.

For the analysis of binary longitudinal outcomes, multivariate logistic regression models have been extensively studied in the literature (see, for example, Zeger et al., 1988; Liang et al., 1992; Glonek and McCullagh, 1995; Fitzmaurice et al., 1996). In this article, we explore a multivariate logistic regression model for analyzing incomplete binary longitudinal data. We consider estimating the model parameters by maximizing the observed data likelihood function for nonignorable missing responses.

The paper is organized as follows. Section 2 introduces the likelihood method for analyzing incomplete binary data. Section 3 investigates the asymptotic bias of the ML estimators when the likelihood is derived under a misspecified missing data model. Section 4 explores the likelihood ratio statistic for testing the missing data mechanism. Section 5 presents an application of the likelihood ratio test using some real data obtained from a longitudinal study. Section 6 gives the conclusions of the paper.

2 Model and Notation

Suppose k individuals are observed at a fixed set of n time points, $t = 1, \ldots, n$. For the ith individual ($i = 1, \ldots, k$), the response $\mathbf{y}_i$ is a vector of n binary outcomes $(y_{i1}, \ldots, y_{in})$, which may be partially observed. The ith individual is assumed to have a $p \times 1$ vector of covariates, $\mathbf{x}_{it}$, at time t, and we assume that all the covariates are fully observed.

The marginal distribution of the tth binary outcome, y_{it}, is Bernoulli with success probability $p_{it} = E[y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}] = P(y_{it} = 1|\mathbf{x}_{it}, \boldsymbol{\beta})$, which is assumed to follow the logistic regression model

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) = \mathbf{x}_{it}^t \boldsymbol{\beta}. \tag{2.1}$$

We are interested in making inferences about the regression parameters, $\boldsymbol{\beta}$, as well as the association parameters, $\boldsymbol{\alpha}$, of the joint distribution of y_{it} and y_{il}, where the joint probability of success is

$$p_{itl} = P(y_{it} = 1, y_{il} = 1|\mathbf{x}_{it}, \mathbf{x}_{il}, \boldsymbol{\beta}, \boldsymbol{\alpha}). \tag{2.2}$$

We consider modelling this joint probability using a Bahadur type (Bahadur, 1961) multivariate binary distribution.

To model bivariate and higher-order correlations in binary data, the Bahadur multivariate binary distribution has been extensively studied in the literature (see, for example, Davidson and Bradley, 1971; Bahadur and Gupta, 1986; Prentice and Zhao, 1991; Sutradhar and Sinha, 2002). When $n = 3$, the Bahadur multivariate density has the form

$$f(y_{i1}, y_{i2}, y_{i3}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \left\{\prod_{t=1}^{3} p_{it}^{y_{it}}(1 - p_{it})^{(1-y_{it})}\right\}\Big\{1 + \rho_{12}z_{i1}z_{i2} + \rho_{13}z_{i1}z_{i3} + \rho_{23}z_{i2}z_{i3} + \rho_{123}z_{i1}z_{i2}z_{i3}\Big\}, \tag{2.3}$$

where

$$z_{it} = \frac{y_{it} - p_{it}}{\sqrt{p_{it}(1-p_{it})}},$$
$$\rho_{tl} = \text{Corr}(y_{it}, y_{il}) = \frac{E\{(y_{it} - p_{it})(y_{il} - p_{il})|\mathbf{x}_{it}, \mathbf{x}_{il}\}}{\sqrt{p_{it}(1-p_{it})p_{il}(1-p_{il})}},$$
$$\rho_{123} = \frac{E\{(y_{i1} - p_{i1})(y_{i2} - p_{i2})(y_{i3} - p_{i3})|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}\}}{\sqrt{p_{i1}(1-p_{i1})p_{i2}(1-p_{i2})p_{i3}(1-p_{i3})}},$$

for $t = 1, 2, 3$.

2.1 Missing Data Mechanism

In a typical longitudinal study, individuals are not observed at all n occasions on account of some stochastic missing data mechanism. Nonignorable models are needed when the missing data mechanism depends on the missing observations. Examples of outcome-dependent dropout are given in Little and Rubin (2002). Diggle and Kenward (1994) and Ibrahim et al. (2001) also discuss nonignorable models with outcome-dependent dropout.

To describe a missing data mechanism, we introduce n binary random variables, r_{it}, $(t = 1, \ldots, n)$, with r_{it} equal to 1 if y_{it} is observed, and 0 if y_{it} is missing. A possible model for the vector $\mathbf{r}_i = (r_{i1}, \ldots, r_{in})^t$ of missing data indicators is the binomial model:

$$f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma}) = \prod_{t=1}^{n} \pi_{it}^{r_{it}}(1-\pi_{it})^{1-r_{it}}, \tag{2.4}$$

where $\mathbf{X}_i$ is the design matrix for individual i, and π_{it} is the probability of response at time t, which may be modeled by a logistic regression in the form

$$\pi_{it} = P(r_{it} = 1|y_{it}, \mathbf{x}_{it}, \boldsymbol{\gamma}) = \frac{\exp(\gamma_0 + \gamma_1 y_{it} + \boldsymbol{\gamma}_2^t \mathbf{x}_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{it} + \boldsymbol{\gamma}_2^t \mathbf{x}_{it})}\ . \tag{2.5}$$

Note that if $\gamma_1 \neq 0$, then the missing data mechanism is nonignorable, since the probability of missingness depends on possibly unobserved data y_{it}.

Model (2.4) assumes independence between the elements in $\mathbf{r}_i$. For a more general form of the missing data mechanism involving multinomial model, see Little (1995) and Little and Rubin (2002). Ibrahim et al. (1999) consider modelling the missing data mechanism $f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma})$ as the product of a sequence of one-dimensional conditional distributions as

$$\begin{aligned} f(r_{i1}, \ldots, r_{in}|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma}) \quad &= \quad f(r_{in}|r_{i1}, \ldots, r_{i,n-1}, \mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma}_n) \\ &\quad \times f(r_{i,n-1}|r_{i1}, \ldots, r_{i,n-2}, \mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma}_{n-1}) \times \cdots \\ &\quad \times f(r_{i2}|r_{i1}, \mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma}_2)\ f(r_{i1}|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\gamma}_1), \end{aligned} \tag{2.6}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^t, \ldots, \boldsymbol{\gamma}_n^t)^t$ in which the tth element $\boldsymbol{\gamma}_t$ represents a vector of parameters for the tth conditional distribution. As r_{it} is binary, one can consider a sequence of logistic regression models for the conditional distributions in (2.6).

Note that the modeling strategy (2.6) allows flexibility in the specification of the missing data mechanism, and provides a natural way to specify the joint distribution of the missing data indicators when the knowledge about missingness of one variable influences the probability of missingness of others. Since each of the univariate distributions on the right side of (2.6) can be modeled as a logistic regression, each objective function is log-concave in the parameters. This property of log-concavity eases the computations of the maximum likelihood estimates.

The choice of appropriate covariates for the missing data model is also an important issue in missing data problems. To compare various models, we can use the likelihood ratio or Akaike information criterion for the fitted models. However, as indicated by Baker and Laird (1988) and Ibrahim et al. (1999), we need to be careful not to build a vary large model for the missing data mechanism, since the model can easily become unidentifiable due to overparameterization. Baker and Laird (1988) also point out that the issue of estimability can often arise in nonignorable missing data mechanism and it is not clear how to characterize the set of estimable parameters for a given class of models.

2.2 Likelihood Function for Nonignorable Missing Data

Let $\{(\mathbf{y}_i, \mathbf{X}_i); i = 1, \ldots, k\}$ denote the data that would occur in the absence of missing values. Also let $\mathbf{y}_{\text{obs},i}$ denote the observed values and $\mathbf{y}_{\text{mis},i}$ the missing values of $\mathbf{y}_i$. Assuming arbitrary, nonmonotone patterns of missing data in $\mathbf{y}_i$, some permutation of the indices of $\mathbf{y}_i$ can be written as $\mathbf{y}_i = (\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i})$, where $\mathbf{y}_{\text{mis},i}$ is the $n_i \times 1$ vector of missing values of $\mathbf{y}_i$. We assume that the missing data mechanism is nonignorable. We consider a parametric model for the missing data mechanism as described in the previous section.

For the ith individual, the actual observed data consist of the values $(\mathbf{y}_{\text{obs},i}, \mathbf{X}_i, \mathbf{r}_i)$. The distribution of the observed data is obtained by summing $\mathbf{y}_{\text{mis},i}$ out of the joint density of $(\mathbf{y}_i, \mathbf{r}_i)$. That is,

$$f(\mathbf{y}_{\text{obs},i}, \mathbf{r}_i | \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{\mathbf{y}_{\text{mis},i}} f(\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i} | \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) f(\mathbf{r}_i | \mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i}, \mathbf{X}_i, \boldsymbol{\gamma}). \tag{2.7}$$

The full likelihood of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is any function of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ proportional to the products of (2.7) for all k individuals:

$$L_{\text{full}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma} | \mathbf{y}_{\text{obs}}, \mathbf{X}, \mathbf{r}) \propto \prod_{i=1}^{k} f(\mathbf{y}_{\text{obs},i}, \mathbf{r}_i | \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}), \tag{2.8}$$

where $\mathbf{y}_{\text{obs}} = \{\mathbf{y}_{\text{obs},i};\ i = 1, \ldots, k\}$, $\mathbf{r} = \{\mathbf{r}_i;\ i = 1, \ldots, k\}$, and $\mathbf{X} = \{\mathbf{X}_i;\ i = 1, \ldots, k\}$ is the design matrix for all k individuals. The above likelihood function cannot be evaluated in a closed form, and numerical methods can be used to maximize the observed data likelihood function.

When the density $f(\mathbf{y}_{\text{obs},i}, \mathbf{r}_i | \mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is in the form (2.7), the ML estimating equa-

tions for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ can be expressed in the form:

$$\sum_{i=1}^{k} E\left\{ \frac{\partial \log f(\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i}|\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\} = \mathbf{0}, \tag{2.9}$$

$$\sum_{i=1}^{k} E\left\{ \frac{\partial \log f(\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i}|\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\} = \mathbf{0}, \tag{2.10}$$

and

$$\sum_{i=1}^{k} E\left\{ \frac{\partial \log f(\mathbf{r}_i|\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i}, \mathbf{X}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\} = \mathbf{0}, \tag{2.11}$$

where the conditional expectations are obtained with respect to $\mathbf{y}_{\text{mis},i}$, given the actual observed data $(\mathbf{y}_{\text{obs},i}, \mathbf{r}_i)$.

The above equations can be solved iteratively using some numerical algorithm. For example, the Newton-Raphson iterative method can be used to solve the above three equations simultaneously for the ML estimates. The initial values of the estimates of the regression parameter $\boldsymbol{\beta}$ and the association parameter $\boldsymbol{\alpha}$ may be obtained from the observed likelihood for the "complete" data, for which the likelihood function has a simple form and it is relatively easy to maximize the likelihood. The initial value of the parameter $\boldsymbol{\gamma}$ of the missing data mechanism may be chosen as the null vector $\mathbf{0}$. However, this initial value for $\boldsymbol{\gamma}$ may not always lead to convergence in the iterative method, and different sets of initial values should be investigated in such cases. Unfortunately, no software package is readily available to obtain the ML estimates. We used the **R** programming language to write small programmes for our numerical computation discussed later.

The asymptotic variance of the ML estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ can be obtained from the observed Fisher information, which can be expressed in the form:

$$\begin{aligned} I_o(\boldsymbol{\theta}) \quad = \quad & -\sum_{i=1}^{k} E\left\{ \partial \mathbf{U}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^t \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\} - \sum_{i=1}^{k} E\left\{ \mathbf{U}(\boldsymbol{\theta})\mathbf{U}^t(\boldsymbol{\theta}) \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\} \\ & + \sum_{i=1}^{k} E\left\{ \mathbf{U}(\boldsymbol{\theta}) \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\} E\left\{ \mathbf{U}^t(\boldsymbol{\theta}) \middle| \mathbf{y}_{\text{obs},i}, \mathbf{r}_i \right\}, \end{aligned} \tag{2.12}$$

where $\mathbf{U}(\boldsymbol{\theta})$ is the likelihood score function: $\mathbf{U}(\boldsymbol{\theta}) = \partial \log f(\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i}, \mathbf{r}_i|\mathbf{X}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$, by treating $(\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i}, \mathbf{r}_i)$ as complete data. In a typical longitudinal study, the main interest is in the estimation of $\boldsymbol{\beta}$, with $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ being viewed as nuisance parameters.

3 Asymptotic Bias Under Misspecified Models

A natural question raised by a missing data analysis concerns the potential bias in the ML estimators for maximizing the observed data likelihood with a misspecified missing data mechanism. In this section, we investigate the asymptotic bias of the regression estimators

obtained by fitting a model under a misspecified MAR assumption, when the "true" model involves a nonignorable missing data mechanism.

We consider a simple binary outcome, y_{it}, observed at three occasions, $t = 1, 2, 3$, which may be only partially observed for a given individual, and a single binary covariate x_i. In a clinical study, such a binary covariate x_i may represent individuals in treatment and control groups. The marginal mean function, $E(y_{it}|x_i, \boldsymbol{\beta}) = P(y_{it} = 1|x_i, \boldsymbol{\beta}) = p_{it}$, of the outcome variable y_{it} is assumed to follow the logistic regression model

$$\text{logit}(p_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right) = \beta_0 + \beta_1 x_i + \beta_2(t - 1), \tag{3.1}$$

for $t = 1, 2, 3$. The correlation between the outcomes y_{it} and y_{il} is assumed to be exchangeable: $\rho_{12} = \rho_{13} = \rho_{23} = \rho$. We also assume $\rho_{123} = 0$, for simplicity. In this setting, the bivariate density of (y_{i1}, y_{i2}, y_{i3}) may be described in a particular form of the Bahadur (1961) distribution

$$f(y_{i1}, y_{i2}, y_{i3}|x_i, \boldsymbol{\beta}, \rho) = \left\{\prod_{t=1}^{3} p_{it}^{y_{it}}(1 - p_{it})^{(1-y_{it})}\right\}\{1 + \rho z_{i1}z_{i2} + \rho z_{i1}z_{i3} + \rho z_{i2}z_{i3}\}, \tag{3.2}$$

where $z_{it} = (y_{it} - p_{it})/\sqrt{p_{it}(1 - p_{it})}$, for $t = 1, 2, 3$.

Further, to define a missing data mechanism, we consider a binary random variable r_{it}, which is 1 if the value of the corresponding outcome y_{it} is observed, and 0 if y_{it} is missing. We assume that given y_{it}, r_{it} is independent with a nonignorable missing data mechanism, and follows the simple logistic regression model

$$\text{logit}\{E(r_{it}|y_{it}, \boldsymbol{\gamma})\} = \text{logit}(\pi_{it}) = \gamma_0 + \gamma_1 y_{it}. \tag{3.3}$$

The parameters $(\boldsymbol{\beta}, \rho, \boldsymbol{\gamma})$ in models (3.1)–(3.3) can be estimated by maximizing the corresponding likelihood function. The inverse of the Hessian matrix can be used to obtain asymptotic variances of the ML estimators. Note that when $\gamma_1 = 0$ in (3.3), the missing data mechanism becomes ignorable, and there is no need to incorporate the missing data model into the likelihood function.

Here we explore the bias of the ML estimators arising from fitting a model under the misspecified MAR assumption, $\gamma_1 = 0$, where the "true" model involves a non-zero γ_1. Without loss of generality, we consider that all individuals are observed at the first time point $t = 1$, that is, $r_{i1} = 1$ for all i. Since the variables $(y_{i1}, y_{i2}, y_{i3}, x_i, r_{i2}, r_{i3})$ considered are all binary, we can find the asymptotic bias of the ML estimators by taking all $2^6 = 64$ combinations of the possible values of these binary variables, and their associated weights for a given set of parameter values. We consider the weights as functions of the "true" joint density obtained under models (3.1)–(3.3). The values of the regression parameters were fixed at $(\beta_0, \beta_1, \beta_2) = (-0.5, 0.5, -0.2)$, the exchangeable correlation at $\rho = 0.4$, and the parameters of the missing data model at $(\gamma_0, \gamma_1) = (0.5, \gamma_1)$.

Figure 1 presents the asymptotic biases of the ML estimators of the regression parameters $(\beta_0, \beta_1, \beta_2)$ obtained under the MAR assumption. Note that when the true value of γ_1 is

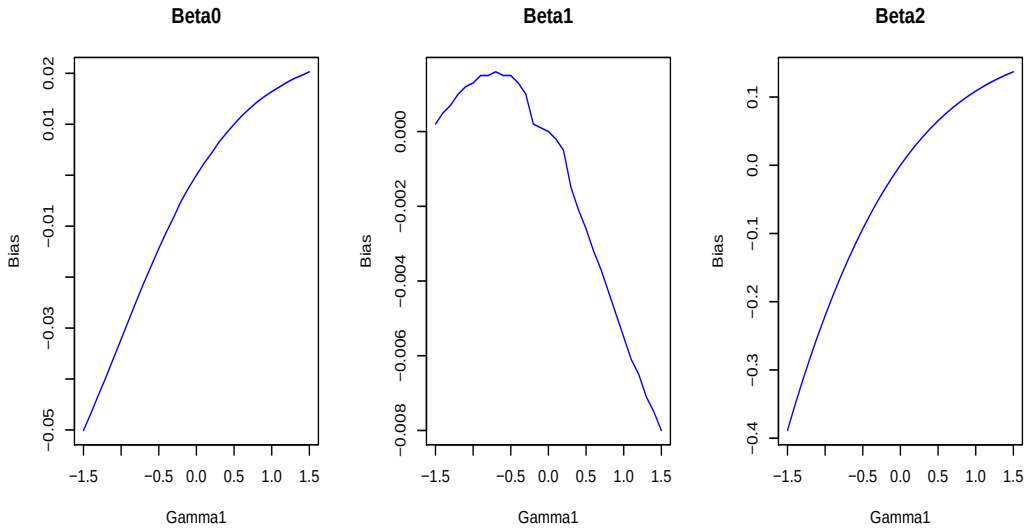

Figure 1: Asymptotic biases of regression estimators under misspecified missing data model. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-0.5, 0.5, -0.2)$.

0, the missing data mechanism is ignorable, and the fitted model is correctly specified. In this case, it is clear from the figure that the regression estimators are unbiased, as expected. For a non-zero γ_1, all estimators are found to be biased to some extent. In particular, the estimator of the time coefficient, β_2, produces huge bias when the value of γ_1 is large in magnitude. The bias of the estimator of β_2 appears to be more severe for a negative value of γ_1 as compared to a positive value of the same magnitude. The estimators of β_0 and β_1 are also found to be biased, but not to the same extent as we observe in the case of the time coefficient, β_2.

From the above simple analysis, it is clear that the ML method is sensitive to a misspecified missing data mechanism. In the next section, we explore a formal hypothesis test for testing the significance of nonignorable missingness in binary longitudinal data.

4 Likelihood Ratio Test of Missing Data Mechanism

Recall model (2.5) for the missing data mechanism. Under the null hypothesis of ignorable missingness, $H_0 : \gamma_1 = 0$, model (2.5) reduces to

$$\pi_{it} = P(r_{it} = 1 | y_{it}, \mathbf{x}_{it}, \boldsymbol{\gamma}) = \frac{\exp(\gamma_0 + \boldsymbol{\gamma}_2^t \mathbf{x}_{it})}{1 + \exp(\gamma_0 + \boldsymbol{\gamma}_2^t \mathbf{x}_{it})} \,. \tag{4.1}$$

In this case, the observed data likelihood function takes a simpler form since it does not require modeling the missing data mechanism.

Given the observed data $(\mathbf{y}_{\text{obs}}, \mathbf{X}, \mathbf{r})$, suppose L_0 is the likelihood of $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ evaluated at the ML estimators $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}})$ obtained under the null hypothesis $H_0 : \gamma_1 = 0$, and L_1 is the likelihood evaluated at the ML estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ obtained under the full likelihood (2.8). Then the likelihood ratio statistic $-2\log(L_0/L_1)$ is asymptotically distributed as chi-square with one degree of freedom. In the next section, we explore the empirical level and power of this likelihood ratio test in a small simulation study.

4.1 Simulation Study

To investigate the finite-sample properties of the likelihood ratio test, we ran a series of simulations using a multivariate logistic regression model for incomplete longitudinal data. Specifically, we generated the data $\{(y_{i1}, y_{i2}, y_{i3});\ i = 1, \ldots, k\}$ from the multivariate binary logistic model (3.2), with

$$p_{it} = P(y_{it} = 1 | x_i, \boldsymbol{\beta}) = \frac{\exp\{\beta_0 + \beta_1 x_i + \beta_2 (t-1)\}}{1 + \exp\{\beta_0 + \beta_1 x_i + \beta_2 (t-1)\}}, \tag{4.2}$$

for $t = 1, 2, 3$. The values of the regression coefficients were fixed at $(\beta_0, \beta_1, \beta_2) = (0.5, 0.5, -0.2)$, and the exchangeable correlation at $\rho = 0.40$. The missing data indicators (r_{i1}, r_{i2}, r_{i3}) were assumed to follow the joint density

$$f(r_{i1}, r_{i2}, r_{i3} | y_{i1}, y_{i2}, y_{i3}, \gamma_0, \gamma_1) = \prod_{t=1}^{3} \pi_{it}^{r_{it}} (1 - \pi_{it})^{1 - r_{it}}, \tag{4.3}$$

where $\pi_{it} = P(r_{it} = 1) = 1$ for $t = 1$, and for $t = 2, 3$,

$$\pi_{it} = P(r_{it} = 1 | y_{it}, \gamma_0, \gamma_1) = \frac{\exp(\gamma_0 + \gamma_1 y_{it})}{1 + \exp(\gamma_0 + \gamma_1 y_{it})}. \tag{4.4}$$

The value of the intercept coefficient γ_0 was fixed at 0. For each combination of $k = 150, 250$ individuals, with each individual being observed at $t = 1, 2, 3$ time points, we performed a simulation study based on 500 replicates of data sets. We set $\gamma_1 = 0$ to examine the level of significance of the likelihood ratio test for nonignorable missingness. The empirical level of significance was obtained as the proportion of samples in which a given p-value was less than $\alpha = 0.05$.

To examine the power of the likelihood ratio test, we used the same simulation configurations as above. We calculated the empirical power of the test for each combination of the values $\gamma_1 = -0.5, -1.0, -2.0$. As before, we used 500 replicates of data sets for each simulation configuration, and found the p-value of the tests.

Table 1 presents the empirical levels of the likelihood ratio tests for all simulation configurations considered. Approximate standard errors of the empirical levels are shown in parentheses. It is clear from the table that the empirical level of significance of the test is generally close to the nominal 0.05 level of significance. The level gets closer to the nominal 0.05 level when the number of individuals, k, increases, as expected. For example, when

Table 1: Empirical power of likelihood ratio test for nonignorable missingness (standard error in parenthesis).

Parameter	$k = 150$	$k = 250$
$\gamma_1 = \quad 0$	$0.042_{(0.0090)}$	$0.050_{(0.0097)}$
$\gamma_1 = -0.5$	$0.168_{(0.0167)}$	$0.304_{(0.0206)}$
$\gamma_1 = -1.0$	$0.592_{(0.0220)}$	$0.770_{(0.0188)}$
$\gamma_1 = -2.0$	$0.956_{(0.0092)}$	$0.996_{(0.0028)}$

$k = 250$, the empirical level appears to be exactly equal to 0.05. For non-zero γ_1's, the empirical powers and their approximate standard errors are also shown in Table 1. The power of the test is found to increase with the increased value of γ_1, as expected. Also, the power increases with the sample size. For example, at $\gamma_1 = -1.0$, the empirical power increased from 0.592 to 0.770 when the value of k increased from 150 to 250.

5 Example: AIDS Data

Kahn et al. (1992) and Gallant et al. (1992) analyzed a data set from two longitudinal clinical trials of HIV-infected patients. Among $N = 1528$ patients considered in the study, 431 patients were diagnosed with AIDS or AIDS-related complex. The two AIDS clinical trials are randomized phase III double-blind trials, designed to compare two therapeutic treatments, zidovudine (AZT) and didanosine (DDI). The outcome of interest at time (in weeks) $t = 0, 1, \ldots, 14$ is the patient's CD4 count sufficiency, with $y_{it} = 1$ if the CD4 count for patient i exceeds 200, and 0 otherwise. The goal was to investigate the effect of treatment on changes in CD4 cell count sufficiency over time. In this study, nonmonotone missing data occurred in the responses since some patients were not available at some follow-up time points. For example, about 79% of the patients have outcomes measured at the first three occasions.

Several predictors were considered in the analysis. The predictor AZT is an indicator variable for treatment AZT, with AZT = 1 if a patient was randomized to AZT and AZT = 0 if he/she was randomized to DDI. AIDS is also an indicator variable, with AIDS = 1 if a patient was diagnosed with AIDS, and AIDS = 0, otherwise. The predictor AGE is defined to be an indicator variable, with AGE = 1 if a patient was 35 or older at baseline period, and AGE = 0, otherwise. The predictor TIME represents the time points t.

We revisit the data to explore the missing data mechanism in the longitudinal outcomes. For illustrative purposes, we just consider the first three time points, and analyze the data with the multivariate logistic regression model discussed earlier. Specifically, we model the

Table 2: Analysis of AIDS data.

Variable	Estimate	Std error	z value
Regression model			
INTERCEPT	–0.8923	0.0724	–12.32
AIDS	–1.1013	0.1207	–9.12
AGE	0.1438	0.0852	1.69
TIME	0.0798	0.0321	2.49
AZT * TIME	–0.0184	0.0476	–0.39
Exchangeable correlation			
ρ	0.4687	0.0119	39.39
Missing data model			
INTERCEPT	1.7669	0.0686	25.76
y	0.8816	0.2952	2.99

logit of $p_{it} = P(y_{it} = 1|\mathbf{x}_{it})$, the probability that CD4 count ≥ 200 at a given time t, as

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 \text{AIDS}_i + \beta_2 \text{AGE}_i + \beta_3 \text{TIME}_t + \beta_4 \text{AZT}_i * \text{TIME}_t, \tag{5.1}$$

for $t = 0, 1, 2$, with an exchangeable correlation structure, $\text{Corr}(y_{it}, y_{il}) = \rho$. Note that model (5.1) does not include the main effects of treatment AZT since the mean number of CD4 counts can be assumed to be equal in the two treatment groups at baseline period $t = 0$. To model the missing data mechanism, we assume that given the outcomes y_{it}, the missing data indicators r_{it} are independent and follow a simple logistic regression model in the form

$$\text{logit}(\pi_{it}) = \text{logit}\{P(r_{it} = 1|y_{it}, \gamma_0, \gamma_1)\} = \gamma_0 + \gamma_1 y_{it}, \tag{5.2}$$

where π_{it} is the probability of response from patient i at time t. The model becomes ignorable under the null hypothesis $H_0 : \gamma_1 = 0$.

Here to assess the significance of the coefficient γ_1, we apply the likelihood ratio test described earlier. The likelihood ratio statistic produced a large value of 12.83. The p-value of the test based on the chi-square distribution with one degree of freedom is obtained as 0.00034. Clearly, the test indicates strong evidence against the null hypothesis $H_0 : \gamma_1 = 0$, that is, the missing data mechanism is nonignorable.

Since γ_1 is significant, we consider fitting the multivariate regression model (5.1) with the nonignorable missing data model (5.2). The ML estimates of the model parameters, and their corresponding standard errors are presented in Table 2. It is clear that the patients who were diagnosed with AIDS had lower CD4 counts, as expected. The CD4 counts tend

to increase over time. But there is no evidence of interaction effects between treatment AZT and TIME, which indicates that the effects of the two treatments AZT and DDI are not significantly different.

6 Conclusions

We have explored the asymptotic bias of regression estimators obtained by maximizing the observed data likelihood function derived under an incorrectly specified missing data model. The ML estimators generally produced large biases under the misspecified MAR assumption. It is, therefore, important to conduct a formal hypothesis test for assessing the significance of nonignorable missingness when analyzing incomplete longitudinal data.

We have studied the performance of the likelihood ratio statistic for testing the missing data mechanism in small simulations. The likelihood ratio test has been found to produce approximately the correct level of significance for moderate to large sample sizes. The power of the test has been found to be consistent in the simulations – the power increased with larger sample size as well as with larger value of the parameter indicating nonignorable missingness.

Acknowledgement

The author is grateful for the support provided by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction, Solomon H. (ed), Stanford University Press, 158–168.*

[2] Bahadur, R. R. and Gupta, J. C. (1986). Distribution optimality and second-order efficiency of test procedures. In *John Van Ryzin Adaptive statistical procedures and related topics, Institute of Mathematical Statistics, Hayward, CA, 315–331.*

[3] Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonresponse. *Journal of the American Statistical Association,* **83***, 62–69.*

[4] Davidson, R. R. and Bradley, R. A. (1971). A regression relationship for multivariate paired comparisons. *Biometrika,* **58***, 555–560.*

[5] Diggle, P. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics,* **43***, 49–94.*

[6] Fitzmaurice, G. M., Molenberghs G. and Lipsitz, S. R. (1996). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society, Series B,* **57***, 691–704.*

[7] Gallant, J. E., Moore, R. D., Richman, D. D., Keruly, J. and Chaisson, R. E. (1992). Incidence and natural history of cytomegalovirus disease in patients with advanced human immunodeficiency virus disease treated with zidovudine. *Journal of Infectious Diseases,* **166***, 1223–1227.*

[8] Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B,* **57(3)***, 533–546.*

[9] Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika,* **88***, 551–564.*

[10] Ibrahim, J. G., Lipsitz, S. R. and Chen, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society, Series B,* **61***, 173–190.*

[11] Kahn, J. O., Lagakos, S. W. and Richman, D. D. (1992). A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New England Journal of Medicine,* **327***, 581–587.*

[12] Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B,* **54***, 2–24.*

[13] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data.* John Wiley & Sons, New Jersey, 2nd Edition.

[14] Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association,* **90***, 1112–1121.*

[15] Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics 1991,* **47***, 825–839.*

[16] Rubin, D. B. (1976). Inference and missing data. *Biometrika,* **63***, 581–592.*

[17] Sutradhar, B. C. and Sinha, S. K. (2002). On pseudo-likelihood inference in the binary longitudinal mixed model. *Communications in Statistics: Theory and Methods,* **31***, 397–417.*

[18] Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics,* **44***, 1049–1060.*