Journal of Statistical Research *ISSN 0256 - 422 X*
2013, Vol. 47, No. 1, pp. 31-50

MARGINAL MODELS FOR BINARY LONGITUDINAL DATA WITH DROPOUTS

SALEHIN K. CHOWDHURY
School of Mathematics and Statistics
Carleton University, Ottawa, Ontario, K1S 5B6, Canada
Email: schowdh2@math.carleton.ca

SANJOY K. SINHA
School of Mathematics and Statistics
Carleton University, Ottawa, Ontario, K1S 5B6, Canada
Email: sinha@math.carleton.ca

SUMMARY

In this paper, we propose and explore a set of weighted generalized estimating equations for fitting regression models to longitudinal binary responses when there are dropouts. Under a given missing data mechanism, the proposed method provides unbiased estimators of the regression parameters and the association parameters. Simulations were carried out to study the robustness properties of the proposed method under both correctly specified and misspecified correlation structures. The method is also illustrated in an analysis of some actual incomplete longitudinal data on cigarette smoking trends, which were used to study coronary artery development in young adults.

Keywords and phrases: Generalized estimating equation, Inverse probability weight, Longitudinal data, Marginal model, Missing response.

AMS Classification: MSC 2000: Primary 62F10; secondary 62F35

1 Introduction

We often collect longitudinal data in biological, medical and environmental studies. The main feature of longitudinal studies is that measurements from the same subjects are taken repeatedly over a given period of time. A common goal of a longitudinal study is to characterize the change in response over time and the factors that influence the change.

Our focus is on regression models for longitudinal binary responses, in which the mean binary response at a given time is related to a set of covariates and a time trend by a known link function. The analysis of longitudinal data is often complicated by the fact that not all outcomes are observed at all occasions. In general, the outcome of a subject can be missing at one follow-up time and be observed at the next follow-up time, resulting in a large class of missing data patterns. We restrict

our attention to monotone missing data patterns resulting from attrition, where a subject may drop out prior to the end of the study and does not return. For example, in a clinical trial, a patient may drop out due to unknown reasons, possibly side effects of a treatment or curing of a disease. Once a subject drops out of the study, no more measurements are taken on that subject. Dropout patterns in longitudinal studies have been studied by many authors in the literature (for example, Little and Rubin, 1987; Diggle and Kenward, 1994; Fitzmaurice et al., 1995; and Touloumi et al., 1999).

We focus on marginal models for analyzing longitudinal binary outcomes with dropouts. A common approach for estimating the regression parameters of marginal models for longitudinal binary responses is the generalized estimating equations (GEEs) approach of Liang and Zeger (1986). Prentice (1988) extended the GEE approach for estimating both regression and association parameters of marginal models. The GEE approach is based on a "working" correlation structure, and provides consistent estimators even under a misspecified correlation structure. Lipsitz et al. (1991), Carey et al. (1993) and Fitzmaurice and Lipsitz (1995) considered analyzing the longitudinal data by modelling the association among repeated responses in terms of the marginal odds ratios.

When there are missing data, the classical GEE approaches of Liang and Zeger (1986) and Prentice (1988) are valid only when the data are missing completely at random (MCAR) (Rubin, 1976); that is, given the covariates, the missing data process is independent of both the observed and unobserved outcomes. Under a weaker assumption of missing at random (MAR), where missingness depends on the observed but not the unobserved outcomes, the classical GEE estimator may be biased (Fitzmaurice et al., 1995). Robins et al. (1995) proposed an inverse probability-weighted first order GEE approach, in which a subject's contribution to the usual GEE is reweighted by the estimated probability of dropout at the time of attrition. This method yields unbiased estimating equations and hence consistent estimators for the mean parameters when the missing data model is MAR and the probability of dropout is correctly specified.

In this paper, we incorporate the inverse probability-weights of Robins et al. (1995) into the GEE approach of Prentice (1988) for analyzing longitudinal binary responses with dropouts. We study the empirical properties of the weighted GEE method in simulations. The paper is organized as follows. Section 2 introduces the model and notation to define the response process and missing data mechanism for incomplete binary longitudinal data. Section 3 reviews the ordinary unweighted GEE and weighted GEE approaches for analyzing the incomplete data. Section 4 presents an application of the proposed method using actual longitudinal data from a health study. Section 5 presents results from a simulation study, which was carried out to investigate the empirical properties of the weighted GEE approach. Section 6 gives the conclusions of the paper.

2 Model and Notation

2.1 Response Model

Suppose K subjects are observed at a fixed set of T time points. Let Y_{it} represent a binary response variable from subject i, $i = 1, \ldots, K$, at visit t, $t = 1, \ldots, T$. For the ith subject, we can form a $T \times 1$ vector, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$, of binary response variables. Let the lower case letters y_{it} and

$\mathbf{y}_i$ denote the realizations of Y_{it} and $\mathbf{Y}_i$, respectively. Also, let $x_{it} = (1, x_{it,1}, \ldots, x_{it,p-1})'$ be a $p \times 1$ vector of covariates from subject i at time t. The covariates may be time-dependent or fixed across the entire observation times. Let $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})'$.

Assume that the marginal distribution of Y_{it} is Bernoulli:

$$Y_{it} \sim \text{Bernoulli}(p_{it}),\ i = 1, \ldots, K;\ t = 1, \ldots, T, \tag{2.1}$$

with the probability of success,

$$p_{it} = E(Y_{it}|\mathbf{x}_i) = P(Y_{it} = 1|\mathbf{x}_i). \tag{2.2}$$

Let $\mathbf{p}_i = (p_{i1}, \ldots, p_{iT})'$. Assuming that the mean of response variable Y_{it} depends only on the covariate vector for subject i at time t, i.e., $p_{it} = E(Y_{it}|\mathbf{x}_i) = E(Y_{it}|x_{it})$ (Pepe and Anderson, 1994), we consider modelling the mean response by the logistic regression:

$$\text{logit}(p_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right) = \mathbf{x}'_{it}\boldsymbol{\beta}, \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is the vector of regression parameters. The marginal variance of the response variable Y_{it} is specified as a function of the marginal mean as

$$v_{it} = \text{var}(Y_{it}|x_{it}) = p_{it}(1 - p_{it}). \tag{2.4}$$

We assume that Y_{it} and $Y_{i't'}$ are uncorrelated when $i \neq i'$. Let

$$\text{corr}(Y_{it}, Y_{it'}) = \alpha_{tt'} \tag{2.5}$$

represent the correlation between Y_{it} and $Y_{it'}$ for given x_{it}, where $\boldsymbol{\alpha} = (\alpha_{12}, \ldots, \alpha_{1T}, \alpha_{23}, \ldots, \alpha_{T-1,T})'$ is the vector of correlation parameters.

2.2 Missing Data Model

Note that attrition due to drop out and staggered entry is common in many longitudinal studies, so the response vector $\mathbf{Y}_i$ may not be completely observed for many subjects. In many cases, the responses are missing due to some stochastic missing data mechanism. To introduce a missing data model, let $\mathbf{R}_i = (R_{i1}, \ldots, R_{iT})'$ denote the response indicators for the vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$, i.e., R_{it} be the indicator variable taking the value 1 if the response Y_{it} is observed and 0 otherwise. Throughout the paper, we assume a monotone missing data pattern, where $R_{i1} \geq \ldots \geq R_{iT}$ and $R_{i1} = 1$ for all subjects.

In general, the missing data mechanism can depend on the full vector of responses $\mathbf{Y}_i$ (including the unobserved components of $\mathbf{Y}_i$) and the matrix of covariates $\mathbf{x}_i$. Let

$$\lambda_{it} = P(R_{it} = 1|R_{i1} = \ldots = R_{i,t-1} = 1, \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau}) \tag{2.6}$$

be the probability that the ith subject is observed at time t, given that the subject is observed at previous $t - 1$ time points and given the response vector $\mathbf{y}_i$ and covariate matrix $\mathbf{x}_i$. Here the

components of $\boldsymbol{\tau}$ are referred to as the "nuisance parameters" of the missing data model. We denote the observed components of the response vector $\mathbf{Y}_i$ by the vector $\mathbf{Y}_i^o$ and unobserved components by the vector $\mathbf{Y}_i^u$. In (2.6), as missingness depends on the unobserved values of the response variables, it is referred to as nonignorable (NI) missingness. Data are called missing at random (MAR) if

$$\lambda_{it} = P(R_{it} = 1 | R_{i1} = \ldots = R_{i,t-1} = 1, \mathbf{y}_i^o, \mathbf{x}_i, \boldsymbol{\tau}). \tag{2.7}$$

Data are called missing completely at random (MCAR) if $\lambda_{it} = P(R_{it} = 1 | R_{i1} = \ldots = R_{i,t-1} = 1, \mathbf{x}_i, \boldsymbol{\tau})$.

Note that in the case of dropouts, the random vector $\mathbf{R}_i = (R_{i1}, \ldots, R_{iT})'$ of binary indicators can be characterized by a single random variable

$$M_i = 1 + \sum_{t=1}^{T} R_{it}, \tag{2.8}$$

which indicates the time of dropout. In this case, the missing data or dropout process can be defined by

$$\nu_{im_i} = f_{M_i}(m_i | \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau}) = P(M_i = m_i | \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau}). \tag{2.9}$$

If we assume that all subjects are observed on the first occasion, then M_i takes on values between 2 and $T+1$, where the maximum value $(T+1)$ corresponds to a complete measurement sequence. It can be shown that

$$\begin{aligned} P(M_i = m | \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau}) &= P(R_{i2} = \cdots = R_{i,m-1} = 1, R_{im} = 0 | y_{i1}, \ldots, y_{im}, \mathbf{x}_i, \boldsymbol{\tau}) \\ &= \left\{ \prod_{t=2}^{m-1} P(R_{it} = 1 | R_{i1} = \cdots = R_{i,t-1} = 1, y_{i1}, \ldots, y_{it}, \mathbf{x}_i, \boldsymbol{\tau}) \right\} \\ &\quad \times \{ P(R_{im} = 0 | R_{i1} = \cdots = R_{i,m-1} = 1, y_{i1}, \ldots, y_{im}, \mathbf{x}_i, \boldsymbol{\tau}) \}^{\Delta} \end{aligned} \tag{2.10}$$

where $\Delta = I\{m \leq T\}$ and $I\{\}$ denotes an indicator variable.

3 Methods of Estimation

3.1 Generalized Estimating Equation

Our primary interest lies in estimating the regression parameters $\boldsymbol{\beta}$ as well as the association parameters $\boldsymbol{\alpha}$, with $\boldsymbol{\tau}$ being viewed as nuisance parameters of the missing data model. Recall that we partitioned $\mathbf{Y}_i$ into the observed components $\mathbf{Y}_i^o$ and the unobserved components $\mathbf{Y}_i^u$. Similarly, we consider partitioning the mean vector $\mathbf{p}_i$ into $\mathbf{p}_i^o$ and $\mathbf{p}_i^u$.

As a naive method, one can consider analyzing the longitudinal data by simply ignoring the missing data pattern and then estimating the model parameters based on the observed data only. In such a case, the generalized estimating equations (GEEs) of Liang and Zeger (1986) can be used for estimating the regression parameters $\boldsymbol{\beta}$:

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Y}_i^o - \mathbf{p}_i^o) = \mathbf{0}, \tag{3.1}$$

where $\mathbf{D}_i = \partial \mathbf{p}_i^o / \partial \boldsymbol{\beta}$, $\mathbf{B}_i = \text{diag}\{p_{i1}(1-p_{i1}), \ldots, p_{iT}(1-p_{iT})\}$, $\mathbf{B}_i^o$ has the same form as $\mathbf{B}_i$, but with $\mathbf{p}_i$ replaced by $\mathbf{p}_i^o$, $\mathbf{V}_i = (\mathbf{B}_i^o)^{1/2}\mathbf{R}^o(\boldsymbol{\alpha})(\mathbf{B}_i^o)^{1/2}$, and $\mathbf{R}^o(\boldsymbol{\alpha})$ is a "working" correlation matrix for $\mathbf{Y}_i^o$ depending on the vector $\boldsymbol{\alpha}$ of correlation parameters. The above equations can be solved numerically for $\hat{\boldsymbol{\beta}}$ using an iterative method.

Liang and Zeger (1986) considered estimating the association parameters $\boldsymbol{\alpha}$ by the method of moments, which uses the Pearson residuals

$$\hat{r}_{it} = \frac{(y_{it} - \hat{p}_{it})}{(1-\hat{p}_{it})^{1/2}}. \tag{3.2}$$

The moment estimators of $\alpha_{tt'}$ may be obtained as

$$\hat{\alpha}_{tt'} = \sum_{i=1}^{K} \hat{r}_{it}\hat{r}_{it'}/(K-p) \tag{3.3}$$

where p is the length of $\boldsymbol{\beta}$.

Prentice (1988) considered an extension of the GEE approach to allow joint estimation of the regression parameters $\boldsymbol{\beta}$ and the association parameters $\boldsymbol{\alpha}$. Specifically, a GEE estimator of the correlation parameter $\boldsymbol{\alpha}$ may be obtained from a second set of estimating equations by noting that the "sample correlation"

$$Z_{itu} = Z_{itu}(\beta) = \frac{(Y_{it} - p_{it})(Y_{iu} - p_{iu})}{(p_{it}q_{it}p_{iu}q_{iu})^{1/2}} \tag{3.4}$$

has mean ρ_{itu} and variance

$$w_{itu} = 1 + (1-2p_{it})(1-2p_{iu})(p_{it}q_{it}p_{iu}q_{iu})^{-1/2}\rho_{itu} - \rho_{itu}^2, \tag{3.5}$$

for $t < u < T$, $i = 1, \ldots, K$, and $t = 1, \ldots, T$. Let $\mathbf{Z}_i = (Z_{i12}, \ldots, Z_{i1T}, Z_{i23}, \ldots, Z_{i,T-1,T})'$ and $\boldsymbol{\rho}_i = (\rho_{i12}, \ldots, \rho_{i1T}, \rho_{i23}, \ldots, \rho_{i,T-1,T})'$. We denote the observed component of $\mathbf{Z}_i$ by $\mathbf{Z}_i^o$ and that of $\boldsymbol{\rho}_i$ by $\boldsymbol{\rho}_i^o$. Let, w_{itu}^o is calculated with respect to the ith and uth elements of $\mathbf{p}_i^o$ and $\boldsymbol{\rho}_i^o$.

Then following Prentice (1988), the GEE estimators of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ may be obtained by solving the estimating equations

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Y}_i^o - \mathbf{p}_i^o) = \mathbf{0}, \tag{3.6}$$

$$\mathbf{U}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \mathbf{G}_i' \mathbf{W}_i^{-1} (\mathbf{Z}_i^o - \boldsymbol{\rho}_i^o) = \mathbf{0}, \tag{3.7}$$

where $\mathbf{G}_i = \partial \boldsymbol{\rho}_i^o / \partial \boldsymbol{\alpha}$ and $\mathbf{W}_i = \text{diag}\{w_{i12}^o, \ldots, w_{i1T}^o, w_{i23}^o, \ldots, w_{i,T-1,T}^o\}$.

Note that under an MAR or NI process, $E(\mathbf{Y}_i^o | \mathbf{x}_i, \boldsymbol{\beta}) \neq \mathbf{p}_i^o$, in general, and consequently, the ordinary GEE approach may provide biased estimators of both the regression parameters $\boldsymbol{\beta}$ and the association parameters $\boldsymbol{\alpha}$ (Fitzmaurice et al., 1995). In the next section, we discuss the use of weighted generalized estimating equations for estimating the regression and association parameters when there are dropouts.

3.2 Weighted Generalized Estimating Equation

Robins et al. (1995) proposed the weighted GEE (WGEE) approach for estimating the regression parameters. In the WGEE approach, a subject's contribution to the ordinary GEEs is weighted by the inverse probability of dropout at the given time. The WGEE estimators of $\boldsymbol{\beta}$ are obtained by solving:

$$\tilde{\mathbf{U}}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \frac{1}{\nu_{im}} \mathbf{D}_i' \mathbf{V}_i^{-1} \left(\mathbf{Y}_i^o - \mathbf{p}_i^o\right) = \mathbf{0}, \tag{3.8}$$

where ν_{im} is introduced in (2.9). The estimating equations (3.8) are unbiased for $\mathbf{0}$ at the true $\boldsymbol{\beta}$ if ν_{im} is correctly specified (Fitzmaurice et al., 1995).

Note that to estimate the association parameters $\boldsymbol{\alpha}$, here we consider extending the GEE approach of Prentice (1988). Specifically, we consider estimating $\boldsymbol{\alpha}$ by solving the weighted GEEs:

$$\begin{aligned} \tilde{\mathbf{U}}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^{K} \frac{1}{\nu_{im}} \mathbf{G}_i' \mathbf{W}_i^{-1} \left(\mathbf{Z}_i^o - \boldsymbol{\rho}_i^o\right) \\ &= \sum_{i=1}^{K} \sum_{m_i=2}^{T+1} \frac{I\{M_i = m_i\}}{\nu_{im_i}} \mathbf{G}_i'(m_i) \mathbf{W}_i^{-1}(m_i) \left\{\mathbf{Z}_i(m_i) - \boldsymbol{\rho}_i(m_i)\right\} = \mathbf{0}, \end{aligned} \tag{3.9}$$

where $\mathbf{Z}_i(m_i)$ and $\boldsymbol{\delta}_i(m_i)$ are the corresponding $m_i - 1$ elements of of $\mathbf{Z}_i$ and $\boldsymbol{\rho}_i$. For instance, if $m_i = m$ then $\mathbf{Z}_i(m) = (Z_{i12}, \ldots, Z_{i1,m-1}, Z_{i23}, \ldots, Z_{i,m-2,m-1})'$ and $\boldsymbol{\rho}_i(m) = (\rho_{i12}, \ldots, \rho_{i1,m-1}, \rho_{i23}, \ldots, \rho_{i,m-2,m-1})'$. We define $\mathbf{G}_i(m_i)$ and $\mathbf{W}_i(m_i)$ analogously. Note that, the number of response patterns in $\mathbf{Z}_i$ is the same as that of in $\mathbf{Y}_i$ since everyone responds to the first observation, and there are only T possible patterns. We can show that

$$\begin{aligned} & E\left[\frac{I\{M_i = m_i\}}{\nu_{im_i}} \mathbf{G}_i'(m_i) \mathbf{W}_i^{-1}(m_i) \left\{\mathbf{Z}_i(m_i) - \boldsymbol{\rho}_i(m_i)\right\}\right] \\ &= E_{\mathbf{Y}_i}\left[\mathbf{G}_i'(m_i) \mathbf{W}_i^{-1}(m_i) \left\{\mathbf{Z}_i(m_i) - \boldsymbol{\rho}_i(m_i)\right\} E_{M_i|Y_i}\left\{\frac{I\{M_i = m_i\}}{\nu_{im_i}}\right\}\right] \\ &= E_{\mathbf{Y}_i}\left[\mathbf{G}_i'(m_i) \mathbf{W}_i^{-1}(m_i) \left\{\mathbf{Z}_i(m_i) - \boldsymbol{\rho}_i(m_i)\right\}\right] = \mathbf{0}. \end{aligned} \tag{3.10}$$

Thus the estimating equations (3.9) are unbiased for $\mathbf{0}$ at the true $\boldsymbol{\alpha}$ if ν_{im} is correctly specified. Since the estimating equations for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unbiased for $\mathbf{0}$, from the standard theory of method of moments, we can argue that the WGEE estimators $\left(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}\right)$ obtained by solving (3.8) and (3.9) are consistent for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. If the drop out probabilities ν_{im_i} are consistently estimated, then the WGEE estimators would still provide consistent estimators of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

The iterative procedure for calculating the WGEE estimators $\left(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}\right)$ begins with some starting values $(\boldsymbol{\beta}_{(0)}, \boldsymbol{\alpha}_{(0)})$ and produces updated values $(\boldsymbol{\beta}_{s+1}, \boldsymbol{\alpha}_{s+1})$ from interim values $(\boldsymbol{\beta}_s, \boldsymbol{\alpha}_s)$ by means of the iterative equations

$$\boldsymbol{\beta}_{s+1} = \boldsymbol{\beta}_s + \left(\sum_{i=1}^{K} \frac{1}{\nu_{im}} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} \mathbf{D}_i' \mathbf{V}_i^{-1} \left(\mathbf{Y}_i^o - \mathbf{p}_i^o\right), \tag{3.11}$$

$$\boldsymbol{\alpha}_{s+1} = \boldsymbol{\alpha}_s + \left(\sum_{i=1}^{K} \frac{1}{\nu_{im}} \mathbf{G}_i' \mathbf{W}_i^{-1} \mathbf{G}_i \right)^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} \mathbf{G}_i' \mathbf{W}_i^{-1} \left(\mathbf{Z}_i^o - \boldsymbol{\rho}_i^o \right), \tag{3.12}$$

for $s = 0, 1, 2, \ldots$, where the second term on the right side of each estimating equation is evaluated at the current estimates $(\boldsymbol{\beta}_s, \boldsymbol{\alpha}_s)$. Note that the association parameters $\boldsymbol{\alpha}$ are often estimated by the method of moments or by the GEE approach of Prentice (1988), without any use of inverse probability weights in the estimating equations for $\boldsymbol{\alpha}$. Here we consider finding the estimators by incorporating the weights into the GEEs for $\boldsymbol{\alpha}$. This weighted GEE is found to improve the efficiency of the estimators of $\boldsymbol{\alpha}$.

3.3 Approximate Variance of the WGEE Estimator

Similarly to White (1982), we consider approximating the variance-covariance matrices of the WGEE estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\alpha}}$ by using sandwich-type estimators. In particular, we approximate the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$ from

$$V(\tilde{\boldsymbol{\beta}}) = \mathbf{M}_{\boldsymbol{\beta}}^{-1} \mathbf{Q}_{\boldsymbol{\beta}} \mathbf{M}_{\boldsymbol{\beta}}^{-1}, \tag{3.13}$$

where the matrices $\mathbf{M}_{\boldsymbol{\beta}}$ and $\mathbf{Q}_{\boldsymbol{\beta}}$ are obtained as $\mathbf{M}_{\boldsymbol{\beta}} = \sum_{i=1}^{K} (1/\nu_{im}) \mathbf{D}_i' \, V_i^{-1} \mathbf{D}_i$ and $\mathbf{Q}_{\boldsymbol{\beta}} = \sum_i^K \mathbf{S}_{\boldsymbol{\beta},i} \mathbf{S}_{\boldsymbol{\beta},i}'$ with $\mathbf{S}_{\boldsymbol{\beta},i} = (1/\nu_{im}) \mathbf{D}_i' \mathbf{V}_i^{-1} \left(\mathbf{Y}_i^o - \mathbf{p}_i^o \right)$.

Similarly, the variance-covariance matrix of the WGEE estimator $\tilde{\boldsymbol{\alpha}}$ is obtained from

$$V(\tilde{\boldsymbol{\alpha}}) = \mathbf{M}_{\boldsymbol{\alpha}}^{-1} \mathbf{Q}_{\boldsymbol{\alpha}} \mathbf{M}_{\boldsymbol{\alpha}}^{-1}, \tag{3.14}$$

where the matrices $\mathbf{M}_{\boldsymbol{\alpha}}$ and $\mathbf{Q}_{\boldsymbol{\alpha}}$ are given by $\mathbf{M}_{\boldsymbol{\alpha}} = \sum_{i=1}^{K} (1/\nu_{im}) \mathbf{G}_i' \mathbf{W}_i^{-1} \mathbf{G}_i$ and $\mathbf{Q}_{\boldsymbol{\alpha}} = \sum_i^K \mathbf{S}_{\boldsymbol{\alpha},i} \mathbf{S}_{\boldsymbol{\alpha},i}'$ with $\mathbf{S}_{\boldsymbol{\alpha},i} = (1/\nu_{im}) \mathbf{G}_i' \mathbf{W}_i^{-1} \left(\mathbf{Z}_i^o - \boldsymbol{\rho}_i^o \right)$. The matrices $\mathbf{M}_{\boldsymbol{\beta}}$, $\mathbf{M}_{\boldsymbol{\alpha}}$, $\mathbf{Q}_{\boldsymbol{\beta}}$, and $\mathbf{Q}_{\boldsymbol{\alpha}}$ are evaluated at the WGEE estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\alpha}}$.

4 Application: Analysis of Smoking Data

We present an analysis of data on cigarette smoking trends from the Coronary Artery Development in Young Adults (CARDIA) study, an epidemiological study that recorded cardiovascular risk factors on five occasions over a 10-year period in black and white males and females (Hughes et al., 1987). This study was conducted in four urban centres (Birmingham, AL; Chicago, IL, Minneapolis, MN; and Oakland, CA) across the United States in which a total of 5,115 young adults aged 18-30 years were followed prospectively and examined up to five times from 1986 to 1996. Recruitment, restricted to blacks and whites, was carried out to achieve approximate balance in sample size with respect to age, race, gender, and education. Study participants were scheduled for visits at years 0, 2, 5, 7, and 10. We consider the first four visits and 5,078 (99.3%) young adults with self reported smoking status (yes/no) known at baseline (year 0). Data from person-exams occurring after a person's first missed exam were omitted to create a data set with monotone missingness. Specifically, 578 person-exams of a total of 17,995 were omitted to create a monotone data set.

Here the goal is to draw inferences on the change in smoking prevalence of young adults in the presence of missing data and intraperson correlation based on our proposed method. The model parameters were estimated and compared using the unweighted GEE , WGEE1 and the proposed WGEE2 methods.

Let the binary response variable $Y_{it} = 1$ if the ith individual is smoker at the tth visit, and 0 if he/she is a nonsmoker. The marginal mean response $p_{it} = E(Y_{it})$ is defined as a function of the covariates in the form

$$\begin{aligned} \text{logit}(p_{it}) = \beta_0 + \beta_1(\text{age/10})_i + \beta_2 x_t + \beta_3 \text{eduh}_i + \beta_4 \text{educ}_i \\ + \beta_5 \text{racebf}_i + \beta_6 \text{racewm}_i + \beta_7 \text{racewf}_i, \end{aligned} \tag{4.1}$$

for $i = 1, \ldots, 5078$ and $t = 1, \ldots, 4$, where age_i = age of individual i in years at baseline time; $x_t =$ year since the baseline measurement = 0,2,5,7; the binary indicators $\text{eduh}_i = 1$ if ith individual's education level is high school or less, and 0 otherwise; $\text{educ}_i = 1$ if education level is up to some college, and 0 otherwise; $\text{racebf}_i = 1$ if the person is black female, and 0 otherwise; $\text{racewm}_i = 1$ if the person is white male, and 0 otherwise; and $\text{racewf}_i = 1$ if the person is white female, and 0 otherwise.

Table 1: ML Estimates and Standard Errors of Missing Data Model Parameters for the CARDIA Study.

	Estimate	S.E	z-value
Intercept	-1.923	0.038	-51.28
Previous smoking status	0.389	0.037	10.57
Race (Black Female)	-0.197	0.046	-4.25
Race (White Male)	-0.809	0.053	-15.18
Race (White Female)	-0.722	0.062	-11.68

To estimate the model parameters in (4.1) by the WGEE methods, we first estimate the inverse probability weights based on the missing data model

$$\text{logit}(p_{it}^*) = \tau_0 + \tau_1 y_{i,t-1} + \tau_2 \text{racebf}_i + \tau_3 \text{racewm}_i + \tau_4 \text{racewf}_i, \tag{4.2}$$

where $p_{it}^* = P(R_{it} = 0 | R_{i1} = \ldots, R_{i,t-1} = 1, \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau})$ is the conditional probability that the ith individual drops out at time t.

Table 5 presents the pseudo-maximum likelihood estimates of the missing data model parameters $\boldsymbol{\tau} = (\tau_0, \tau_1, \tau_2, \tau_3, \tau_4)$, their standard errors, and the corresponding z-values. Results in this table suggest that the dropout probabilities vary across the race and gender as well as the smoking status of an individual at the previous visit. Young adults are likely to have $\exp(0.389) = 1.47$ times higher odds to miss a visit if they were a smoker (vs. nonsmoker) at the previous visit. Also the study suggests that the black males are more likely to miss a visit than any other race-gender combinations. The seven-year follow-up rates for the CARDIA study were 62%, 68%, 81% and 79% for black males, black females, white males and white females, respectively. The fitted missing data model appears to reflect this scenario.

Table 2: Estimates and Standard Errors of Regression and Association Parameters for the CARDIA Study.

Fitted Model	Covariates	GEE			WGEE1			WGEE2		
		Estimate	S.E	z-value	Estimate	S.E	z-value	Estimate	S.E	z-value
Exchangeable	Intercept	-2.544	0.221	-11.51	-2.781	0.326	-8.52	-2.782	0.327	-8.51
	Age/10	0.320	0.082	3.91	0.481	0.119	4.04	0.481	0.119	4.04
	Year Followup	-0.014	0.004	-3.35	-0.017	0.006	-2.74	-0.017	0.006	-2.76
	Education (High School or less)	1.924	0.085	22.61	1.730	0.129	13.40	1.728	0.129	13.37
	Education (Some college)	1.169	0.079	14.77	0.903	0.123	7.35	0.902	0.123	7.33
	Race (Black Female)	-0.201	0.080	-2.52	-0.285	0.103	-2.77	-0.284	0.103	-2.75
	Race (White Male)	-0.111	0.091	-1.22	-0.079	0.135	-0.58	-0.077	0.135	-0.57
	Race (White Female)	-0.104	0.089	-1.18	-0.088	0.126	-0.70	-0.086	0.126	-0.68
	α	0.709	0.017	40.92	0.706	0.016	43.56	0.727	0.015	49.43
Serial	Intercept	-2.621	0.223	-11.77	-2.773	0.323	-8.58	-2.773	0.323	-8.58
	Age/10	0.343	0.082	4.17	0.472	0.117	4.02	0.472	0.117	4.02
	Year Followup	-0.015	0.004	-3.58	-0.021	0.008	-2.69	-0.021	0.008	-2.69
	Education (High School or less)	1.943	0.086	22.61	1.737	0.128	13.55	1.737	0.128	13.55
	Education (Some college)	1.177	0.080	14.71	0.910	0.122	7.48	0.910	0.122	7.47
	Race (Black Female)	-0.207	0.080	-2.57	-0.274	0.102	-2.70	-0.274	0.102	-2.69
	Race (White Male)	-0.115	0.092	-1.25	-0.060	0.134	-0.45	-0.060	0.134	-0.45
	Race (White Female)	-0.095	0.089	-1.06	-0.082	0.126	-0.65	-0.082	0.126	-0.65
	α	0.809	0.013	62.05	0.805	0.012	66.09	0.786	0.012	67.74

We finally find the GEE, WGEE1, and WGEE2 estimators of the regression parameters in the marginal model (4.1) assuming both exchangeable and serial correlation structures. The estimates of the model parameters, their standard error and the corresponding z-values are presented in Table 6. From these results, the ordinary GEE approach appears to give somewhat different results as compared to the WGEE1 and WGEE2 approaches. For example, under the exchangeable correlation structure, the covariate age has the coefficient 0.320 by the GEE method, whereas this coefficient is 0.481 by both WGEE1 and WGEE2 methods. The results are generally very close by the two weighted GEE methods, except for the correlation coefficient α. For example, the estimate of α under the exchangeable correlation structure is 0.727 by the WGEE2 method, whereas the estimate is 0.706 by the WGEE1 method and 0.709 by the unweighted GEE method.

It is clear from Table 6 that the covariates Age, Year Followup, and Level of Education have significant influence on the smoking trend in young adults. The older individuals are more likely to be a smoker, but there is an overall trend of quitting this habit over time. These results also suggest that the level of education has strong influence on the smoking status of the subjects. For example, the young adults are estimated to have $\exp(1.728) = 5.64$ times higher odds to be a smoker if their level of education is up to high school or less than those who have a college degree or more. The results are somewhat similar under the serial and exchangeable correlation structures. However, the serial correlation structure may be preferable here, as this appears to give slightly smaller standard errors of the estimates as compared to the exchangeable correlation structure.

5 Simulation Study

To study the empirical properties of the GEE and WGEE estimators under incomplete longitudinal data, we ran two sets of simulations. In the first set, the estimators were studied under correctly specified MAR models. In the second set, the robustness properties of the estimators were studied under misspecified nonignorable missing data models. In each set of simulations, data were generated under both exchangeable and serial correlation structures among the responses. The following three methods were compared in the simulations:

i. GEE: Estimators of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained by solving the unweighted GEEs (3.6) and (3.7) following Prentice (1988).

ii. WGEE1: Estimators of $\boldsymbol{\beta}$ are obtained by solving the weighted GEEs (3.8), but estimators of $\boldsymbol{\alpha}$ are obtained by solving the unweighted GEEs (3.7).

iii. WGEE2: Estimators of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained by solving the weighted GEEs (3.8) and (3.9).

Note that although the GEE approach of Prentice (1988) has been studied extensively for complete data, but little is known about its properties under incomplete data with a stochastic missing data mechanism. We are not aware of any work that studies the performance of the WGEE1 and WGEE2 methods for incomplete longitudinal data.

5.1 Response Models for Simulations

For the simulation study, we consider a two-group design configuration with a binary response measured on four occasions. The marginal model for the mean response $E(Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})$ is given by

$$\text{logit}\,\{E(Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})\} = \beta_0 + \beta_1 x_i + \beta_2 t, \quad t = 1, \ldots, 4, \tag{5.1}$$

where x_i is a dichotomous covariate indicating the group membership for the ith individual ($i = 1, \ldots, K$) observed over a fixed set of $T = 4$ time-points, $t = 1, 2, 3, 4$. Throughout the simulations, we consider $P(X_i = 1) = 0.2$. The marginal mean $E(Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})$ is defined as a function of both x_i and time-point t.

The simulated data were generated using two types of correlation structures: exchangeable and serial, with the correlation parameter α. For the exchangeable correlation, we chose corr$(Y_{it}, Y_{it'}) = \alpha$ and for the serial correlation, we chose corr$(Y_{it}, Y_{it'}) = \alpha^{|t-t'|}$ for all (t, t'). We employ Bahadur (1961) model to generate the longitudinal data. We estimate the model parameters assuming both exchangeable and serial correlation structures. Under the "true" correlation structure, the "fitted" model assumes the same correlation as that of the true model, whereas under the "misspecified" correlation structure, these two correlations are different.

Throughout the simulations, the regression and association parameters were fixed at $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (-1, 1, 0.2)'$ and $\alpha = 0.5$, respectively. The regression parameters $\boldsymbol{\beta}$ were chosen so that the marginal means $E(Y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})$ ranged from 0.3 to 0.7. Also, the correlation parameter α was chosen so that the conditional probabilities from the Bahadur (1961) model were within the range $(0, 1)$. Each simulation run was based on 1000 replications of data sets, with each data set containing $K = 500$ subjects and a maximum of $T = 4$ observations per subject.

5.2 Dropout Models for Simulations

The dropout model was assumed to be functionally independent of the group membership, but was assumed to be dependent on the current and previous values of the response variable. That is, we assumed that

$$P(M_i = m_i|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau}) = P(M_i = m_i|y_{i1}, \ldots, y_{im_i}, \boldsymbol{\tau}). \tag{5.2}$$

We assumed that all subjects were measured at the first time-point. Since we observe the individuals at a fixed set of $T = 4$ time-points, the values of M_i can vary between 2 and 5, $m_i = 2, \ldots, 5$.

To calculate $P(M_i = m_i|y_{i1}, \ldots, y_{im_i}, \boldsymbol{\tau})$ in (5.2), the individual conditional probabilities of the missing data indicators R_{it} were obtained from

$$P(R_{it} = 0|R_{i1} = \ldots = R_{i,t-1} = 1, y_{i1}, \ldots, y_{im_i}, \boldsymbol{\tau}) = \frac{\exp(\tau_0 + \tau_P y_{i,t-1} + \tau_C y_{it})}{1 + \exp(\tau_0 + \tau_P y_{i,t-1} + \tau_C y_{it})}, \tag{5.3}$$

for $t = 2, 3, 4$, i.e., the probability of being observed at a given time is entirely determined by the previous and the current, possibly unobserved, responses. Note that the choice $\tau_C = 0$ leads a MAR model, whereas the choice $\tau_P = \tau_C = 0$ leads to the assumption that the data are missing completely at random (MCAR). For $\tau_C \neq 0$, missingness depends on a current value of the response variable Y, and the missing data become nonignorable (NI).

Using (2.10) and (5.3) , the probability $P(M_i = m_i|y_{i1}, \ldots, y_{im_i}, \boldsymbol{\tau})$ in (5.2) is obtained as

$$\nu_{im_i} = P(M_i = m_i|y_{i1}, \ldots, y_{im_i}, \boldsymbol{\tau}) \tag{5.4}$$

5.3 Estimating Dropout Probabilities

For calculating the WGEE estimators, prior to solving the iterative equations (3.11) and (3.12), we estimate the response probability weights ν_{im_i} using (2.10). From (5.4) we find the pseudo-likelihood function for $\boldsymbol{\tau}$ as

$$L(\boldsymbol{\tau}) = \prod_{i=1}^{K} P(M_i = m_i|y_{i1}, \ldots, y_{im_i}, \boldsymbol{\tau}) \tag{5.5}$$

Let $\text{logit}\{p^*_{it}(\boldsymbol{\tau})\} = \tau_0 + \tau_P y_{i,t-1} + \tau_C y_{it}$. We find the pseudo-ML estimator of $\boldsymbol{\tau}$ by maximizing the above likelihood function. From (5.5), the pseudo-score equations for $\boldsymbol{\tau}$ takes the form

$$S(\boldsymbol{\tau}) = \sum_{i=1}^{K} \left\{ - \sum_{t=2}^{m_i-1} p^*_{it}(\boldsymbol{\tau}) \mathbf{y}^*_{it} + I(m_i \leq 4)\{1 - p^*_{im}(\boldsymbol{\tau})\}\mathbf{y}^*_{im} \right\} = \mathbf{0}, \tag{5.6}$$

where $\mathbf{y}^*_{it} = (1, y_{i,t-1}, y_{it})'$. The approximate variance of the pseudo-ML estimator of $\boldsymbol{\tau}$ is obtained from the Information matrix

$$I(\boldsymbol{\tau}) = \sum_{i=1}^{K} \sum_{t=2}^{\min(m_i,4)} p^*_{it}(\boldsymbol{\tau})(1 - p^*_{it}(\boldsymbol{\tau}))\mathbf{y}^*_{it}\mathbf{y}^{*'}_{it}. \tag{5.7}$$

We apply Newton-Rapson iterative algorithm to solve the estimating equations for the pseudo-ML estimator $\hat{\boldsymbol{\tau}} = (\hat{\tau_0}, \hat{\tau_P}, \hat{\tau_C})'$. The predicted probabilities of response for individual i at time t is obtained as

$$1 - \hat{p}^*_{it} = \frac{1}{1 + \exp(\mathbf{y}^{*'}_{it}\hat{\boldsymbol{\tau}})}. \tag{5.8}$$

Then we estimate the probability of dropping out for individual i at time m_i by

$$\hat{\nu}_{im_i} = P(M_i = m_i|y_{i1}, \ldots, y_{im_i}, \hat{\boldsymbol{\tau}}) = \left\{ \prod_{i=2}^{m_i-1} (1 - \hat{p}^*_{it}) \right\} \times \left\{ \hat{p}^*_{i,m_i} \right\}^{I\{m_i \leq 4\}}. \tag{5.9}$$

We replace ν_{im}'s in equations (3.11) and (3.12) by $\hat{\nu}_{im}$ and then solve these equations iteratively for the WGEE estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$.

In the first set of simulations, the data were generated using a MAR model. The parameter values of the dropout model (5.3) were chosen as $\boldsymbol{\tau} = (\tau_0, \tau_P, \tau_C)' = (-2, 2, 0)'$ and $(-2, 3, 0)'$. For these two choices of $\boldsymbol{\tau}$, the data contained roughly 30% and 40% missing values, respectively. Table 1 presents the empirical percentage relative biases, mean squared errors, and coverage probabilities of the GEE, WGEE1 and WGEE2 estimators of the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$

and the correlation parameter α for $(\tau_0, \tau_P, \tau_C) = (-2, 2, 0)$. Table 2 repeats these results for $(\tau_0, \tau_P, \tau_C) = (-2, 3, 0)$.

In the second set of simulations, we study the effects of misspecified missing data models on the estimators of the regression and association parameters. As before, the data were generated from the binary longitudinal model (5.1), but with a nonignorable missing data model (5.3) with non-zero $\tau_C = 0.5, 1.0$. Table 3 presents the empirical percentage relative biases, MSEs and coverage probabilities of the estimators for $\boldsymbol{\tau} = (\tau_0, \tau_P, \tau_C) = (-2, 2, 0.5)$ for which the data contained roughly 32% missing observations. Table 4 repeats the results for $\boldsymbol{\tau} = (\tau_0, \tau_P, \tau_C) = (-2, 1, 1)$ for which the data contained roughly 40% missing observations. In both cases, we estimated the model parameters under the misspecified MAR model.

5.4 Diagnostic methods

We compare three methods based on empirical biases, mean squared errors and coverage probabilities of the estimators. Specifically, the bias of an estimator $\hat{\theta}$ of θ is estimated by

$$\text{bias}(\hat{\theta}) \approx \sum_{s=1}^{S} \frac{(\hat{\theta}_s - \theta)}{S}, \tag{5.10}$$

where $\hat{\theta}_s$ is the estimate of θ obtained from the sth simulated data set and S is the simulation size. We calculate the percentage relative bias as

$$\frac{\text{bias}(\hat{\theta})}{\theta} \times 100. \tag{5.11}$$

Note that as the values of model parameters chosen in the simulations are different in magnitude, we consider calculating the percentage relative biases rather than the absolute biases of the estimators.

The mean squared error (MSE) of $\hat{\theta}$ is estimated by

$$\text{MSE}(\hat{\theta}) \approx \sum_{s=1}^{S} \frac{(\hat{\theta}_s - \theta)^2}{S}. \tag{5.12}$$

We also find the coverage probabilities of an estimator $\hat{\theta}$ for 95% confidence intervals, $\hat{\theta} \pm 1.96 \times \text{SE}(\hat{\theta})$, where $\text{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$. The empirical coverage probability (CP) is obtained from

$$\text{CP}(\hat{\theta}) = \frac{1}{S} \sum_{s=1}^{S} I\left\{ |\hat{\theta}_s - \theta| \leq 1.96 \times \text{SE}(\hat{\theta}) \right\}, \tag{5.13}$$

where $I\{\}$ is an indicator variable.

5.5 Results

It is clear from Table 1 that both WGEE1 and WGEE2 methods provide approximately unbiased estimates of the regression parameters under all simulation configurations considered. The biases of

these estimators increase slightly when the proportion of missing observations increases, as shown in Table 2. We expect such small increase in bias for the larger proportion of missing data, as fewer observations are available to model the missing data pattern in this case. We have noticed in further simulations (not shown here) that these biases decrease with larger sample sizes.

On the other hand, the unweighted GEE approach, generally provides large biases for both the regression parameters and the correlation parameter, as expected. For example, as shown in Table 2, under the correctly specified serial correlation structure, the GEE estimator of β_2 gives -46.39% relative bias and that of α gives -28.48% relative bias. When comparing the WGEE1 and WGEE2 methods, our proposed WGEE2 method generally provides smaller bias for the correlation parameter α. For example, from Table 2, under the correctly specified serial correlation, the WGEE1 estimator of α gives -14.14% relative bias, whereas the proposed WGEE2 estimator of α gives a much smaller relative bias of -1.99%.

The WGEE2 estimators also give better coverage probabilities; in particular, for the correlation parameter α. For example, as shown in Table 1, under the correctly specified exchangeable correlation structure, the WGEE2 estimator of α gives an empirical coverage probability of 97%, which is close to the nominal 95% confidence level. On the other hand, the GEE and WGEE1 estimators of α give empirical coverage probabilities of 35% and 49%, respectively.

The mean squared errors of the WGEE2 estimators are also smaller as compared to the WGEE1 and GEE estimators. Under misspecified correlation structures, although the WGEE2 method provides slightly larger biases, but these biases are still smaller than those obtained by the GEE and WGEE1 methods. In this sense, the WGEE2 method is considered to be more robust than the other two methods.

It is clear from Tables 3 and 4 that all three methods provide biased estimators of the model parameters under the misspecified missing data model. However, the extent of the bias from the WGEE2 method is less severe as compared to the GEE and WGEE1 methods. The WGEE2 method also provides smaller MSEs and better coverage probabilities as compared to the other two methods.

6 Conclusions

The purpose of this research was to provide a better alternative to the unweighted GEE models of Prentice (1988) for analyzing incomplete longitudinal data. Our simulation study demonstrates that the proposed WGEE2 approach generally provides unbiased and efficient estimators when the missing data mechanism follows a MAR model. When the missing data are nonignorable, all three methods give biased estimators of the model parameters. However, the extent of the bias is generally less severe as compared to the unweighted GEE and WGEE1 methods.

Acknowledgements

This research is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Table 3: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under MAR dropout (True parameter values: $\boldsymbol{\beta} = (-1, 1, .2)'$, $\alpha = 0.5$ and $\boldsymbol{\tau} = (-2, 2, 0)'$).

True Model	Fitted Model	Method	% Relative Bias				MSE				Coverage Probability			
			β_0	β_1	β_2	α	β_0	β_1	β_2	α	β_0	β_1	β_2	α
Exchangeable	Exchangeable	GEE	3.02	0.35	-15.42	-15.98	0.016	0.038	0.003	0.008	0.95	0.96	0.89	0.35
		WGEE1	-0.08	0.93	0.05	-12.63	0.027	0.065	0.003	0.006	0.95	0.95	0.94	0.49
		WGEE2	-0.04	0.99	-0.09	-0.58	0.027	0.065	0.003	0.002	0.95	0.95	0.94	0.97
	Serial	GEE	9.67	-0.32	-37.41	28.67	0.025	0.039	0.008	0.002	0.88	0.95	0.61	0.23
		WGEE1	-0.07	0.96	-0.11	9.23	0.030	0.066	0.004	0.004	0.94	0.95	0.93	0.64
		WGEE2	-0.02	1.01	-0.20	4.68	0.010	0.065	0.004	0.011	0.94	0.95	0.93	0.78
Serial	Exchangeable	GEE	-5.27	1.21	6.53	-34.58	0.024	0.040	0.003	0.031	0.93	0.94	0.92	0.01
		WGEE1	-0.68	0.87	0.60	-33.82	0.036	0.067	0.005	0.030	0.95	0.93	0.95	0.01
		WGEE2	-0.66	0.87	0.62	-23.27	0.036	0.067	0.005	0.016	0.95	0.93	0.95	0.35
	Serial	GEE	2.20	0.77	-13.26	-12.46	0.022	0.038	0.004	0.005	0.93	0.94	0.88	0.52
		WGEE1	-0.50	0.83	0.02	-9.86	0.037	0.065	0.006	0.004	0.94	0.93	0.94	0.64
		WGEE2	-0.51	0.84	0.16	-0.95	0.038	0.065	0.006	0.002	0.94	0.93	0.94	0.95

Table 4: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under MAR dropout (True parameter values: $\boldsymbol{\beta} = (-1, 1, .2)'$, $\alpha = 0.5$ and $\boldsymbol{\tau} = (-2, 3, 0)'$).

True Model	Fitted Model	Method	% Relative Bias				MSE				Coverage Probability			
			β_0	β_1	β_2	α	β_0	β_1	β_2	α	β_0	β_1	β_2	α
Exchangeable	Exchangeable	GEE	9.46	-1.89	-52.83	-31.59	0.023	0.041	0.014	0.028	0.90	0.95	0.36	0.04
		WGEE1	-1.63	-1.63	1.07	-16.31	0.068	0.136	0.008	0.013	0.93	0.92	0.92	0.39
		WGEE2	-1.65	-1.34	1.07	-3.57	0.069	0.132	0.008	0.009	0.93	0.93	0.92	0.97
	Serial	GEE	23.69	-3.70	-105.83	-24.66	0.073	0.042	0.049	0.019	0.57	0.94	0.05	0.16
		WGEE1	-1.14	-1.64	-0.91	3.39	0.067	0.145	0.009	0.009	0.93	0.92	0.91	0.49
		WGEE2	-1.27	-1.25	-0.48	15.00	0.069	0.141	0.009	0.015	0.93	0.92	0.91	0.80
Serial	Exchangeable	GEE	-4.12	-1.46	-6.67	-41.35	0.020	0.039	0.004	0.046	0.94	0.96	0.91	0.01
		WGEE1	-2.06	-2.27	3.39	-34.93	0.087	0.140	0.014	0.035	0.93	0.91	0.91	0.08
		WGEE2	-1.98	-1.97	2.97	-24.30	0.088	0.137	0.014	0.023	0.93	0.91	0.91	0.63
	Serial	GEE	8.55	-2.76	-46.39	-28.48	0.027	0.038	0.013	0.023	0.92	0.95	0.61	0.09
		WGEE1	-1.78	-2.38	2.54	-14.14	0.084	0.142	0.014	0.011	0.93	0.90	0.91	0.43
		WGEE2	-1.87	-2.00	2.82	-1.99	0.085	0.138	0.014	0.008	0.93	0.91	0.91	0.93

Table 5: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under NI dropout (True parameter values: $\boldsymbol{\beta} = (-1, 1, .2)'$, $\alpha = 0.5$ and $\boldsymbol{\tau} = (-2, 2, 0.5)'$).

True Model	Fitted Model	Method	% Relative Bias				MSE				Coverage Probability			
			β_0	β_1	β_2	α	β_0	β_1	β_2	α	β_0	β_1	β_2	α
Exchangeable	Exchangeable	GEE	9.29	-1.54	-56.64	-29.11	0.024	0.041	0.015	0.023	0.88	0.95	0.26	0.03
		WGEE1	0.18	1.00	-0.62	-15.51	0.032	0.084	0.003	0.009	0.94	0.94	0.95	0.40
		WGEE2	0.23	1.06	-0.75	-1.38	0.032	0.082	0.003	0.004	0.94	0.94	0.96	0.94
	Serial	GEE	20.31	-2.75	-93.58	-15.38	0.057	0.041	0.038	0.009	0.65	0.94	0.03	0.43
		WGEE1	0.20	0.95	-0.82	5.18	0.033	0.086	0.004	0.005	0.94	0.94	0.95	0.63
		WGEE2	0.24	1.03	-0.85	17.75	0.033	0.085	0.004	0.012	0.94	0.94	0.95	0.65
Serial	Exchangeable	GEE	3.60	-0.92	-37.46	-44.81	0.018	0.037	0.009	0.052	0.93	0.97	0.73	0.00
		WGEE1	9.12	-1.45	-44.94	-44.01	0.053	0.071	0.014	0.051	0.92	0.97	0.82	0.00
		WGEE2	9.28	-1.56	-45.23	-30.93	0.053	0.071	0.014	0.027	0.92	0.97	0.81	0.27
	Serial	GEE	12.73	-1.25	-61.98	-25.44	0.033	0.035	0.018	0.018	0.86	0.97	0.38	0.08
		WGEE1	9.48	-0.83	-45.27	-21.95	0.055	0.067	0.015	0.015	0.91	0.97	0.82	0.22
		WGEE2	9.69	-0.85	-45.51	-8.79	0.056	0.067	0.015	0.005	0.92	0.97	0.83	0.87

Table 6: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under NI dropout (True parameter values: $\boldsymbol{\beta} = (-1, 1, .2)'$, $\alpha = 0.5$ and $\boldsymbol{\tau} = (-2, 2, 1)'$).

True Model	Fitted Model	Method	% Relative Bias				MSE				Coverage Probability			
			β_0	β_1	β_2	α	β_0	β_1	β_2	α	β_0	β_1	β_2	α
Exchangeable	Exchangeable	GEE	18.82	-3.29	-114.36	-43.73	0.051	0.038	0.055	0.050	0.69	0.95	0.00	0.00
		WGEE1	-0.18	-0.59	-2.62	-15.78	0.045	0.127	0.004	0.012	0.94	0.92	0.96	0.39
		WGEE2	-0.21	-0.26	-2.64	-3.98	0.045	0.124	0.004	0.007	0.94	0.93	0.96	0.87
	Serial	GEE	33.42	-5.03	-165.40	-36.47	0.130	0.039	0.113	0.036	0.31	0.95	0.00	0.01
		WGEE1	0.30	-0.49	-4.78	3.49	0.042	0.136	0.005	0.009	0.94	0.92	0.96	0.49
		WGEE2	0.20	-0.05	-4.49	14.74	0.043	0.132	0.005	0.013	0.95	0.92	0.96	0.81
Serial	Exchangeable	GEE	10.77	-3.49	-91.95	-57.84	0.032	0.038	0.038	0.085	0.87	0.94	0.12	0.00
		WGEE1	19.07	-13.67	-98.61	-41.80	0.089	0.111	0.048	0.049	0.92	0.89	0.51	0.28
		WGEE2	18.95	-13.90	-98.16	-56.53	0.088	0.115	0.048	0.082	0.92	0.89	0.49	0.00
	Serial	GEE	21.08	-4.30	-120.31	-42.37	0.066	0.038	0.062	0.047	0.68	0.94	0.03	0.00
		WGEE1	20.21	-13.59	-99.95	-37.82	0.093	0.113	0.049	0.039	0.91	0.88	0.48	0.02
		WGEE2	20.61	-13.27	-100.63	-20.98	0.095	0.109	0.050	0.017	0.91	0.89	0.48	0.67

References

[1] Bahadur R. T. (1961). A representation of the joint distribution of response to n dichotomous items. *Studies in Item Analysis and Prediction, Stanford University Press, 158-168.*

[2] Carey, V., Zeger, S. L., and Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regression. *Biometrika,* **80***, 517-526.*

[3] Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistic,* **43***, 49-93.*

[4] Fitzmaurice, G. M., and Lipsitz, S. R. (1995). A model for binary time series data with serial odds ratio pattern. *Applied Statistics,* **44***, 51-61.*

[5] Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society, Ser.B,* **57***, 691-704.*

[6] Hughes, G. H., Cutter, G, Donahue, R., Freidman, G. D.,Hully, S., Hunkeler, E., Jacobs, D. R., Liu, K., Orden, S.,Pirie, P., Tucker, B., and Wagenknecht, L. (1987). Recruitment in coronary artery risk development in young adults (CARDIA) study. *Controlled Clinical Trials,* **8***, 68S-73S.*

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

[8] Lipsitz, S. R., Laird,N. M., and Harrington,D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika,* **78***, 153-160.*

[9] Little, R.J.A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* Wiley, New York

[10] Pepe, M. S., and Anderson,G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics, Simulation and Computation,* **23***, 939-951.*

[11] Preisser, J. S., Galecki, A. T., Lohman, K. K., and Wagenknecht, L. E. (2000). Analysis of smoking trends with incomplete longitudinal binary responses. *Journal of American Statistical Association,* **73***, 1021-1031.*

[12] Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics,* **44***, 1033-1048.*

[13] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association,* **90***, 106-121.*

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

[15] Sinha, S. K., Laird, N. M., and Fitzmaurice, G. M. (2010). Multivariate logistic regression with incomplete covariate and auxiliary information. *Journal of Multivariate Analysis,* **101***, 2389-2397.*

[16] Touloumi, G., Pocock, S. J., Babiker, A. G, and Darbyshire, J. H. (1999). Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. *Statistics in Medicine,* **18***, 1215-33.*

[17] Sinha, S. K., Troxel, A. B., Lipsitz, S. R., Sinha, D., Fitzmaurice, G. M., Molenberghs, G., and Ibrahim, J. G. (2011). A bivariate pseudo-likelihood for incomplete longitudinal binary data with nonignorable non-monotone missingness. *Biometrics,* **67***, 1119-1126.*

[18] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica,* **50***, 1-26.*

[19] Yi, G. Y., and Cook R. J. (2002). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of American Statistical Association,* **97***, 1071-1080.*