*Journal of Statistical Research* *https://doi.org/10.47302/jsr.2020540102*
*2020, Vol. 54, No. 1, pp. 27-42* *ISSN 0256 - 422 X*

# MARGINAL MODELS FOR LONGITUDINAL COUNT DATA WITH DROPOUTS

SEEMA ZUBAIR[⋆]

*School of Mathematics and Statistics, Carleton University*
*Ottawa, ON K1S 5B6 Canada*
*Email: seemazubair@cmail.carleton.ca*

SANJOY K. SINHA

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

SUMMARY

In this article, we investigate marginal models for analyzing incomplete longitudinal count data with dropouts. Specifically, we explore commonly used generalized estimating equations and weighted generalized estimating equations for fitting log-linear models to count data in the presence of monotone missing responses. A series of simulations were carried out to examine the finite-sample properties of the estimators in the presence of both correctly specified and misspecified dropout mechanisms. An application is provided using actual longitudinal survey data from the Health and Retirement Study (HRS) (HRS, 2019)

*Keywords and phrases:* Count data. Generalized estimating equation. Longitudinal study. Missing response. Weighted generalized estimating equation.

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

## 1 Introduction

We often encounter longitudinal data in surveys and clinical trials, where individuals or units are monitored repeatedly for a specified study period and data are recorded for each individual or unit. Longitudinal studies allow researchers to investigate the change in a response variable over time along with changes in available covariates. Repeated measurements in a longitudinal study are correlated by nature and proper statistical methods are needed for analyzing such data by taking the correlations into account.

The most challenging and common problem of longitudinal studies is the presence of missing observations in the data. In a longitudinal study, an individual's response may be missing during one follow-up time and observed at the next follow-up time, resulting in a large class of missing

[⋆] Corresponding author

data patterns. Many authors studied missingness patterns in longitudinal data, which include Little (1995), Little and Rubin (1987), Diggle et al. (1994), Fitzmaurice et al. (1995), Pantazis and Touloumi (2010), Diggle et al. (2002), and Fitzmaurice et al. (2012).

Generalized estimating equations (GEEs) introduced by Liang and Zeger (1986) are widely used for analyzing longitudinal data. The technique of GEE, which is a multivariate analog of the quasi-likelihood approach, is useful for fitting marginal mean response models to dependent data. The GEE is grounded on a "working" correlation structure and is attractive in the sense that it does not require any distributional assumption and can provide consistent estimators of regression coefficients even under a misspecified correlation structure.

Although we are primarily interested in regression parameters of a marginal model, but in recent years there has been a growing interest in estimating the association parameters efficiently for an improved statistical inference. The GEE approach of Liang and Zeger (1986) can be extended for simultaneously estimating both regression and associations parameters, as suggested by Prentice (1988). Fitzmaurice et al. (1995), Lipsitz et al. (1991), and Carey et al. (1993) studied longitudinal data by modeling the association among responses in terms of pairwise odds ratios. When data are missing, the standard GEE approaches of Liang and Zeger (1986) and Prentice (1988) are valid only when the data are missing completely at random (MCAR) (Robins et al., 1995), i.e., given the covariates, the missing data process is independent of both observed and unobserved values of the response variable. The standard GEE estimator may be biased under a weaker assumption of missing at random (MAR) mechanism, in which missingness depends on the observed values of the response variable, but not on the unobserved values (Fitzmaurice et al., 1995).

Robins et al. (1995) suggested the inverse probability of first-order weighted GEE (WGEE) method in which the traditional GEE is weighted by estimated response probabilities. The probabilities creating these weights are attained by modeling missing data indicators as a function of the response variable and associated covariates. The WGEE method produces unbiased estimating equations and hence consistent estimators of the mean response parameters when the missing data follow a correctly specified MAR mechanism (Robins et al., 1995).

In this paper, we focus on studying incomplete longitudinal count data, where we limit our attention to monotone missing data patterns resulting from attrition or dropout. Our research was motivated by an actual longitudinal household survey data obtained from the Health and Retirement Study (HRS) (HRS, 2019). The survey was conducted by the Institute for Social Research at the University of Michigan. The survey data contain variables on demographics, health, social security, pensions, family structure, retirement plans, and employment history for individuals over age 50 and their spouses. The data were collected from several waves of interviews across fifteen survey years ranging from 1992–2016. In our analysis, the outcome variable of interest is the number of doctor visits by a respondent over a two-year period prior to an interview. The goal is to determine subgroups of respondents with similar behaviors in terms of the number of doctor visits and to identify predictors that affect the number of visits especially in the presence of missing data, intraperson correlations and overdispersion. Details about the data analysis are given in the application section.

The paper is organized as follows. Section 2 introduces the model and notation to describe the mean response and dropout mechanism for missing data. Section 3 describes the standard GEE and weighted GEE methods for analyzing incomplete longitudinal data. Section 4 presents results from a simulation study that was carried out to investigate the finite-sample properties of the estimators. Section 5 provides an application using actual longitudinal data from the Health and Retirement Study (HRS). Section 6 provides conclusions of the paper.

# 2 Model and Notation

## 2.1 Log-Linear Model

Assume that there are $N$ subjects in a study, where each subject is measured at a fixed set of $T$ time points. Let $y_{it}$ represent a count response from the $i$th subject at time $t$ and $\mathbf{x}_{it} = (x_{it,1}, \ldots, x_{it,p})^{'}$ represent a $p$-dimensional vector of covariates associated with $y_{it}$. The covariates may be fixed or time dependent throughout the observation times and may also include the intercept term in a regression model. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})^{'}$ denote the vector of longitudinal responses from the $i$th subject.

Assume that the marginal mean response $E(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})$ is given by

$$\log\{E(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})\} = \mathbf{x}_{it}^{'}\boldsymbol{\beta}, \tag{2.1}$$

for $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $\boldsymbol{\beta}$ is a $p$-dimensional vector of regression coefficients. The corresponding marginal variance is given by

$$\text{Var}(y_{it}) = \phi\mu_{it}, \tag{2.2}$$

where $\phi$ is a dispersion parameter that needs to be estimated. Responses from different subjects are assumed independent. However, repeated responses from a given subject $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})^{'}$ are assumed correlated with a correlation structure

$$\text{Corr}(y_{it}, y_{it'}) = \rho_{tt'}(\boldsymbol{\alpha}), \tag{2.3}$$

depending on a $q$-dimensional vector of association parameters $\boldsymbol{\alpha}$, for $t \neq t^{'} = 1, \ldots, T$.

Here the extra variance assumption of the response model permits the variance to be inflated by a factor $\phi$ ($\phi > 1$). The excess variability in count data can be accounted for by including the dispersion parameter $\phi$ (Fitzmaurice et al., 2012). We are interested in estimating the regression parameters $\boldsymbol{\beta}$, dispersion parameter $\phi$ and association parameters $\boldsymbol{\alpha}$ using a suitable robust method without making any distributional assumptions about the response variable $y_{it}$. Several methods are available for estimating these model parameters, which include the GEE method of Liang and Zeger (1986) and an extended GEE method of Prentice (1988).

## 2.2 Dropout Model

We often encounter attrition in a longitudinal study due to dropouts and delayed enrollments, where participants drop out before the end of the study and do not return. For example, members of a panel may drop out in panel surveys because they have moved to a place that is inconvenient to the researchers, or, in a clinical study, some participants may drop out due to side effects of a drug used for curing a disease, or for other unknown reasons. The pattern of attrition is an example of monotonous missing data, where follow-up measurements $y_{i,t+1}, \ldots, y_{iT}$ are missing and all previous measurements $y_{i1}, \ldots, y_{i,t-1}$ are observed. The missing data pattern is rarely monotonous in practice, but it is often close to monotonous (Little and Rubin, 1987).

To define the missing data mechanism, we consider an indicator variable $v_{it}$ that takes the value 1 if the response $y_{it}$ is observed and 0 if $y_{it}$ is missing. Let $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iT})'$ denote the vector of missing data indicators for the $i$th subject. Suppose we have a monotone missingness pattern, so that $v_{i1} \geq \cdots \geq v_{iT}$ and $v_{i1} = 1$ for all subjects. In general, the missing data mechanism $f_{v_i}(\mathbf{v}_i|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau})$ depends on the complete vector of responses $\mathbf{y}_i$ and design matrix $\mathbf{X}_i$ for the $i$th subject.

In the case of dropouts, the missing data indicators $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iT})'$ may be defined by a single random variable

$$m_i = 1 + \sum_{t=1}^{T} v_{it}, \tag{2.4}$$

indicating the dropout time. Then the dropout or missing data process can be redefined by

$$\pi_{im} = f_{m_i}(m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}) = P(m_i = m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}). \tag{2.5}$$

The value of $m$ lies between 2 and $T + 1$ if all subjects are observed at the first visit, where the maximum value of $T + 1$ corresponds to a full sequence of measurements.

Let $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$ denote the observed and missing components of the response vector $\mathbf{y}_i$. In general, there are three types of dropout mechanism that we encounter in longitudinal studies. The first is called missing completely at random (MCAR) mechanism, where missingness (dropout) does not depend on any of the observed or missing components of the response vector, i.e., $P(m_i = m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}) = P(m_i = m|\mathbf{X}_i, \boldsymbol{\tau})$. The second is called missing at random (MAR) mechanism, where missingness depends only on the observed components of the response vector $\mathbf{y}_i^o$, i.e., $P(m_i = m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}) = P(m_i = m|\mathbf{y}_i^o, \mathbf{X}_i, \boldsymbol{\tau})$. The third is called nonignorable (NI) mechanism, where missingness depends on both observed or missing components of the response vector, i.e., $P(m_i = m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}) = P(m_i = m|\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{X}_i, \boldsymbol{\tau})$. It is often assumed that the NI dropout probability depends on the missing components $\mathbf{y}_i^m$ only through the current response $y_{im}$. In this case, the dropout probability is given by

$$\begin{aligned} P(m_i = m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}) &= P(v_{i2} = \ldots = v_{i,m-1} = 1, v_{im} = 0|y_{i1}, \ldots, y_{im}, \mathbf{X}_i, \boldsymbol{\tau}) \\ &= \prod_{t=2}^{m-1} P(v_{it} = 1|v_{i1} = \ldots = v_{i,t-1} = 1, y_{i1}, \ldots, y_{it}, \mathbf{X}_i, \boldsymbol{\tau}) \times \\ &\quad P(v_{im} = 0|v_{i1} = \ldots = v_{im-1} = 1, y_{i1}, \ldots, y_{im}, \mathbf{X}_i, \boldsymbol{\tau})^{I\{m \leq T\}}, \end{aligned} \tag{2.6}$$

for an indicator variable $I\{\}$.

# 3 Methods of Estimation

## 3.1 Generalized Estimating Equations

For estimating the regression and association parameters, the methods of generalized estimating equation (GEE) and weighted generalized estimating equation (WGEE) are discussed in this section. Our main interest is in the estimation of regression parameters $\boldsymbol{\beta}$, association parameters $\boldsymbol{\alpha}$ and dispersion parameter $\phi$, whereas $\boldsymbol{\tau}$ is considered as a vector of nuisance parameters of the missing data mechanism. Partitioning the mean response vector $\boldsymbol{\mu}_i$ into its observed and missing components, we can write $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_i^o, \boldsymbol{\mu}_i^m)$.

Ordinary GEE estimates are obtained from available data by ignoring the missing data pattern. Following Liang and Zeger (1986), the GEE estimates of $\boldsymbol{\beta}$ for given $(\boldsymbol{\alpha}, \phi)$ may be obtained by solving the equations

$$\mathbf{U}_\beta(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \sum_{i=1}^{N} \mathbf{D}_i^{'} \mathbf{V}_i^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o(\boldsymbol{\beta})) = \mathbf{0}, \tag{3.1}$$

where $\mathbf{D}_i = \partial \boldsymbol{\mu}_i^o(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and $\mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$ with $\mathbf{A}_i = \text{diag}\{\boldsymbol{\mu}_i^o\}$ and $\mathbf{R}_i(\boldsymbol{\alpha})$ being a working correlation matrix for the observed response vector $\mathbf{y}_i^o$.

If $\boldsymbol{\alpha}$ and $\phi$ are known, the solutions to the above equations are asymptotically efficient (Liang and Zeger, 1986). Liang and Zeger (1986) consider estimating the correlation and dispersion parameters by the method of moments. Prentice (1988) extended the GEE technique to allow simultaneous estimation of the vector of regression parameters $\boldsymbol{\beta}$ and association parameters $(\boldsymbol{\alpha}, \phi)$. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{'}, \phi)^{'}$ be the vector of association and dispersion parameters. Following Prentice (1988), we can estimate $\boldsymbol{\theta}$ by solving a second set of estimating equations

$$\mathbf{U}_\theta(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{L}_i^{'} \mathbf{W}_i^{-1} (\mathbf{z}_i^o - \boldsymbol{\eta}_i^o(\boldsymbol{\beta}, \boldsymbol{\theta})) = \mathbf{0}, \tag{3.2}$$

where $\mathbf{z}_i^o$ represents the observed components of $\mathbf{z}_i = (r_{i1}r_{i2}, \ldots, r_{i,T-1}r_{iT}, r_{i1}^2, \ldots, r_{iT}^2)^{'}$ with $r_{it} = (y_{it} - \mu_{it})/\sqrt{\mu_{it}}$, $\boldsymbol{\eta}_i^o(\boldsymbol{\beta}, \boldsymbol{\theta}) = E(\mathbf{z}_i^o | \boldsymbol{\beta}, \boldsymbol{\theta})$, $\mathbf{L}_i = \partial \boldsymbol{\eta}_i^o(\boldsymbol{\beta}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and $\mathbf{W}_i$ is a working covariance matrix of $\mathbf{z}_i^o$. One can set $\mathbf{W}_i$ to be a identity matrix to avoid estimating additional parameters involving higher-order moments and to reduce sampling variation (Diggle et al., 2002).

## 3.2 Weighted Generalized Estimating Equations

In the case of dropouts, ordinary GEE estimators of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ are generally biased. To obtain unbiased estimators, the weighted GEE (WGEE) method may be used (Robins et al., 1995). The WGEE estimates of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ may be obtained by solving the equations

$$\mathbf{U}_{1\beta}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{1}{\pi_{im}} \mathbf{D}_i^{'} \mathbf{V}_i^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o(\boldsymbol{\beta})) = \mathbf{0}, \tag{3.3}$$

$$\mathbf{U}_{1\theta}(\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{1}{\pi_{im}} \mathbf{L}_i^{'} \mathbf{W}_i^{-1} (\mathbf{z}_i^o - \boldsymbol{\eta}_i^o(\boldsymbol{\beta},\boldsymbol{\theta})) = \mathbf{0}, \tag{3.4}$$

with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. The above estimating equations are unbiased and hence the WGEE estimators are consistent, as can be shown from the standard theory of method of moments. If the dropout probabilities $\pi_{im}$ are estimated consistently, then the WGEE method would still provide consistent estimators of $(\boldsymbol{\beta},\boldsymbol{\theta})$ (Robins et al., 1995).

To solve Eqs. (3.3) and (3.4), we use an iterative method, which begins with some initial values $(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)$ and then produces updated values $(\boldsymbol{\beta}_{s+1},\boldsymbol{\theta}_{s+1})$ by means of the iterative equations

$$\boldsymbol{\beta}_{s+1} = \boldsymbol{\beta}_s + \left( \sum_{i=1}^{N} \frac{1}{\pi_{im}} \mathbf{D}_i^{'} \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \sum_{i=1}^{N} \frac{1}{\pi_{im}} \mathbf{D}_i^{'} \mathbf{V}_i^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o), \tag{3.5}$$

$$\boldsymbol{\theta}_{s+1} = \boldsymbol{\theta}_s + \left( \sum_{i=1}^{N} \frac{1}{\pi_{im}} \mathbf{L}_i^{'} \mathbf{W}_i^{-1} \mathbf{L}_i \right)^{-1} \sum_{i=1}^{N} \frac{1}{\pi_{im}} \mathbf{L}_i^{'} \mathbf{W}_i^{-1} (\mathbf{z}_i^o - \boldsymbol{\eta}_i^o), \tag{3.6}$$

for $s = 0, 1, 2, \ldots$, where the second term on the right side of each equation is evaluated at the current estimates $(\boldsymbol{\beta}_s,\boldsymbol{\theta}_s)$. The estimates at convergence are called the WGEE estimates $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})$.

### 3.3 Approximate Variance of WGEE Estimators

Following White (1982), the variance-covariance matrix of the WGEE estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ can be approximated by using sandwich type estimators. The variance of $\hat{\boldsymbol{\beta}}$ can be approximated from

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \mathbf{H}_\beta^{-1} \mathbf{Q}_\beta \mathbf{H}_\beta^{-1}, \tag{3.7}$$

where $\mathbf{H}_\beta = \sum_{i=1}^{N} (1/\pi_{im}) \mathbf{D}_i^{'} \mathbf{V}_i^{-1} \mathbf{D}_i$ and $\mathbf{Q}_\beta = \sum_{i=1}^{N} \mathbf{S}_{\beta,i} \mathbf{S}_{\beta,i}^{'}$ with the score function for $\hat{\boldsymbol{\beta}}$, $\mathbf{S}_{\beta,i} = (1/\pi_{im}) \mathbf{D}_i^{'} \mathbf{V}_i^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)$, for the $i$th subject.

The variance of $\hat{\boldsymbol{\theta}}$ can be approximated from

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{H}_\theta^{-1} \mathbf{Q}_\theta \mathbf{H}_\theta^{-1}, \tag{3.8}$$

where $\mathbf{H}_\theta = \sum_{i=1}^{N} (1/\pi_{im}) \mathbf{L}_i^{'} \mathbf{W}_i^{-1} \mathbf{L}_i$ and $\mathbf{Q}_\theta = \sum_{i=1}^{N} \mathbf{S}_{\theta,i} \mathbf{S}_{\theta,i}^{'}$ with the score function for $\hat{\boldsymbol{\theta}}$, $\mathbf{S}_{\theta,i} = (1/\pi_{im}) \mathbf{L}_i^{'} \mathbf{W}_i^{-1} (\mathbf{z}_i^o - \boldsymbol{\eta}_i^o)$, for the $i$th subject.

In the next section, we study the finite-sample properties of the estimators based on Monte Carlo simulations.

## 4 Simulation Study

We ran a series of simulations using incomplete longitudinal count data to study the empirical properties of the proposed weighted GEE estimators. In particular, we investigate the three methods below:

i) GEE: Estimates of the regression parameters $\boldsymbol{\beta}$, association parameters $\boldsymbol{\alpha}$ and dispersion parameter $\phi$ are obtained by using the unweighted GEE approach of Prentice (1988).

ii) WGEE1: Estimates of $\boldsymbol{\beta}$ are obtained from the weighted GEEs, but estimates of $(\boldsymbol{\alpha}, \phi)$ are obtained from the unweighted GEEs.

iii) WGEE2: Estimates of all parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)$ are obtained from the weighted GEEs.

## 4.1 Response model

To produce correlated count data, we first created a population using a Poisson mixed model $y_{it}|u_i \sim$ ind. Poisson$(\mu_{it}^*)$, with $\log(\mu_{it}^*) = \beta_0^* + \beta_1^* x_i + \beta_2^* t + \beta_3^* x_i t + u_i$ and $u_i \sim$ ind. $N(0, 0.05^2)$, for $(\beta_0^*, \beta_1, \beta_2^*, \beta_3^*) = (1.5, -0.5, -0.5, -0.50)$, $i = 1, \ldots, N_0$ and $t = 1, \ldots, T$. The covariate $x_i$ was chosen as the binary indicator of a treatment with $P(x_i = 1) = 0.5$. The "population data" were generated from the given model for a large group of $N_0 = 500,000$ subjects with each subject being measured at $T = 3$ time points. We then fitted a marginal mean response model

$$\log\{E((y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta})\} = \log(\mu_{it}) = \beta_0 + \beta_1 x_i + \beta_2 t + \beta_3 x_i t$$

to these data assuming Var$(y_{it}) = \phi\mu_{it}$ and Corr$(y_{it}, y_{it'}) = \alpha$. The GEE estimates from this fit were treated as the "true" values of the parameters, which were obtained as $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)' = (1.60, -0.51, -0.50, -0.50)'$ and $\boldsymbol{\theta} = (\alpha, \phi)' = (0.20, 1.30)'$.

We then generated random samples from the above population by drawing $N$ subjects randomly for each combination of $N = 300$, $N = 500$ and $N = 1000$. Each simulation run was based on 1000 replicates of data sets. The numerical study requires intensive computations. To reduce the computation time, we used 1000 replicates of data sets for each set of simulations, which were found sufficient to produce empirical results with negligible simulation variations.

## 4.2 Dropout Model

To generate data with missing responses, we use a dropout model, where the response probability depends on previous and current responses, but is independent of the covariate $x_i$, so that

$$P(m_i = m|\mathbf{y}_i, \mathbf{X}_i, \boldsymbol{\tau}) = P(m_i = m|y_{i1}, \ldots, y_{im}, \boldsymbol{\tau}). \tag{4.1}$$

We assume that all subjects are observed at the first time point. We denote $v_{it} = 1$ if $y_{it}$ is observed and 0 if $y_{it}$ is missing. We further assume

$$P(v_{it} = 0|v_{i1} = \cdots = v_{it-1} = 1, y_{i1}, \ldots, y_{it}, \boldsymbol{\tau}) = \frac{\exp(\psi_{it})}{1 + \exp(\psi_{it})}, \tag{4.2}$$

where $\psi_{it} = \tau_0 + \tau_1 y_{it-1} + \tau_2 y_{it}$, for $t = 2, 3$. Note that $\tau_2 = 0$ leads to MAR mechanism, whereas $\tau_2 \neq 0$ leads to NI (nonignorable) missing data mechanism. The dropout probability is given by

$$\pi_{im} = P(m_i = m|y_{i1}, \ldots, y_{im}, \boldsymbol{\tau}) = \left\{\prod_{t=2}^{m-1} \frac{1}{1 + \exp(\psi_{it})}\right\} \left\{\frac{\exp(\psi_{im})}{1 + \exp(\psi_{im})}\right\}^{I\{m \leq T\}}. \tag{4.3}$$

To assess the performance of the three methods (GEE, WGEE1 and WGEE2), we ran two sets of simulations. In the first set, data were produced under the MAR mechanism with $\boldsymbol{\tau} = (-1.5, 0.3, 0)^{'}$ and the model parameters were also estimated under the correctly specified MAR mechanism. In the second set, data were generated under the nonignorable (NI) missing data mechanism with $\boldsymbol{\tau} = (-1.5, 0.3, 0.3)^{'}$, whereas the model parameters were estimated under the misspecified MAR mechanism. Both sets produced roughly 30% missing data on the response variable, with about 12% missing at the second time point and 18% at the third time point.

## 4.3 Estimating Dropout Probabilities

To estimate the dropout probabilities $\pi_{im} = P(m_i = m|y_{i1}, \ldots, y_{im}, \boldsymbol{\tau})$, we consider estimating $\boldsymbol{\tau}$ under the MAR mechanism. The pseudo-likelihood function for $\boldsymbol{\tau}$ is given by

$$\begin{aligned} L(\boldsymbol{\tau}) &= \prod_{i=1}^{N} P(m_i = m|y_{i1}, \ldots, y_{im}, \boldsymbol{\tau}) \\ &= \prod_{i=1}^{N} \left\{ \prod_{t=2}^{m-1} \frac{1}{1+\exp(\psi_{it}^*)} \right\} \left\{ \frac{\exp(\psi_{im}^*)}{1+\exp(\psi_{im}^*)} \right\}^{I\{m \le T\}}, \end{aligned} \tag{4.4}$$

where $\psi_{it}^* = \tau_0 + \tau_1 y_{i,t-1}$. Let $p_{it}^*(\boldsymbol{\tau}) = \exp(\psi_{it}^*)/(1+\exp(\psi_{it}^*))$. Then the score equation for $\boldsymbol{\tau}$ takes the form

$$S(\boldsymbol{\tau}) = \sum_{i=1}^{N} \left\{ -\sum_{t=2}^{m-1} p_{it}^*(\boldsymbol{\tau}) \mathbf{y}_{it}^* + I\{m \le T\}\{1 - p_{im}^*(\boldsymbol{\tau})\} \mathbf{y}_{im}^* \right\} = \mathbf{0}, \tag{4.5}$$

where $\mathbf{y}_{it}^* = (1, y_{i,t-1})^{'}$ and $\boldsymbol{\tau} = (\tau_0, \tau_1)^{'}$. The pseudo-ML estimator $\hat{\boldsymbol{\tau}}$ is obtained by solving (4.5) using an iterative method.

The variance of $\hat{\boldsymbol{\tau}}$ may be approximated from the information matrix

$$I(\boldsymbol{\tau}) = \sum_{i=1}^{N} \sum_{t=2}^{\min(m,T)} p_{it}^*(\boldsymbol{\tau})(1 - p_{it}^*(\boldsymbol{\tau})) \mathbf{y}_{it}^* \mathbf{y}_{it}^{*}{}^{'}. \tag{4.6}$$

For the $i$th individual at time $t$, the dropout probability is estimated by

$$\hat{\pi}_{im} = P(m_i = m|y_{i1}, \ldots, y_{im}, \hat{\boldsymbol{\tau}}) = \left\{ \prod_{t=2}^{m-1} (1 - \hat{p}_{it}^*) \right\} \times \{\hat{p}_{im}^*\}^{I\{m \le T\}}, \tag{4.7}$$

where $\hat{p}_{it}^* = p_{it}^*(\hat{\boldsymbol{\tau}})$. We find the WGEE estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from the iterative equations (3.5) and (3.6) by replacing $\pi_{im}$ with $\hat{\pi}_{im}$.

## 4.4 Results

Table 1 presents empirical percentage relative biases, mean squared errors and coverage probabilities of the GEE, WGEE1 and WGEE2 estimators of the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{'}$ and association parameters $(\alpha, \phi)$ under the correctly specified MAR mechanism with $\boldsymbol{\tau} = (-1.5, 0.3, 0)'$. Table 2 repeats the results under a misspecified missing data mechanism with $\boldsymbol{\tau} = (-1.5, 0.3, 0.3)'$. It is clear from Table 1 that under the MAR model, both WGEE1 and WGEE2 methods provide similar results for estimating the regression parameters. For estimating the correlation parameter $\alpha$ and dispersion parameter $\phi$, the proposed WGEE2 method appears to be the most efficient in terms of small percentage relative biases and good coverage probabilities of the estimators. For example, when estimating $\alpha$ at $N = 500$, the WGEE1 estimator provides a percentage relative bias of -17.40% and a coverage probability of 81.1%, whereas the WGEE2 estimator provides a much smaller relative bias of 6.25% and a good coverage probability of 94.6% that is close to the nominal 95% confidence level. Also, for estimating $\phi$ at $N = 500$, the WGEE1 method provides a percentage relative bias of -13.24% and a poor coverage probability of 40.9%, whereas the WGEE2 estimator provides a very small relative bias of 0.65% and a good coverage probability of 95.1%. Under the misspecified dropout model, it is evident from Table 2 that all three methods provide biased estimators of both regression and association parameters. However, the extent of the bias from the WGEE2 method is generally less as compared to the other two methods. For example, when estimating $\alpha$ at $N = 1000$, the WGEE2 method provides a bias of -24.7%, whereas the WGEE1 and GEE methods provide larger biases of -43.6% and -44.6%, respectively. For the overdispersion parameter $\phi$, unlike the GEE and WGEE1 methods, the WGEE2 method generally provides unbiased estimates and a coverage probability that is close to the nominal confidence level. The WGEE2 estimates are generally more robust than those obtained from the GEE and WGEE1 methods under a misspecified dropout model.

Table 3 presents empirical relative biases and mean squared errors of the pseudo-ML estimators of dropout model parameters $(\tau_o, \tau_1)$ under both correctly specified MAR and misspecified NI dropout models. As expected, the pseudo-ML method provides unbiased estimates under the correctly specified MAR model. However, under the misspecified (NI) model, it is evident from Table 3 that the ML method generally provides biased estimates.

## 5 Application: HRS Longitudinal Data

Here we consider analyzing longitudinal count data obtained from the Health and Retirement Study (HRS) (HRS, 2019), which is a longitudinal household survey conducted by the Institute for Social Research at the University of Michigan. The RAND HRS Longitudinal File (RAND HRS, 2019) contains variables on demographics, health, health insurance, social security, pensions, family structure, retirement plans, expectations, and employment history. The HRS is a national panel survey of individuals over age 50 and their spouses. The data were collected from thirteen waves of interviews across fifteen survey years (1992, 1993, 1994, 1995, and biennially 1996–2016). The data are available at *https://hrs.isr.umich.edu/data-products*. We consider analyzing a subset of the HRS data obtained from the most recent four waves of surveys carried out in the years 2010, 2012, 2014 and 2016, where the year 2010 was considered as the baseline.

In our analysis, the response variable of interest is the number of doctor visits by a respondent over a two-year period prior to an interview. The goal is to determine subgroups of respondents

Table 1: Empirical percentage relative biases, mean squared errors (MSEs) and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under correctly specified MAR dropout model. True parameter values: $\boldsymbol{\beta} = (1.60, -0.51, -0.50, -0.50)^{'}$, $\alpha = 0.20$, $\phi = 1.30$ and $\boldsymbol{\tau} = (-1.5, 0.3, 0)^{'}$.

| | Relative bias (%) | | | | | | MSE | | | | | | Coverage probability (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ | $\phi$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ | $\phi$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ | $\phi$ |
| $N = 300$ | | | | | | | | | | | | | | | | | | |
| GEE | 7.72 | 6.16 | 5.70 | –7.18 | –20.65 | –13.85 | 0.026 | 0.039 | 0.005 | 0.018 | 0.005 | 0.038 | 73.7 | 93.7 | 89.6 | 89.6 | 82.3 | 52.0 |
| WGEE1 | 5.97 | –1.74 | –1.16 | –0.16 | –19.05 | -13.51 | 0.026 | 0.062 | 0.006 | 0.027 | 0.005 | 0.038 | 89.7 | 93.0 | 92.6 | 92.6 | 82.9 | 54.1 |
| WGEE2 | 6.05 | –1.33 | –1.24 | –0.32 | 1.30 | –0.92 | 0.026 | 0.052 | 0.006 | 0.027 | 0.007 | 0.027 | 89.7 | 93.3 | 92.7 | 92.7 | 92.8 | 93.4 |
| $N = 500$ | | | | | | | | | | | | | | | | | | |
| GEE | 7.47 | 6.70 | 4.98 | –6.96 | –19.40 | –13.40 | 0.021 | 0.023 | 0.003 | 0.011 | 0.003 | 0.004 | 62.8 | 94.2 | 89.9 | 89.9 | 80.3 | 39.7 |
| WGEE1 | 5.92 | –0.56 | 0.20 | 0.40 | –17.40 | –13.24 | 0.019 | 0.034 | 0.003 | 0.014 | 0.003 | 0.034 | 86.0 | 94.6 | 93.0 | 93.0 | 81.1 | 40.9 |
| WGEE2 | 5.97 | –0.40 | 0.32 | 0.26 | 6.25 | 0.65 | 0.021 | 0.033 | 0.003 | 0.014 | 0.006 | 0.022 | 86.1 | 94.9 | 93.1 | 93.2 | 94.6 | 95.1 |
| $N = 1000$ | | | | | | | | | | | | | | | | | | |
| GEE | 7.61 | 7.20 | 4.96 | –7.12 | –18.20 | –13.14 | 0.018 | 0.012 | 0.002 | 0.006 | 0.002 | 0.031 | 38.3 | 93.0 | 86.0 | 86.0 | 74.2 | 13.6 |
| WGEE1 | 6.24 | 0.08 | 0.38 | 0.40 | –16.15 | –13.19 | 0.016 | 0.019 | 0.002 | 0.007 | 0.002 | 0.032 | 75.4 | 95.3 | 94.4 | 94.4 | 75.4 | 13.4 |
| WGEE2 | 6.28 | 0.24 | 0.40 | –0.09 | 8.75 | 2.26 | 0.016 | 0.019 | 0.002 | 0.007 | 0.004 | 0.025 | 75.7 | 95.1 | 94.7 | 94.7 | 96.5 | 96.6 |

Table 2: Empirical percentage relative biases, mean squared errors (MSEs) and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under misspecified NI dropout model. True parameter values: $\boldsymbol{\beta} = (1.60, -0.51, -0.50, -0.50)^{'}, \alpha = 0.20, \phi = 1.30$ and $\boldsymbol{\tau} = (-1.5, 0.3, 0.3)^{'}$.

| | Relative bias (%) | | | | | | MSE | | | | | | Coverage probability (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ | $\phi$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ | $\phi$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ | $\phi$ |
| $N = 300$ | | | | | | | | | | | | | | | | | | |
| GEE | 15.84 | 19.88 | 32.94 | –20.44 | –45.40 | –17.70 | 0.076 | 0.053 | 0.032 | 0.029 | 0.011 | 0.059 | 36.2 | 88.7 | 33.8 | 33.8 | 54.0 | 35.1 |
| WGEE1 | 13.83 | 2.96 | 29.44 | –10.52 | –45.10 | –16.90 | 0.083 | 0.087 | 0.033 | 0.036 | 0.012 | 0.055 | 73.2 | 92.9 | 64.9 | 64.9 | 54.7 | 38.1 |
| WGEE2 | 13.92 | 2.84 | 29.74 | –10.40 | –29.15 | –3.20 | 0.084 | 0.088 | 0.033 | 0.036 | 0.012 | 0.035 | 73.5 | 93.7 | 65.0 | 65.0 | 75.5 | 86.2 |
| $N = 500$ | | | | | | | | | | | | | | | | | | |
| GEE | 15.70 | 19.20 | 32.54 | –19.76 | –46.70 | –17.35 | 0.071 | 0.033 | 0.029 | 0.022 | 0.011 | 0.054 | 15.0 | 89.9 | 14.0 | 14.0 | 38.6 | 18.3 |
| WGEE1 | 14.46 | 4.32 | 30.04 | –10.58 | –46.40 | –16.52 | 0.083 | 0.058 | 0.032 | 0.025 | 0.011 | 0.051 | 63.9 | 92.7 | 53.0 | 53.0 | 38.4 | 24.2 |
| WGEE2 | 14.42 | 3.66 | 30.00 | –10.00 | –30.75 | –0.777 | 0.085 | 0.060 | 0.035 | 0.027 | 0.010 | 0.033 | 63.9 | 92.7 | 52.6 | 52.6 | 71.6 | 85.8 |
| $N = 1000$ | | | | | | | | | | | | | | | | | | |
| GEE | 15.85 | 18.52 | 32.58 | –19.24 | –44.60 | –17.42 | 0.068 | 0.021 | 0.028 | 0.015 | 0.009 | 0.053 | 1.5 | 91.0 | 1.5 | 1.5 | 17.4 | 2.2 |
| WGEE1 | 14.45 | 2.92 | 29.16 | –8.94 | –43.60 | –16.93 | 0.066 | 0.028 | 0.025 | 0.012 | 0.008 | 0.051 | 32.8 | 93.5 | 24.9 | 24.9 | 19.1 | 4.0 |
| WGEE2 | 13.04 | 2.06 | 24.7 | –4.02 | –24.70 | –0.062 | 0.065 | 0.028 | 0.030 | 0.029 | 0.007 | 0.019 | 32.2 | 93.6 | 23.8 | 23.8 | 70.4 | 91.7 |

Table 3: Empirical percentage relative biases and mean squared errors of pseudo-ML estimators under MAR and nonignorible (NI) dropout models.

| True dropout model | No. of subjects ($N$) | Parameter | True value | Relative bias(%) | MSE |
|---|---|---|---|---|---|
| MAR | 300 | $\tau_0$ | –1.5 | 0.53 | 0.0219 |
| | | $\tau_1$ | 0.3 | 0.60 | 0.0027 |
| | 500 | $\tau_0$ | –1.5 | 0.14 | 0.0138 |
| | | $\tau_1$ | 0.3 | 0.41 | 0.0027 |
| | 1000 | $\tau_0$ | –1.5 | –0.08 | 0.0064 |
| | | $\tau_1$ | 0.3 | 0.20 | 0.0008 |
| NI | 300 | $\tau_0$ | –1.5 | –8.48 | 0.0388 |
| | | $\tau_1$ | 0.3 | 32.67 | 0.0132 |
| | 500 | $\tau_0$ | –1.5 | –8.33 | 0.0289 |
| | | $\tau_1$ | 0.3 | 32.16 | 0.0113 |
| | 1000 | $\tau_0$ | –2.0 | –14.04 | 0.0924 |
| | | $\tau_1$ | 0.3 | 30.63 | 0.0096 |

with similar behaviors in terms of the number of doctor visits and to identify factors that affect the number of visits to a medical doctor. The following baseline covariates were considered for the analysis: Age (age of a respondent in years at baseline), Smoke (1, if the respondent ever smoked cigarettes and 0, if not), Cancer (1, if the respondent is diagnosed with a cancer and 0, if not), Heart (1, if the respondent has a heart condition and 0, if not), Lung (1, if the respondent has a lung condition and 0, if not), Sex (1, if male and 0, if female), Hospital (1, if the respondent reports any overnight stay in a hospital in the reference period and 0, if not), and BMI (body mass index).

We retained subjects for whom complete data were available on the covariates. The data exhibited a number of extreme outliers in the response variable. Any response with the number of doctor visits 41 or above over a two-year period was treated as an outlier, and was removed from the analysis to avoid any potential influence of outliers on the model fit. The discarded data accounted for only about 1% respondents.

Let $y_{it}$ be the number of doctor visits by the $i$th respondent ($i = 1, \ldots, N$) reported at time $t$ ($t = 1, \ldots, T$), with $N = 4814$ subjects and $T = 4$ time points. The marginal mean response $\mu_{it} = E(y_{it}|\mathbf{x}_i, \boldsymbol{\beta})$ is given by

$$\begin{aligned}\log(\mu_{it}) = \beta_0 + \beta_1\text{Lung}_i + \beta_2\text{Hospital}_i + \beta_3\text{Cancer}_i + \beta_4(\text{Age}/10)_i +& \\ \beta_5\text{Sex}_i + \beta_6\text{Heart}_i + \beta_7\text{Time}_t + \beta_8\text{BMI}_i + \beta_9\text{Smoke}_i.& \quad (5.1)\end{aligned}$$

Table 4: ML estimates, their standard errors (SEs) and $z$-values of the dropout model parameters in HRS study.

| Coefficient | Estimate | SE | $z$-value |
|---|---|---|---|
| Intercept | –6.4989 | 0.3174 | –20.48 |
| $y_{i,t-1}$ | 0.0282 | 0.0033 | 8.63 |
| Age | 0.5021 | 0.0415 | 12.10 |
| Cancer | 0.1417 | 0.0640 | 2.21 |
| Heart | 0.2557 | 0.0552 | 4.63 |
| Lung | 0.5328 | 0.0706 | 7.55 |
| Time | 0.2994 | 0.0306 | 9.79 |

The marginal variance and correlations are given by

$$\text{Var}(y_{it}) = \phi\mu_{it}, \quad \text{Corr}(y_{it}, y_{it'}) = \alpha,$$

for $t \neq t^{'} = 1, \ldots, T$. We estimate both regression and association parameters using the ordinary GEE and weighted GEE methods discussed earlier. For the weighted GEEs, we estimate the dropout probability $\pi_{im}$ based on the logistic model

$$\text{logit}(p^*_{it}) = \tau_0 + \tau_1 y_{i,t-1} + \tau_2(\text{Age/10})_i + \tau_3\text{Cancer}_i + \tau_4\text{Heart}_i + \tau_5\text{Lung}_i + \tau_6\text{Time}_t, \quad (5.2)$$

where $p^*_{it} = P(v_{it} = 0|v_{i1} = \ldots = v_{i,t-1} = 1, \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\tau})$.

The pseudo-ML estimates of the dropout model parameters $\boldsymbol{\tau}$, their standard errors and corresponding $z$-values are presented in Table 4. It is evident from the table that the rates of dropout differ across all covariates as well as the number of doctor visits by respondents in previous years. In particular, the dropout rate is higher among older respondents; the rate also increases with time as well as with an increased number of doctor visits by the respondent in the previous year. In addition, the dropout rate is higher among respondents with cancer as well as with heart or lung conditions. For example, a respondent with a lung condition has an odds of dropout that is $\exp(0.5328) = 1.7$ times higher than that for a respondent without any lung condition.

Table 5 presents the estimates of the regression and association parameters, their standard errors, and corresponding $z$-values obtained by the three methods GEE, WGEE1, and WGEE2 discussed earlier. Here the GEE estimates of the regression parameters appear to be somewhat different than those obtained by the WGEE1 and WGEE2 methods. The estimates under the WGEE1 and WGEE2 methods are very similar.

From the WGEE2 model fit, it is evident that females, patients who had recent overnight hospital stays, and patients with cancer, lung or heart conditions tend to visit doctors more frequently, as compared to others. The number of visits tends to decrease over time. The overdispersion and correlation parameters also appear to be significant by all methods.

Table 5: Estimates, standard errors (SEs) and $z$-values of regression, association and dispersion parameters of the count response model in HRS study.

| | GEE | | | WGEE1 | | | WGEE2 | | |
|---|---|---|---|---|---|---|---|---|---|
| Coefficient | Estimate | SE | $z$-value | Estimate | SE | $z$-value | Estimate | SE | $z$-value |
| Intercept | 1.720 | 0.128 | 13.361 | 1.778 | 0.208 | 8.525 | 1.779 | 0.208 | 8.533 |
| Lung | 0.183 | 0.029 | 6.202 | 0.177 | 0.032 | 5.598 | 0.176 | 0.032 | 5.596 |
| Hospital | 0.244 | 0.022 | 11.270 | 0.265 | 0.028 | 9.463 | 0.265 | 0.028 | 9.463 |
| Cancer | 0.144 | 0.024 | 5.807 | 0.168 | 0.032 | 5.325 | 0.168 | 0.032 | 5.327 |
| Age | 0.034 | 0.016 | 2.083 | 0.031 | 0.027 | 1.124 | 0.031 | 0.027 | 1.122 |
| Sex | –0.051 | 0.020 | –2.223 | –0.059 | 0.026 | –2.218 | –0.059 | 0.026 | –2.218 |
| Heart | 0.227 | 0.022 | 10.233 | 0.226 | 0.026 | 8.532 | 0.226 | 0.026 | 8.539 |
| Time | –0.001 | 0.005 | –0.228 | –0.023 | 0.007 | –2.880 | –0.022 | 0.007 | –2.874 |
| BMI | –0.000 | 0.002 | –0.020 | –0.000 | 0.002 | –0.124 | –0.000 | 0.002 | –0.127 |
| Smoke | –0.007 | 0.019 | –0.387 | –0.022 | 0.025 | –0.863 | –0.022 | 0.025 | –0.861 |
| $\alpha$ | 0.338 | 0.014 | 22.818 | 0.337 | 0.015 | 22.231 | 0.331 | 0.016 | 20.184 |
| $\phi$ | 6.140 | 0.145 | 42.313 | 6.279 | 0.153 | 41.114 | 6.591 | 0.175 | 37.472 |

# 6 Discussion

The aim of this research was to provide a better alternative to the GEE approach of Prentice (1988) for analyzing incomplete longitudinal count data with dropouts. Our simulation study reveals that the proposed WGEE2 method offers unbiased and efficient estimators under the MAR dropout mechanism. All methods, however, provide biased estimators under the misspecified nonignorable (NI) dropout mechanism, but the extent of the bias from the WGEE2 method appears to be less as compared to the GEE and WGEE1 methods.

We have studied the aforementioned three methods under both large and small proportions of missing data at different follow-up times in a longitudinal setting. As expected, the loss of efficiency by the ordinary GEE approach was not so dramatic when the proportion of missing responses was small (e.g., 10% or less). But for large proportions of missing responses (e.g., 50% or 60%), the GEE approach often provides large systematic biases in the estimation and hence invalid inference on the model parameters, especially when the missing responses are not missing completely at random. In such cases, attempts should be made to explore possible reasons for the missingness and to use a suitable missing data model in the proposed weighted GEE approach for a valid statistical inference.

For the analysis of missing data, the approach of multiple imputation (MI) has been extensively studied in the literature. There is software available for the MI in the context of linear and logistic regression models. However, we are not aware of any software for MI with correlated count responses. Multiple imputation for incomplete longitudinal count data is beyond the scope of this paper. We intend to study this in a future research.

### Acknowledgements

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

## References

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

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002), *Analysis of longitudinal data*, New York: Oxford University Press.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994), *Longitudinal data analysis*, New York: Oxford University Press.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012), *Applied longitudinal analysis*, New Jersey: John Wiley & Sons.

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: Series B (Methodological)*, 57, 691–704.

HRS (2019), *Health and Retirement Study, (RAND HRS Longitudinal File 2016) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740)*, Ann Arbor, MI.

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

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.

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

Little, R. J. A. and Rubin, D. B. (1987), "Statistical analysis with missing data," *Hoboken, NJ: Wiley*.

Pantazis, N. and Touloumi, G. (2010), "Analyzing longitudinal data in the presence of informative dropout: The jmre1 command," *The Stata Journal*, 10, 226–251.

Prentice, R. L. (1988), "Correlated binary regression with covariates specific to each binary observation," *Biometrics*, 44, 1033–1048.

RAND HRS (2019), *RAND HRS Longitudinal File 2016. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration.*, Santa Monica, CA.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 90, 106–121.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, 50, 1–25.

Received: *May 26, 2020*

Accepted: *July 31, 2020*