Inference Progress in Missing Data Analysis from Independent to Longitudinal Setup

Brajendra C. Sutradhar

Abstract In the independent setup with multivariate responses, the data become incomplete when partial responses, such as responses on some variables as opposed to all variables, are available from some individuals. The main challenge here is obtaining valid inferences such as unbiased and consistent estimates of mean parameters of all response variables by using available responses. Typically, unbalanced correlation matrices are formed and moments or likelihood analysis based on the available responses are employed for such inferences. Various imputation techniques also have been used. In the longitudinal setup, when a univariate response is repeatedly collected from an individual, these repeated responses become correlated and the responses form a multivariate distribution. In this setup, it may happen that a portion of responses are not available from some individuals under study. These non-responses may be monotonic or intermittent. Also the response may be missing following a mechanism such as missing completely at random (MCAR), missing at random (MAR), or missing non-ignorably. In a longitudinal regression setup, the covariates may also be missing, but typically they are known for all time periods. Obtaining unbiased and consistent regression estimates specially when longitudinal responses are missing following MAR or ignorable mechanism becomes a challenge. This happens because one requires to accommodate both longitudinal correlations and missing mechanism to develop a proper inference tool. Over the last three decades some progress has been made toward this mainly by taking partial care of missing mechanism in developing estimation techniques. But overall, they fall short and may still produce biased and hence inconsistent estimates. The purpose of this paper is to outline these perspectives in a comprehensive manner so that real progress and challenges are understood in order to develop proper inference techniques.

B.C. Sutradhar (✉)
Memorial University, St. John's, Canada, A1C5S7
e-mail: bsutradh@mun.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_5, © Springer Science+Business Media New York 2013

1 Introduction

Missing data analysis in the independent setup with multivariate responses has a long history. For example, for an early work, we refer to Lord (1995) who considered a set of incomplete trivariate normal responses collected from K independent individuals. But all components of all three variables were not available from the K individuals. To estimate the mean parameters consistently, instead of dropping out the individuals with incomplete information, Lord (1995) has utilized the available information and constructed unbalanced (bivariate and trivariate) probability functions for individuals toward writing a likelihood function for the desired inference. Note that this technique for consistent estimation of the parameters and other similar inferences by using incomplete data have been used by many researchers over the last six decades. See, for example, Mehta and Gurland (1973), Morrison (1973), Naik (1975), Little (1988), and Krishnamoorthy and Pannala (1999), among others.

In the independent setup, techniques of imputation and multiple imputation (Rubin 1976; Rubin and Schenker 1986; Meng 1994) have also been widely used. Some authors such as Paik (1997) used this imputation technique in repeated measure (longitudinal) setup. The imputation at a given time point is done mainly by averaging over the responses of other individuals at that time who has the same covariates history as that of the individual concerned. Once the missing values are estimated, they are used as data with necessary adjustments to construct complete data based estimating equations for the desired parameters.

In a univariate longitudinal response setup, when T repeated measures are taken they become correlated and hence they jointly follow a T-dimensional multivariate distribution. However, unlike in the Gaussian setup for linear data, the multivariate distributions for repeated binary and count data become complex or impractical. However if a portion of individuals do not provide responses for all T time points, then adopting likelihood approach by blending missing mechanism and correlation structure of the repeated data would naturally become extremely complicated or impossible. As a remedy, either imputation or estimating equation approaches became popular which, however, work well if the missing data occur following the simplest MCAR mechanism. When the missing data occur following the MAR mechanism, writing a proper estimating equation by accommodating both longitudinal correlations and missing mechanism becomes difficult. Robins et al. (1995) proposed an inverse probability weights based generalized estimating equations (WGEE) approach as an extension of the GEE approach proposed by Liang and Zeger (1986) to the incomplete setup. Remark that as demonstrated by Sutradhar and Das (1999) and Sutradhar (2010), for example, the GEE approach can produce less efficient regression estimates than the well-known simpler moments or quasi-likelihood (QL) estimates, in the complete data setup. Thus, to be realistic, there is no reason how WGEE approach can be more efficient in the incomplete longitudinal setup as compared to simpler moments and QL estimates. In fact in the incomplete longitudinal setup, the WGEE approach constructed based on working correlations as opposed to the use of

MAR based correlation matrix may yield biased and hence inconsistent regression estimates (Sutradhar and Mallick 2010). Further remark that this inconsistency issue was, however, not adequately addressed in the literature including the studies by Robins et al. (1995), Paik (1997), Rotnitzky et al. (1998), and Birmingham et al. (2003). One of the main reasons is this that none of the studies used any stochastic correlation structure in conjunction with the missing mechanism to model the longitudinal count and binary data in the incomplete longitudinal setup. Details on this inconsistency problem are given in Sect. 3, whereas in Sect. 2 we provide a detailed discussion on missing data analysis in independent setup.

Without realizing the aforementioned inconsistency problems that can be caused because of the use of working correlations in the estimating equations under the MAR based longitudinal setup, some authors such as Wang (1999) and Rotnitzky et al. (1998) used similar estimating equations approach in non-ignorable missing mechanism-based incomplete longitudinal setup. Some authors such as Troxel et al. (1998) (see also Troxel et al. 1997) and Ibrahim et al. (2001) (see also Ibrahim et al. 1999) have used random effects based generalized linear mixed model to accommodate the longitudinal correlations and certain binary logistic models to generate the non-ignorable mechanism based response indicator variables. In general expectation-maximization (EM) techniques are used to estimate the likelihood based parameters. These approaches appear to encounter similar difficulties as the existing MAR based approaches in generating first the response indicator and then the responses so that underlying longitudinal correlation structure is satisfied. Thus the inference validity of these approaches is not yet established. This problem becomes more complicated when longitudinal correlations are not generated through random effects and writing a likelihood such as for repeated count data becomes impossible. For clarity, in this paper we discuss in detail the successes and challenges with the inferences for MAR based incomplete longitudinal models only. The non-ignorable missing data based longitudinal analysis will therefore be beyond the scope of the paper.

2 Missing Data Analysis in Independent Setup

Missing data analysis in the independent setup with multivariate responses has a long history. For example, for an early work, we refer to Lord (1995) who considered a set of incomplete trivariate normal responses collected from K independent individuals. To be specific, suppose that $y = (y_1, y_2, y_3)'$ represents a trivariate response, but all components of y were not available from K individuals. Suppose that y_3 was recorded from all K individuals, and either y_1 or y_2 was recorded for all individuals, but not both. For $j = 1, \ldots, 3$, let K_j denote the number of individuals having the response y_j. It then follows that

$$K_1 + K_2 = K,\ K3 = K.$$

Further suppose that the K_1 individuals for whom y_1 is recorded will be denoted collectively as group 1 (G_1); and the K_2 individuals with y_2 will be denoted as group 2 (G_2). Now because y_1 and y_2 are correlated, it is obvious that the data for G_2 contain some information relevant for estimating the parameters of variable y_1, and that the data for G_1 contain some information relevant for estimating the parameters of y_2. The problem is to use the available data as efficiently as possible for estimating the parameters concerned. Denote the distribution of $y = [y_1,\, y_2,\, y_3]'$ as

$$y \sim N(\mu, \Sigma),$$

with $\mu = [\mu_1, \mu_2, \mu_3]'$ and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \rho_{12}[\sigma_{11}\sigma_{22}]^{\frac{1}{2}} & \rho_{13}[\sigma_{11}\sigma_{33}]^{\frac{1}{2}} \\ & \sigma_{22} & \rho_{23}[\sigma_{22}\sigma_{33}]^{\frac{1}{2}} \\ & & \sigma_{33} \end{pmatrix}.$$

Note that in this setup, there are no data available to estimate ρ_{12}. For the likelihood estimation of all the other parameters, define

$$\begin{aligned}
\bar{y}_1^* &= \frac{1}{K_1}\sum_{i=1}^{K_1} y_{1i},\ \bar{y}_2^* = \frac{1}{K_2}\sum_{i=1}^{K_2} y_{2i},\ \bar{y}_3 = \frac{1}{K}\sum_{i=1}^{K} y_{3i},\ \bar{y}_3^* = \frac{1}{K_1}\sum_{i=1}^{K_1} y_{3i},\ \bar{y}_3^{**} = \frac{1}{K_2}\sum_{i=1}^{K_2} y_{3i} \\
s_{11}^* &= \frac{1}{K_1}\sum_{i=1}^{K_1}[y_{1i} - \bar{y}_1^*]^2,\ s_{22}^* = \frac{1}{K_2}\sum_{i=1}^{K_2}[y_{2i} - \bar{y}_2^*]^2,\ s_{33} = \frac{1}{K}\sum_{i=1}^{K}[y_{3i} - \bar{y}_3]^2, \\
s_{33}^* &= \frac{1}{K_1}\sum_{i=1}^{K_1}[y_{3i} - \bar{y}_3^*]^2,\ s_{33}^{**} = \frac{1}{K_2}\sum_{i=1}^{K_2}[y_{3i} - \bar{y}_3^{**}]^2 \\
r_{13} &= \frac{1}{K_1}\sum_{i=1}^{K_1}[(y_{1i} - \bar{y}_1^*)(y_{3i} - \bar{y}_3^*)]/[s_{11}^* s_{33}^*]^{\frac{1}{2}}, \\
r_{23} &= \frac{1}{K_2}\sum_{i=1}^{K_2}[(y_{2i} - \bar{y}_2^*)(y_{3i} - \bar{y}_3^{**})]/[s_{22}^* s_{33}^{**}]^{\frac{1}{2}}.
\end{aligned} \tag{1}$$

The maximum likelihood estimators for the means are then given by

$$\hat{\mu}_1 = \bar{y}_1^* - b_{13}[\bar{y}_3^* - \bar{y}_3],\ \hat{\mu}_2 = \bar{y}_2^* - b_{23}[\bar{y}_3^{**} - \bar{y}_3],\ \text{and } \hat{\mu}_3 = \bar{y}_3, \tag{2}$$

where

$$b_{13} = r_{13}\frac{s_{11}^*}{s_{33}^*},\ \text{and } b_{23} = r_{23}\frac{s_{22}^*}{s_{33}^{**}}.$$

These estimators in (2) are unbiased and consistent for μ_1, μ_2, and μ_3, respectively. The remaining parameters may also be estimated similarly (Lord 1995).

Note that the aforementioned technique for consistent estimation of the parameters and for other similar inferences by using incomplete data has been subsequently used by many researchers over the last six decades. See, for example, Mehta and Gurland (1973), Morrison (1973), Naik (1975), Little (1988), and Krishnamoorthy and Pannala (1999), among others. This idea of making inferences about the underlying model parameters such that the missing data (assuming a small proportion of missing) may not to any major extent negatively influence the inferences has also been extended to the analysis of incomplete repeated measure data. For example, one may refer to Little (1995), Robins et al. (1995), and Paik (1997), as some of the early studies. This inference procedure for incomplete longitudinal data is discussed in detail in the next section.

In the independent setup, techniques of imputation and multiple imputation (Rubin 1976; Rubin and Schenker 1986; Meng 1994) have also been widely used. Later on some authors also used this imputation technique in repeated measure (longitudinal) setup. For example, here we illustrate an imputation formula from Paik (1997) in repeated measure setup. The imputation at a given time point is done mainly by averaging over the responses of other individuals at that time who has the same covariates history as that of the individual concerned. Once the missing values are estimated, they are used as data with necessary adjustments to construct complete data based estimating equations for the desired parameters.

In a univariate longitudinal response setup, when T repeated measures are taken they become correlated and hence they jointly follow a T-dimensional multivariate distribution. Now suppose that T_i responses are observed for the ith $(i = 1,\ldots,K)$ individual. So, one requires to impute $T - T_i$ missing values which may be done following Paik (1997), for example. Interestingly, a unified recursive relation can be developed as follows to obtain the imputed value $\tilde{y}_{i,T_i+k_i}$ at time point $T_i + k_i$ for all $k_i = 1,\ldots,T - T_i$. For this, first define

$$\tilde{y}^{(0)}_{j,T_i+k_i} = y_{j,T_i+k_i} \tag{3}$$

for the jth individual where $j \neq i$, $j = 1,\ldots,K$. Also, let D_{iT_i} denote the covariate history up to time point T_i for the ith individual, and

$$D^*_{i,T_i+k_i} = (x_{i,T_i+1},\ldots,x_{i,T_i+k_i})$$

is the covariate information for the ith individual from time $T_i + 1$ up to $T_i + k_i$ for $k_i = 1,\ldots,T - T_i$. Further let, $r_{jw} = 1, \text{or}, 0$, for example, indicates the response status of the jth individual at wth time. One may then obtain $\tilde{y}_{i,T_i+k_i}$ by computing $\tilde{y}^{(k_i)}_{i,T_i+k_i}$, that is, $\tilde{y}_{i,T_i+k_i} \equiv \tilde{y}^{(k_i)}_{i,T_i+k_i}$, where

$$\tilde{y}^{(k_i)}_{i,T_i+k_i} = \left[\sum_{j=1}^{K} \tilde{y}^{(0)}_{j,T_i+k_i} \Pi^{k_i}_{u=1} r_{j,T_i+u} I(D_{jT_i} = D_{iT_i}, D^*_{j,T_i+k_i} = D^*_{i,T_i+k_i})\right.$$

$$+ \sum_{m_i=1}^{k_i-1} \sum_{j=1}^{K} \tilde{y}^{(m_i)}_{j,T_i+k_i} \Pi^{k_i}_{u=k_i-(m_i-1)} (1 - r_{j,T_i+u})$$

$$\left. \times\, \Pi^{k_i-m_i}_{u=1} r_{j,T_i+u} I(D_{jT_i} = D_{iT_i}, D^*_{j,T_i+k_i} = D^*_{i,T_i+k_i})\right]$$

$$\times \left[\sum_{j=1}^{K} r_{j,T_i+1} I(D_{jT_i} = D_{iT_i}, D^*_{j,T_i+k_i} = D^*_{i,T_i+k_i})\right]^{-1}. \tag{4}$$

Note that $\tilde{y}_{i.T_i+k_i} \equiv \tilde{y}^{(k_i)}_{i,T_i+k_i}$ is an unbiased estimate of μ_{i,T_i+k_i} as the individuals used to impute the missing value of the ith subject has the same covariate history up to time point $T_i + k_i$, unlike the covariate history up to time point T_i (Paik 1997).

3 Missing Data Models in Longitudinal Setup

Let Y_{it} be the potential response from the ith $(i = 1, \ldots, K)$ individual at time point t which may or may not be observed, and $x_{it} = (x_{it1}, \ldots, x_{itp})'$ be the corresponding p-dimensional covariate vector which is assumed to be available for all times $t = 1, \ldots, T$. In this setup, K is large $(K \to \infty)$ and T is small such as 3 or 4. Suppose that $\beta = (\beta_1, \ldots, \beta_p)'$ denote the effect of x_{it} on y_{it}. Irrespective of the situation whether Y_{it} is observed or not, it is appropriate in the longitudinal setup to assume that the repeated responses follow a correlation model with known functional forms for the mean and the variance, but the correlation structure may be unknown. Recall that in the independent setup, Lord (1995) considered multivariate responses having a correlation structure and incompleteness arose because of missing information on some response variables, whereas in the present longitudinal setup, repeated responses from an individual form a multivariate response with a suitable mean, variance, and correlation structures, but it remains a possibility that one individual may not provide responses for the whole duration of the study. As indicated in the last section, suppose that for the ith $(i = 1, \ldots, K)$ individual T_i responses $(1 < T_i \leq T)$ are collected. Also suppose that the remaining $T - T_i$ potential responses are missing and the non-missing responses occur in a monotonic pattern.

As far as the mean, variance, and correlation structure of the potential responses are concerned, it is convenient to define them for the complete data. Let $y_i{}^c = (y_{i1}, \cdots, y_{it}, \cdots, y_{iT})'$ and $X_i{}^c = (x_{i1}, \cdots, x_{it}, \cdots, x_{iT})'$ denote the $T \times 1$ complete outcome vector and $T \times p$ covariate matrix, respectively, for the i-th $(i = 1, \cdots, K)$ individual over T successive points in time. Also, let

$$E(Y_i{}^c|x_i{}^c) = \mu_i{}^c(\beta) = (\mu_{i1}(\beta), \cdots, \mu_{it}(\beta), \cdots, \mu_{iT}(\beta))' \tag{5}$$

where $\mu_{it}(\beta) = h^{-1}(\eta_{it})$ with $\eta_{it} = x'_{it}\beta$, h being a suitable link function. For example, for linear models, a linear link function is used so that $\mu_{it}(\beta) = x'_{it}\beta$; whereas for the binary data a logistic link function is commonly used so that $\mu_{it}(\beta) = \exp(\eta_{it})/[1+\exp(\eta_{it})]$, and for count data a log linear link function is used so that $\mu_{it}(\beta) = \exp(\eta_{it})$. Further let

$$\Sigma_i^c(\beta,\rho) = A_i^{c\frac{1}{2}}(\beta)\tilde{C}_i(\rho, x_i^c)A_i^{c\frac{1}{2}}(\beta) \tag{6}$$

be the true covariance matrix of y_i^c, where $A_i^c(\beta) = \text{diag}[\sigma_{i11}(\beta), \cdots, \sigma_{itt}(\beta), \cdots, \sigma_{iTT}(\beta)]$ with $\sigma_{itt}(\beta) = var(Y_{it})$, and $\tilde{C}_i(\rho, x_i^c)$ is the correlation matrix for the ith individual with ρ as a suitable vector of correlation parameters, for example, $\rho \equiv (\rho_1, \ldots, \rho_\ell, \ldots, \rho_{T-1})'$, where ρ_ℓ is known to be the lag ℓ auto-correlation. Note that when covariates are time dependent, the true correlation matrix is free from time-dependent covariates in linear longitudinal setup, but it depends on the time-dependent covariates through X_i^c in the discrete longitudinal setup (Sutradhar 2010). In the stationary case, that is, when covariates are time independent, we will denote the correlation matrix by $\tilde{C}(\rho)$ in the complete longitudinal setup, and similar to Sutradhar (2010, 2011), this matrix satisfies the auto-correlation structure given by

$$\tilde{C}(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots \rho_{T-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots \quad 1 \end{bmatrix}, \tag{7}$$

where for $\ell = 1, \ldots, T$, ρ_ℓ is known to be the ℓth lag auto-correlation. Note that when this correlation structure (7) will be used in the incomplete longitudinal setup, it would be denoted by $\tilde{C}_i(\rho)$ as it will be constructed for T_i available responses.

As far as the missing mechanism is concerned, it is customary to assume that a longitudinal response may be missing completely at random (MCAR), or missing at random (MAR), or the missing can be non-ignorable. Under the MCAR mechanism, the missing-ness does not depend on any present, past, or future responses. Under the MAR mechanism, the missing-ness depends only on the past responses but not on the present or future responses, whereas under the non-ignorable mechanism the missing-ness depends on the past, present, and future possible responses. In notation, let R_{it} be a response indicator variable at time t $(t = 1, \cdots, T)$ for the i-th $(i = 1, \cdots, K)$ individual, so that

$$R_{it} = \begin{cases} 1, \text{ if } Y_{it} \text{ is observed} \\ 0, \text{ otherwise.} \end{cases} \tag{8}$$

Note that all individuals provide the responses at the first time point $t = 1$. Thus, we set $R_{i1} = 1$ with $P(R_{i1} = 1) = 1.0$ for all $i = 1, \cdots, K$. Further we assume that the response indicators satisfy the monotonic relationship

$$R_{i1} \geq R_{i2} \geq \cdots \geq R_{it} \geq \cdots \geq R_{iT}. \tag{9}$$

Next suppose that r_{it} denote the observed value for R_{it}. For $t = 2, \ldots, T$, one may then describe the aforementioned three missing mechanisms as

$$\begin{aligned}
\text{MCAR Model} \ &: \ Pr(R_{it} = 1 \mid y_i{}^c, x_i, r_{i,t-1} = 1) = Pr(R_{it} = 1 \mid r_{i,t-1} = 1) \\
\text{MAR Model} \ &: \ Pr(R_{it} = 1 \mid y_i{}^c, x_i, r_{i,t-1} = 1) \\
&= Pr(R_{it} = 1 \mid y_{i1}, \cdots, y_{i,t-1}, x_i, r_{i,t-1} = 1) \\
\text{Non-ignorable Model} \ &: \ Pr(R_{it} = 1 \mid y_i{}^c, x_i, r_{i,t-1} = 1) \\
&= Pr(R_{it} = 1 \mid y_{i1}, \cdots, y_{i,t-1}, y_{it}, \ldots, y_{iT}, x_i, r_{i,t-1} = 1)
\end{aligned}$$

(Little and Rubin 1987; Laird 1988; Fitzmaurice et al. 1996). Furthermore, it follows under the monotonic missing pattern (9) that $Pr(R_{it} = 1 | y_i^c, x_i, r_{i,t-1} = 0) = 0$ irrespective of the missing mechanism. Note that the inferences based on the non-ignorable missing mechanism may be quite complicated, and we do not include this complicated mechanism in the current paper.

3.1 Inferences When Longitudinal Responses Are Subject to MCAR

When the longitudinal responses are MCAR, R_{it} does not depend on the past, present, or future responses. In such a situation, R_{it} and Y_{it} are independent, implying that

$$E[R_{it}(Y_{it} - \mu_{it}(\beta))] = E[R_{it}]E[Y_{it} - \mu_{it}(\beta)] = 0, \tag{10}$$

because $E[Y_{it} - \mu_{it}(\beta)] = 0$. It is then clear that the inference for β involved in $\mu_{it}(\beta)$ is not affected by the MCAR mechanism. Thus, one may estimate the regression effects β consistently and efficiently by solving the GQL estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta} \Sigma_i^{-1}(\beta, \hat{\rho})(y_i - \mu_i(\beta)) = 0, \tag{11}$$

where for T_i-dimensional observed response vector $y_i = (y_{i1}, \ldots, y_{iT_i})'$,

$$\begin{aligned}
\mu_i(\beta) = E[Y_i] &= (\mu_{i1}(\beta), \cdots, \mu_{it}(\beta), \cdots, \mu_{iT_i}(\beta))' \\
\Sigma_i(\beta, \hat{\rho}) &= A_i^{1/2}(\beta)\tilde{C}_i(\rho, x_i)A_i^{1/2}(\beta),
\end{aligned}$$

with $A_i(\beta) = \text{diag}(\sigma_{i,11}(\beta), \cdots, \sigma_{i,tt}(\beta), \cdots, \sigma_{i,T_iT_i}(\beta))$, where $\sigma_{i,tt}(\beta) = \text{var}[Y_{it}]$. Note that the incomplete data based estimating equation (11) can be written in terms of pretended complete data. To be specific, by using the available responses $y_i = (y_{i1}, \ldots, y_{iT_i})'$ corresponding to the known response indicators

$$R_i^c = r_i^c = \begin{bmatrix} I_{T_i} & 0 \\ 0 & 0, \end{bmatrix}$$

one may write the GQL estimating equation (11) under the MCAR mechanism as

$$\sum_{i=1}^{K} \frac{\partial \mu_i^{c'}(\beta)}{\partial \beta} \left[\{I - r_i^c\} + r_i^c \Sigma_i^c(\beta, \hat{\rho}) r_i^{c'} \right]^{-1} r_i^c \left(y_i^c - \mu_i^c(\beta) \right) = 0, \tag{12}$$

where $y_i^c = (y_i^{'}, y_{im}^{'})'$ with y_{im} representing the $T - T_i$ dimensional missing responses which are unobserved but for the computational purpose in the present approach one can use it as a zero vector, for convenience, without any loss of generality. Let $\hat{\beta}_{GQL,MCAR}$ denote the solution of (11) or (12). This estimator is asymptotically unbiased and hence consistent for β.

Note that the computation of $\tilde{C}_i(\hat{\rho}, x_i)$ matrix in (11) in general, i.e., when covariates are time dependent, depends on the specific correlation structure (Sutradhar 2010). In stationary cases as well as in linear longitudinal model setup, one may, however, compute the stationary correlation matrix $\tilde{C}_i(\hat{\rho})$, by first computing a larger $\tilde{C}(\hat{\rho})$ matrix for $\ell = 1, \ldots, T-1$, and then using the desired part of this large matrix for $t = 1, \ldots, T_i$. Turning back to the computation for the larger matrix with dimension $T = \max_{1 \le i \le K} T_i$ for $T_i \geq 2$, we exploit the observed response indicator r_{it} given by

$$r_{it} = \begin{cases} 1 \text{ if } t \leq T_i \\ 0 \text{ if } T_i < t \leq T. \end{cases}$$

for all $t = 1, \ldots, T$. For known β and σ_{itt}, the ℓth lag correlation estimate $\hat{\rho}_\ell$ for the larger $\tilde{C}(\hat{\rho})$ matrix may be computed as

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K} \sum_{t=1}^{T-\ell} r_{it} r_{i,t+\ell} [\left(\frac{y_{it} - x_{it}'\beta}{\sigma_{itt}} \right) \left(\frac{y_{i,t+\ell} - x_{it,t+\ell}'\beta}{\sigma_{i,t+\ell,t+\ell}} \right)] / \sum_{i=1}^{K} \sum_{t=1}^{T-\ell} r_{it} r_{i,t+\ell}}{\sum_{i=1}^{K} \sum_{t=1}^{T} r_{it} [\frac{y_{it} - x_{it}'\beta}{\sigma_{itt}}]^2 / \sum_{i=1}^{K} r_{it}}, \tag{13}$$

(cf. Sneddon and Sutradhar 2004, eqn. (16)) for $\ell = 1, \ldots, T-1$. Note that as this estimator contains $\hat{\beta}_{GQL,MCAR}$, both (11) and (13) have to be computed iteratively until convergence.

Further note that in the existing GEE approach, instead of (11), one solves the estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta} V_i^{-1}(\beta, \hat{\alpha})(y_i - \mu_i(\beta)) = 0, \tag{14}$$

[Liang and Zeger 1986] where $V_i(\beta,\hat{\alpha}) = A_i^{1/2}(\beta)Q_i(\alpha)A_i^{1/2}(\beta)$, with $Q_i(\alpha)$ as the $T_i \times T_i$ "working" correlation matrix of y_i. It is, however, known that this GEE approach may sometimes encounter consistency breakdown (Crowder 1995) because of the difficulty in estimating the "working" correlation or covariance structure, leading to the failure of estimation of β or the non-convergence of β estimator to β. Furthermore, even if GEE β estimate becomes consistent, it may produce inefficient estimate than simpler independence assumption based moment or quasi-likelihood (QL) estimate (Sutradhar and Das 1999; Sutradhar 2011). Thus, one should be clear from these points that the GEE approach even if corrected for missing mechanism may encounter similar consistency and inefficiency in estimating the regression parameters.

We also remark that even though the non-response probability is not affected by the past history under the MCAR mechanism, the respective efficiency of GQL and GEE estimators will decrease if T_i is very small as compared to the attempted complete duration T, that is, if $T - T_i$ is large. As far as the value of T_i is concerned, it depends on the probability, $P[R_{it} = 1]$ which in general decreases due to the monotonic condition (9). This is because under this monotonic property (9) and following MCAR mechanism, one writes

$$\begin{aligned} Pr[R_{it} = 1] &\equiv Pr[R_{i1} = 1, R_{i2} = 1, \ldots, R_{it} = 1] \\ &= \Pi_{j=2}^{t} P[R_{ij} = 1], \end{aligned} \tag{15}$$

which gets smaller as t gets larger, implying that T_i can be small as compared to T if $P[R_{ij} = 1]$ is far away down from 1 such as $P[R_{ij} = 1] = 0.90$, say.

3.2 Inferences When Longitudinal Responses Are Subject to MAR

Unlike in the MCAR case, R_{it} and y_{it} are not independent under the MAR mechanism. That is

$$E[R_{it}(Y_{it} - \mu_{it}(\beta))] \neq 0 \text{ under MAR.} \tag{16}$$

This is because

$$\begin{aligned} & E\left[R_{it}(Y_{it} - \mu_{it}(\beta)) \mid H_{i,t-1}(y)\right] \\ &= E_{Y_{it}} E\left[R_{it}(Y_{it} - \mu_{it}(\beta)) \mid Y_{it}, H_{i,t-1}(y)\right] \\ &= E_{Y_{it}}\left[\{(Y_{it} - \mu_{it}(\beta))|H_{i,t-1}(y)\} E\{R_{it}|Y_{it}, H_{i,t-1}(y)\}\right] \\ &= E_{Y_{it}}\left[\{(Y_{it} - \mu_{it}(\beta))|H_{i,t-1}(y)\} E\{R_{it}|H_{i,t-1}(y)\}\right] \end{aligned} \tag{17}$$

as R_{it} does not depend on Y_{it} by the definition of MAR.

Next due to the monotonic property (9) of the response indicators

$$\begin{aligned}
&E\left[R_{it} \mid H_{i,t-1}(y)\right] \\
&= P\left[R_{i1} = 1, R_{i2} = 1, \cdots, R_{i,t-1} = 1, R_{it} = 1 | H_{i,t-1}(y)\right] \\
&= P(R_{i1} = 1) P\left[R_{i2} = 1 \mid R_{i1} = 1; H_{i1}(y)\right] \cdots \\
&\times P\left[R_{it} = 1 \mid R_{i1} = 1, \cdots, R_{i,t-1} = 1; H_{i,t-1}(y)\right] \\
&= \prod_{j=1}^{t} g_{ij}(y_{i,j-1}, \cdots, y_{i,j-q}; \gamma) \\
&= w_{it}\{H_{i,t-1}(y); \gamma\}, \qquad (18)
\end{aligned}$$

and

$$E_{Y_{it}}\left[(Y_{it} - \mu_{it}(\beta)) | H_{i,t-1}(y)\right] = (\lambda_{it}(H_{i,t-1}(y), \beta, \rho) - \mu_{it}(\beta)), \qquad (19)$$

where $\lambda_{it}(H_{i,t-1}(y), \beta, \rho)$ is the conditional mean of Y_{it}. In (18), one may, for example, use $g_{ij}(\gamma)$ as

$$\begin{aligned}
g_{ij}(\gamma) &= Pr[(R_{ij} = 1) | R_{i1} = 1, \ldots, R_{i,j-1} = 1, H_{i,j-1}(y)] \\
&= \frac{\exp(1 + \gamma y_{i,j-1})}{1 + \exp(1 + \gamma y_{i,j-1})}. \qquad (20)
\end{aligned}$$

Now because both $w_{it}\{H_{i,t-1}(y); \gamma\}$ and $\lambda_{it}(H_{i,t-1}(y), \beta, \rho)$ are functions of the past history of responses $H_{i,t-1}(y)$, and because

$$E_{H_{i,t-1}(y)}[\lambda_{it}(H_{i,t-1}(y), \beta, \rho) - \mu_{it}(\beta)] = 0, \qquad (21)$$

it then follows from (17), by (18) and (19), that

$$E[R_{it}(Y_{it} - \mu_{it}(\beta))] = E_{H_{i,t-1}(y)} E\left[R_{it}(Y_{it} - \mu_{it}(\beta)) \mid H_{i,t-1}(y)\right] \neq 0, \qquad (22)$$

unless $w_{it}\{H_{i,t-1}(y); \gamma\}$ is a constant free of $H_{i,t-1}(y)$, which is, however, impossible under MAR missing mechanism as opposed to the MCAR mechanism. Thus, $E[R_{it}\{Y_{it} - \mu_{it}(\beta)\}] \neq 0$.

3.2.1 Existing Partially Standardized GEE Estimation for Longitudinal Data Subject to MAR

Note, however, that

$$\begin{aligned}
&E\left\{\frac{R_{it}}{w_{it}\{H_{i,t-1}(y); \gamma\}}(Y_{it} - \mu_{it}(\beta))\right\} \\
&= E_{H_{i,t-1}(y)}\left[(\lambda_{it}(H_{i,t-1}(y), \beta, \rho) - \mu_{it}(\beta))\right] = 0. \qquad (23)
\end{aligned}$$

Now suppose that

$$\Delta_i = \text{diag}[\delta_{i1}, \delta_{i2}, \cdots, \delta_{iT_i}] \text{ with } \delta_{it} = R_{it}/w_{it}\{H_{i,t-1}(y); \gamma\}$$

implying that $E[\Delta_i | H_i(y)] = I_{T_i}$, and where $H_i(y)$ is used to denote appropriate past history showing that the response indicators are generated based on observed responses only.

By observing the unconditional expectation property from (23), in the spirit of GEE [Liang and Zeger 1986], Robins et al. (1995, eqn. (10), p. 109) proposed a conditional inverse weights based PSGEE for the estimation of β which has the form

$$\sum_{i=1}^{K} \frac{\partial E_{H_i(y)} E[\{\Delta_i \mu_i(\beta)\}' | H_i(y)]}{\partial \beta} V_i^{-1}(\hat{\alpha}) \{\Delta_i (y_i - \mu_i(\beta)) | H_i(y)\}$$

$$= \sum_{i=1}^{K} \frac{\partial \{\mu_i(\beta)\}'}{\partial \beta} V_i^{-1}(\hat{\alpha}) \{\Delta_i (y_i - \mu_i(\beta))\} = 0, \qquad (24)$$

(see also Paik 1997, eqn. (1), p. 1321). Note that we refer to the GEE in (23) as a partly or partially standardized GEE (PSGEE) because $V_i(\alpha) = \hat{\text{cov}}(Y_i)$ used in this GEE is a partial weight matrix which ignores the missing mechanism, whereas $\text{cov}[\Delta_i (y_i - \mu_i(\beta))]$ would be a full weight matrix.

Note that over the last decade many researchers have used this PSWGEE approach for studying various aspects of longitudinal data subject to non-response. See, for example, the studies by Rotnitzky et al. (1998), Preisser et al. (2002), and Birmingham et al. (2003), among others. However, even if the MAR mechanism is accommodated to develop an unbiased estimating function $\Delta_i (y_i - \mu_i(\beta))$ (for 0) to construct the fully standardized GEE (FSGEE), the consistency of the estimator of β may break down (see Crowder 1995 for complete longitudinal models) because of the use of "working" covariance matrix $V_i(\alpha)$, whereas the true covariance matrix for y_i is given by $\text{cov}[Y_i] = \Sigma_i(\rho)$. This can happen for those cases where α is not estimable. To be more clear, $V_i(\alpha)$ is simply a "working" covariance matrix of y_i, whereas a proper estimating equation must use the correct variance (or its consistent estimate) matrix of $\{\Delta_i (y_i - \mu_i(\beta))\}$.

To understand the roles of both missing mechanism and longitudinal correlation structure in constructing a proper estimating equation, we now provide following three estimating equations for β. The difficulties and/or advantages encountered by these equations are also indicated.

3.2.2 Partially Standardized GQL (PSGQL) Estimation for Longitudinal Data Subject to MAR

When $V_i(\alpha)$ matrix in (24) is replaced with the true $T_i \times T_i$ covariance matrix of the available responses, that is, $\Sigma_i(\rho) = \text{cov}[Y_i]$, one obtains the PSGQL estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial\{\mu_i(\beta)\}'}{\partial\beta} \Sigma_i^{-1}(\hat{\rho})\{\Delta_i(y_i - \mu_i(\beta))\} = 0, \tag{25}$$

which also may produce biased and hence inconsistent estimate. This is because $\Sigma_i(\rho)$ may still be very different than the covariance matrix of the actual variable $\{\Delta_i(y_i - \mu_i(\beta))\}$. Thus, if the proportion of missing values is more, one may not get convergent solution to the estimating equation (25) and the consistency for β would break down (Crowder 1995). The convergence problems encountered by (24) would naturally be more severe as even in the complete data case $V_i(\alpha)$ may not be estimable.

3.2.3 Partially Standardized Conditional GQL (PSCGQL) Estimation for Longitudinal Data Subject to MAR

Suppose that one uses conditional (on history) variance

$$\text{cov}\{\Delta_i(Y_i - \mu_i(\beta))\}\,|H_i(y) = \Sigma^*_{ich}(H_i(y), \beta, \rho, \gamma), \tag{26}$$

to construct the estimating equation. Then following (25), one may write the PSCGQL estimating equation given by

$$\sum_{i=1}^{K} \frac{\partial\{\mu_i(\beta)\}'}{\partial\beta} {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)\{\Delta_i(y_i - \mu_i(\beta))\} = 0 \tag{27}$$

It is, however, seen that

$$\begin{aligned} & {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)[\Delta_i(Y_i - \mu_i(\beta))] \\ \to\ & {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)E[\{\Delta_i(Y_i - \mu_i(\beta))\}|H_i(y)] \\ =\ & {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma)[\lambda_i(H_i(y)) - \mu_i(\beta)] \end{aligned} \tag{28}$$

But,

$$\begin{aligned} E_{H_i(y)} \Bigg[& \frac{\partial\{\mu_i(\beta)\}'}{\partial\beta} {\Sigma^*_{ich}}^{-1}(H_i(y), \beta, \rho, \gamma) \\ \times\ & [\lambda_i(H_i(y)) - \mu_i(\beta)]\Bigg] \neq 0, \end{aligned} \tag{29}$$

even though

$$E_{H_i(y)}[\lambda_i(H_i(y)) - \mu_i(\beta)] = 0.$$

Thus, the PSCGQL estimating equation (27) is not an unbiased equation for 0, and may produce bias estimate.

Computational formula for $\Sigma^*_{ich}(\beta,\rho,\gamma)$

For convenience, we first write

$$\begin{aligned}\Delta_i &= W_i^{-1}R_i, \text{ with}\\ W_i &= \text{diag}[w_{i1},w_{i2},\ldots,w_{iT_i}], \text{ and } R_i = \text{diag}[R_{i1},\ldots,R_{iT_i}].\end{aligned}$$

It then follows that

$$\begin{aligned}\Sigma^*_{ich}(\beta,\rho) &= \text{cov}[\{\Delta_i\{(y_i-\mu_i(\beta))\}\}|H_i(y)]\\ &= W_i^{-1}\text{cov}[\{R_i(y_i-\mu_i(\beta))\}|H_i(y)]W_i^{-1}. \qquad (30)\end{aligned}$$

Now to compute the covariance matrix in the middle term in the right-hand side of (30), we first re-express $R_i(y_i-\mu_i(\beta))$ as

$$R_i(y_i-\mu_i(\beta)) = [R_{i1}(y_{i1}-\mu_{i1}),\ldots,R_{it}(y_{it}-\mu_{it}),\ldots,R_{iT_i}(y_{iT_i}-\mu_{iT_i})]',$$

and compute the variances for its components as

$$\text{var}[\{R_{i1}(y_{i1}-\mu_{i1})\}|y_{i1}] = 0, \qquad (31)$$

because $R_{i1}=1$ always and y_{i1} is random . In the Poisson case $\sigma_{i,11}=\mu_{i1}$ and in the binary case $\sigma_{i,11}=\mu_{i1}(1-\mu_{i1})$, with appropriate formula for μ_{i1} in a given case. Next for $t=2,\ldots,T_i$,

$$\begin{aligned}&\text{var}[R_{it}(y_{it}-\mu_{it})|H_{i,t-1}(y)] = \text{var}[R_{it}|H_{i,t-1}(y)]\text{var}[y_{it}|H_{i,t-1}(y)]\\ &+E^2[R_{it}|H_{i,t-1}(y)]\text{var}[(y_{it})|H_{i,t-1}(y)]+\text{var}[R_{it}|H_{i,t-1}(y)]E^2[(y_{it}-\mu_{it})|H_{i,t-1}(y)]\\ &= w_{it}(1-w_{it})\sigma_{ic,tt}+w_{it}^2\sigma_{ic,tt}+w_{it}(1-w_{it})\{\lambda_{it}-\mu_{it}\}^2\\ &= w_{it}[\sigma_{ic,tt}+(\lambda_{it}-\mu_{it})^2]-w_{it}^2(\lambda_{it}-\mu_{it})^2, \qquad (32)\end{aligned}$$

where, given the history, λ_{it} and $\sigma_{ic,tt}$ are the conditional mean and variance of y_{it}, respectively.

Furthermore, all pairwise covariances conditional on the history $H_{i,t-1}(y)$ may be computed as follows. For $u<t$,

$$\begin{aligned}&\text{cov}[\{R_{iu}(y_{iu}-\mu_{iu}),R_{it}(y_{it}-\mu_{it})\}|H_{i,t-1}(y)]\\ &= E[\{R_{iu}R_{it}(y_{iu}-\mu_{iu})(y_{it}-\mu_{it})\}|H_{i,t-1}(y)]-E[\{R_{iu}(y_{iu}-\mu_{iu})\}|H_{i,t-1}(y)]\\ &\times E[\{R_{it}(y_{it}-\mu_{it})\}|H_{i,t-1}(y)]\\ &= (y_{iu}-\mu_{iu})E[\{R_{it}(y_{it}-\mu_{it})\}|H_{i,t-1}(y)]-[w_{iu}(y_{iu}-\mu_{iu})][w_{it}(\lambda_{it}-\mu_{it})]\\ &= [(y_{iu}-\mu_{iu})(1-w_{iu})][w_{it}(\lambda_{it}-\mu_{it})] \qquad (33)\end{aligned}$$

3.2.4 A Fully Standardized GQL (FSGQL) Approach

All three estimating equations, namely PSGEE (24), PSGQL (25), and PSCGQL (27) may produce bias estimates, PSGEE being the worst. The reasons for the poor performance of PSGEE are two fold. This is because it completely ignores the missing mechanism and uses a working correlation matrix to accommodate the longitudinal nature of the available data. As opposed to the PSGEE approach, PSGQL approach uses the true correlation structure under a class of auto-correlations but similar to the PSGEE approach it also ignores the missing mechanism. As far as the PSCGQL approach it uses a correct conditional covariance matrix which accommodates both missing mechanism and correlation structure. However, the resulting estimating equation may not unbiased for zero as the history of the responses involved in covariance matrix make a weighted distance function which is not unbiased.

To remedy the aforementioned problems, it is therefore important to use the correct covariance matrix or its consistent estimate to construct the weight matrix by accommodating both missing mechanism and longitudinal correlations of the repeated data. For this to happen, because the distance function is unconditionally unbiased for zero, i.e.,

$$E_{H_i(y)}E[\{\Delta_i(Y_i-\mu_i(\beta))\}\,|H_i(y)]=0,$$

one must use the unconditional covariance matrix of $\{\Delta_i(Y_i-\mu_i(\beta))\}$ to compute the incomplete longitudinal weight matrix, for the construction of a desired unbiased estimating equation. Let $\Sigma_i^*(\beta,\rho,\gamma)$ denote this unconditional covariance matrix which is computed by using the formula

$$\begin{aligned}\Sigma_i^*(\beta,\rho,\gamma)&=\text{cov}\{\Delta_i(Y_i-\mu_i(\beta))\}=E_{H_i(y)}[\text{cov}\{\Delta_i(Y_i-\mu_i(\beta))\}\,|H_i(y)]\\&\quad+\text{cov}_{H_i(y)}[E\{\Delta_i(Y_i-\mu_i(\beta))\}\,|H_i(y)].\end{aligned}\tag{34}$$

In the spirit of Sutradhar (2003), we propose the FSGQL estimating equation for β given by

$$\begin{aligned}&\sum_{i=1}^{K}\frac{\partial E_{H_i(y)}E[\{\Delta_i\mu_i(\beta)\}'|H_i(y)]}{\partial\beta}[\text{cov}\{\Delta_i(y_i-\mu_i)\}]^{-1}\{\Delta_i(y_i-\mu_i(\beta))\}\\&=\sum_{i=1}^{K}\frac{\partial\mu_i'}{\partial\beta}[\Sigma_i^*(\beta,\rho,\gamma)]^{-1}\{\Delta_i(y_i-\mu_i(\beta))\}=0,\end{aligned}\tag{35}$$

where $\Sigma_i^*(\beta,\rho,\gamma)$ is yet to be computed. This estimating equation is solved iteratively by using

$$\hat{\beta}_{FSGQL}(m+1) = \hat{\beta}_{FSGQL}(m) + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'(\beta)}{\partial \beta}[\Sigma_i^*(\beta,\rho,\gamma)]^{-1}\frac{\partial \mu_i(\beta)}{\partial \beta'}\right]_m^{-1}$$
$$\times \left[\sum_{i=1}^{K} \frac{\partial \mu_i^{'}}{\partial \beta}[\Sigma_i^*(\beta,\rho,\gamma)]^{-1}\Delta_i(y_i - \mu_i(\beta))\right]_m \quad (36)$$

Computation of $\Sigma_i^*(\beta,\rho,\gamma) = cov[\Delta_i(y_i - \mu_i)]$

Rewrite (34) as

$$\begin{aligned}\Sigma_i^*(\beta,\rho,\gamma) &= E_{H_i(y)}[\text{cov}\{\Delta_i(Y_i - \mu_i(\beta))\}|H_i(y)] \\ &+ \text{cov}_{H_i(y)}[E\{\Delta_i(Y_i - \mu_i(\beta))\}|H_i(y)] \\ &= E_{H_i(y)}[\Sigma_{ich}^*(\beta,\rho)] + \text{cov}_{H_i(y)}[E_{ich}(\beta,\rho)],\end{aligned} \quad (37)$$

where $\Sigma_{ich}^*(\beta,\rho)$ is constructed by (30) by using the formulas from (31) to (33), and $E_{ich}(\beta,\rho)$ has the form $E_{ich}(\beta,\rho) = [(y_{i1} - \mu_{i1}), (\lambda_{i2} - \mu_{i2}), \ldots, (\lambda_{iT_i} - \mu_{iT_i})]'$.

It then follows that the components of the $T_i \times T_i$ unconditional covariance matrix $\Sigma_i^*(\beta,\rho,\gamma)$ are given by

$$\text{cov}[\delta_{iu}(y_{iu} - \mu_{iu}), \delta_{it}(y_{it} - \mu_{it})] \quad (38)$$
$$= \begin{cases} \text{var}_{y_{i1}}(y_{i1} - \mu_{i1}) = \sigma_{i11} & \text{for u=t=1} \\ E_{H_i(y)}[w_{it}^{-1}\{\sigma_{ic,tt} + (\lambda_{it} - \mu_{it})^2\} - (\lambda_{it} - \mu_{it})^2] + E_{H_i(y)}(\lambda_{it} - \mu_{it})^2, & \text{for u=t=2,...} \\ E_{H_i(y)}[(y_{i1} - \mu_{i1})(\lambda_{it} - \mu_{it})] & \text{for u=1,t=2,...} \\ E_{H_i(y)}[(w_{iu}^{-1} - 1)\{(y_{iu} - \mu_{iu})(\lambda_{it} - \mu_{it})\}] & \\ \quad + E_{H_i(y)}[(\lambda_{iu} - \mu_{iu})(\lambda_{it} - \mu_{it})], & \text{for u=2,...; } u < t \end{cases}$$

(a). Example of $\Sigma_i^*(\beta,\rho,\gamma)$ under linear longitudinal models with $T = 2$

Note that $R_{i1} = r_{i1} = 1$ always. But R_{i2} can be 1 or 0 and under MAR, its probability depends on y_{i1}. Consider

$$Pr[R_{i1} = 1] = g_{i1} = w_{i1} = 1.0$$

$$P[R_{i2} = 1|r_{i1} = 1, y_{i1}] = g_{i2}(\gamma) = \frac{\exp\{1 + \gamma y_{i1}\}}{1 + \exp\{1 + \gamma y_{i1}\}} \quad (39)$$

by (20), yielding

$$\begin{aligned} w_{i2} &= E[R_{i2}|H_{i1(y)}] = P[R_{i1}=1, R_{i2}=1|H_{i1}(y)] \\ &= P[R_{i1}=1]P[R_{i2}=1|H_{i1}(y)] = g_{i1}g_{i2}(y_{i1}), \end{aligned}$$

(see also (18)). With regard to the longitudinal model for potential responses y_{i1}, y_{i2}, along with their non-stationary (time dependent covariates), consider the model as:

$$y_{it} \sim (x'_{it}\beta, \frac{\sigma^2}{1-\rho^2}),\ \mathrm{corr}(Y_{it}, Y_{i,t+\ell}) = \rho^{\ell}. \tag{40}$$

Assuming normal distribution, one may write

$$\begin{aligned} E[Y_{it}|H_{i,t-1}] &= x'_{it}\beta + [\mathrm{cov}(y_{it}, y_{i,t-\ell})][\mathrm{var}(y_{i,t-\ell})]^{-1}(y_{i,t-\ell} - x'_{i,t-\ell}\beta) \\ &= x'_{it}\beta + [\frac{\sigma^2\rho^{\ell}}{1-\rho^2}] = x'_{it}\beta + \rho^{\ell}[y_{i,t-\ell} - x'_{i,t-\ell}\beta]. \end{aligned} \tag{41}$$

When the response y_{it} depends on its immediate history, the conditional mean has the formula

$$E[Y_{it}|y_{i,t-1}] = \lambda_{it} = x'_{it}\beta + \rho(y_{i,t-1} - x'_{i,t-1}\beta),$$

implying that the unconditional mean is given by $\mu_{it} = E[Y_{it}] = x'_{it}\beta$, which is the same as the mean in (40), as expected.

Now following (38), we provide the elements of the 2×2 matrix $\Sigma^*_i(\beta, \rho, \gamma)$ as

$$\begin{aligned} \sigma^*_{i11} &= \frac{\sigma^2}{1-\rho^2} \\ \sigma^*_{i12} &= \sigma^*_{i21} = \rho\,\mathrm{var}[Y_{i1} - x_{i11}\beta] = \rho\frac{\sigma^2}{1-\rho^2} \\ \sigma^*_{i22} &= E_{y_{i1}}[w_{i2}^{-1}\{\mathrm{var}(Y_{i2}|y_{i1}) + (\lambda_{i2} - \mu_{i2})^2\}] \\ &= E_{y_{i1}}[\{1 + \frac{1}{\exp(1+\gamma y_{i1})}\}\{\sigma^2 + \rho^2(y_{i1} - x_{i11}\beta)^2\}] \\ &= \frac{\sigma^2}{1-\rho^2} + \sigma^2 E_1 + \rho^2 E_2, \end{aligned} \tag{42}$$

where

$$E_1 = \int [\frac{1}{\exp(1+\gamma y_{i1})}] g_N(y_{i1}) dy_{i1}, \text{ and}$$

$$E_2 = \int [\frac{\{y_{i1} - x_{i11}\beta\}^2}{\exp(1+\gamma y_{i1})}] g_N(y_{i1}) dy_{i1},$$

$g_N(y_{i1})$ being the normal (say) density of y_{i1}. Thus, $\Sigma_i^*(\beta,\rho,\gamma)$ has the formula

$$\Sigma_i^*(\beta,\rho,\gamma) = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho \\ \rho & \{1+(1-\rho^2)E_1 + \frac{\rho^2(1-\rho^2)}{\sigma^2}E_2\} \end{bmatrix}, \tag{43}$$

Note that in the complete longitudinal case w_{i2} would be 1 and σ^*_{i22} would reduce to $\frac{\sigma^2}{1-\rho^2}$, leading to

$$\Sigma_i^*(\beta,\rho,\gamma) = \Sigma_i(\beta,\rho) = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{44}$$

which is free from β in this linear model case, and the *PSGEE* (24) uses a "working" version of (44), namely

$$V_i(\alpha) = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \tag{45}$$

whereas the FSGQL estimating equation (35) would use $\Sigma_i^*(\beta,\rho,\gamma)$ from (43). This shows the effect of missing mechanism in the construction of the weight matrix for the estimating equation.

(b). Example of $\Sigma_i^*(\beta,\rho,\gamma)$ under binary longitudinal AR(1) model with $T = 2$

Consider a binary AR(1) model with

$$\lambda_{it} = E[Y_{it}|y_{i,t-1}] = \mu_{it} + \rho(y_{i,t-1} - \mu_{i,t-1}),\ t = 2,\ldots,T, \tag{46}$$

where $\mu_{it} = \frac{\exp(x'_{it}\beta)}{1+\exp(x'_{it}\beta)}$, for all $t = 1,\ldots,T$.

Now considering y_{i1} as fixed, by using (31)–(33) we first compute the history-dependent conditional covariance matrix $\Sigma_{ich}(\beta,\rho) = \text{cov}[\{\Delta_i(y_i - \mu_i)\}|H_i(y)]$ as:

$$\begin{aligned} &\text{var}[\delta_{i1}(y_{i1} - \mu_{i1})] = 0 \\ &\text{var}[\{\delta_{i2}(y_{i2} - \mu_{i2})\}|y_{i1}] = \frac{1}{w_{i2}}[\lambda_{i2}(1-\lambda_{i2}) + \rho^2(y_{i1} - \mu_{i1})^2] - \rho^2(y_{i1} - \mu_{i1})^2 \\ &\text{cov}[\{\delta_{i1}(y_{i1} - \mu_{i1}), \delta_{i2}(y_{i2} - \mu_{i2})\}|y_{i1}] = 0, \end{aligned} \tag{47}$$

yielding

$$\begin{aligned} &E_{H_i(y)}[\Sigma_{ich}(\beta,\rho)] \\ &= \begin{cases} E_{y_{i1}}[\sigma_{ich,11}] = E_{y_{i1}}[0] = 0 \\ E_{y_{i1}}[\sigma_{ich,22}] = E_{y_{i1}}[\frac{1}{w_{i2}}[\lambda_{i2}(1-\lambda_{i2}) + \rho^2(y_{i1} - \mu_{i1})^2] - \rho^2(y_{i1} - \mu_{i1})^2] \\ E_{y_{i1}}[\sigma_{ich,12}] = E_{y_{i1}}[0] = 0 \end{cases} \end{aligned} \tag{48}$$

Next because

$$E_{ich}(\beta,\rho) = E\{\Delta_i(y_i - \mu_i)\}|H_i(y)] = [(y_{i1} - \mu_{i1}),(\lambda_{i2} - \mu_{i2})]',$$

one obtains

$$\text{cov}_{H_i(y)}[E_{ich}(\beta,\rho)] \tag{49}$$
$$= \begin{cases} \text{var}_{y_{i1}}[y_{i1} - \mu_{i1}] = \mu_{i1}[1 - \mu_{i1}] \\ \text{cov}_{y_{i1}}[(y_{i1} - \mu_{i1}),(\lambda_{i2} - \mu_{i2})] = \rho\,\text{var}_{y_{i1}}[y_{i1} - \mu_{i1}] = \rho\mu_{i1}[1 - \mu_{i1}] \\ \text{var}_{y_{i1}}[\lambda_{i2} - \mu_{i2}] = \text{var}_{y_{i1}}[\rho(y_{i1} - \mu_{i1})] = \rho^2[\mu_{i1}(1 - \mu_{i1})]. \end{cases}$$

By combining (48) and (49), it follows from (38) that the 2 × 2 unconditional covariance matrix $\Sigma_i^*(\beta,\rho,\gamma)$ has the form

$$\begin{aligned} \text{var}[\delta_{i1}(y_{i1} - \mu_{i1})] &= \mu_{i1}[1 - \mu_{i1}] \\ \text{var}[\delta_{i2}(y_{i2} - \mu_{i2})] &= E_{y_{i1}}[\frac{1}{w_{i2}}\{\lambda_{i2}(1 - \lambda_{i2}) + \rho^2(y_{i1} - \mu_{i1})^2\}] \\ &= [\mu_{i2}(1 - \mu_{i2})]E[w_{i2}^{-1}] \qquad (50) \\ &\quad +\rho(1 - 2\mu_{i2})E[w_{i2}^{-1}(y_{i1} - \mu_{i1})] \\ &= [\mu_{i2}(1 - \mu_{i2})]E_{1y} + \rho(1 - 2\mu_{i2})[E_{2y} - \mu_{i1}E_{1y}] \\ \text{cov}[\delta_{i1}(y_{i1} - \mu_{i1}),\delta_{i2}(y_{i2} - \mu_{i2})] &= \rho[\mu_{i1}\{1 - \mu_{i1}\}], \qquad (51) \end{aligned}$$

where

$$\begin{aligned} E_{1y} = E[w_{i2}^{-1}] &= \{1 + \exp(-1) + \mu_{i1}\exp(-1)(\exp(-\gamma) - 1)\} \\ E_{2y} = E[\frac{y_{i1}}{w_{i2}}] &= \mu_{i1}\{1 + \exp(-\gamma - 1)\}. \end{aligned}$$

General formula for $\Sigma_i^*(\beta,\rho,\gamma)$ under the binary AR(1) model

In general, it follows from (38) that the elements of the $T_i \times T_i$ unconditional covariance matrix $\Sigma_i^*(\beta,\rho,\gamma)$ under AR(1) binary model are given by

$$\text{cov}[\delta_{iu}(y_{iu} - \mu_{iu}),\delta_{it}(y_{it} - \mu_{it})] \tag{52}$$
$$\equiv \begin{cases} \sigma^*_{i,11} = \mu_{i1}[1 - \mu_{i1}] \\ \sigma^*_{i,tt} = E_{H_i(y)}[w_{it}^{-1}\{\mu_{it}(1 - \mu_{it}) + \rho(1 - 2\mu_{it})(y_{i,t-1} - \mu_{i,t-1})\}, (\text{for } t = 2,\ldots,T_i) \\ \sigma*_{i,ut} = \rho\rho^{t-1-u}\mu_{iu}(1 - \mu_{iu}),\ (\text{for } u = 1 < t) \\ \sigma^*_{i,ut} = \rho^2\rho^{t-u}\mu_{i(u-1)}(1 - \mu_{i(u-1)}),\ (\text{for } 1 < u < t). \end{cases}$$

3.3 An Empirical Illustration

First, to illustrate the performance of the existing PSGEE (24) approach, we refer to some of the simulation results reported by Sutradhar and Mallick (2010). It was shown that this approach may produce highly biased and hence inconsistent regression estimates. In fact these authors also demonstrated that PSGEE(I) (independence assumption based) approach produces less biased estimates than any "working" correlation structures based PSGEE approaches. For example, we consider here their simulation design chosen as

Simulation Design

$K = 100$, $T = 4$, $p = 2$, $q = 1$, $\gamma = 4$, $\rho = 0.4, 0.8$, $\beta_1 = \beta_2 = 0$ along with two time-dependent covariates:

$$x_{it1} = \begin{cases} \frac{1}{2} & \text{for } i = 1, \cdots, \frac{K}{4}; t = 1,2 \\ 0 & \text{for } i = 1, \cdots, \frac{K}{4}; t = 3,4 \\ -\frac{1}{2} & \text{for } i = \frac{K}{4}+1, \cdots, \frac{3K}{4}; t = 1 \\ 0 & \text{for } i = \frac{K}{4}+1, \cdots, \frac{3K}{4}; t = 2,3 \\ \frac{1}{2} & \text{for } i = \frac{K}{4}+1, \cdots, \frac{3K}{4}; t = 4 \\ \frac{t}{2T} & \text{for } i = \frac{3K}{4}+1, \cdots, K; t = 1, \cdots, 4 \end{cases}$$

and

$$x_{it2} = \begin{cases} \frac{t-2.5}{2T} & \text{for } i = 1, \cdots, \frac{K}{2}; t = 1, \cdots, 4 \\ 0 & \text{for } i = \frac{K}{2}+1, \cdots, K; t = 1,2 \\ \frac{1}{2} & \text{for } i = \frac{K}{2}+1, \cdots, K; t = 3,4 \end{cases}$$

Details on the MAR based incomplete binary data generation, one may be referred to Sutradhar and Mallick (2010, Sect. 2.1). Based on 1,000 simulations, the PSGEE estimates obtained from (24) and PSGEE (I) obtained from (24) by using zero correlation are displayed in Table 1.

These results show that the PSGEE estimates for $\beta_1 = 0$ and $\beta_2 = 0$ are highly biased. For example, when $\rho = 0.8$, the estimates of β_1 and β_2 are -0.213 and -0.553, respectively. These estimates are inconsistent and unacceptable. Note that these biases are caused by the wrong correlation matrix used to construct the PSGEE (24), whereas this PSGEE provides almost unbiased estimates when data are treated to be independent even if truly they are not so. However the standard errors of the PSGEE(I) estimates appear to be large and hence it may provide inefficient estimates. In fact when the proportion of missing values is large, the PSGEE(I) will also encounter estimation breakdown or it will produce biased estimates. This

Table 1 Simulated means (SMs), simulated standard errors (SSEs), and simulated mean squared errors (SMSEs) for "working" correlations based PSGEE (24) estimates, when the incomplete longitudinal responses were generated based on MAR mechanism (20) with $\gamma = 4.0$ and a longitudinal AR(1) correlation structure with correlation index parameter ρ; $\beta_1 = \beta_2 = 0$; based on 1,000 simulations

		Estimation approach				
		PSGEE(AR(1))			PSGEE (I)	
ρ	Statistic	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$	$\hat{\beta}_1$	$\hat{\beta}_2$
0.4	SM	−0.076	−0.224	0.404	0.015	0.015
	SSE	0.361	0.544	0.062	0.384	0.587
	SMSE	0.136	0.346	0.004	0.148	0.344
0.8	SM	−0.213	−0.553	0.802	0.007	0.017
	SSE	0.257	0.381	0.038	0.378	0.614
	SMSE	0.112	0.450	0.001	0.143	0.377

is verified by a simulation study reported by Mallick et al. (2013). The reason for this inconsistency encountered by PSGEE and PSGEE(I) is the failure of accommodating MAR mechanism in the covariance matrix used as the longitudinal weights.

As a remedy to this inconsistency, we have developed a FSGQL (35) estimating equation by accommodating both MAR mechanism and longitudinal correlation structure in constructing the weight matrix $\Sigma_i^*(\beta, \rho, \gamma)$. This FSGQL equation would provide consistent and efficient regression estimates. For simplicity, Mallick et al. (2013) have demonstrated through a simulation study that FSGQL(I) approach by using $\rho = 0$ in $\Sigma_i^*(\beta, \rho = 0, \gamma)$ produces almost unbiased estimates with small variances. This provides a guidance that ignoring missing mechanism in constructing the weight matrix would provide detrimental results, whereas ignoring longitudinal correlations does not appear to cause any significant loss.

Acknowledgment The author fondly acknowledges the stimulating discussion by the audience of the symposium and wishes to thank for their comments and suggestions.

References

Birmingham, J., Rotnitzky, A., Fitzmaurice, G.M.: Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. J. R. Stat. Soc. Ser. B **65**, 275–297 (2003)

Crowder, M.: On the use of a working correlation matrix in using generalized linear models for repeated measures. Biometrika **82**, 407–410 (1995)

Fitzmaurice, G.M., Laird, N.M., Zahner, G.E.P.: Multivariate logistic models for incomplete binary responses. J. Am. Stat. Assoc. **91**, 99–108 (1996)

Ibrahim, J.G., Lipsitz, S.R., Chen, M.H.: Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. J. R. Stat. Soc. Ser. B **61**, 173–190 (1999)

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R.: Missing responses in generalized linear mixed models when the missing data mechanism is non-ignorable. Biometrika **88**, 551–564 (2001)

Krishnamoorthy, K., Pannala, M.K.: Confidence estimation of a normal mean vector with incomplete data. Can. J. Stat. **27**, 395–407 (1999)

Laird, N.M.: Missing data in longitudinal studies. Stat. Med. **7**, 305–315 (1988)

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

Little, R.J.A.: A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc. **83**, 1198–1202 (1988)

Little, R.J.A.: Modeling the drop-out mechanism in repeated-measures studies. J. Am. Stat. Assoc. **90**, 1112–1121 (1995)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (1987)

Lord, F.M.: Estimation of parameters from incomplete data. J. Am. Stat. Assoc. **50**, 870–876 (1995)

Mallick, T., Farrell, P.J., Sutradhar, B.C.: Consistent estimation in incomplete longitudinal binary models. In: Sutradhar, B.C. (ed.) ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers. Springer Lecture Notes Series, pp. 125146. Springer, New York, (2013)

Mehta, J.S., Gurland, J.: A test of equality of means in the presence of correlation and missing values. Biometrika **60**, 211–213 (1973)

Meng, X.L.: Multiple-imputation inferences with uncongenial sources of input. Stat. Sci. **9**, 538–573 (1994)

Morrison, D.F.: A test for equality of means of correlated variates with missing data on one response. Biometrika **60**, 101–105 (1973)

Naik, U.D.: On testing equality of means of correlated variables with incomplete data. Biometrika **62**, 615–622 (1975)

Paik, M.C.: The generalized estimating equation approach when data are not missing completely at random. J. Am. Stat. Assoc. **92**, 1320–1329 (1997)

Preisser, J.S., Lohman, K.K., Rathouz, P.J.: Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. Stat. Med. **21**, 3035–3054 (2002)

Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data . J. Am. Stat. Assoc. **90**, 106–121 (1995)

Rotnitzky, A., Robins, J.M., Scharfstein, D.O.: Semi-parametric regression for repeated outcomes with nonignorable nonresponse. J. Am. Stat. Assoc. **93**, 1321–1339 (1998)

Rubin, D.B.: Inference and missing data (with discussion). Biometrika **63**, 581–592 (1976)

Rubin, D.B., Schenker, N.: Multiple imputation for interval estimation from simple random sample with ignorable nonresponses. J. Am. Stat. Assoc. **81**, 366–374 (1986)

Sneddon, G., Sutradhar, B.C.: On semi-parametric familial longitudinal models. Statist. Prob. Lett. **69**, 369–379 (2004)

Sutradhar, B.C.: An overview on regression models for discrete longitudinal responses. Stat. Sci. **18**, 377–393 (2003)

Sutradhar, B.C.: Inferences in generalized linear longitudinal mixed models. Can. J. Stat. **38**, 174–196 (2010), Special issue

Sutradhar, B.C.: Dynamic Mixed Models for Familial Longitudinal Data. Springer, New York (2011)

Sutradhar, B.C., Das, K.: On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika **86**, 459–65 (1999)

Sutradhar, B.C., Mallick, T.S.: Modified weights based generalized quasilikelihood inferences in incomplete longitudinal binary models. Can. J. Stat. **38**, 217–231 (2010), Special issue

Troxel, A.B., Lipsitz, S.R., Harrington, D.P.: Marginal models for the analysis of longitudinal measurements subject to non-ignorable and non-monotonic missing data. Biometrika **85**, 661–672 (1988)

Troxel, A.B., Lipsitz, S.R., Brennan, T.A.: Weighted estimating equations with nonignorably missing response data. Biometrics **53**, 857–869 (1997)

Wang, Y.-G.: Estimating equations with nonignorably missing response data. Biometrics **55**, 984–989 (1999)