Statistics and Its Interface Volume 6 (2013) 19–26

Modeling of mean-covariance structures in generalized estimating equations with dropouts

Jianxin Pan*, Tapio Nummi and Kun Liu

Within the framework of joint mean-covariance models, we study the effects of dropout missing at random (MAR) on the estimation of mean and covariance structures for longitudinal data using generalized estimating equations (GEE). It is evidential that the MAR dropout has more severe influences on the estimation of variance-covariances relative to the mean estimation, as the former involves the estimation of the second moments. We propose to use the inverse probability weighted generalized estimating equation (WGEE) method to model the mean and covariance structures, simultaneously, in order to accommodate the effects of MAR dropout. The proposed WGEE approach produces unbiased estimators of parameters in both the mean and covariances for longitudinal data with MAR dropout. Simulation studies are conducted to assess the performance of the proposed approach and a real data analysis for the PANSS data [7] is made to illustrate the effectiveness of the proposed method.

Keywords and phrases: Dropout, Joint mean and covariance model, Longitudinal data, Missing at random, Weighted generalized estimating equation.

1. INTRODUCTION

Longitudinal studies frequently involve dropouts in response data. A subject is called dropout when the response variable is not observed through a certain visit and is then missing for all the subsequent visits [6]. Different factors may have impact on the accessibility of the observations of response at a certain occasion, such as happenstance, an adverse event, lack of efficacy on the drug, etc. Problems arise if the mechanism of the dropout is not independent of the observations of response that are either observed or not observed. It is well known that statistical inference based only on complete cases can lead to a very biased result. The generalized estimating equations (GEEs) approach of Zeger and Liang [22] extends the generalized linear model and yields more efficient and unbiased regression parameter estimators relative to ordinary least squares regression. Under a missing completely at random (MCAR) mechanism [11], consistent and unbiased estimators of parameters in the mean are produced by GEEs. However, when the dropout is missing at random (MAR) or missing not at random (MNAR), GEEs may produce very biased results and lead to inefficient estimation [1]. Robins and Rotnitzky [17] and Robins, Rotnitzky and Zhao [18] proposed a class of weighted generalized estimating equations (WGEEs) to handle longitudinal or clustered data with MAR, and their approach provides consistent estimators of the mean parameters.

The MAR issue in the GEE framework has been studied for a long time period. However, the literature research has primarily focused on the mean estimation and little research was done to study the effects of MAR dropout on the variance or covariance estimation. It is anticipated that such an influence may be more severe relative to the mean estimation as it involves the estimation of the second moments. Any small deviation of the model may lead to a substantial impact on the estimation of variance or covariance components. Recently there is an increasing number of research work in the area of joint mean and covariance models for longitudinal data, including Pourahmadi [14, 15], Ye and Pan [21] and Leng, Zhang and Pan [10], among others. However, most existing work does not readily accommodate the effects of MAR dropout. In this article, we concentrate on studying the effects of MAR dropout on variance-covariance parameter estimators, and aim to develop a new approach which is able to eliminate the bias, not only in the mean estimation, but also the variance-covariance estimation due to MAR dropout. The proposed method introduces three estimating equations that accommodate the effects of MAR dropout on the mean, generalized autoregressive parameters and log-innovation variances, simultaneously. The key idea is to use the inverse probability weighted GEE approach [18] to estimate parameters, not only in the mean, but also the variance-covariances. It is shown that the proposed method can significantly improve the mean and covariance estimation in contrast to the GEE method of Ye and Pan [21] which does not adjust the effects of MAR dropout.

The rest of this paper is organized as follows. In Section 2 we first introduce the notation, and describe some concepts of joint mean and covariance models and the reparameterization of the covariance matrix through a modified Cholesky decomposition. In Section 3 we briefly review the conventional joint mean-covariance model introduced by Pourahmadi [14] and the GEE approach for joint mean-covariance models by Ye and Pan [21]. We then provide an empirical

*Corresponding author.

analysis based on Kenward's [8] cattle data to illustrate the impact of MAR dropout on Ye and Pan's [21] GEE method. In Section 4 we propose a joint mean and covariance model using inverse probability weighted GEE and provide the details of the estimation method. Section 5 assesses the performance of the methodology by simulation studies. In Section 6 we conduct a real data analysis for the PANSS data [7] to illustrate the proposed method. Section 7 provides further discussions of the proposed method.

2. NOTATION AND MODIFIED CHOLESKY DECOMPOSITION

Let y_{ij} be the jth of m_i measurements on the ith of n subjects. Assume t_{ij} is the time at which the measurement y_{ij} is made. Denoted by $Y_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})'$ and $t_i = (t_{i1}, t_{i2}, \ldots, t_{im_i})'$, the $(m_i \times 1)$ vectors of responses and times of the ith subject. It is assumed that

$$E(Y_i) = \mu_i \quad \text{and} \quad \operatorname{Var}(Y_i) = \Sigma_i$$

where $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im_i})'$ is the $(m_i \times 1)$ mean vector and Σ_i is the $(m_i \times m_i)$ positive definite covariance matrix of the responses Y_i.

The expectation μ_i may depend on some covariates X_i $(m_i \times p)$ of interest, for example, $g(\mu_i) = X_i\beta$ where β is a $(p \times 1)$ vector of parameters and $g(\cdot)$ is the link function. It is usually assumed that the covariance matrix Σ_i has the form $\Sigma_i = \Sigma_i(t_i, \theta)$ where θ is a low-dimensional parameter vector characterizing the dependence on t_i. The 'best' covariance structure may be selected from a class of candidate structures containing, for example, compound symmetry and AR(1), using certain information criteria such as AIC or BIC [13]. However, a potential problem is that it may select a wrong covariance structure, even if the class of candidate structures is broad. For example, the true covariance structure may not be included in the class for various reasons. A typical example is that the true covariances may depend on covariates of interest in addition to times. As a result, misspecification of the covariance structure occurs which in turn leads to inefficient estimation of the mean. In some circumstances, for example, when missing data (MAR) are present, it may severely bias the estimators of regression coefficients [4].

Using the modified Cholesky decomposition (e.g., Pourahmadi [14, 15]), for any i $(1 \le i \le n)$, Σ_i can be diagonalized by a unique lower triangular matrix T_i with 1's as diagonal elements, i.e.,

$$T_i\Sigma_i T_i' = D_i, \tag{1}$$

where D_i is a unique diagonal matrix with positive diagonal elements. The elements of T_i and D_i have a very nice statistical interpretation in terms of least squares regressions. In fact, the lower-diagonal entries of $T_i = (-\phi_{ijk})$ are the negatives of the regression coefficients of $\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik})$, the linear least squares predictor of y_{ij} based on its predecessors $y_{i1}, \ldots, y_{i,j-1}$, and the diagonal entries of $D_i = \operatorname{diag}(\sigma_{i1}^2, \ldots, \sigma_{im_i}^2)$ are the prediction error variances $\sigma_{ij}^2 = \operatorname{Var}(y_{ij} - \widehat{y}_{ij})$ for $1 \le i \le n, 1 \le j \le m_i$. The new parameters ϕ_{ijk}'s and σ_{ij}^2's are called generalized autoregressive parameters (GARP) and innovation variances (IV), respectively. Thus the decomposition (1) converts the constrained entries of $\{\Sigma_i : i = 1, \ldots, n\}$, due to the positive definiteness constraints, into two sets of unconstrained 'regression' and 'variance' parameters given by $\{\phi_{ijk} : i = 1, \ldots, n; j = 2, \ldots, m_i; k = 1, \ldots, j-1\}$ and $\{\log\sigma_{i1}^2, \ldots, \log\sigma_{im_i}^2 : i = 1, \ldots, n\}$.

3. JOINT MEAN-COVARIANCE MODEL

The unconstrained parameters μ_{ij}, ϕ_{ijk} and $\log\sigma_{ij}^2$ can be modelled using linear regression models:

$$g(\mu_{ij}) = x_{ij}'\beta, \qquad \phi_{ijk} = z_{ijk}'\gamma, \qquad \log\sigma_{ij}^2 = h_{ij}'\lambda,$$

where β, γ and λ are p-, q- and d-dimensional vectors of parameters associated with the covariates x_{ij}, z_{ijk} and h_{ij}, respectively, and $g(\cdot)$ is the link function. When modeling stationary growth curve data using polynomials in time, for example, the covariates may take the form:

$$\begin{aligned} x_{ij} &= \left(1, t_{ij}, t_{ij}^2, \ldots, t_{ij}^{p-1}\right)', \\ z_{ijk} &= \left(1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, \ldots, (t_{ij} - t_{ik})^{q-1}\right)', \\ h_{ij} &= \left(1, t_{ij}, t_{ij}^2, \ldots, t_{ij}^{d-1}\right)'. \end{aligned}$$

These models were considered by Pourahmadi [14, 15] and Pan and Mackenzie [13] for balanced and unbalanced longitudinal data, respectively. Note that the resulting estimators $\widehat{\Sigma}_i$ are guaranteed to be positive definite, see [14] for details.

Under the normality assumption, Pourahmadi [14, 15] and Pan and Mackenzie [13] obtained the maximum likelihood estimators of the parameters β, γ and λ. Without any distributional assumptions, Ye and Pan [21] proposed three generalized estimating equations

$$\begin{aligned} S_1(\beta) &= \sum_{i=1}^{n} \left[\frac{\partial\mu_i'}{\partial\beta}\right] \Sigma_i^{-1}(Y_i - \mu_i), \\ S_2(\gamma) &= \sum_{i=1}^{n} \left[\frac{\partial\hat{r}_i'}{\partial\gamma}\right] D_i^{-1}(r_i - \hat{r}_i), \\ S_3(\lambda) &= \sum_{i=1}^{n} \left[\frac{\partial\sigma_i^{2\prime}}{\partial\lambda}\right] W_i^{-1}(\varepsilon_i^2 - \sigma_i^2) \end{aligned}$$

to estimate β, γ and λ. In the 2nd equation above, r_i and $\hat{r}_i$ are $(m_i \times 1)$ vectors with the jth components $r_{ij} = y_{ij} - \mu_{ij}$ and $\hat{r}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$ $(j = 1, \ldots, m_i)$, respectively, and $D_i = \operatorname{Var}(r_i - \hat{r}_i) = \operatorname{diag}(\sigma_{i1}^2, \ldots, \sigma_{im_i}^2)$. In the 3rd equation above, ε_i^2 and σ_i^2 are $(m_i \times 1)$ vectors with the jth components ε_{ij}^2 and σ_{ij}^2 $(j = 1, \ldots, m_i)$, respectively, where

$\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$. Furthermore, $E(\varepsilon_i^2) = \sigma_i^2$ and $\mathrm{Var}(\varepsilon_i^2) \equiv W_i$. When Y_i's are normally distributed, it can be shown that $W_i = 2\ \mathrm{diag}(\sigma_{i1}^4, \sigma_{i2}^4, \ldots, \sigma_{im_i}^4)$. In this case, these equations reduce to the estimating equations of Pourahmadi [14] for balanced data and Pan and Mackenzie [13] for unbalanced data. In general, the variance matrix W_i of ε_i^2 is no longer diagonal and remains unknown. Ye and Pan [21] proposed to use a sandwich 'working' covariance structure to approximate W_i, i.e., $W_i = A_i^{1/2} R_i(\rho) A_i^{1/2}$ where $A_i = 2\ \mathrm{diag}(\sigma_{i1}^4, \ldots, \sigma_{im_i}^4)$ and $R_i(\rho)$ mimics the correlation between ε_{ij}^2 and ε_{ik}^2 $(i \neq k)$ by introducing a new parameter ρ. Typical examples for $R_i(\rho)$ include compound symmetry and AR(1). It was shown that the parameter ρ has little impact on the estimators of γ and λ [21].

Note that the method of Ye and Pan [21] is very different from the GEE2, a second-order extension of generalized estimating equations proposed by Zhao and Prentice [23] and Prentice and Zhao [16]. In the approach of Ye and Pan [21], the estimating equations $S_2(\gamma)$ and $S_3(\lambda)$ avoid the use of the cross-product terms $(y_{ij} - \mu_{ij})(y_{ij'} - \mu_{ij'})$ that are used by the GEE2, due to the appealing property of the modified Cholesky decomposition in terms of the GARP and IV. More importantly, GEE2 can yield substantial bias in the estimators of parameters in the mean and covariances if assumptions about second moments are misspecified [16, 23]. In contrast, the approach of Ye and Pan [21] produces consistent estimators of parameters in both the mean and covariances. See Ye and Pan [21] for more details.

Missing data problems arise very often in longitudinal studies. A particular type of missingness that is quite common in longitudinal studies is dropout. To assess the effects of MAR dropout on the GEE method of Ye and Pan [21], we conduct an empirical study using Kenward's cattle data. Kenward [8] analyzed an experiment in which cattle were assigned randomly to two treatment groups A and B, and their weights were recorded over time. Thirty animals received treatment A and another thirty received treatment B. The animals were weighed 11 times over a 133-day period at 0, 14, 28, 42, 56, 70, 84, 98, 112, 126 and 133 days and thus the longitudinal data are balanced. Zimmerman and Núñez Antón [24] rejected the equality of the two within treatment-group covariance matrices using the classical likelihood ratio test. Thus, it is advisable to study each treatment group's covariance matrix separately.

To observe the impact of MAR dropout on the joint mean covariance model, we conduct a data analysis based on group B of Kenward's cattle data. We set up a MAR dropout process as follows. 100% of cattle have their first four repeated measurements of weight. And then we assume that there is a certain chance that one particular cow can quit the study (dropout) after the fourth measurement is taken, for example, if the weight at the fourth measurement time is below a certain threshold. The threshold value of weights is chosen such that a fixed rate of MAR dropout is achieved. Table 1 below gives such threshold values of weights and the associated subject dropout rates ranging from 10% to 90%. We aim to study how the MAR dropout affects the estimators of the mean, GARP and IV within the framework of GEE approach by Ye and Pan [21], where each parameter is modeled by a cubic polynomial in time and $R_i(\rho)$ in W_i is set as an AR(1) structure with $\rho = 0.2$.

Table 1. Dropout details

Dropout rate (in %)	Threshold weights	Dropout cattle	Total data left
10	210	3	309
20	220	6	288
30	230	9	267
40	240	12	246
50	250	15	225
60	260	18	204
70	270	21	183
80	280	24	162
90	290	27	141

In Figure 1, we display the estimated polynomials in lag for the GARP (Panel (a)) and in time for log-IV (Panel (b)) for the cattle data, by varying the rate of dropout from 10% to 90%. To save space, we choose not to display the estimated curve for the mean here. In the meantime, it is already well known that the GEE-based mean estimation is affected by the MAR dropout. For the GARP, the GEE approach for the data with a small rate of dropout may give an estimation with mild bias. However, the bias increases with the dropout rate and the estimated curve with a relatively large rate of dropout does not follow the same pattern as the one without dropout. For the log-IV, a small rate of dropout may lead to a substantial bias of the estimation. The estimated curve for the log-IV does not display the same pattern as that based on the complete data, in particular, when the rate of dropout is relatively large. It indicates that the MAR dropout has a substantial impact on the estimation of the innovation variance.

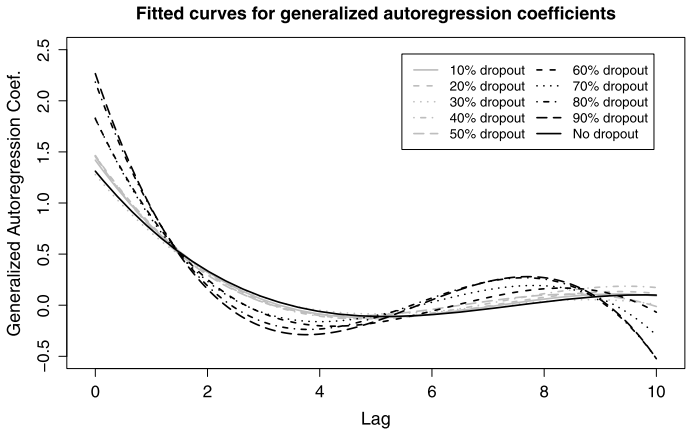
4. WEIGHTED GENERALIZED ESTIMATING EQUATIONS

4.1 Dropout model estimation

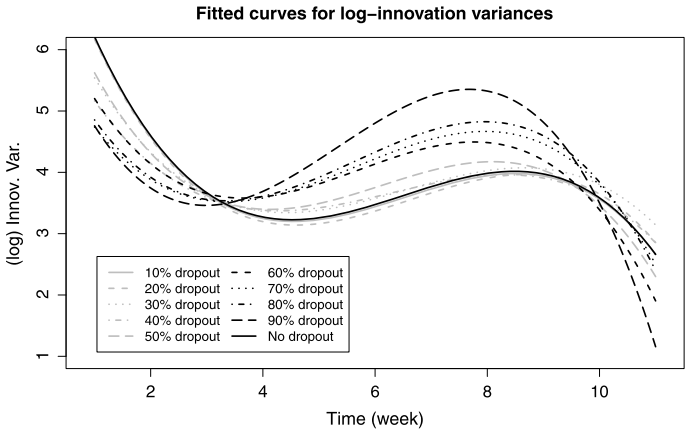
To introduce a model for dropout, we define a missing value indicator R_{ij} as

$$R_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Dropout gives rise to a monotone missing data pattern in the sense that if y_{ij} is missing, then $y_{i(j+1)}, \ldots, y_{im_i}$ are also missing. Equivalently, when expressed in terms of the missing value indicator, dropout refers to the case where if $R_{ij} = 0$ then $R_{i(j+1)} = \cdots = R_{im_i} = 0$. Let $p_{ij}(\alpha) = Pr[R_{ij} = 0 | R_{i(j-1)} = 1, y_{i1}, \ldots, y_{i(j-1)}, X_i; \alpha]$ denote the ith subject's probability of dropout at occasion j, given the history of all



(a) GEE fitted values for generalized autoregressive coefficients.



(b) GEE fitted values for log-innovation variances.

Figure 1. Panels (a) and (b) are the GEE estimated curves of the GARP and log-IV for the cattle data with different rates of dropout.

available data of response observed up to occasion $(j-1)$, where X_i and α are vectors of covariates and regression coefficients, respectively. We usually assume that all subjects are observed at their first occasion, in other words, $R_{i1}=1$, which in turn implies $p_{i1}(\alpha)=0$. In general $p_{ij}(\alpha)$ is not known but can be estimated from observed data by fitting a logistic regression model $\text{logit}\{p_{ij}(\alpha)\}=Z_{ij}'\alpha$. Equivalently,

$$p_{ij}(\alpha)=\frac{\exp(Z_{ij}'\alpha)}{1+\exp(Z_{ij}'\alpha)} \tag{3}$$

where Z_{ij} is a vector of covariates which may contain X_i and the observed responses before the dropout. The log partial likelihood function has the form

$$\sum_{i=1}^{n}\sum_{j=2}^{m_i} R_{i(j-1)}\log\big\{p_{ij}(\alpha)^{R_{ij}}[1-p_{ij}(\alpha)]^{1-R_{ij}}\big\}. \tag{4}$$

Differentiating (4) with respect to α leads to the estimating equations

$$S_0(\alpha)=\sum_{i=1}^{n}\sum_{j=2}^{m_i} R_{i(j-1)}[R_{ij}-p_{ij}(\alpha)]Z_{ij}. \tag{5}$$

Setting (5) equal to zero yields an estimator $\hat\alpha$. We therefore obtain the estimator $p_{ij}(\hat\alpha)$ of $p_{ij}(\alpha)$. The asymptotic variance of $n^{1/2}(\hat\alpha-\alpha)$ is given by $[\text{Var}\{S_0\}]^{-1}$. More details may refer to Robins, Rotnitzky and Zhao [18]. Under the MAR dropout assumption, the probability of remaining in the study at occasion j can be calculated through

$$\begin{aligned}\pi_{ij}(\alpha)&=Pr[R_{ij}=1|R_{i(j-1)}=1,y_{i1},\ldots,y_{i(j-1)},X_i,\alpha]\\&=\prod_{j=1}^{m_i}\{1-p_{ij}(\alpha)\}.\end{aligned}$$

Accordingly, the corresponding estimator of $\pi_{ij}(\alpha)$ can be obtained by $\pi_{ij}(\hat\alpha)=\prod_{j=1}^{m_i}\{1-p_{ij}(\hat\alpha)\}$.

4.2 Weighted generalized estimating equations

We apply the inverse probability weights to the above three GEEs in order to correct the bias caused by MAR dropout. The idea behind this is that, if the observation y_{ij} is observed with the probability π_{ij}, then this observation should be given a certain of weight w_{ij} in the GEE in order to reduce the bias due to MAR dropout. The weight w_{ij} for the ith subject at occasion j can be assigned as the inverse of the cumulative product of the fitted probabilities, i.e.,

$$w_{ij}(\alpha)=(\pi_{i1}(\alpha)\times\pi_{i2}(\alpha)\times\cdots\times\pi_{ij}(\alpha))^{-1}$$

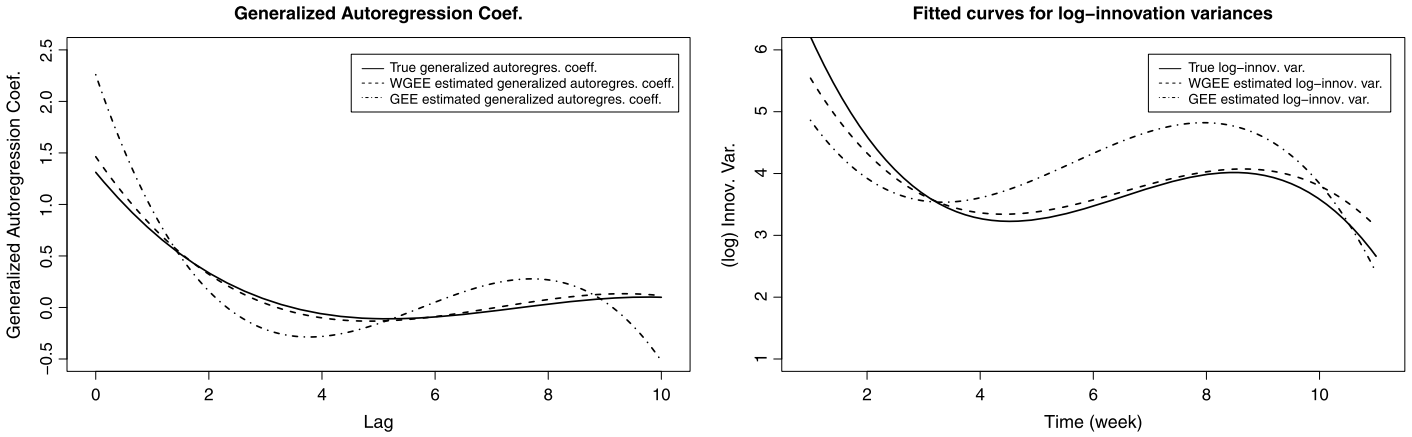
for $i=1,2,\ldots,n$ and $j=2,\ldots,m_i$. Note that we assume $w_{i1}=1$. It is clear that $w_{ij}(\alpha)$ can be estimated by $\hat w_{ij}=w_{ij}(\hat\alpha)$. Let $W_i^*=\text{diag}(R_{i1}\hat w_{i1},R_{i2}\hat w_{i2},\ldots,R_{im_i}\hat w_{im_i})$. We then propose the following inverse probability weighted generalized estimating equations

$$\begin{aligned}S_1^*(\beta)&=\sum_{i=1}^{n}\left(\frac{\partial\mu_i'}{\partial\beta}\right)\Sigma_i^*(Y_i-\mu_i),\\S_2^*(\gamma)&=\sum_{i=1}^{n}\left(\frac{\partial\hat r_i'}{\partial\gamma}\right)D_i^*(r_i-\hat r_i),\\S_3^*(\lambda)&=\sum_{i=1}^{n}\left(\frac{\partial\sigma_i^{2\prime}}{\partial\lambda}\right)\widetilde W_i^*(\varepsilon_i^2-\sigma_i^2),\end{aligned}$$

to estimate the parameters β, γ and λ in the mean, GARP and log-IV, where $\Sigma_i^*=W_i^{*1/2}\Sigma_i^{-1}W_i^{*1/2}$, $D_i^*=W_i^{*1/2}D_i^{-1}W_i^{*1/2}$ and $\widetilde W_i^*=W_i^{*1/2}W_i^{-1}W_i^{*1/2}$.

Solving the equations

$$S_1^*(\beta)=0,\quad S_2^*(\gamma)=0,\quad\text{and}\quad S_3^*(\lambda)=0$$



(a) WGEE fitted values for generalized autoregressive coefficients.

(b) WGEE fitted values for log-innovation variances.

Figure 2. Panels (a) and (b) are the estimation for the GARP and log-IV for the simulated data sets by WGEE (dashed curve) and GEE (dot-dashed curve). The solid curve is the true curve.

gives the estimators of $\theta = (\beta', \gamma', \lambda')'$. In the spirit of Ye and Pan [21], it can be shown that the quasi-Fisher information matrix of θ is of block diagonal, so that the parameter estimators must be of the forms

$$\hat{\beta} = \left[\sum_{i=1}^{n}\left(\frac{\partial \mu_i'}{\partial \beta}\right)\Sigma_i^*\left(\frac{\partial \mu_i'}{\partial \beta}\right)'\right]^{-1}\left[\sum_{i=1}^{n}\left(\frac{\partial \mu_i'}{\partial \beta}\right)\Sigma_i^*\tilde{y}_i\right],$$

$$\hat{\gamma} = \left[\sum_{i=1}^{n} E\left[\left(\frac{\partial \hat{r}_i'}{\partial \gamma}\right)D_i^*\left(\frac{\partial \hat{r}_i'}{\partial \gamma}\right)'\right]\right]^{-1}\left[\sum_{i=1}^{n}\left(\frac{\partial \hat{r}_i'}{\partial \gamma}\right)D_i^* r_i\right],$$

$$\hat{\lambda} = \left[\sum_{i=1}^{n}\left(\frac{\partial \sigma_i^{2'}}{\partial \lambda}\right)\widetilde{W}_i^*\left(\frac{\partial \sigma_i^{2'}}{\partial \lambda}\right)'\right]^{-1}\left[\sum_{i=1}^{n}\left(\frac{\partial \sigma_i^{2'}}{\partial \lambda}\right)\widetilde{W}_i^*\tilde{\varepsilon}_i^2\right]$$

where $\tilde{y}_i = (y_i - \mu_i) + (\partial \mu_i'/\partial \beta)\beta$, $\tilde{\varepsilon}_i^2 = (\varepsilon_i^2 - \sigma_i^2) + D_i \log \sigma_i^2$ and $\log \sigma_i^2 = (\log \sigma_{i1}^2, \ldots, \log \sigma_{im_i}^2)$. If $g(\cdot)$ is the identity link, then we have $\tilde{y}_i \equiv y_i$.

The consistency and asymptotic normality of the generalized estimating equations estimators $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$ can be proved in a similar manner to Ye and Pan [21], and Robins, Rotnitzky and Zhao [18]. It is noted that in order to ensure the consistency and asymptotic normality, the regularity conditions provided by Ye and Pan [21], such as unbiasedness of the estimating equations and existence of the solution to the estimating equations, are required. In addition, the WGEE approach also needs to assume that the dropout process is a MAR mechanism and the dropout probability $p_{ij}(\alpha)$ in (3) satisfies $p_{ij}(\alpha) > c > 0$ for all α and some $c > 0$. Further regularity conditions on the dropout model can be found in Robins, Rotnitzky and Zhao [18].

5. SIMULATION STUDIES

To assess the performance of the proposed approach, we conduct simulation studies. The simulation setup we adopt has the same design protocol as that used in the real cattle data analysis. We generate $N = 100$ random samples from a Normal distribution, where each sample has 30 cattle and each subject has 11 repeated measures of bodyweight taken at the same observation times as these in the real data. The Normal distribution we used is $N(\mu_i, \Sigma_i)$ $(i = 1, \ldots, 30)$, where μ_i and Σ_i are formed using the parameter estimators obtained by Ye and Pan [21] for the real data, involving the use of three polynomials in time with the triple degrees $(p, q, d) = (4, 4, 4)$. We also use an AR(1) covariance structure for $R_i(\rho)$ in W_i with $\rho = 0.2$. We assume all the cattle have complete observations for their first four measurements of weight, so that the dropout can only happen from the fifth repeated measurement. It implies that if one has the fifth measurement observed then there is no dropout anymore. Note that each simulated data set with dropout uses the missing value indicators $R_{i1}, \ldots, R_{im_i}$, where we assume $R_{i1} = R_{i2} = R_{i3} = R_{i4} = 1$ and set $R_{i,4+k} = 0$ $(k > 1)$ if $R_{i5} = 0$, so that an intermittent missing data pattern does not occur. The MAR dropout model we used is of the form

$$\text{logit}(p_{ij}) = \alpha_0 + \alpha_1 y_{i,j-1}$$

for $i = 1, \ldots, 30$ and $j = 5, \ldots, 11$, where $\alpha = (\alpha_0, \alpha_1) = (-11, 0.05)$. The values of α_0 and α_1 are chosen such that the maximum probability of dropout at the 5th repeated measurement, among the 100 simulated sets, is about 50%. The actual rate of subject dropout for the 100 simulated data ranges from 30% to 80% with a mean 55.90%.

We now use the proposed WGEE method with the estimated dropout probability $\hat{p}_{ij} = p_{ij}(\hat{\alpha})$ to analyze each simulated data set, where $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1)'$. We compare the proposed WGEE method to the ordinary GEE approach by Ye and Pan [21]. In Figure 2 (a) and (b) we report the average of the trajectory estimation for the GARP and log-IV over the 100 simulations, represented by the dashed curve. We also present the ordinary GEE estimation (the dot-dashed

curve) along with the true structure (the solid black curve). From Figure 2, it is clear that the ordinary GEE estimation is substantially biased due to the lack of consideration of the effects of the dropout. The impacts of dropout on the estimation of the GARP and log-IV are even more substantial than those on the estimation of the mean. The participation of the fourth moments in the parameter estimation leads to more challenges than the standard GEE estimation for the mean [21], which explains the large discrepancy from the true curve for both the GARP and log-IV estimation. In contrast, the proposed WGEE method can successfully alleviate and even eliminate the bias of estimation in both the GARP and log-IV. The estimation for the mean has a similar explanation and has been discussed in the vast amount of literature, and hence is not presented here.

We note that the average of the estimated regression coefficients in the dropout model, over the 100 simulations, is given by $(\hat{\alpha}_0, \hat{\alpha}_1)' = (-11.068, 0.047)'$, and the associated simulated standard errors are 0.984 and 0.007, respectively.

6. ANALYSIS OF PANSS SCHIZOPHRENIA DATA

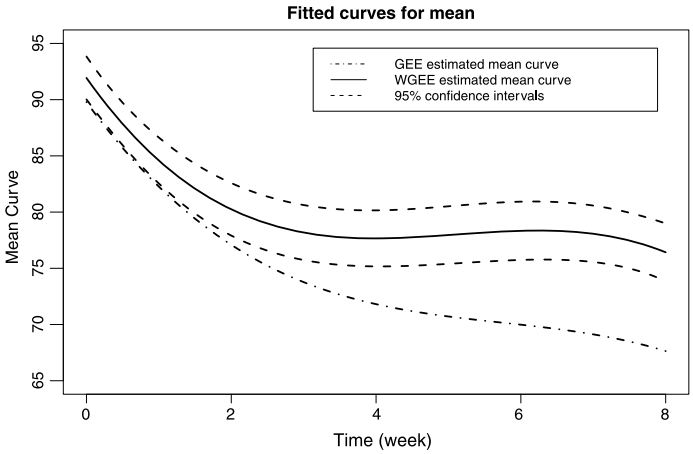
The Positive and Negative Syndrome Scale (PANSS) is a measurement for schizophrenia patients with a range from 30 to 210. PANSS reflects the severity of someone's condition. The larger PANSS is, the poorer the mental status the patient is in. The PANSS schizophrenia data was collected in the process of a clinical trial and was first studied by Kay, Flszbein and Opfer [7] and Chouinard et al. [3]. The aim of the trial was to compare different treatments for schizophrenia. 517 schizophrenia patients were randomly assigned to six different treatment groups: placebo, haloperidol and four different dose levels of risperidone. PANSS of patients were measured at −1, 0, 1, 2, 4, 6 and 8 weeks from the start of treatment. A particular patient was chosen to enter the trail if their PANSS at week −1 (before they enter the randomized trail) exceeded the level of 90. Of the 517 participants, 248 did not complete the trial because of dropout. Although risperidone was dosed at four different levels, Diggle [5] combined the data for all different dose levels because dosage was seen to have little effect on the response. The resulting numbers of the patients who were randomized to placebo, haloperidol and risperidone became 85, 87 and 345 respectively.

In our analysis, we are interested in the performance of the proposed model in dealing with dropout. Risperidone treatment data at week 0, 1, 2, 4, 6 and 8 are analyzed for this purpose since this arm contains the most patient information and also because risperidone treatment is the novel treatment in the trial. There are 345 patients in the risperidone treatment arm contributing 1,696 PANSS measurements.

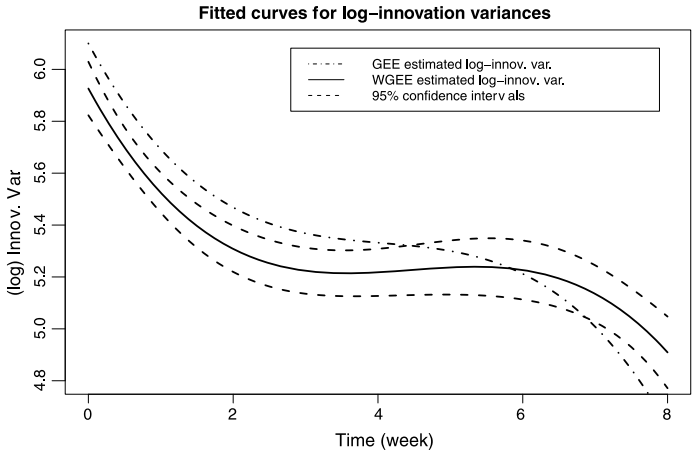
Table 2 summarizes the details of the dropout information. Kurland and Heagerty [9] checked the missing data mechanism and confirmed that it is a MAR dropout process. In fact, Kurland and Heagerty [9] used the logistic model $\text{logit}(p_{ij}) = \alpha_0 + \alpha_1 y_{i,j-1} + \alpha_2 y_{i,j-2}$ to model the dropout probability. In our analysis of the WGEE, we also use the same MAR dropout probability model as the one by Kurland and Heagerty [9]. Based on the regressogram of [14], we propose to model the mean, GARP and log-IV using three cubic polynomials in time. In other words, the covariates x_{ij}, z_{ijk}, h_{ij} take the same forms as those in Section 3 with $p = q = d = 4$. We also set an AR(1) as the correlation

Table 2. The details of the dropout information

Treatment	Number of non-dropouts at week					
	0	1	2	4	6	8
Risperidone	345	340	307	276	229	199



(a) WGEE fitted values for mean.



(b) WGEE fitted values for log-innovation variances.

Figure 3. Estimation for the mean and log-IV for the PANSS schizophrenia data using the ordinary GEE and WGEE methods.

structure of $R_i(\rho)$ in the working covariance matrices W_i. Since ρ has little impact on the estimators of β, γ and λ [21], we choose $\rho = 0.5$, which is helpful for producing reasonable confidence interval estimators for the parameters γ and λ.

In Figure 3, we display the estimated curves for the mean and log-IV for the PANSS schizophrenia data, by the ordinary GEE (the dot-dashed curve) and the proposed WGEE (the solid black curve) methods. This time we choose to present the estimated mean curve so that the effects of MAR dropout on the mean estimation can be observed. In fact, the mean estimation by the WGEE approach, see Figure 3(a), is very similar to that found by Diggle [5] and Kurland and Heagerty [9] where the within-subject covariance structure was pre-specified. It is clear that the ordinary GEE method by Ye and Pan [21] provides a quite low level of PANSS measures during the observational time period. In contrast, the WGEE approach leads to a clearly high level of PANSS measures, which drops sharply in the first two weeks and then varies almost constantly after the second week. In Figure 3(b), the estimated curve for the log-IV by the WGEE approach is observed to have a decreasing trend. It is clear that there is a sharp decrease from week 0 to week 2, and again from week 6 to week 8 with leveling off in between. A confidence interval for the WGEE estimated curve is also provided. The ordinary GEE method yields an estimated curve for the log-IV that differs substantially from the WGEE estimation. The estimation for the GARP has a similar explanation and is not presented here to save space.

7. CONCLUSION

The GEE estimators of the parameters in the mean and particularly in the covariances are affected substantially by dropout that is missing at random. The proposed inverse probability weighted GEE method works well to handle the MAR dropout for both the mean and covariance parameter estimation. It can help to eliminate the bias of the GEE-based estimation of the mean and covariances due to MAR dropout. Note that the proposed method requires a correctly specified dropout model, and this then substantially improves the effectiveness of the proposed method. If the dropout model is misspecified, it may not only cost the effectiveness of the method but also affect the asymptotic normality and consistency of the parameter estimators [2, 12, 19, 20]. In principle, as long as the model for missing data is correctly specified, the proposed method is applicable to the case of intermittent missing data. However, there must be additional challenges to correctly specify the missing mechanism as MAR.

ACKNOWLEDGEMENTS

Pan's research was supported by a grant from the Royal Society of the UK (International Exchange Scheme), and Liu's research was funded by a scholarship from the University of Manchester.

Received 14 December 2011

REFERENCES

[1] Afifi, A. and Elashoff, R. (1966). Missing observations in multivariate statistics: I. Review of the literature. *Journal of the American Statistical Association* **61** 595–604. MR0203865

[2] Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973. MR2216189

[3] Chouinard, G., Jones, B., Remington, G., Bloom, D. et al. (1993). A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients. *Journal of Clinical Psychopharmacology* **13** 25–40.

[4] Daniels, M. J. and Zhao, Y. D. (2003). Modelling the random effects covariance matrix in longitudinal data. *Statistics in medicine* **22** 1631–1647.

[5] Diggle, P. (1998). Dealing with missing values in longitudinal studies. *Recent Advances in the Statistical Analysis of Medical Data* 203–228.

[6] Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press, USA. MR2049007

[7] Kay, S. R., Flszbein, A. and Opfer, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin* **13** 261.

[8] Kenward, M. (1987). A method for comparing profiles of repeated measurements. *Applied Statistics-Journal of the Royal Statistical Society Series C* **36** 296–308.

[9] Kurland, B. F. and Heagerty, P. J. (2004). Marginalized transition models for longitudinal binary data with ignorable and nonignorable drop-out. *Statistics in medicine* **23** 2673–2695.

[10] Leng, C., Zhang, W. and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *Journal of the American Statistical Association* **105** 181–193. MR2656048

[11] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data (Second Edition)*. New York: Wiley. MR1925014

[12] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics* **4** 2111–2245. MR1315971

[13] Pan, J. and Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika* **90** 239–244. MR1966564

[14] Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

[15] Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87** 425–435. MR1782488

[16] Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47** 825–839. MR1141951

[17] Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 122–129. MR1325119

[18] 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* 106–121. MR1325118

[19] Seaman, S. and Copas, A. (2009). Doubly robust generalized estimating equations for longitudinal data. *Statistics in medicine* **28** 937–955. MR2518358

[20] Vansteelandt, S., Rotnitzky, A. and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94** 841–860. MR2416795

[21] Ye, H. and Pan, J. (2006). Modelling of covariance structures in generalised estimating equations for longitudinal data. *Biometrika* **93** 927–941. MR2285080

[22] Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.
[23] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77** 642–648. MR1087856
[24] Zimmerman, D. L. and Núñez Antón, V. (1997). Structured antedependence models for longitudinal data. *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions* **122** 63–76.

Jianxin Pan
School of Mathematics
The University of Manchester
M13 9PL
UK
E-mail address: jianxin.pan@manchester.ac.uk

Tapio Nummi
Tampere School of Public Health
FI-33014 University of Tampere
Finland
E-mail address: tapio.nummi@uta.fi

Kun Liu
School of Mathematics
The University of Manchester
M13 9PL
UK
E-mail address: kun.liu@manchester.ac.uk