Biometrics 73, 1132–1139
December 2017

DOI: 10.1111/biom.12703

Doubly Robust Estimation of Generalized Partial Linear Models for Longitudinal Data with Dropouts

Huiming Lin,[1,2] Bo Fu,[3,4] Guoyou Qin,[1,2] and Zhongyi Zhu[5]

[1]Department of Biostatistics, School of Public Health and Key Laboratory of Public Health Safety, Fudan University, Shanghai 200032, China
[2]Collaborative Innovation Center of Social Risks Governance in Health, Fudan University, Shanghai 200032, China
[3]School of Data Science, Fudan University, Shanghai 200433, China
[4]Centre for Biostatistics & Arthritis Research UK Centre for Epidemiology, The University of Manchester, Manchester, U.K.
[5]Department of Statistics, Fudan University, Shanghai 200433, China
email: gyqin@fudan.edu.cn

Summary. We develop a doubly robust estimation of generalized partial linear models for longitudinal data with dropouts. Our method extends the highly efficient aggregate unbiased estimating function approach proposed in Qu et al. (2010) to a doubly robust one in the sense that under missing at random (MAR), our estimator is consistent when either the linear conditional mean condition is satisfied or a model for the dropout process is correctly specified. We begin with a generalized linear model for the marginal mean, and then move forward to a generalized partial linear model, allowing for nonparametric covariate effect by using the regression spline smoothing approximation. We establish the asymptotic theory for the proposed method and use simulation studies to compare its finite sample performance with that of Qu's method, the complete-case generalized estimating equation (GEE) and the inverse-probability weighted GEE. The proposed method is finally illustrated using data from a longitudinal cohort study.

Key words: Doubly robust; Dropouts; Generalized partial linear models; Missing at random.

1. Introduction

Missing data induced by dropouts are common in longitudinal studies due to discontinued participation or other loss to follow-up. In our motivating example in Section 6, we use data from a longitudinal cohort study of rheumatoid arthritis patients (Symmons et al., 1994) to investigate how the repeatedly measured outcome, the Health Assessment Questionnaire (HAQ) score, is associated with baseline covariates and disease duration. At baseline, 994 subjects were included in the study sample. During the following five scheduled visits, dropouts occurred due to subject withdrawals or loss to follow-up, resulting in reduced numbers of reported HAQ measurements over follow-up. At the end of 5 years, there remained only 672 subjects. It is well recognized that in general the GEE approach is in its basic form valid only under missing completely at random (MCAR), which means that missingness is independent of both observed and unobserved data. To relax this strong assumption, advanced statistical methods concerning the modifications of GEE have been developed to deal with missing data under less restrictive assumptions such as missing at random (MAR), which means that the missingness does not depend on unobserved data (Rubin, 1976). Among them, Robins et al. (1995) proposed an inverse probability weighted (IPW) generalized estimating equations approach for repeated outcomes in the presence of missing response data and further discussed augmentation terms in their Section 6 to construct more efficient estimators; Rotnitzky et al. (1998) proposed augmented inverse probability weighted (AIPW) estimators in missing data models to improve estimation efficiency over the IPW methods; Paik (1997) proposed the mean imputation and the multiple imputation methods for longitudinal data with dropout.

Some extensions of the AIPW method were also developed including doubly robust (DR) estimators (Bang and Robins, 2005; Seaman and Copas, 2009), and their properties have been well studied (Tan, 2010; Rotnitzky et al., 2012). The implementation of DR estimation usually requires estimating the conditional mean of the outcome given both observed responses and covariates in the augmentation term. One popular way to do this is to postulate a parametric model for its conditional distribution and the conditional mean can be then obtained analytically (Carpenter et al., 2006; Shardell et al., 2014). Tan (2010) reviewed several DR estimators involving parametric modeling of the conditional mean, and considered DR estimation based on the restricted nonparametric likelihood. All of the aforementioned methods require either parametric or nonparametric modeling of the conditional mean or the conditional distribution. However, it may be difficult in practice to decide which variables should be included in the conditional model and to decide their proper functional forms (e.g., linear or quadratic) for parametric

modeling. For nonparametric modeling, technical challenges such as curse of dimensionality usually arise.

Recently, Qu et al. (2010) developed a new method to handle MAR dropouts based on the best linear approximation of the full data efficient scores. Their method does not require modeling the missing probability or imputing the missing response based on assumed models if their linear conditional mean (LCM) condition holds. As Qu et al. (2010) noted, the difference between the LCM condition and the imputation method is that while the latter just plugs in predicted values based on an assumed imputation model, the former calculates the conditional mean by using the intrinsic correlation relation between observed and missing responses. Rather than filling in the missing values and treating them as if they were observed, the method proposed in Qu et al. (2010) is closer to an expectation–maximization (EM) type of algorithms. The use of the LCM condition is particularly advantageous in certain situations where the correct specification for the imputation model is difficult. Moreover, Qu et al. (2010)'s method has wide applications as the LCM condition holds or approximately holds for a large class of response distributions such as elliptical distributions and bivariate binary distributions. However, situations do exist when the LCM is not satisfied, motivating us to extend their method to allow violation of the LCM condition.

On the other hand, we consider the generalized partial linear models (GPLM) framework for longitudinal data as GPLM is more flexible than generalized linear models (GLM) to capture nonlinear association between the response and covariates. In our motivating example, previous research indeed suggested that the relationship between the mean HAQ score and the disease duration is nonlinear. The nonparametric function can be approximated by a linear combination of regression splines that is then included as a part of the covariate vector to obtain statistical inference as in GLM; see, for example, He et al. (2005). However, so far there is only limited work on GPLM for incomplete longitudinal data. Among them, Chen and Zhou (2013) incorporated population-level information through an empirical likelihood-based method and approximated the nonparametric part using local linear method; Qin et al. (2015) considered robust estimation in the presence of outliers and used regression splines to estimate the nonparametric function.

In this article, we aim to develop DR estimator for GPLM in the analysis of longitudinal data with monotone missing responses due to dropout. The proposed estimator is doubly robust in that it is consistent when either Qu et al. (2010)'s LCM condition holds or a propensity score model for the dropout process is correctly specified. The basic idea in constructing DR AIPW estimators is similar to many existing methods in the literature, such as Tsiatis et al. (2011), in the sense that the augmentation term is based on the conditional expectation of residuals given both covariates and the observed response history. Tsiatis et al. (2011) constructed a special doubly robust estimating equation that is optimal when missing data probability is modeled correctly and the conditional model in the augmented term may be incorrect. They suggested a possible way (i.e., specifying a model for the part of the joint distribution of the full data) to obtain the conditional expectation involved in their augmentation term. Specifically, in their numerical studies, the calculation of the conditional expectation is actually based on the assumption of multivariate normal distribution. Our method, on the other hand, focuses on enhancing robustness of Qu et al. (2010)'s method against the violation of the LCM condition and gives technical details in the calculation of the augmentation term based on the LCM condition. Moreover, to the best of our knowledge, there is no existing research considering doubly robust estimators for GPLM with incomplete longitudinal data.

The article is organized as follows. Section 2 introduces the models and proposes a doubly robust estimating equation approach. We start with the GLM for the marginal mean of response and then extend it to the GPLM framework. Section 3 shows the asymptotic properties of the proposed estimators. Simulation studies and a sensitivity analysis are presented in Section 4 and 5 to evaluate the performance of the proposed method. Real data analysis is given in Section 6 and we end with concluding remarks in Section 7.

2. Model

2.1. *Generalized Linear Model and Dropout Model*

Consider a longitudinal study consisting of n subjects with m observations over time for each subject. Let Y_{ij} be the response for the ith subject at the jth observation, X_{ij} be the p-dimensional covariate vector. For simplicity, let $Y_i = (Y_{i1}, \ldots, Y_{im})^T$, $X_i = (X_{i1}, \ldots, X_{im})^T$, and Σ_i denote the covariance matrix of Y_i. Let $E(Y_{ij}|X_i) = \mu_{ij}$, $var(Y_{ij}) = \phi v(\mu_{ij})$, where ϕ is a scale parameter and $v(\cdot)$ is a known variance function. First, we introduce the following generalized linear marginal mean model

$$E(Y_{ij}|X_i) = f(X_{ij}^T\beta_0), \quad i = 1, \ldots, n, j = 1, \ldots, m, \qquad (1)$$

where β_0 is a p-dimensional vector of regression parameters and $f(\cdot)$ is the inverse of the link function between the response and covariates. We assume $E(Y_{ij}|X_i) = E(Y_{ij}|X_{ij})$, which is a necessary condition for the GEE estimator to be consistent in the context of longitudinal data (Pepe and Anderson, 1994).

Next, we describe the model for the dropout process. Let the missing indicator R_{ij} be 1 if Y_{ij} is observed, and 0 otherwise. Without loss of generality, we assume $R_{i1} = 1$ for each subject. Consider the MAR mechanism for the dropout process. Here, MAR means that for given covariates, the conditional distribution of the missing data indicator $R_i = (R_{i1}, \ldots, R_{im})^T$, $f_{R_i}(r_i|X_i, Y_i)$, only depends on X_i and the observed component of Y_i denoted by Y_i^o. Dropout means the missing process is monotone, that is, $R_{ij} = 0$ implies $R_{ik} = 0$ for all $k \geq j$.

Let $p_{ij} = P(R_{ij} = 1|R_{i,j-1} = 1, X_i, Y_i)$, $j \geq 2$, and $\omega_{ij} = P(R_{ij} = 1|X_i, Y_i)$. It is obvious that $\omega_{ij} = \prod_{k=2}^{j} p_{ik}$. We denote the response history up to (but not including) time point j by $H_{ij}^y = \{Y_{i1}, \ldots, Y_{i,j-1}\}$. A common logistic

regression model for the dropout process is

$$ln\frac{p_{ij}}{1-p_{ij}} = Z_{ij}^T\,\gamma_0, \tag{2}$$

where Z_{ij} is the vector consisting of the covariates X_i and the observed response history H_{ij}^y; and γ_0 is the regression parameter. Let Q_i denote the random dropout time for subject i, and q_i be its observed value for $i = 1, \ldots, n$. Define $L_i(\gamma) = (1-p_{iq_i})\prod_{k=2}^{q_i-1} p_{ik}$, if $q_i \le m$; otherwise, $L_i(\gamma) = \prod_{k=2}^{m} p_{ik}$, where p_{ik} is determined by model (2). Then the estimator $\hat{\gamma}$ of γ_0 can be obtained by solving

$$G_\gamma(\gamma) = \sum_{i=1}^{n} G_{\gamma,i}(\gamma) = 0, \tag{3}$$

where $G_{\gamma,i}(\gamma) = \frac{\partial log L_i(\gamma)}{\partial \gamma}$.

2.2. *The Proposed Method under GLM*

Let $Y_i = (Y_i^{o^T}, Y_i^{m^T})^T$ be a decomposition of the response vector into observed, Y_i^o, and missing, Y_i^m, variables. Qu et al. (2010) proposed the following quasilikelihood equations under MAR,

$$\sum_{i=1}^{n} \dot{\mu}_i \Sigma_i^{-1} E(Y_i - \mu_i | Y_i^o, X_i) = 0, \tag{4}$$

where $\mu_i = f(X_i\beta)$ is the mean of Y_i, and $\dot{\mu}_i = \partial\mu_i/\partial\beta$. They considered the LCM condition, which means that $E(Y_i^m|Y_i^o)$ is a linear function of Y_i^o. Denote $\Sigma_i = \begin{pmatrix} \Sigma_i^{11} & \Sigma_i^{12} \\ \Sigma_i^{21} & \Sigma_i^{22} \end{pmatrix}$, where Σ_i^{11}, Σ_i^{22} are the variance of Y_i^o and Y_i^m, and $\Sigma_i^{12} = (\Sigma_i^{21})^T$ is the covariance of these two. Let $(Y_i - \mu_i) = (Y_i^{o^T} - \mu_i^{o^T}, Y_i^{m^T} - \mu_i^{m^T})^T$ be a decomposition of the residuals into observed and unobserved components. Under the LCM condition, the following equation can be obtained

$$E(Y_i^m - \mu_i^m | Y_i^o, X_i) = \Sigma_i^{21}(\Sigma_i^{11})^{-1}(Y_i^o - \mu_i^o). \tag{5}$$

Note that under the LCM condition, it is unnecessary to specify the conditional distribution of Y_i^m given Y_i^o and X_i, which may be difficult to be correctly specified or to be calculated, particularly in the context of longitudinal data. In practice, Σ_i needs to be estimated by using observed data. Qu et al. (2010) suggested specifying the covariance matrices using regression parameters β and additional covariance parameters ρ that are common to all Σ_i. They proposed an unbiased complete data score for ρ in the GEE model as $e_i(\rho) = \sum_{a\le b} \frac{\partial}{\partial\rho}\sigma_i^{ab}(\rho)\left[\sigma_i^{ab}(\rho) - (Y_{ia} - \mu_{ia})(Y_{ib} - \mu_{ib})\right]$, where σ_i^{ab} is the abth element of Σ_i. They also noted that if one adds an additional constant conditional variance (CCV) assumption, which means that the conditional covariance of any two unobserved observations y_{ia} and y_{ib} given an observed y_{id} is constant over the choice of y_{id}, the projection of this estimating function onto the observed data can be calculated directly. In other words, if both the LCM and CCV conditions hold,

$$E\{(Y_{ia} - \mu_{ia})(Y_{ib} - \mu_{ib})|Y_i^o, X_i, \Sigma_i\}$$
$$= \begin{cases} (Y_{ia} - \mu_{ia})(Y_{ib} - \mu_{ib}) & \text{if both } Y_{ia} \text{ and } Y_{ib} \text{ are observed,} \\ (Y_{ia}^* - \mu_{ia})(Y_{ib} - \mu_{ib}) & \text{if } Y_{ia} \text{ is missing and } Y_{ib} \text{ is observed,} \\ (Y_{ia} - \mu_{ia})(Y_{ib}^* - \mu_{ib}) & \text{if } Y_{ia} \text{ is observed and } Y_{ib} \text{ is missing,} \\ \left[\Sigma_i^{*22}\right]_{ab} + (Y_{ia}^* - \mu_{ia})(Y_{ib}^* - \mu_{ib}) & \text{if both } Y_{ia} \text{ and } Y_{ib} \text{ are missing} \end{cases} \tag{6}$$

where Y_{ia}^*, Y_{ib}^* are predicted responses based on the LCM, and $\Sigma_i^{*22} = \Sigma_i^{22} - \Sigma_i^{21}(\Sigma_i^{11})^{-1}\Sigma_i^{12}$.

To widen the application, we extend Qu et al. (2010)'s method to be DR by introducing a propensity score model. Let $\tilde{Y}_i^j = (Y_{i1}, \ldots, Y_{i,j-1})^T$ and $\tilde{\mu}_i^j$ be its expectation, for $j = 2, \ldots, m$. Denote $\tilde{E}(Y_i - \mu_i) = (E^*(Y_i^o - \mu_i^o)^T, E_{Y_i^m|X_i,Y_i^o}(Y_i^m - \mu_i^m)^T)^T$, where $E^*(Y_i^o - \mu_i^o) = (Y_{i1} - \mu_{i1}, E_{Y_{i2}|X_i,\tilde{Y}_i^2}(Y_{i2} - \mu_{i2}), \ldots, E_{Y_{iv_i}|X_i,\tilde{Y}_i^{v_i}}(Y_{iv_i} - \mu_{iv_i}))^T$, and v_i is the number of observed values in Y_i. We propose doubly robust estimating equations as follows

$$\sum_{i=1}^{n}\{\dot{\mu}_i\Sigma_i^{-1}W_i(Y_i - \mu_i) + \dot{\mu}_i\Sigma_i^{-1}(I_i - W_i)\tilde{E}(Y_i - \mu_i)\} = 0, \tag{7}$$

where W_i is the diagonal matrix with elements R_{ij}/ω_{ij}, and I_i is the identity matrix of the same dimension as W_i. It is clear that $\tilde{E}(Y_i - \mu_i)$ is in fact the conditional expectation of residuals given both covariates and the observed response history up to, but not including, the current time point.

Denote $cov\begin{pmatrix} \tilde{Y}_i^j - \tilde{\mu}_i^j \\ Y_{ij} - \mu_{ij} \end{pmatrix} = \tilde{\Sigma}_i^j = \begin{pmatrix} \tilde{\Sigma}_i^{j11} & \tilde{\Sigma}_i^{j12} \\ \tilde{\Sigma}_i^{j21} & \tilde{\Sigma}_i^{j22} \end{pmatrix}$, where $\tilde{\Sigma}_i^{j11}$, $\tilde{\Sigma}_i^{j22}$ are the variance of $\tilde{Y}_i^j$ and Y_{ij}, and $\tilde{\Sigma}_i^{j12}$ (which is also the transpose of $\tilde{\Sigma}_i^{j21}$) is the covariance of these two. Under the LCM condition, we have $E_{Y_{ij}|X_i,\tilde{Y}_i^j}(Y_{ij} - \mu_{ij}) = \tilde{\Sigma}_i^{j21}(\tilde{\Sigma}_i^{j11})^{-1}(\tilde{Y}_i^j - \tilde{\mu}_i^j)$. The double robustness of the proposed estimator is shown in the Web Appendix.

Remark 1: If we simply use equation (5) to construct doubly robust estimating equations, that is, $\sum_{i=1}^{n}\{\dot{\mu}_i\Sigma_i^{-1}W_i(Y_i - \mu_i) + \dot{\mu}_i\Sigma_i^{-1}(I_i - W_i)E(Y_i - \mu_i|Y_i^o, X_i)\} = 0$, the resulting estimating equations would reduce to equation (4) and the double robustness property can not be achieved.

2.3. *Extension to Generalized Partial Linear Model*

We now extend the proposed DR estimator for GLM to the GPLM framework, allowing nonlinearity in the marginal mean model. We consider a case similar to that in Section 2.1 and assume that both covariates X_{ij} and T_{ij} are always observed. For simplicity, we denote $T_i = (T_{i1}, \ldots, T_{im})^T$. We add a nonparametric covariate effect to the original mean model (1)

$$E(Y_{ij}|X_i, T_i) = f(X_{ij}^T\beta_0 + g_0(T_{ij})), i = 1, \ldots, n, j = 1, \ldots, m, \tag{8}$$

where $g_0(\cdot)$ is an unknown smooth function. Now our goal is to estimate both β_0 and $g_0(\cdot)$.

Assume that the domain of T_{ij} is confined to the interval $[0, 1]$. Let $0 = s_0 < s_1, \ldots, s_{k_n} < s_{k_n+1} = 1$ be a partition of $[0, 1]$. Taking $\{s_i\}$ as knots, we can get $N_k = k_n + l$ normalized B-spline basis functions of order l, denoted by $\{B_1(t), \ldots, B_{N_k}(t)\}$. Then $g_0(t)$ can be approximated by $\pi(t)^T\alpha_0$, where $\pi(t) = (B_1(t), \ldots, B_{N_k}(t))^T$ and $\alpha_0 \in R^{N_k}$ is the vector of spline coefficients. This linearizes the regression model (8) and we have

$$\eta_{ij}(\theta_0) = f^{-1}(\mu_{ij}(\theta_0)) = X_{ij}^T\beta_0 + \pi_{ij}^T\alpha_0 = D_{ij}^T\theta_0,$$

where $D_{ij} = (X_{ij}^T, \pi_{ij}^T)^T$, $\pi_{ij} = \pi(T_{ij})$, and $\theta_0 = (\beta_0^T, \alpha_0^T)^T$ is the combined regression parameters. Following He et al. (2005), we use cubic splines of order 4 and calculate the sample quantiles of $\{T_{ij}\}$ as the knots. The number of the internal knots k_n is taken to be the integer part of $F_n^{1/5}$, where F_n is the number of distinct values of $\{T_{ij}\}$. This choice is consistent with the asymptotic results given in Section 3. Now the nonparametric function is linearized so that any algorithm designed for the linear models can be directly applied to the partial linear models.

The construction of the dropout model is similar: $ln\frac{p_{ij}}{1-p_{ij}} = \tilde{Z}_{ij}^T\gamma_0$, where $\tilde{Z}_{ij}$ is the vector consisting of the information of X_i, T_i, and the observed response history H_{ij}^y. Using the likelihood method, the estimator $\hat{\gamma}$ of γ_0 can be obtained by solving equation (3).

Following the same idea in constructing doubly robust estimating equations under GLM, we propose doubly robust estimating equations for GPLM as follows:

$$\sum_{i=1}^{n} D_i\Delta_i^T(\mu_i(\theta))\Sigma_i^{-1}h_i(\mu_i(\theta), \gamma) = 0,$$

where $D_i = (D_{i1}^T, \ldots, D_{im}^T)$, $\Delta_i(\mu_i(\theta)) = diag\{\dot{\mu}_{i1}(\theta), \ldots, \dot{\mu}_{im}(\theta)\}$, $\dot{\mu}$ denotes the first derivative of $f(\cdot)$ evaluated at $D_i\theta$, $h_i(\mu_i(\theta), \gamma) = W_i(Y_i - \mu_i) + (I_i - W_i)\tilde{E}(Y_i - \mu_i)$. The arguments for the double robustness are similar to those in Section 2.2 and are omitted here. The estimating equations can be solved through Newton–Raphson iterative algorithm. Denote the final estimator by $\hat{\theta} = (\hat{\beta}^T, \hat{\alpha}^T)^T$. Then the estimated regression coefficients and nonparametric function are $\hat{\beta}$ and $\hat{g}(t) = \pi^T(t)\hat{\alpha}$ respectively.

3. Asymptotic Properties

We first introduce some notations and then establish the asymptotic properties of the proposed estimator. Let $\Sigma_i = \Sigma_i(\mu_i(\theta))$, $\Delta_i = \Delta_i(\mu_i(\theta))$, and $h_i = h_i(\mu_i(\theta), \gamma)$. Let $M_i = (\pi_{i1}, \ldots, \pi_{im})^T$ and $M = (M_1^T, \ldots, M_n^T)^T$, $\Omega_i = \Delta_i^T\Sigma_i^{-1}E\{\frac{\partial}{\partial\mu_i}h_i\}\Delta_i$ and $\Omega = diag\{\Omega_1, \ldots, \Omega_n\}$. Let $P = M(M^T\Omega M)^{-1}M^T\Omega$, $X = (X_1^T, \ldots, X_n^T)^T$, and $X^* = (X_1^{*T}, \ldots, X_n^{*T})^T = (I - P)X$, where I is the identity matrix. Let $B_i = X_i^{*T}\Delta_i^T\Sigma_i^{-1}h_i - [\sum_{i=1}^n X_i^{*T}\Delta_i^T\Sigma_i^{-1}\frac{\partial}{\partial\gamma}h_i(\mu_i, \gamma)] \cdot [\frac{\partial}{\partial\gamma}G_\gamma(\gamma)]^{-1}G_{\gamma,i}(\gamma)$. We denote the value of Ω_i evaluated at the true μ_i and γ_0 by $\Omega_{0,i}$ and use notations $B_{0,i}$ and $X_{0,i}^*$ in a similar fashion.

Under the regularity conditions (C.1)–(C.9) given in the Web Appendix, the asymptotic properties of the proposed estimators can be established. Specifically, Theorem 1 shows the asymptotic normality of the proposed estimator for regression coefficients $\hat{\beta}$ and shows that the proposed estimator for the nonparametric function can achieve the optimal rate of convergence under the smoothing condition (C.4).

THEOREM 1. *Assume that conditions (C.1)–(C.9) hold. If the number of knots $k_n \approx n^{1/(2r+1)}$ where r is defined in (C.4) in the Appendix, then*

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(t_i) - g_0(t_i))^2 = O_p(n^{-\frac{2r}{2r+1}}),$$

$$\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, K^{-1}BK^{-1}),$$

where K and B are some positive definite matrix such that $\frac{1}{n}K_n \to K$, $\frac{1}{n}B_n \to B$ in probability, with $K_n = \sum_{i=1}^n X_{0,i}^{*T}\Omega_{0,i}X_{0,i}^*$ and $B_n = \sum_{i=1}^n B_{0,i}B_{0,i}^T$.

The proof of Theorem 1 is given in the Web Appendix. Following Theorem 2 in Qin and Zhu (2007), the covariance matrix can be consistently estimated by $\hat{K}^{-1}\hat{B}\hat{K}^{-1}$, where $\hat{K} = \frac{1}{n}\sum_{i=1}^n X_i^{*T}\Omega_i X_i^*$, $\hat{B} = \frac{1}{n}\sum_{i=1}^n B_i B_i^T$ with all the involved quantities evaluated at $\hat{\theta}$ and $\hat{\gamma}$.

4. Simulation Study

We carried out a simulation study to compare the performances of the proposed method with those of the complete-case GEE, the IPW GEE (Robins et al., 1995) and Qu's method (Qu et al., 2010). The covariance or correlation matrix involved in the proposed method was estimated using the same technique as mentioned in Section 2.2. We calculated the bias, the standard error (SE), and the mean square error (MSE) for $\hat{\beta}$, as well as the integrated mean square error (IMSE) for $\hat{g}(\cdot)$ through Monte Carlo simulations.

We considered a true partial linear model

$$\mu_{ij} = X_{ij}^T\beta_0 + 0.5\exp(0.1T_{ij}), i = 1, \ldots, n, j = 1, \ldots, m,$$

where $\beta_0 = 1$. The covariates were generated as follows: $X_{ij} = u_{ij} + b_{1,ij}$, $T_{ij} = u_{ij} + b_{2,ij}$ with u_{ij}, $b_{1,ij}$, and $b_{2,ij}$ being independently drawn from a uniform distribution on $(-0.5, 0.5)$. The random error $e_i = (e_{i1}, \ldots, e_{im})^T$ was generated from a multivariate normal distribution with mean zero and covariance matrix $R_i(\rho)\sigma^2$, where $R_i(\rho)$ is the correlation matrix chosen to be first order autoregressive (AR1) with $\rho = 0.6$ and $\sigma^2 = 1$. The sample size is $n = 600$ with $m = 6$.

The missing indicators R_{ij} were generated from the true dropout model

$$ln\frac{p_{ij}}{1 - p_{ij}} = \gamma_0 + \gamma_1 Y_{ij-1} + \gamma_2 X_{ij}, \quad (9)$$

where $(\gamma_0, \gamma_1, \gamma_2)^T$ is taken to be $(2.6, 1.0, -1.0)^T$, indicating that 15% of subjects have missing data.

We considered the following five scenarios:

S1: there is no misspecification.

S2: the LCM is satisfied but the dropout model is misspecified by excluding X_{ij} when estimating propensity scores.

S3: the dropout model is correct but the LCM is violated with the random errors e_i generated from a log-normal distribution.

S4: the LCM is violated as in S3 and the dropout model is misspecified as in S2.

S5: the dropout process is missing not at random (MNAR), that is, $ln\frac{p_{ij}}{1-p_{ij}} = \gamma_0 + \gamma_1 Y_{i,j-1} + \gamma_2 X_{ij} + \gamma_3 Y_{ij}$ with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)^T$ taken to be $(2.6, 1.0, -1.0, 0.5)^T$. The working dropout model for estimation is model (9).

More specifically, we considered the violation of LCM condition as follows. We first generated $e_i^{'}$ such that $\log\left(e_i^{'}\right) = \left(\log(e_{i1}^{'}), \ldots, \log(e_{im}^{'})\right)^T \sim N\left(0_{m\times 1}, R_i(\rho)\sigma^2\right)$, where $R_i(\rho)$ is AR1 with $\rho = 0.6$ and $\sigma^2 = 1.3$, and then obtained e_i by centering $e_i^{'}$ to be zero mean.

Table 1 summarizes the simulation results based on 500 replications for continuous outcomes. It shows that the proposed method enjoys a DR property. When the LCM condition holds, Qu's method and the proposed method have almost equally good performances, except that the latter has a slightly larger standard error. This may be due to the fact that the proposed approach involves estimating the weight, which will induce more variability. In the case that the LCM is violated but the dropout model is correct, the proposed method gives much smaller bias, standard error and MSE for $\hat{\beta}$ and much smaller IMSE for $\hat{g}(\cdot)$ than Qu's method. Interestingly, when the LCM condition is violated and the dropout model is misspecified (S4), the DR estimator has the best performance among all the estimators in terms of the bias, the standard error, the MSE, and the IMSE. It is also seen that the complete-case GEE, which completely ignores the impact of missing data, gives the largest biases, MSEs and IMSEs in all scenarios. Besides, the estimated standard errors for $\hat{\beta}$ based on 500 replications are close to those obtained from asymptotic approximations, which indicates that the large-sample estimate of the variance is satisfactory. Simulation results of S5 shows that when MAR is violated, estimates from the four methods all deviate from the truth, as is indicated by large estimation biases and MSEs for $\hat{\beta}$ and large IMSEs for $\hat{g}(\cdot)$. We also conducted simulations for binary outcomes and the findings are similar to the continuous case. See Web Appendix for detailed descriptions.

Table 1
Simulation results for continuous outcome

	$\beta_0 = 1$				
	BIAS	ESE	SE	MSE	IMSE
S1					
GEE-C	0.0291	0.0388	0.0403	0.0025	0.0047
GEE-W	0.0010	0.0404	0.0420	0.0018	0.0027
Qu	0.0001	0.0371	0.0388	0.0015	0.0024
DR-PLM	0.0007	0.0390	0.0403	0.0016	0.0026
S2					
GEE-C	0.0291	0.0388	0.0403	0.0025	0.0047
GEE-W	0.0221	0.0404	0.0422	0.0023	0.0027
Qu	0.0001	0.0371	0.0388	0.0015	0.0024
DR-PLM	0.0007	0.0387	0.0403	0.0016	0.0026
S3					
GEE-C	0.0848	0.1486	0.1559	0.0314	0.0508
GEE-W	0.0021	0.1344	0.1395	0.0194	0.0270
Qu	0.0258	0.1369	0.1418	0.0207	0.0279
DR-PLM	0.0044	0.1221	0.1253	0.0157	0.0248
S4					
GEE-C	0.0848	0.1486	0.1559	0.0314	0.0508
GEE-W	0.0629	0.1355	0.1409	0.0238	0.0272
Qu	0.0258	0.1369	0.1418	0.0207	0.0279
DR-PLM	0.0238	0.1229	0.1263	0.0165	0.0248
S5					
GEE-C	0.0306	0.0387	0.0392	0.0025	0.0081
GEE-W	0.0098	0.0406	0.0410	0.0018	0.0038
Qu	0.0103	0.0369	0.0379	0.0015	0.0031
DR-PLM	0.0117	0.0392	0.0400	0.0017	0.0034

Note: SE, standard error; ESE, estimated standard error from asymptotic theory; MSE, mean square error; IMSE, integrated MSE; GEE-C, complete-case GEE; GEE-W, IPW GEE; Qu, Qu's method; DR-PLM, the proposed method.

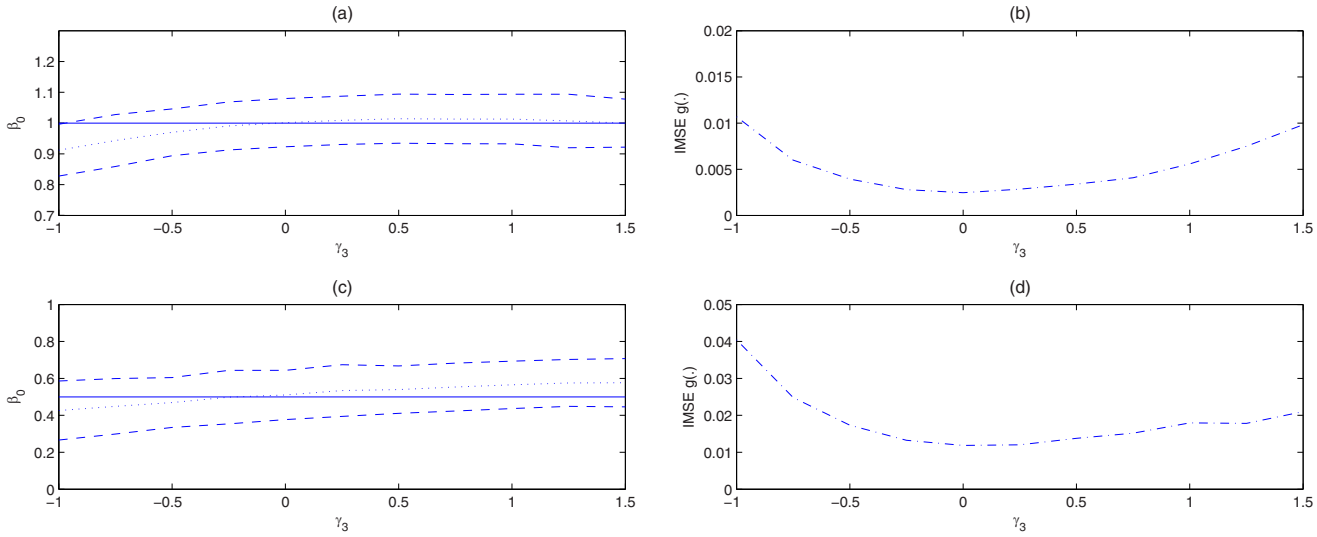

Figure 1. Sensitivity analysis of the proposed method. Under MNAR, (a), (c) display the average estimate of β_0 (dotted curves) and its 95% CI as γ_3 changes, for continuous and discrete outcome, respectively. The true value of β_0 is indicated by the solid lines. (b), (d) display the IMSE of $\hat{g}(\cdot)$ for continuous and discrete outcome, respectively.

5. Sensitivity Analysis

Exploring sensitivity to unverifiable missing data assumptions is important in the analysis of incomplete data (Daniels and Hogan, 2008). In this section, we carried out a sensitivity analysis to assess the robustness of the proposed method when the underlying MAR assumption is violated. Similar to Yi et al. (2012), we considered the missing indicator R_{ij} generated from $ln\frac{p_{ij}}{1-p_{ij}} = \gamma_0 + \gamma_1 Y_{i,j-1} + \gamma_2 X_{ij} + \gamma_3 Y_{ij}$ and examined the sensitivity of estimation $\hat{\beta}$ and $\hat{g}(\cdot)$ with the change of the sensitivity parameter γ_3. Note that $\gamma_3 = 0$ indicates a MAR scenario and $\gamma_3 \neq 0$ corresponds to MNAR scenarios. We used model (9) as our working propensity score model in the estimation.

Let γ_3 vary from $(-1, 1.5)$, representing a wide range of scenarios. The estimation results for $\hat{\beta}$ and the IMSE of $\hat{g}(\cdot)$ for continuous outcomes are displayed in Figure 1a and b, respectively. Similar analyses are applied to binary outcomes and the results are shown in Figure 1c and d. They show that as γ_3 gradually moves away from 0 (the MAR assumption is violated), the estimates tend to deviate from the true β_0 and the true $g_0(\cdot)$, as is indicated by the increasing bias of $\hat{\beta}$ and the increasing IMSE of $\hat{g}(\cdot)$. Meanwhile, it is also seen that within a proper range of γ_3, the estimates are close to the true ones, demonstrating the robustness of the proposed method.

6. Application to Real Data

We apply the proposed method to a longitudinal cohort study of rheumatoid arthritis patients (Symmons et al., 1994). The study sample includes 994 patients recruited to the cohort between 1990 and 1994 who had early rheumatoid arthritis at baseline (Norton et al., 2014). The response variable is the HAQ score, which is a widely used measure of functional disability in rheumatoid arthritis patients and ranges from 0 (best) to 3 (worst). For each subject, HAQ was repeatedly measured at baseline (year 0), year 1, year 2, year 3, year 4, and year 5. All subjects ($N = 994$) reported their HAQ at baseline, but during the follow up, dropouts occurred, resulting to a decreasing number, namely 943 after 1 year, 864 after 2 years, 828 after 3 years, 716 after 4 years, and 672 after 5 years, of reported HAQ scores.

As for the covariates, we included gender ($X1 = 1$ if female; or 0 if male), age at disease onset ($X2$), c-reaction protein ($X3$), number of swollen joints ($X4$), number of tender joints ($X5$), rheumatic factor ($X6 = 1$ if positive; or 0 if negative), anti-cyclic citrullinated peptide antibody (anti-CCP) ($X7 = 1$ if positive; or 0 if negative), and disease duration (in years) from symptom onset to current assessment (T). Among these covariates, $X1$–$X7$ are time-independent baseline measurements and T is time-varying. We investigate how the mean of the longitudinal response HAQ is associated with baseline covariates $X1, X2, \ldots, X7$ and changes with time variable T. Previous epidemiological research suggested that the mean HAQ score over time is J-shaped with an initial improvement after registration to the primary care based cohort shortly followed by starting treatment in patients with early rheumatoid arthritis (Norton et al., 2014). It motivates us to consider a partial linear model by incorporating a nonlinear relationship between the response and disease duration T,

$$E(HAQ|X1, X2, X3, X4, X5, X6, X7, T) = \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_7 X7 + g(T) \quad (10)$$

where $g(\cdot)$ is an unknown smooth function. We used a four-order regression spline with the knots-selection procedure given in Section 2.3 to approximate $g(\cdot)$. More specifically, we select the knots using the sample quantiles of T. Since T has 2938 distinct values, the number of knots is chosen to be 4, which is the integer part of $2938^{1/5}$.

Table 2
Regression coefficient estimates in analysis of rheumatoid arthritis data

	X1	X2	X3	X4	X5	X6	X7
GEE-C							
EST	0.2832	0.0113	0.0029	0.0032	0.0251	0.0096	0.1846
ESE	0.0388	0.0012	0.0008	0.0030	0.0021	0.0458	0.0504
95%CI_L	0.2071	0.0090	0.0013	−0.0027	0.0210	−0.0802	0.0858
95%CI_U	0.3593	0.0136	0.0046	0.0091	0.0292	0.0995	0.2835
GEE-W							
EST	0.2906	0.0117	0.0027	0.0030	0.0252	0.0068	0.2009
ESE	0.0416	0.0013	0.0009	0.0032	0.0022	0.0482	0.0528
95%CI_L	0.2090	0.0092	0.0010	−0.0032	0.0208	−0.0877	0.0974
95%CI_U	0.3722	0.0142	0.0044	0.0092	0.0296	0.1013	0.3045
Qu							
EST	0.2475	0.0110	0.0039	0.0035	0.0253	0.0087	0.1538
ESE	0.0370	0.0011	0.0008	0.0029	0.0020	0.0447	0.0488
95%CI_L	0.1750	0.0087	0.0024	−0.0021	0.0213	−0.0791	0.0581
95%CI_U	0.3200	0.0132	0.0055	0.0091	0.0293	0.0964	0.2495
DR-PLM							
EST	0.2488	0.0111	0.0039	0.0034	0.0254	0.0105	0.1574
ESE	0.0371	0.0011	0.0008	0.0029	0.0020	0.0450	0.0490
95%CI_L	0.1760	0.0088	0.0023	−0.0022	0.0214	−0.0776	0.0613
95%CI_U	0.3216	0.0133	0.0055	0.0090	0.0294	0.0986	0.2535

Note: EST, parameter estimation; ESE, estimated standard error; 95%CI_U, upper bound of 95% confidence interval; 95%CI_L, lower bound of 95% confidence interval.

We modeled the dropout process by $ln\frac{p_{ij}}{1-p_{ij}} = \gamma_0 + \gamma_1 Z_{ij}$, where Z_{ij} could include variables observed up to time point j. Determining which variables to include is somewhat subjective and should be in consultation with subject-matter experts. We included baseline covariates and $HAQ_{i,j-1}$ in Z_{ij} as this is the only time-varying variable measured in this study. We find that a subject with a higher HAQ score at an earlier time point or a male may be more likely to dropout, although their coefficient estimators are not statistically significant.

Table 2 presents a comparison of estimators for the regression parameters in model (10) along with their standard errors and 95% confidence intervals among the complete-case GEE, the weighted GEE, Qu's method, and the proposed method. All four methods select X1, X2, X3, X5, X7 as significant predictors for the longitudinal functional disability outcome HAQ. The proposed doubly robust method and Qu's method give similar estimates, but the complete-case GEE and the weighted GEE give different estimates for regression parameters. It implies that the latter two methods may overestimate the impact of gender ($X1$) and anti-CCP ($X7$) but underestimate the effect of c-reactive protein ($X3$). Figure 2 shows a comparison of the estimated smooth function $g(\cdot)$. We see that the estimated effects of disease duration from the four methods are all decreasing within the first 2 years from disease onset and are then increasing for a couple of years before it reaches a peak. This is clinically reasonable as patients usually get better after registering to the primary care cohort and starting with treatments but their outcomes then get worse as the disease progresses.

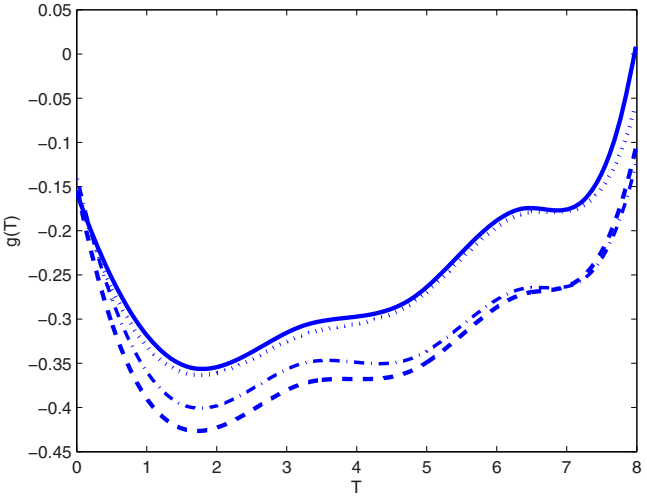

Figure 2. The estimated function on T. The dot-dashed, dashed, solid, dotted lines represent the curves estimated by GEE-C, GEE-W, Qu and the proposed DR-PLM methods respectively.

7. Conclusions

We extend Qu's method and propose a DR estimation for GPLM in the analysis of longitudinal data with monotone missing responses due to dropout. Similar to many other existing DR methods, our proposed estimator also has an AIPW estimating equation form and involves in the augmentation the conditional expectation of residuals given both

covariates and observed responses. However, our method differs from existing literature on several grounds. First, we use the LCM condition in the construction of doubly robust estimating equations. Instead of modeling the conditional distribution or plugging in predicted values based on an assumed imputation model for missing responses for the augmentation construction, the LCM condition incorporates the intrinsic correlation between observed and missing responses and enables us to simplify the procedure of constructing the augmentation term. More appealingly, the LCM condition holds or approximately holds under a large class of response distributions. Second, we consider the GPLM framework for longitudinal data and use B splines to approximate the nonlinear function. Although it brings challenges in establishing the asymptotic theory due to an infinite-dimensional problem, it is computationally easy to implement as the nonparametric part is linearized. To the best of our knowledge, so far there is no work concerning DR estimators for GPLM for longitudinal data with dropouts.

Further extensions of the proposed method may be of interest. One may consider a robust DR estimator against outliers in the analysis of longitudinal data with dropouts. Other future work may be concerned with measurement errors in the covariates, which are common in some longitudinal studies and would introduce bias to the analysis.

8. Supplementary Materials

Example data and code and Web Appendix referenced in Sections 2.2, 3, and 4, are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

We thank the editor, the associate editor, and the referee for their constructive suggestions that largely improved the presentation of this article. This work was partially supported by the National Natural Science Foundation of China (11371100, 11271080), China Medical Board Collaborating Program in Health Technology Assessment (Grant 16-251), and Shanghai Leading Academic Discipline Project, Project number: B118. We thank the PI in the Arthritis Research UK Centre for Epidemiology for allowing us to use the NOAR data. Bo Fu was partially supported by a UK MRC grant (MR/M025152/1).

References

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

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **169**, 571–584.

Chen, B. and Zhou, X. (2013). Generalized partially linear models for incomplete longitudinal data in the presence of population-level information. *Biometrics* **69**, 386–395.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton: CRC Press.

He, X., Fung, W. K., and Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association* **100**, 1176–1184.

Norton, S., Fu, B., Scott, D. L., Deighton, C., Symmons, D. P., Wailoo, A. J., et al. (2014). Health assessment questionnaire disability progression in early rheumatoid arthritis: Systematic review and analysis of two inception cohorts. *Seminars in Arthritis and Rheumatism* **44**, 131–144.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association* **92**, 1320–1329.

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

Qin, G. and Zhu, Z. (2007). Robust estimation in generalized semiparametric mixed models for longitudinal data. *Journal of Multivariate Analysis* **98**, 1658–1683.

Qin, G., Zhu, Z., and Fung, W. K. (2015). Robust estimation of generalized partially linear model for longitudinal data with dropouts. *Annals of the Institute of Statistical Mathematics* 1–24.

Qu, A., Lindsay, B. G., and Lu, L. (2010). Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random. *Journal of the American Statistical Association* **105**, 194–204.

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

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–456.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

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

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

Shardell, M., Hicks, G. E., and Ferrucci, L. (2014). Doubly robust estimation and causal inference in longitudinal studies with dropout and truncation by death. *Biostatistics* kxu032.

Symmons, D., Barrett, E., Bankhead, C., Scott, D., and Silman, A. (1994). The incidence of rheumatoid arthritis in the united kingdom: Results from the norfolk arthritis register. *Rheumatology* **33**, 735–739.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–682.

Tsiatis, A. A., Davidian, M., and Cao, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics* **67**, 536–545.

Yi, G. Y., Ma, Y., and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika* **99**, 151–165.

Received October 2015. *Revised March* 2017. *Accepted March* 2017.