Computational Statistics and Data Analysis 101 (2016) 174–185

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Using link-preserving imputation for logistic partially linear models with missing covariates

Qixuan Chen [a], Myunghee Cho Paik [b,*], Minjin Kim [b], Cuiling Wang [c]

[a] Department of Biostatistics, Columbia University, New York, NY, United States

[b] Department of Statistics, Seoul National University, Seoul, Republic of Korea

[c] Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, United States

ARTICLE INFO

Article history:
Received 9 July 2015
Received in revised form 3 March 2016
Accepted 6 March 2016
Available online 15 March 2016

Keywords:
Doubly robust estimator
Kernel-assisted estimating equation
Logistic partially linear models
Inverse probability weighting
Link-preserving imputation
Missing covariates

ABSTRACT

To handle missing data one needs to specify auxiliary models such as the probability of observation or imputation model. Doubly robust (DR) method uses both auxiliary models and produces consistent estimation when either of the model is correctly specified. While the DR method in estimating equation approaches could be easy to implement in the case of missing outcomes, it is computationally cumbersome in the case of missing covariates especially in the context of semiparametric regression models. In this paper, we propose a new kernel-assisted estimating equation method for logistic partially linear models with missing covariates. We replace the conditional expectation in the DR estimating function with an unbiased estimating function constructed using the conditional mean of the outcome given the observed data, and impute the missing covariates using the so called link-preserving imputation models to simplify the estimation. The proposed method is valid when the response model is correctly specified and is more efficient than the kernel-assisted inverse probability weighting estimator by Liang (2008). The proposed estimator is consistent and asymptotically normal. We evaluate the finite sample performance in terms of efficiency and robustness, and illustrate the application of the proposed method to the health insurance data using the 2011–2012 National Health and Nutrition Examination Survey, in which data were collected in two phases and some covariates were partially missing in the second phase.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Recently generalized partially linear models (GPLM) draw a lot of attention (Severini and Staniswalis, 1994; Carroll et al., 1997; Liang et al., 2004). The GPLMs include a nonparametric covariate effect in an otherwise generalized linear model. The logistic partially linear models (LPLM), as a special case of the GPLM for binary data, relax the structure of the mean in a logistic regression to be partially linear. Specifically, let Y be the binary outcome, $\mathbf{X}$ be parametrically modeled covariates and Z be a nonparametrically modeled covariate. The conditional mean of Y is assumed to be a twice differentiable function of linear predictor $\mathbf{X}^T\boldsymbol{\beta} + \nu(Z)$ where $\boldsymbol{\beta}$ are unknown parameters and $\nu(.)$ is a smooth unknown function of Z. In this paper, we investigate the estimation of the LPLM when Y and Z are fully observed but some of $\mathbf{X}$ are partially missing.

When there are missing data, a likelihood method can naturally handle the problem by integrating over the missing data and maximizing the integrated marginal likelihood function. However for non-likelihood methods, the same technique

* Corresponding author. Tel.: +82 2 880 6764; fax: +82 2 883 6144.
E-mail address: myungheechopaik@snu.ac.kr (M.C. Paik).

http://dx.doi.org/10.1016/j.csda.2016.03.004
0167-9473/© 2016 Elsevier B.V. All rights reserved.

cannot be used. There are two paradigms in handling missing data in estimating equation approaches to construct unbiased estimating functions, namely, imputation (e.g. Reilly and Pepe, 1995; Paik, 1997) and inverse probability weighting (IPW, e.g. Robins et al., 1994, 1995). The imputation method fills in missing statistics by its 'best' guess, the conditional expectation. The IPW weights the observed records by the inverse of the observation probability to properly represent the whole data, and has been very popular in various settings since it is easy to implement. Validity of the inference in both paradigms depends on correctness of assumptions on auxiliary models, the imputation model in the case of the imputation approach or the response model in the case of the IPW approach. The imputation method is generally more efficient than the IPW especially when there is a potent predictor for missing data (Wang and Paik, 2006). The efficiency of the IPW method can be effectively improved by subtracting projection onto the nuisance tangent space (Robins et al., 1994, 1995), but the projection term involves the conditional mean of the estimating function. The projection method requires assumptions on both auxiliary models but the inference is valid when either one of the assumptions is correct. Because of this property, this method is called doubly robust (DR) method. In the case of missing outcomes, simple implementation of the DR method is discussed in Bang and Robins (2005), Scharfstein et al. (1999), and Little and An (2004). Although the same principles apply for missing outcomes and missing covariates, the imputation method and the DR method, in the case of missing covariates, require evaluation of the conditional expectation of the product of missing covariates and the conditional mean of outcomes given observed data, which is a main hurdle for computation.

Missing data problem becomes even more computationally demanding in the context of semiparametric regression models. When outcomes are missing, Chen et al. (2006) and Wang et al. (2010) proposed weighted kernel estimating equations for the GPLMs. Wang et al. (1998) is one of the first work tackling missing covariate problem in a nonparametric regression model using the IPW approach. Liang et al. (2004) considered estimation of a partially linear model with missing covariates using the IPW-type kernel based method. Liang (2008) proposed a kernel-assisted IPW method for the GPLMs with missing covariates and derived asymptotic properties of the DR estimator, but discouraged using the DR estimator due to the complexity of implementation. Qin et al. (2012) also considered an IPW-type approach for robust GPLMs in the sense of Huber with missing covariates using a regression spline.

In this paper we propose a new kernel-assisted estimating equation approach to handle missing covariates in the context of LPLMs. The proposed method modifies the DR estimating function by replacing the conditional expectation with an unbiased estimating function constructed using the mean of the outcome conditioning on the observed covariates but marginalizing out the missing covariates. This marginal mean usually is not easy to evaluate. To overcome this, we introduce the concept of link-preserving imputation. We call imputation models link-preserving if the part of the linear predictor concerning completely observed covariates is preserved under the same link function. Under link-preserving imputation, the marginal mean can be easily obtained by replacing the missing covariate with some imputation value, which allows simple implementation of the proposed method via data augmentation. Use of the marginal mean coupled with link-preserving imputation greatly reduces the computational difficulty in solving the estimating equations for both the parametric and the nonparametric parts. The proposed estimator is more efficient than the kernel-assisted IPW estimator by Liang (2008).

The rest of the paper is organized as follows. In Section 2, we briefly describe the notation and framework. We propose new methods in Section 3. Simulation studies follow in Section 4. In Section 5, we show application to the health insurance coverage problem using the data of the 2011–2012 National Health and Nutrition Examination Survey. Concluding remarks follow in Section 6.

2. Notation and framework

Suppose that there are n independently identically distributed observations $\{(Y_i, \mathbf{X}_i^T, Z_i)^T, i = 1, \ldots, n\}$. Let Y_i denote a binary outcome variable for the ith subject, Z_i denote a single nonparametrically modeled covariate associated with the ith subject, and $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \mathbf{X}_{i2}^T)^T$ where $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ denote a vector of parametrically modeled covariates for the ith subject with p and q elements, respectively. We consider the following logistic partially linear model,

$$\text{logit}\,\{E(Y_i|\mathbf{X}_i, Z_i)\} = \log \frac{P(Y_i = 1|\mathbf{X}_i, Z_i)}{P(Y_i = 0|\mathbf{X}_i, Z_i)} = \mathbf{X}_{i1}^T\boldsymbol{\beta}_1 + \mathbf{X}_{i2}^T\boldsymbol{\beta}_2 + \nu(Z_i), \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ are unknown parameters of interest associated with parametrically modeled covariates $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$, respectively, and $\nu(.)$ is an unknown smooth function of Z_i. Suppose that $\mathbf{X}_{i2}$ and Z_i are fully observed, but $\mathbf{X}_{i1}$ are missing for some cases. We assume that all elements of $\mathbf{X}_{i1}$ are observed or missing together, with applications in such as two-phase studies where $\mathbf{X}_{1i}$ are collected only among the sub-sample of the second phase. The observation indicator for $\mathbf{X}_{i1}$ is denoted by R_i; if $R_i = 1$, $\mathbf{X}_{i1}$ are observed and, if $R_i = 0$, $\mathbf{X}_{i1}$ are missing. We assume that $\mathbf{X}_{i1}$'s are missing at random (MAR), i.e., $P(R_i = 1|Y_i, \mathbf{X}_i, Z_i) = P(R_i = 1|Y_i, \mathbf{X}_{i2}, Z_i) \equiv \pi_i(Y_i, \mathbf{X}_{i2}, Z_i; \boldsymbol{\alpha}) \equiv \pi_i(\boldsymbol{\alpha})$.

Liang (2008) proposed an IPW estimator in the context of GPLMs by solving

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i(\boldsymbol{\alpha})} \mathbf{Q}_i V_i^{-1}\{Y_i - E(Y_i|\mathbf{X}_i, Z_i)\} = 0, \tag{2}$$

where $\mathbf{Q}_i \equiv \frac{\partial}{\partial \boldsymbol{\beta}} E(Y_i|\mathbf{X}_i, Z_i)$ and $V_i \equiv \text{Var}(Y_i|\mathbf{X}_i, Z_i)$, coupled with kernel method for estimating $\nu(Z_i)$, and showed that the resulting estimator for $\boldsymbol{\beta}$ is consistent.

The IPW estimator can be improved to have a smaller variance by subtracting the projection term. The DR estimating equation has a form of

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i(\boldsymbol{\alpha})} \mathbf{Q}_i V_i^{-1}\{Y_i - E(Y_i|\mathbf{X}_i, Z_i)\} - \sum_{i=1}^{n} \frac{R_i - \pi_i(\boldsymbol{\alpha})}{\pi_i(\boldsymbol{\alpha})} E[\mathbf{Q}_i V_i^{-1}\{Y_i - E(Y_i|\mathbf{X}_i, Z_i)\}|Y_i, \mathbf{X}_{i2}, Z_i].$$

Note that the first term is the IPW estimating function, but if $\pi_i(\boldsymbol{\alpha})$ is replaced by 1, the above estimating function reduces to that of the imputation method (Reilly and Pepe, 1995; Paik, 1997). When the outcome is missing and covariates are completely observed, the second term is linear in Y, which makes implementation simple by replacing missing Y with its conditional expectation given all available information. However in the missing covariate case, the second term is not a linear function of $\mathbf{X}_1$. Under the canonical link, the conditional expectation $E[\mathbf{Q}_i V_i^{-1}\{Y_i - E(Y_i|\mathbf{X}_i, Z_i)\}|Y_i, \mathbf{X}_{i2}, Z_i]$ requires estimating $E\{\mathbf{X}_{i1}E(Y_i|\mathbf{X}_i, Z_i)|Y_i, \mathbf{X}_{i2}, Z_i\}$ as well as $E(\mathbf{X}_{i1}|Y_i, \mathbf{X}_{i2}, Z_i)$. In partially linear models, Liang et al. (2004) estimated the conditional mean of $\mathbf{X}_i\mathbf{X}_i^T$ as well as $\mathbf{X}_i$. In the GPLM, the approach of Liang et al. (2004) cannot be directly applied since it requires modeling $\mathbf{X}_{i1}E(Y_i|\mathbf{X}_i, Z_i)$ where $E(Y_i|\mathbf{X}_i, Z_i)$ is a nonlinear function of $\mathbf{X}_i$, $\boldsymbol{\beta}$ and $\nu(Z_i)$. Liang (2008) derived the asymptotic properties of the DR estimator for GPLMs with missing covariates but stated that '*we do not recommend this alternative (DR) because its implementation is complex*'.

Our goal of this paper is to propose an alternative estimator to the kernel-assisted IPW estimator by Liang (2008) for GPLMs with missing covariates that is easier to implement than the DR method but is more efficient than the IPW estimator. We focus on binary outcomes with logit link function. As in the IPW method, we assume that the response model is correctly specified. Our strategy to achieve easy implementation is by facilitating link-preserving imputation models, which are defined in Section 3.1. We first present a motivating easy-to-implement estimator in Section 3.2. In Section 3.3 we extend the proposed method so that the asymptotic variance is guaranteed to be smaller than that of the IPW estimator.

3. Method

3.1. Link-preserving imputation model

The main idea of the proposed method starts from the simple fact that we can still estimate $E(Y|\mathbf{X}_2, Z)$ even when $\mathbf{X}_1$ is missing in the context of LPLMs. Evaluating $E(Y|\mathbf{X}_2, Z)$ is not always straightforward, but under certain class of imputation models it could be manageable. We define link-preserving imputation models as follows. Let $\eta = \text{logit}\{E(Y|\mathbf{X}, Z; \boldsymbol{\beta}, \nu)\} = \mathbf{X}_1^T\boldsymbol{\beta}_1 + \mathbf{X}_2^T\boldsymbol{\beta}_2 + \nu(Z)$. We call that imputation models for $\mathbf{X}_1$ are *link-preserving* if they produce a form of $E(Y|\mathbf{X}_2, Z; \boldsymbol{\beta}, \nu, \boldsymbol{\gamma})$ such that $\eta^* = \text{logit}\{E(Y|\mathbf{X}_2, Z; \boldsymbol{\beta}, \nu, \boldsymbol{\gamma})\} = \mathbf{X}_1^{*T}\boldsymbol{\beta}_1 + \mathbf{X}_2^T\boldsymbol{\beta}_2 + \nu(Z)$, where $\mathbf{X}_1^*$ is a function of $\mathbf{X}_2$, Y, Z, and the imputation model parameters $\boldsymbol{\gamma}$. Under link-preserving imputation models, the part of the linear predictors involving the completely observed covariates, $\mathbf{X}_2^T\boldsymbol{\beta}_2 + \nu(Z)$, is preserved when the same link function, logit(.), is applied to the marginal mean, $E(Y|\mathbf{X}_2, Z; \boldsymbol{\beta}, \nu, \boldsymbol{\gamma})$.

The idea of link-preserving imputations was explored before in a parametric setup by Paik and Sacco (2000). We show below the derivation of the link-preserving imputations in the setting of LPLMs (1). Using the Bayes' rule, we have

$$\log \frac{P(Y_i = 1|\mathbf{X}_{i2}, Z_i)}{P(Y_i = 0|\mathbf{X}_{i2}, Z_i)} = \log \frac{P(Y_i = 1|\mathbf{X}_{i1}, \mathbf{X}_{i2}, Z_i)}{P(Y_i = 0|\mathbf{X}_{i1}, \mathbf{X}_{i2}, Z_i)} - \log \frac{f(\mathbf{X}_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = 1)}{f(\mathbf{X}_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = 0)}. \tag{3}$$

Let $p = 1$, we assume that $f(X_{i1}|\mathbf{X}_{i2}, Z_i, Y_i)$ belongs to an exponential family with the canonical parameter $\theta(Y, \mathbf{X}_2, Z)$:

$$f(X_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = y) = \exp\left\{\frac{\theta(y, \mathbf{X}_{i2}, Z_i)X_{i1} - b(\theta(y, \mathbf{X}_{i2}, Z_i))}{\varphi} + c(\varphi, X_{i1})\right\}, \tag{4}$$

and obtain

$$\log \frac{f(X_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = 1)}{f(X_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = 0)} = \frac{\theta(1, \mathbf{X}_{i2}, Z_i) - \theta(0, \mathbf{X}_{i2}, Z_i)}{\varphi} X_{i1} - \frac{b(\theta(1, \mathbf{X}_{i2}, Z_i)) - b(\theta(0, \mathbf{X}_{i2}, Z_i))}{\varphi}. \tag{5}$$

From (3) and (5) we can derive the following identity,

$$\begin{aligned}
\exp(\beta_1 x) &= \frac{P(Y_i = 1|X_{i1} = x, \mathbf{X}_{i2}, Z_i)P(Y_i = 0|X_{i1} = 0, \mathbf{X}_{i2}, Z_i)}{P(Y_i = 0|X_{i1} = x, \mathbf{X}_{i2}, Z_i)P(Y_i = 1|X_{i1} = 0, \mathbf{X}_{i2}, Z_i)} \\
&= \frac{f(X_{i1} = x|\mathbf{X}_{i2}, Z_i, Y_i = 1)f(X_{i1} = 0|\mathbf{X}_{i2}, Z_i, Y_i = 0)}{f(X_{i1} = x|\mathbf{X}_{i2}, Z_i, Y_i = 0)f(X_{i1} = 0|\mathbf{X}_{i2}, Z_i, Y_i = 1)} \\
&= \exp\left[(\theta(1, \mathbf{X}_{i2}, Z_i) - \theta(0, \mathbf{X}_{i2}, Z_i))\, x/\varphi\right],
\end{aligned}$$

which gives $\beta_1 = (\theta(1, \mathbf{X}_{i2}, Z_i) - \theta(0, \mathbf{X}_{i2}, Z_i))/\varphi$. Plugging this identity into (5), we obtain

$$\log \frac{f(X_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = 1)}{f(X_{i1}|\mathbf{X}_{i2}, Z_i, Y_i = 0)} = \beta_1 X_{i1} - \beta_1 \frac{b(\theta(1, \mathbf{X}_{i2}, Z_i)) - b(\theta(0, \mathbf{X}_{i2}, Z_i))}{\theta(1, \mathbf{X}_{i2}, Z_i) - \theta(0, \mathbf{X}_{i2}, Z_i)}. \tag{6}$$

Combining (1), (3) and (6) we have

$$\eta_i^* = \log \frac{P(Y_i = 1|\mathbf{X}_{i2}, Z_i)}{P(Y_i = 0|\mathbf{X}_{i2}, Z_i)} = \beta_1 X_{i1}^* + \mathbf{X}_{i2}^T \boldsymbol{\beta}_2 + \nu(Z_i),$$

where

$$X_{i1}^* = \frac{b\left(\theta(1, \mathbf{X}_{i2}, Z_i)\right) - b\left(\theta(0, \mathbf{X}_{i2}, Z_i)\right)}{\theta(1, \mathbf{X}_{i2}, Z_i) - \theta(0, \mathbf{X}_{i2}, Z_i)}.$$

As an example, when X_1 is a binary variable, according to (4) the canonical parameter $\theta(Y, \mathbf{X}_2, Z) = \text{logit}\{\Pr(X_1 = 1|Y, \mathbf{X}_2, Z)\}$ and $b(\theta(Y, \mathbf{X}_2, Z)) = \log(1 + e^{\theta(Y, \mathbf{X}_2, Z)})$. With $\text{logit}\{\Pr(X_1 = 1|Y, \mathbf{X}_2, Z)\} = \gamma_0 + \mathbf{X}_2^T \boldsymbol{\gamma}_1 + \gamma_2 Y + \gamma_3 Z$, we can obtain $X_1^* = \gamma_2^{-1}\{\log(1 + e^{\gamma_0 + \mathbf{X}_2^T \boldsymbol{\gamma}_1 + \gamma_2 + \gamma_3 Z}) - \log(1 + e^{\gamma_0 + \mathbf{X}_2^T \boldsymbol{\gamma}_1 + \gamma_3 Z})\}$. The derivation is also valid when $p > 1$. When $f(\mathbf{X}_1|Y, \mathbf{X}_2, Z)$ is multivariate normal with mean $\boldsymbol{\mu}(Y, \mathbf{X}_2, Z)$, we can obtain $\mathbf{X}_1^* = \frac{1}{2}\{\boldsymbol{\mu}(1, \mathbf{X}_2, Z) + \boldsymbol{\mu}(0, \mathbf{X}_2, Z)\}$. With the link-preserving imputation models, we can easily calculate η^* and thus $E(Y|\mathbf{X}_2, Z)$, which can then be used in the estimating equations for LPLMs in Section 3.2.

3.2. Easy-to-implement estimation procedure

As in the doubly robust method, we specify three models in the proposed estimating function: (i) the analysis model, $E(Y|\mathbf{X}, Z; \boldsymbol{\beta}, \nu)$, (ii) the response model, $\pi(\boldsymbol{\alpha})$, and (iii) the imputation model, $f(\mathbf{X}_1|Y, \mathbf{X}_2, Z; \boldsymbol{\gamma})$, which is link-preserving. Note that (i) is the analysis model, which we need to specify even if data are fully observed, and (ii) and (iii) are auxiliary models, whose specification is required to handle missing data. As for the IPW method, we assume that (i) and (ii) are correctly specified, while (iii) can be subject to model misspecification. Estimation of the two auxiliary models are not intertwined with estimation of the analysis model. Given the estimate of the auxiliary models, the proposed estimating procedure alternates estimating $\boldsymbol{\beta}$ and $\nu(Z)$. For $\boldsymbol{\beta}$, the estimating equation has a form of

$$\mathbf{0} = \mathbf{U}_\mathbf{n} = \sum_{i=1}^{n} \frac{R_i}{\pi_i(\boldsymbol{\alpha})} \mathbf{s}_A(Y_i, \mathbf{X}_i, Z_i) - \sum_{i=1}^{n} \frac{R_i - \pi_i(\boldsymbol{\alpha})}{\pi_i(\boldsymbol{\alpha})} \mathbf{s}_B(Y_i, \mathbf{X}_{i2}, Z_i), \tag{7}$$

where

$$\mathbf{s}_A(Y_i, \mathbf{X}_i, Z_i; \boldsymbol{\beta}, \nu) = (\mathbf{X}_i - \boldsymbol{\mu}_\mathbf{X}) \frac{\partial}{\partial \eta_i} E(Y_i|\mathbf{X}_i, Z_i) V_i^{-1}\{Y_i - E(Y_i|\mathbf{X}_i, Z_i)\},$$

$$\mathbf{s}_B(Y_i, \mathbf{X}_{i2}, Z_i; \boldsymbol{\beta}, \nu) = (\mathbf{X}_i^* - \boldsymbol{\mu}_\mathbf{X}) \frac{\partial}{\partial \eta_i^*} E(Y_i|\mathbf{X}_{i2}, Z_i) V_i^{*-1}\{Y_i - E(Y_i|\mathbf{X}_{i2}, Z_i)\},$$

with $\mathbf{X}_i^* = (\mathbf{X}_{i1}^{*T}, \mathbf{X}_{i2}^T)^T$, $V_i = \text{Var}(Y_i|\mathbf{X}_i, Z_i)$, and $V_i^* = \text{Var}(Y_i|\mathbf{X}_{i2}, Z_i)$. The kth element of $\boldsymbol{\mu}_\mathbf{X}$ is $\mu_{X_k} = E\{w(\mathbf{X}, Z)X_k\}/E\{w(\mathbf{X}, Z)\}$ with $w(\mathbf{X}, Z) = \left[\frac{\partial E(Y|\mathbf{X}, Z)}{\partial \eta}\right]^2 \text{Var}(Y|\mathbf{X}, Z)^{-1}$, and can be estimated as follows:

$$\hat{\mu}_{X_k} = \frac{\sum_{i=1}^{n} R_i X_{ik} \hat{w}(\mathbf{X}_i, Z_i)/\pi_i(\hat{\boldsymbol{\alpha}})}{\sum_{i=1}^{n} R_i \hat{w}(\mathbf{X}_i, Z_i)/\pi_i(\hat{\boldsymbol{\alpha}})}.$$

We can show that centering $\mathbf{X}$ around $\boldsymbol{\mu}_\mathbf{X}$ renders orthogonality between the estimators of $\boldsymbol{\beta}$ and $\nu(Z)$.

For $\nu(Z)$, we can set $\mu_i(z_0) = \text{logit}^{-1}(\mathbf{X}_i^T \boldsymbol{\beta} + \nu(z_0))$ and $\mu_i^*(z_0) = \text{logit}^{-1}(\mathbf{X}_i^{*T} \boldsymbol{\beta} + \nu(z_0))$, and solve the following Nadaraya–Watson kernel-assisted estimating equation for $\nu(z_0)$,

$$\begin{aligned} 0 = &\sum_{i=1}^{n} K_h(Z_i - z_0) \frac{R_i}{\pi_i(\boldsymbol{\alpha})} \frac{\partial \mu_i(z_0)}{\partial \eta_i} V_i(\mu_i(z_0))^{-1}(Y_i - \mu_i(z_0)) \\ &- \sum_{i=1}^{n} K_h(Z_i - z_0) \frac{R_i - \pi_i(\boldsymbol{\alpha})}{\pi_i(\boldsymbol{\alpha})} \frac{\partial \mu_i^*(z_0)}{\partial \eta_i^*} V_i^*(\mu_i^*(z_0))^{-1}(Y_i - \mu_i^*(z_0)), \end{aligned} \tag{8}$$

where $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function and h is the bandwidth. The subtracted term on the right hand side of (8) leads to an improved estimation. The advantage of the Nadaraya–Watson estimator lies in its ease in the implementation using built-in functions of the statistical software. Alternatively, we also consider local linear kernel estimator below, which requires more intensive computation. We set $\mu_i(z_0) = \text{logit}^{-1}(\mathbf{X}_i^T \boldsymbol{\beta} + c_0 + c_1(Z_i - z_0)/h)$ and $\mu_i^*(z_0) = \text{logit}^{-1}(\mathbf{X}_i^{*T} \boldsymbol{\beta} + c_0 +$

$c_1(Z_i - z_0)/h)$, and solve the following local linear weighted kernel estimating equation for $\mathbf{c} = (c_0, c_1)^T = (\nu(z_0), h\nu'(z_0))^T$,

$$\mathbf{0} = \sum_{i=1}^{n} K_h(Z_i - z_0)\frac{R_i}{\pi_i(\boldsymbol{\alpha})}\begin{bmatrix}1\\(Z_i - z_0)/h\end{bmatrix}\frac{\partial \mu_i(z_0)}{\partial \eta_i}V_i(\mu_i(z_0))^{-1}(Y_i - \mu_i(z_0))$$
$$- \sum_{i=1}^{n} K_h(Z_i - z_0)\frac{R_i - \pi_i(\boldsymbol{\alpha})}{\pi_i(\boldsymbol{\alpha})}\begin{bmatrix}1\\(Z_i - z_0)/h\end{bmatrix}\frac{\partial \mu_i^*(z_0)}{\partial \eta_i^*}V_i^*(\mu_i^*(z_0))^{-1}(Y_i - \mu_i^*(z_0)). \quad (9)$$

Bandwidth selection is required to solve (8) and (9). The optimal bandwidth that minimizes the conditional mean integrated squared error of $\nu(Z)$ is of order $n^{-1/5}$ excluding the boundary points. In the simulation study, we use a fixed bandwidth $h = n^{-1/5}$, and conduct sensitivity analyses using different bandwidths around this choice but the results are similar. This is expected because bandwidths with the same rate of convergence lead to the same limit distribution (Liang et al., 2004, 2009). In the application to the health insurance data, we use both the method of empirical bias bandwidth selection (Ruppert, 1997) and the method of $n^{-1/5}$, which yield similar bandwidth selection and similar estimates of $\nu(Z)$.

We can easily show that estimating functions in (7), (8) and (9) are unbiased when all models are correctly specified. If the imputation model (iii) is misspecified, we obtain an incorrect form of $E(Y|\mathbf{X}_2, Z)$, but unbiasedness is retained as long as (i) and (ii) are correct. On the other hand, when (ii) is misspecified but (iii) is correctly specified, the estimating equations are unbiased only when R does not depend on $(Y, \mathbf{X}_1)$ given $(\mathbf{X}_2, Z)$ or when $\boldsymbol{\beta}_1 = \mathbf{0}$. This suggests that the test statistic under the null hypothesis $\mathrm{H}_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ has correct Type I error although the estimator is not doubly robust.

Estimation of the auxiliary models can be conducted in the usual manner. To estimate $\boldsymbol{\alpha}$, we assume that $\pi(\boldsymbol{\alpha})$ is a known function indexed by unknown parameter $\boldsymbol{\alpha}$, and $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\alpha}}$ can be obtained by solving

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial \pi_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\{\pi_i(\boldsymbol{\alpha})(1 - \pi_i(\boldsymbol{\alpha}))\}^{-1}(R_i - \pi_i(\boldsymbol{\alpha})).$$

To estimate $\boldsymbol{\gamma}$, we specify link-preserving imputation model $f(\mathbf{X}_1|Y, \mathbf{X}_2, Z; \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ can be consistently estimated using completely observed data alone due to $R \perp\!\!\!\perp \mathbf{X}_1|(Y, \mathbf{X}_2, Z)$ as follows:

$$\mathbf{0} = \sum_{i=1}^{n} R_i \frac{\partial E(\mathbf{X}_{i1}|Y_i, \mathbf{X}_{i2}, Z_i)}{\partial \boldsymbol{\gamma}}\mathrm{Var}(\mathbf{X}_{i1}|Y_i, \mathbf{X}_{i2}, Z_i)^{-1}(\mathbf{X}_{i1} - E(\mathbf{X}_{i1}|Y_i, \mathbf{X}_{i2}, Z_i)).$$

By examining the estimating equations for $\boldsymbol{\beta}$, we observe that for records with $R_i = 0$, contribution to the estimating function is just $\mathbf{s}_B(Y_i, \mathbf{X}_{i2}, Z_i)$, but for records with $R_i = 1$, contribution comes from both $\mathbf{s}_A(Y_i, \mathbf{X}_i, Z_i)$ and $\mathbf{s}_B(Y_i, \mathbf{X}_{i2}, Z_i)$. This suggests that solving Eqs. (7)–(9) can be implemented via the following data expansion. We first start by estimating $\boldsymbol{\alpha}$ then $\boldsymbol{\gamma}$ and obtain $\hat{\mathbf{X}}_1^*$, and computing $\hat{\boldsymbol{\mu}}_{\mathbf{X}}$ using the records with $R_i = 1$. We then estimate $\boldsymbol{\beta}$ and $\nu(Z)$ using the following four steps:

STEP 1 : Fill in $\hat{\mathbf{X}}_1^*$ for record with $R = 0$.

STEP 2 : Expand the dataset by duplicating records with $R = 1$. In the duplicated record, replace $\mathbf{X}_1$ with $\hat{\mathbf{X}}_1^*$.

STEP 3 : Create weights: for original records with $R = 1$, weight is π^{-1}; for duplicated record with $R = 1$, weight is $1 - \pi^{-1}$; and for records with $R = 0$, weight is 1.

STEP 4 : With the expanded dataset, conduct LPLM analysis using $\mathbf{X} - \hat{\boldsymbol{\mu}}_{\mathbf{X}}$ or $\hat{\mathbf{X}}^* - \hat{\boldsymbol{\mu}}_{\mathbf{X}}$ as the parametric terms and Z as the nonparametric term.

Asymptotic properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\nu}(Z)$ are given in Section 1 of the Supplementary Materials (see Appendix A). The variance estimation of $\hat{\boldsymbol{\beta}}$ can be computed either using the asymptotic variance estimation or the jackknife estimate of variance. Both are shown in the simulation study. The method proposed in this section can be extended to the scenario with monotone missing covariates, by expanding the estimating equation (7) to allow multiple summations with one summation for each monotone missing-data pattern, in which the mean of the outcome is modeled conditioning on the observed covariates but marginalizing out the missing covariates of that missing-data pattern.

3.3. Improving efficiency under the alternative

The asymptotic variance of the proposed estimator $\hat{\boldsymbol{\beta}}$ is guaranteed to have a smaller variance than that of the IPW estimator when $\boldsymbol{\beta}_1 = 0$. However there is no guarantee that the asymptotic variance is smaller than that of the IPW in general, although it tends to be smaller in practical cases. In this section we propose an estimator of $\boldsymbol{\beta}$ guaranteed to have a smaller variance than that of the IPW estimator. This approach is described in Wang et al. (2010) and van der Laan and Robins (2003). As in Section 3.2, we first introduce the improved estimator, say, $\hat{\boldsymbol{\beta}}_{\pi,\kappa}$ as the solution of the following equation when all other quantities are known:

$$\mathbf{0} = \mathbf{U}_n^{\kappa} = \sum_{i=1}^{n} \frac{R_i}{\pi_i(\boldsymbol{\alpha})}\mathbf{s}_A(Y_i, \mathbf{X}_i, Z_i) - \boldsymbol{\kappa}\sum_{i=1}^{n} \frac{R_i - \pi_i(\boldsymbol{\alpha})}{\pi_i(\boldsymbol{\alpha})}\mathbf{s}_B(Y_i, \mathbf{X}_{i2}, Z_i), \quad (10)$$

where

$$\boldsymbol{\kappa}^T = E\left[\frac{1-\pi(\boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})}\mathbf{s}_B\mathbf{s}_B^T\right]^{-1} E\left[\frac{1-\pi(\boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})}\mathbf{s}_B E\{\mathbf{s}_A|Y, \mathbf{X}_2, Z\}^T\right].$$

The difference between (7) and (10) is that we multiply $\boldsymbol{\kappa}$ to the second term. Due to $\boldsymbol{\kappa}$, the covariance between the first and the second terms in (10) equals to the variance of the second term, which guarantees that the variance of the estimating function is smaller than that of the IPW estimator.

We can show that $\mathbf{U}_n^{\kappa} = \mathbf{U}_n^{\kappa}\{\boldsymbol{\beta}, \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{R}\}$ has the same asymptotic distribution with $\mathbf{U}_n^{\kappa}\{\boldsymbol{\beta}, \hat{\boldsymbol{\mu}}_{\mathbf{X}}, \hat{\boldsymbol{\nu}}, \boldsymbol{\alpha}, \hat{\boldsymbol{\gamma}}, \boldsymbol{\kappa}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{R}\}$. We can also show that estimating $\boldsymbol{\kappa}$ and $\boldsymbol{\alpha}$ does not affect the asymptotic distribution of $\mathbf{U}_n^{\kappa}$. Let $\hat{\boldsymbol{\beta}}_{\hat{\pi},\hat{\boldsymbol{\kappa}}}$ be the solution of $\mathbf{U}_n^{\kappa}\{\boldsymbol{\beta}, \hat{\boldsymbol{\mu}}_{\mathbf{X}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\kappa}}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{R}\} = \mathbf{0}$. Section 2 of the Supplementary Materials shows the asymptotic property of $\hat{\boldsymbol{\beta}}_{\hat{\pi},\hat{\boldsymbol{\kappa}}}$ (see Appendix A).

Computation of $\hat{\boldsymbol{\beta}}_{\hat{\pi},\hat{\boldsymbol{\kappa}}}$ can be carried out by adding an extra step from computing $\hat{\boldsymbol{\beta}}$. We can regress $\frac{R_i}{\hat{\pi}_i}\hat{\mathbf{s}}_A$ on $\frac{R_i-\hat{\pi}_i}{\hat{\pi}_i}\hat{\mathbf{s}}_B$ through the origin to obtain $\hat{\boldsymbol{\kappa}}$. Then we can solve for $\boldsymbol{\beta}_{\pi,\boldsymbol{\kappa}}$ by plugging in $\hat{\boldsymbol{\kappa}}$ and other estimates obtained in Section 3.2.

4. Simulation study

4.1. Design

We perform a simulation study to evaluate the finite sample performance of the proposed methods. We generate data from model (1) with $\nu(Z) = Z$ or $\nu(Z) = \cos(2\pi Z)$, and generate $Z \sim$ Uniform(0, 1) and $X_2 \sim$ Bernoulli(0.3). Note that $\mathbf{X}_1^*$ can be computed from X_2 and Z as described in Section 3.1. Given $\mathbf{X}_1^*$, X_2 and Z, Y is generated. We consider two scenarios:

(S1): $p = 1$; $\beta_1 = 1$; $\beta_2 = -0.7$; and $\text{logit}\{E(X_1|X_2, Y, Z)\} = -1 - 0.2X_2 + Y + 0.5Z$.
(S2): $p = 2$; $\boldsymbol{\beta}_1 = (-1, 0.2)^T$; $\beta_2 = 0.75$; and $(\mathbf{X}_1|X_2, Y, Z) \sim \text{MVN}_2(\boldsymbol{\mu}(X_2, Y, Z), \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu}(X_2, Y, Z) = \begin{pmatrix} 0.5 - 0.5X_2 - Y + 0.5Z \\ -1 - 0.2X_2 + 0.2Y - 0.5Z \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We assume that X_2 is fully observed but $\mathbf{X}_1$ is missing at random. The missing data indicator for $\mathbf{X}_1$ is denoted by $R_{\mathbf{X}_1}$ and is generated from the following response model

$$\text{logit}\{E(R_{\mathbf{X}_1}|X_2, Y, Z)\} = 1 - X_2 - 0.5Y + 0.75Z.$$

The missingness in $\mathbf{X}_1$ results in a dataset with about two thirds of complete cases.

We denote our easy-to-implement robust estimating equation estimator in Section 3.2 and the improved estimator in Section 3.3 using REE and REE$^{\kappa}$, respectively. We consider the kernel-assisted estimating equation (8) using a Nadaraya–Watson local constant approximation of $\nu(Z)$, which can be easily implemented using the built-in package "gplm" in R for GPLMs. For comparison, we also consider the kernel estimating equation (9) using a local linear approximation of $\nu(Z)$, which requires more intensive computation and is only implemented in S1. To examine the performance of REE and REE$^{\kappa}$ in the scenario when the response model is correct but the imputation model is misspecified, we also estimate X_1^* in S1 using the following two misspecified imputation models:

(S1-M1): $\text{logit}\{E(X_1|X_2, Y, Z)\} = -1 + Y + 0.5Z$.
(S1-M2): $\log[-\log\{1 - E(X_1|X_2, Y, Z)\}] = -1 - 0.2X_2 + Y + 0.5Z$.

Five hundred replicates of simulation are performed with a sample size of $n = 500$. REE and REE$^{\kappa}$ are compared to: (i) FULL, LPLMs applied to the full data before imposing any missingness; (ii) COMP, LPLMs applied to the complete cases by discarding observations with missing covariates; and (iii) IPW, Liang's IPW method for GPLMs (Liang, 2008). We compare our proposed estimators and the IPW estimator with $\pi(\boldsymbol{\alpha})$ known or estimated, using subscript π or $\hat{\pi}$ to indicate whether $\pi(\boldsymbol{\alpha})$ is estimated.

For each method and each β coefficient, we compute empirical bias, mean squared error (MSE), average estimated standard errors (SE) using asymptotic variance estimation, and the associated nominal 95% confidence coverage rate. For comparison, we also calculate the average estimated SE using Jackknife resampling when $\nu(Z)$ is estimated using the local constant approximation (8). To compare the performance in estimating $\nu(Z)$, we first calculate the mean squared error of $\nu(Z)$ across all n observations in a single replicate of simulation,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\nu}(Z_i) - \nu(Z_i)\right)^2,$$

and then take an average of the above MSE estimates across the 500 replicates of simulation.

Table 1
Simulation results of S1: $\boldsymbol{\beta}$ coefficients of the parametrically modeled covariates, with $\nu(Z)$ estimated using both the Nadaraya–Watson kernel regression and local linear kernel regression ($n = 500$). All results are multiplied by 100.

		Nadaraya–Watson kernel					Local linear kernel			
Method	$\boldsymbol{\beta}$	Bias	MSE	Ave. Asymp. SE	Ave. Jackk. SE	Asymp. Cov. Rate	Bias	MSE	Ave. Asymp. SE	Asymp. Cove. Rate
		$\nu(z) = z$								
FULL	X_1	1.4	4.5	21.6	21.1	95.2	1.5	4.5	20.8	94.6
	X_2	−0.3	4.3	21.4	21.5	95.6	−0.4	4.3	21.2	95.2
COMP	X_1	0.9	6.4	26.4	25.8	96.2	1.0	6.4	25.2	94.4
	X_2	−11.8	9.8	27.9	28.6	90.8	−12.1	9.9	27.9	91.0
IPW_{π}	X_1	1.4	6.5	27.8	26.1	97.6	1.5	6.6	25.6	95.6
	X_2	−1.2	8.5	28.3	28.8	94.0	−1.4	8.6	27.9	93.4
$\text{IPW}_{\hat{\pi}}$	X_1	1.3	6.5	27.8	26.4	97.8	1.4	6.6	25.7	95.8
	X_2	−0.8	8.1	28.3	28.9	94.2	−1.1	8.1	27.9	94.0
REE_{π}	X_1	1.6	6.6	27.9	26.2	97.6	1.8	6.7	25.7	96.0
	X_2	−0.5	4.6	22.2	22.2	95.8	−0.6	4.6	21.9	94.8
$\text{REE}_{\hat{\pi}}$	X_1	1.4	6.6	27.7	26.2	97.4	1.6	6.6	25.6	95.8
	X_2	−0.6	4.6	22.1	22.2	95.8	−0.7	4.6	21.8	95.2
$\text{REE}^{\kappa}_{\pi}$	X_1	1.8	6.6	27.8	26.4	97.0	1.3	6.8	25.1	94.4
	X_2	−1.4	4.8	22.0	22.8	95.6	−2.3	4.2	21.7	96.4
$\text{REE}^{\kappa}_{\hat{\pi}}$	X_1	1.6	6.6	27.7	26.4	97.2	1.6	6.6	25.6	95.8
	X_2	−1.0	4.7	22.0	22.6	95.6	−1.0	4.7	21.7	94.6
		$\nu(z) = \cos(2\pi z)$								
FULL	X_1	0.02	4.2	20.4	20.0	94.2	0.5	4.3	20.1	93.8
	X_2	0.7	4.1	20.9	21.3	95.6	0.4	4.2	21.4	95.6
COMP	X_1	−0.3	6.5	24.9	24.5	93.6	0.4	6.6	24.4	92.6
	X_2	−11.2	8.8	27.6	28.8	94.4	−11.7	9.1	28.5	93.8
IPW_{π}	X_1	0.1	6.7	25.6	24.7	94.2	0.7	6.8	24.6	92.4
	X_2	−0.8	7.6	27.1	29.0	96.2	−1.2	7.8	28.0	96.4
$\text{IPW}_{\hat{\pi}}$	X_1	0.03	6.8	25.6	25.0	93.8	0.6	6.9	24.7	92.8
	X_2	−0.3	7.5	27.0	29.1	96.2	−0.7	7.6	28.0	96.6
REE_{π}	X_1	0.20	6.7	25.6	24.9	94.4	0.8	6.8	24.6	92.2
	X_2	0.7	4.3	21.0	21.9	95.6	0.4	4.3	21.8	95.8
$\text{REE}_{\hat{\pi}}$	X_1	0.1	6.8	25.6	24.9	93.8	0.7	6.9	24.6	93.0
	X_2	0.7	4.3	20.8	21.9	95.8	0.3	4.3	21.6	96.4
$\text{REE}^{\kappa}_{\pi}$	X_1	0.4	6.8	25.6	25.1	94.2	0.7	6.8	24.6	92.2
	X_2	−1.0	4.5	20.9	22.9	95.4	−0.5	4.5	21.6	96.0
$\text{REE}^{\kappa}_{\hat{\pi}}$	X_1	0.2	6.8	25.5	25.1	94.0	0.6	6.9	24.6	92.8
	X_2	−0.4	4.4	20.7	22.6	95.0	−0.2	4.5	21.5	96.2

4.2. Simulation results

Tables 1 and 2 present the simulation results in estimating $\boldsymbol{\beta}$ for scenarios S1 and S2, respectively, where both the imputation model and the response model are correctly specified. For $\boldsymbol{\beta}_1$, the regression coefficients of the covariates that are partially observed, the two REE estimators perform similarly to the IPW and the COMP estimators. However, for β_2, the regression coefficient of the covariate that is fully observed, the REE estimators yield smaller MSE and SE than the IPW counterparts. The REE and the IPW estimators also yield much smaller bias than the COMP estimator. Compared to the estimates from the full data, when $\nu(z) = z$ and is estimated using the local constant approximation, the relative efficiency of the IPW_{π} for β_2 is 57% ($=21.36^2/28.34^2$) in S1 and 60% ($=24.37^2/31.56^2$) in S2, while the relative efficiency of the REE_{π} for β_2 is 93% ($=21.36^2/22.18^2$) in S1 and 85% ($=24.37^2/26.39^2$) in S2. The average SE is similar between the asymptotic and jackknife variance estimators. The MSE and average SE are similar between the two REE estimators. This suggests that the easy-to-implement method performs closely as well as the improved method in finite sample examples, although theoretically only the improved method is guaranteed to be more efficient than the IPW. By estimating $\pi(\boldsymbol{\alpha})$, both the IPW and the REE estimators yield reduced MSE and SE as anticipated, although the reduction is small. The performance of the REE estimators in estimating $\boldsymbol{\beta}$ is similar regardless of the choice of $\nu(Z)$ possibly because of the orthogonality between the estimators of $\boldsymbol{\beta}$ and $\nu(Z)$. Using the local linear approximation to estimate $\nu(Z)$ yields similar estimates of $\boldsymbol{\beta}$ as compared to the local constant approximation.

Table 3 shows the simulation results for S1 using the misspecified imputation models S1-M1 and S1-M2. The findings are similar to those in Tables 1 and 2. This suggests that the REE estimators are valid when the response model is correct even if the imputation model is misspecified.

Table 2
Simulation results of S2: $\boldsymbol{\beta}$ coefficients of the parametrically modeled covariates, with $\nu(Z)$ estimated using the Nadaraya–Watson kernel regression ($n = 500$). All results are multiplied by 100.

		$\nu(z) = z$					$\nu(z) = \cos(2\pi z)$				
Method	$\boldsymbol{\beta}$	Bias	MSE	Ave. Asymp. SE	Ave. Jackk. SE	Asymp. Cov. Rate	Bias	MSE	Ave. Asymp. SE	Ave. Jackk. SE	Asymp. Cov. Rate
FULL	$\mathbf{X}_1$	−2.2	1.3	11.6	12.1	94.8	−0.8	1.3	11.8	12.1	94.4
		0.2	1.2	10.6	10.9	95.0	−0.6	1.2	10.8	11.0	95.6
	X_2	1.1	6.5	24.4	25.9	95.0	−2.6	5.9	23.9	24.6	95.0
COMP	$\mathbf{X}_1$	−3.0	2.1	14.2	14.8	94.6	−1.8	2.2	14.6	15.0	95.0
		−0.03	1.7	13.0	13.3	96.0	−0.1	1.9	13.4	13.6	96.4
	X_2	−8.6	12.3	32.2	34.6	92.4	−12.6	12.2	31.6	32.7	92.8
IPW_{π}	$\mathbf{X}_1$	−3.2	2.2	14.5	15.0	95.2	−1.9	2.2	14.7	15.3	94.2
		0.1	1.7	13.3	13.5	96.0	0.05	2.0	13.5	13.8	96.4
	X_2	2.1	11.7	31.6	34.8	94.8	−2.0	10.6	31.0	33.0	95.0
$\text{IPW}_{\hat{\pi}}$	$\mathbf{X}_1$	−3.2	2.2	14.5	15.0	94.8	−2.0	2.2	14.7	15.3	94.4
		0.1	1.7	13.3	13.5	96.2	0.1	2.0	13.5	13.9	97.0
	X_2	2.0	11.3	31.6	34.9	94.4	−2.3	10.1	31.0	33.0	95.0
REE_{π}	$\mathbf{X}_1$	−3.3	2.2	14.5	15.1	94.2	−2.1	2.3	14.7	15.3	94.4
		0.2	1.7	13.3	13.4	95.6	0.01	2.0	13.5	13.8	96.2
	X_2	0.8	7.5	26.4	28.5	94.0	−2.6	7.1	26.2	27.3	94.6
$\text{REE}_{\hat{\pi}}$	$\mathbf{X}_1$	−3.3	2.2	14.5	15.1	95.0	−2.2	2.3	14.7	15.3	94.2
		0.1	1.7	13.3	13.5	96.2	0.03	2.0	13.5	13.8	96.4
	X_2	0.7	7.6	26.3	28.6	93.8	−2.6	7.2	26.0	27.4	94.6
$\text{REE}^{\kappa}_{\pi}$	$\mathbf{X}_1$	−3.6	2.3	14.2	15.2	93.6	−2.4	2.3	14.3	15.4	93.6
		0.2	1.7	12.8	13.6	93.6	0.05	2.0	12.9	13.9	94.6
	X_2	3.5	8.2	27.6	30.1	94.0	−1.0	7.1	25.7	28.2	94.4
$\text{REE}^{\kappa}_{\hat{\pi}}$	$\mathbf{X}_1$	−3.5	2.3	14.2	15.2	92.6	−2.2	2.3	14.3	15.4	93.8
		0.1	1.7	12.8	13.6	94.2	0.04	2.0	12.9	13.9	95.2
	X_2	2.9	7.9	27.4	29.2	93.8	−1.7	7.1	25.6	27.8	94.2

Table 4 shows the average MSE of the $\nu(Z)$ estimates. Similar to the $\boldsymbol{\beta}$ estimates, the REE estimators perform better than the IPW estimator under both $\nu(Z)$ functions. The local linear approximation results in smaller MSE than the local constant approximation when $\nu(z) = \cos(2\pi z)$, while the local constant approximation works better when $\nu(z) = z$.

5. Analysis of the health insurance data

In this application, we study the association between ethnicity and health insurance coverage while controlling for the effect of age, gender, country of birth, and general health condition using the 2011–2012 National Health and Nutrition Examination Survey data (NHANES, Centers for Disease Control and Prevention). Our study sample contains individuals who were 18 years or older at screening, where individuals 80 and over were topcoded at 80 years of age. Survey participants were asked in questionnaires their age, gender, race, country of birth, health insurance status, and general health condition. Health insurance is a binary outcome variable, with 1 for individuals with "health insurance obtained through employment or purchased directly as well as government programs like Medicare and Medicaid that provide medical care or help pay medical bills" and 0 for those without any kind of health insurance (NHANES 2011–2012 Questionnaire). Age is measured in years. Race is a categorical variable, with 1 for "Mexican American", 2 for "Other Hispanic", 3 for "Non-Hispanic White", 4 for "Non-Hispanic Black", and 5 for "Other Race". We combine the categories 1–2 and 3–5 to create a new ethnicity variable with 1 for Hispanic and 0 for non-Hispanic. Country of birth is a binary variable, with 1 for "Born in 50 US states or Washington, DC" and 0 for "others". Current general health condition is a 1–5 scale variable: "excellent", "very good", "good", "fair", and "poor". A new health condition dummy variable is created by combining the first two categories with 1 for healthy individuals and the last three categories with 0 for less healthy individuals. To simplify the analysis, we remove 11 individuals who either refused to answer or did not know whether they had health insurance and 5 individuals without information on country of birth, which results in a dataset of 5848 observations. Among the 5848 participants, 1381 (24%) reported no health insurance, and 1199 (21%) were Hispanics.

The NHANES data were collected in two phases. The health insurance coverage and demographic information were collected in the home interview phase by trained interviewers and were fully observed for our study sample. Upon completion of the home interview, the interviewed persons were requested to report to the Mobile Examination Center (MEC) for physical examination. The Current Health Status questionnaire was administered in the examination phase and only individuals who reported to the MEC answered this question, which resulted in 870 (15%) lines of missing data for the health condition question. For demonstration purpose, we ignore the complex survey design feature of the study and treat the sample participated in the home interview phase as our target population of interest.

Table 3
Simulation results of S1 with the misspecified imputation models S1-M1 and S1-M2: $\boldsymbol{\beta}$ coefficients of the parametrically modeled covariates, with $\nu(Z)$ estimated using the Nadaraya–Watson kernel regression ($n = 500$). All results are multiplied by 100.

		$\nu(z) = z$					$\nu(z) = \cos(2\pi z)$				
Method	$\boldsymbol{\beta}$	Bias	MSE	Ave. Asymp. SE	Ave. Jackk. SE	Asymp. Cov. Rate	Bias	MSE	Ave. Asymp. SE	Ave. Jackk. SE	Asymp. Cov. Rate
		(S1-M1)									
FULL	X_1	0.8	4.3	21.5	20.9	95.0	−1.0	4.1	20.3	19.9	95.0
	X_2	−0.3	4.5	21.5	21.6	95.4	0.8	4.2	20.9	21.3	94.8
COMP	X_1	0.3	6.7	26.3	21.1	94.6	−0.6	6.5	24.8	20.2	93.8
	X_2	−11.9	10.3	28.0	23.5	90.6	−11.2	9.1	27.6	23.8	93.4
IPW_{π}	X_1	0.7	6.9	27.9	26.2	95.0	−0.4	6.8	25.6	24.9	94.6
	X_2	−1.3	9.3	28.6	29.0	93.8	−0.8	7.9	27.1	29.0	95.8
$\text{IPW}_{\hat{\pi}}$	X_1	0.5	6.9	27.9	26.3	94.8	−0.4	6.9	25.6	24.9	93.8
	X_2	−0.9	8.8	28.5	29.0	94.6	−0.3	7.8	27.1	29.0	95.8
REE_{π}	X_1	0.9	6.9	27.9	26.1	95.0	−0.3	6.8	25.6	24.8	94.8
	X_2	−0.7	5.1	22.4	22.3	95.0	0.7	4.4	21.1	22.0	95.0
$\text{REE}_{\hat{\pi}}$	X_1	0.7	7.0	27.8	26.1	95.0	−0.4	6.9	25.6	24.8	94.6
	X_2	−0.7	5.1	22.3	22.3	95.2	0.6	4.4	20.9	22.0	95.0
$\text{REE}^{\kappa}_{\pi}$	X_1	1.2	7.0	27.9	26.3	95.2	−0.1	6.9	25.6	25.0	94.2
	X_2	−1.5	5.4	22.3	23.0	94.2	−0.9	4.6	20.9	22.9	94.2
$\text{REE}^{\kappa}_{\hat{\pi}}$	X_1	0.9	7.0	27.8	26.3	94.8	−0.2	6.9	25.5	25.0	93.8
	X_2	−1.1	5.3	22.2	22.7	94.8	−0.4	4.6	20.8	22.6	94.2
		(S1-M2)									
FULL	X_1	3.3	4.0	21.8	20.9	96.6	2.4	3.9	20.5	19.9	95.6
	X_2	−0.05	4.5	21.7	21.8	96.6	1.5	4.3	20.9	21.4	94.4
COMP	X_1	3.3	6.5	26.7	21.1	96.6	3.2	6.3	25.1	20.2	95.8
	X_2	−11.7	10.1	28.2	23.7	90.4	−10.5	9.1	27.6	23.8	93.6
IPW_{π}	X_1	2.9	6.8	28.7	26.2	96.0	2.8	6.6	26.3	24.8	96.2
	X_2	−1.0	8.9	29.0	29.1	93.8	0.0	8.1	27.1	29.1	95.4
$\text{IPW}_{\hat{\pi}}$	X_1	2.8	6.8	28.7	26.2	96.2	2.7	6.6	26.3	24.9	95.8
	X_2	−0.7	8.4	28.9	29.2	94.8	0.5	7.9	27.1	29.1	95.4
REE_{π}	X_1	3.1	6.9	28.8	26.0	96.6	2.9	6.6	26.3	24.7	95.8
	X_2	−0.4	5.0	22.8	22.4	95.6	1.3	4.6	21.2	22.0	95.2
$\text{REE}_{\hat{\pi}}$	X_1	3.0	6.9	28.6	26.1	96.8	2.8	6.6	26.3	24.8	96.0
	X_2	−0.5	5.0	22.7	22.4	96.2	1.3	4.6	21.0	22.0	95.2
$\text{REE}^{\kappa}_{\pi}$	X_1	3.3	6.9	28.7	26.3	96.4	3.1	6.6	26.2	25.0	96.0
	X_2	−1.4	5.2	22.6	23.2	95.4	−0.3	4.7	21.0	22.9	94.4
$\text{REE}^{\kappa}_{\hat{\pi}}$	X_1	3.1	6.9	28.6	26.3	96.4	2.9	6.6	26.2	25.0	96.2
	X_2	−0.9	5.1	22.6	22.9	95.6	0.3	4.8	20.9	22.7	94.0

Table 4
Simulation results: average mean squared errors of the $\nu(z)$ estimates ($n = 500$) using the Nadaraya–Watson local approximation (N–W) and the local linear approximation (linear). All results are multiplied by 100.

Scenario	Kernel	$\nu(z)$	FULL	COMP	IPW_{π}	$\text{IPW}_{\hat{\pi}}$	REE_{π}	$\text{REE}_{\hat{\pi}}$	$\text{REE}^{\kappa}_{\pi}$	$\text{REE}^{\kappa}_{\hat{\pi}}$
S1	N–W	z	4.3	3.7	5.4	5.0	4.4	4.3	4.3	4.3
		$\cos(2\pi z)$	10.7	9.7	11.5	11.2	10.7	10.7	10.7	10.7
S1	Linear	z	5.3	9.1	7.8	7.4	5.7	5.7	5.8	5.7
		$\cos(2\pi z)$	7.6	9.4	10.2	10.0	8.2	8.2	8.2	8.2
S2	N–W	z	8.1	21.1	9.3	9.0	8.6	8.5	8.7	8.6
		$\cos(2\pi z)$	10.2	16.3	11.3	10.0	10.6	10.4	10.6	10.5
S1-M1	N–W	z	4.8	3.9	6.0	5.6	4.9	4.9	4.9	4.9
		$\cos(2\pi z)$	11.1	9.9	11.9	11.7	11.2	11.1	11.2	11.1
S1-M2	N–W	z	6.0	4.7	7.3	6.8	6.1	6.1	6.1	6.1
		$\cos(2\pi z)$	12.8	10.8	13.5	13.3	12.8	12.8	12.8	12.8

We assume that the general health condition data were missing at random. To allow a flexible age effect on the health insurance coverage, we fit a logistic partially linear model with ethnicity, gender, country of birth, and general health condition as parametrically modeled covariates and age as a nonparametrically modeled covariate, with bandwidth of 0.15 selected using the method of empirical bandwidth selection (Ruppert, 1997). We compare the COMP, the $\text{IPW}_{\hat{\pi}}$, the $\text{REE}_{\hat{\pi}}$ and the $\text{REE}^{\kappa}_{\hat{\pi}}$ estimators. In the link-preserving imputation model, the general health condition is modeled using a logistic

Table 5
Health insurance coverage application: regression coefficient estimates and 95% confidence intervals of the parametrically modeled covariates.

	X_1 = Health Condition	X_2 = Hispanic	X_3 = Female	X_4 = US Born
COMP	1.430 (1.213, 1.685)	0.468 (0.392, 0.558)	1.422 (1.213, 1.666)	1.742 (1.472, 2.063)
$\text{IPW}_{\tilde{\pi}}$	1.434 (1.194, 1.722)	0.457 (0.383, 0.546)	1.417 (1.191, 1.686)	1.726 (1.421, 2.096)
$\text{REE}_{\tilde{\pi}}$	1.432 (1.198, 1.713)	0.438 (0.374, 0.513)	1.397 (1.196, 1.632)	1.668 (1.407, 1.978)
$\text{REE}^{\kappa}_{\tilde{\pi}}$	1.436 (1.201, 1.717)	0.439 (0.374, 0.514)	1.394 (1.194, 1.627)	1.670 (1.408, 1.979)

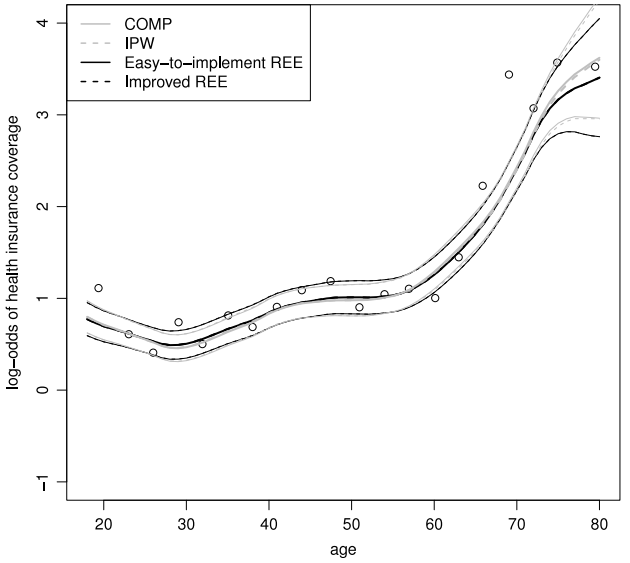

Fig. 1. Log-odds estimates and 95% confidence intervals (CIs) of health insurance coverage as a non-linear function of age using the 2011–2012 National Health and Nutrition Examination Survey. Circles denote the observed log-odds of health insurance coverage in each of the twenty equally-spaced age groups. The solid and dashed gray are the estimates of $\nu(\text{age})$ using the complete cases and IPW method with point-wise CIs respectively, and the solid and dashed black curves are the estimates using the easy-to-implement and the improved REEs with point-wise CIs respectively.

regression on age, gender, race, country of birth, and health insurance coverage. A Nadaraya–Watson kernel regression is considered. Table 5 shows that all four methods yield similar point estimates of β_1 associated with the general health condition, although the REE estimators yield a slightly shorter 95% confidence interval (CI) than the IPW. For β_2 to β_4, the coefficients for ethnicity, gender, and country of birth, the REE methods estimate a slightly larger ethnicity effect but smaller gender and country of birth effects. The REE estimators also yield shorter 95% CIs than the COMP and the IPW. Our analysis shows that non-Hispanic, female, healthy, and born in US people had a higher health insurance coverage rate than Hispanic, male, and people born in other countries with poor general health condition.

To check the model fit of $\nu(\text{age})$, we create 20 equally-sized age groups and calculate the observed log-odds of having health insurance for each age group. Fig. 1 shows that the four methods yield similar estimates of $\nu(\text{age})$ and 95% CIs except when age is older than 75, where the estimates of log-odds are lower using the REEs than the COMP and the IPW. It also shows that the odds of health insurance coverage reach the lowest level at 25–35 years and increase dramatically after 65. This significant increase in the proportion of health insurance coverage after 65 might be explained by the eligibility of Medicare after that age. However, ethnic disparity in health insurance coverage exists, and this disparity is significant. The R code together with a complete documentation that illustrates the application of the proposed methods to the health insurance data are in Section 3 of the Supplementary Materials (see Appendix A).

The assumption of MAR or nonignorable missingness cannot be verified. Instead, we conduct a sensitivity analysis to examine how the coefficient estimates in Table 5 change as the degree of nonignorableness changes. We consider a response model $\text{logit}(\pi_i) = \mathbf{D}_i^T\boldsymbol{\alpha} + \alpha^* x_{1i}$, with $\mathbf{D}_i = (1, x_{2i}, x_{3i}, x_{4i}, y_i, z_i)^T$. We fix α^* to be a value in $(-3, 3)$. At each fixed value of α^*, we first estimate $\boldsymbol{\alpha}$ by solving $\sum_{i=1}^n \mathbf{D}_i(R_i/\pi_i - 1) = 0$ and then estimate $\boldsymbol{\beta}$ in the LPLM using the REE estimators. When $\alpha^* = 0$, the missing mechanism corresponds to MAR. A larger absolute value of α^* indicates more severe degree of nonignorable missingness. Fig. 2 shows that the $\boldsymbol{\beta}$ coefficient estimates are robust with respect to severity of nonignorable missingness, and thus are robust against the MAR assumption.

6. Conclusion

We propose two kernel-assisted estimating equation estimators using link-preserving imputation for logistic partially linear models with missing covariates. The first estimator under this approach is easy to implement by modifying built-in

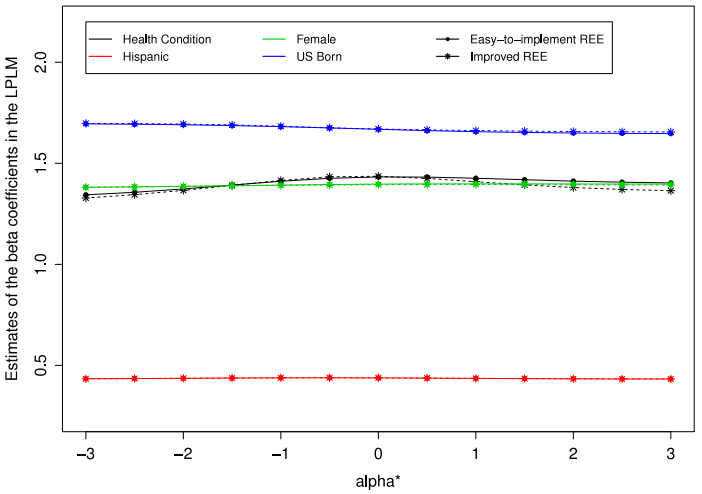

Fig. 2. Health insurance coverage application: Sensitivity analysis to examine how the estimates of the $\boldsymbol{\beta}$ coefficients in the LPLM change as the degree of nonignorableness changes. We consider a response model $\text{logit}(\pi_i) = \mathbf{D}_i^T\boldsymbol{\alpha} + \alpha^* x_{1i}$, with $\mathbf{D}_i = (1, x_{2i}, x_{3i}, x_{4i}, y_i, z_i)^T$. At each fixed value of α^*, we first estimate $\boldsymbol{\alpha}$ by solving $\sum_{i=1}^n \mathbf{D}_i(R_i/\pi_i - 1) = 0$ and then estimate $\boldsymbol{\beta}$ in the LPLM.

functions for complete data in statistical software via data augmentation. The second estimator is an extension of the first estimator but is guaranteed to be more efficient than the IPW. Our proposed estimators are valid when the response model is correct, no matter if the imputation model is correctly specified. When the missingness of $\mathbf{X}_1$ is independent of $(\mathbf{X}_1, Y)$ given $(\mathbf{X}_2, Z)$ or the parametrically modeled missing covariates have no effects on the outcome variable, i.e. $\boldsymbol{\beta}_1 = 0$, the proposed estimators are also doubly robust.

Our simulation study shows that the proposed estimating equation approach can greatly improve the efficiency of the regression coefficients estimates of fully observed covariates, parametrically or nonparametrically modeled, upon the IPW. The easy-to-implement approach performs closely as well as the improved approach in finite sample examples, although the improved estimator tends to yield slightly smaller asymptotic standard errors. We show the application of our proposed methods to NHANES where data were collected in two phases, with demographics and insurance data collected in the home interview phase and the health condition collected in the examination phase. Although in our application only the general health condition collected in the examination phase was considered, the logistic partially linear model can be extended to include multiple covariates collected in the examination phase that are missing together for persons who did not report to the MEC. The proposed methods have important applications to missing covariate problems in logistic partially linear models especially for those arise in two-phase sampling designs.

Acknowledgments

This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013R1A2A2A01067262) and the Seoul National University Research Grant.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.03.004.

References

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.
Carroll, R.J., Fan, J., Gjbels, I., Wand, M.P., 1997. Generalized partially linear single-index models. J. Amer. Statist. Assoc. 92, 477–489.
Chen, J., Fan, J., Li, K.H., Zhou, H., 2006. Local quasi-likelihood estimation with data missing at random. Statist. Sinica 16, 1071–1100.
Liang, H., 2008. Generalized partially linear models with missing covariates. J. Multivariate Anal. 99, 880–895.
Liang, H., Qin, Y., Zhang, X., Ruppert, D., 2009. Empirical likelihood-based inferences for generalized partially linear models. Scand. J. Statist. 36, 433–443.
Liang, H., Wang, S., Robins, J.M., Carroll, R.J., 2004. Estimation in partially linear models with missing covariates. J. Amer. Statist. Assoc. 99, 357–367.
Little, R.J.A., An, H., 2004. Robust likelihood-based analysis of multivariate data with missing values. Statist. Sinica 14, 949–968.
Paik, M.C., 1997. The generalized estimating equation approach when data are not missing completely at random. J. Amer. Statist. Assoc. 92, 1320–1329.
Reilly, M., Pepe, M.S., 1995. A mean score method for missing and auxiliary covariate data in regression models. Biometrika 82, 299–314.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Amer. Statist. Assoc. 90, 106–121.
Scharfstein, D.O., Rotnitzky, A., Robins, J.M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. (with discussion and rejoinder). J. Amer. Statist. Assoc. 94, 1096–1146.

Severini, T.A., Staniswalis, J.G., 1994. Quasi-likelihood estimation in semiparametric models. J. Amer. Statist. Assoc. 89, 501–511.
Paik, M.C., Sacco, R.L., 2000. Matched case-control data analyses with missing covariates. Appl. Stat. 49, 145–156.
Qin, G., Zhu, Z., Fung, W.K., 2012. Robust estimation of the generalised partial linear model with missing covariates. J. Nonparametr. Stat. 24, 517–530.
Ruppert, D., 1997. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. J. Amer. Statist. Assoc. 92, 1049–1062.
van der Laan, M.J., Robins, J.M., 2003. Unified Methods for Censored Longitudinal Data and Causality. Springer, New York.
Wang, C., Paik, M.C., 2006. Efficiencies of methods dealing with missing covariates in regression analysis. Statist. Sinica 16, 1169–1192.
Wang, L., Rotnitzky, A., Lin, X., 2010. Nonparametric regression with missing outcomes using weighted kernel estimating equations. J. Amer. Statist. Assoc. 105, 1135–1146.
Wang, C.Y., Wang, S., Gutierrez, R.G., Carroll, R.J., 1998. Local linear regresion for generalized linear models with missing data. Ann. Statist. 26, 1028–1050.