Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

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 xxxx

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.

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/

Please cite this article in press as: Chen, Q., et al., Using link-preserving imputation for logistic partially linear models with missing covariates. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.03.004

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.

Please cite this article in press as: Chen, Q., et al., Using link-preserving imputation for logistic partially linear models with missing covariates. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.03.004