Journal of Multivariate Analysis 128 (2014) 33–50

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Semiparametric efficient estimation for partially linear single-index models with responses missing at random

Peng Lai [a], Qihua Wang [b,c,*]

[a] School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[b] Institute of Statistical Science, Shenzhen University, Shenzhen 518060, China
[c] Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China

ARTICLE INFO

Article history:
Received 17 July 2012
Available online 17 March 2014

AMS subject classification:
62J99

Keywords:
Efficient score function
Estimating equations
Heteroscedasticity
Missing at random
Partially linear single-index model

ABSTRACT

In this paper, we establish the semiparametric efficient bound for the heteroscedastic partially linear single-index model with responses missing at random, and develop an efficient estimating equation method. By solving the estimating equation, we obtain estimators for the parameter vectors in the linear part and the single index part simultaneously. The estimators are asymptotically semiparametrically efficient when the propensity score function is specified correctly. It should be noted that the inverse probability weighted efficient estimating equation cannot be obtained directly from the full data efficient estimating equation by the inverse probability weighted approach. We establish the estimating equation by deriving the observed data efficient score function. Some simulation studies and a real data application were conducted to evaluate and illustrate the proposed methods.

1. Introduction

How to avoid "curse of dimensionality" is always a challenging problem in the high dimensional data analysis. Some dimension reduction methods and variable selection methods are suggested to overcome these difficulties. Another way to avoid "curse of dimensionality" is to consider dimension reduction models, such as the additive model [1,16], partially linear model [18,23], single-index model [14], generalized partially linear single-index models [2], extended partially linear single-index models [37], semiparametric varying-coefficient partially linear model [11]. Partially linear single-index model is an important alternative to the classical linear model by putting a nonparametric component into the model. That is,

$$Y = \theta^\top Z + g(X^\top \beta) + \varepsilon, \tag{1.1}$$

where Y is a scalar response variate, $(X, Z) \in \mathbb{R}^p \times \mathbb{R}^q$, $g(\cdot)$ is an unknown univariate link function, ε is random error with $E(\varepsilon|X, Z) = 0$ and $\mathrm{var}(\varepsilon|X, Z) = \sigma^2(X, Z)$. (β, θ) is an unknown vector in $\mathbb{R}^p \times \mathbb{R}^q$ with $\|\beta\| = 1$ (where $\|\cdot\|$ denotes the Euclidean metric). The true value of (β, θ) is (β_0, θ_0).

Model (1.1) has been paid considerable attention in recent years. See, e.g., [2,40,36,41,35,22]. Carroll et al. [2] develop a quasi-likelihood method to estimate an unknown link function and unknown parameters. Yu and Ruppert [40] propose the penalized spline estimation procedure. Xia and Härdle [36] extend the method of Härdle et al. [14] for a single index model to model (1.1), and develop the minimum average variance estimation (MAVE) method. Wang et al. [35] consider the estimating problem and discuss the efficiency of the estimators. Wang et al. [35] compare the asymptotic covariance of the proposed estimators with that of Carroll et al. [2], and show that their estimators are more efficient. Liang et al. [22] develop

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China.
E-mail address: qhwang@amss.ac.cn (Q. Wang).

http://dx.doi.org/10.1016/j.jmva.2014.03.001
0047-259X/

the profile least-squares approach to define efficient parameter estimators. But, they consider the special case of $\sigma^2(x,z) = \sigma^2$, where σ^2 is a constant. This method, however, cannot be extended to the case of the heteroscedastic variance directly.

Model (1.1) reduces to the partial linear models when $p = 1$ and $\beta = 1$. For the partially linear models, Härdle et al. [15] propose a direct weighted extension of the estimator given in [21], with the weights being inversely proportional to the variances. However, as Ma et al. [25] point out, this estimator is still not efficient. For defining efficient estimator, Ma et al. [25] consider the heteroscedastic partially linear models, and utilize a weighted estimating equation method to define an efficient semiparametric estimator for the parameter vector in the linear part. This motivates us to develop efficient estimating equation method to define asymptotically efficient semiparametric estimators of the global parameter vectors β and θ for the heteroscedastic partially linear single-index model (1.1), and consider a more complex case where responses are missing. The estimating equations for the heteroscedastic partially linear models due to Ma et al. [25] cannot be applied to model (1.1) directly since model (1.1) concerns a single-index term and we consider missing response problems. Clearly, some more complex techniques for our estimating problem are needed. In the absence of missing response, semiparametric efficiency problem for other models has been studied extensively in literatures, such as [28,32,12,5,3,4,39], and so on.

In practice, response variable may be missing, by design or by happenstance. For example, in the well known two-stage sampling scheme, it is not feasible to obtain enough observations of Y perhaps because it is expensive to obtain more Y's, instead one may take more observations on the covariates such that the information contained in covariates can make up for the lack of information in Y. Missing responses occur commonly in opinion polls, market research surveys, social investigations, medical studies and other disciplines. The existing methods of analyzing data with responses missing at random can be roughly classified as parametric likelihood method [24], imputation method [38,34,33] and inverse probability weighting method [17,30,27]. As a form of inverse probability weighted method, Tsiatis [32] discusses the inverse probability weighted complete case (IPWCC) estimator and augmented inverse probability weighted complete case (AIPWCC) estimator of parametric models for missing data.

We consider the missing response problems and assume that we obtain the following incomplete observations

$$(Y_i, \delta_i, X_i, Z_i), \quad i = 1, 2, \ldots, n,$$

from model (1.1), where $\delta_i = 0$ if Y_i is missing, $\delta_i = 1$ otherwise. Throughout this paper we assume that Y is missing at random (MAR). That is,

$$p(\delta = 1|Y, X, Z) = p(\delta = 1|X, Z).$$

It is a commonly used as missing mechanism for statistical analysis with missing data and is reasonable in many practical situations; see [24].

Naturally, one may consider the estimating problem by first deriving the full-data efficient score function and then use the inverse probability weighted approach to construct an estimating equation to compute estimators. Unfortunately, this method does not define efficient estimators. The main purpose of this article is to establish the semiparametric efficient bound and construct an efficient estimating equation to obtain efficient estimators for heteroscedastic partially linear single-index model with responses missing at random. Our method is to find the optimal element in the orthogonal complement of the nuisance tangent space to obtain the observed data efficient score function such that the semiparametric bound can be calculated, and then obtain an estimated observed-data efficient score function and use it to construct an efficient estimating equation, from which we can define asymptotically efficient estimators of both β and θ with responses missing at random.

The rest of this article is organized as follows. In Section 2, the observed-data efficient score function is presented, and the observed-data semiparametric efficiency bound is proposed when response variable is missing at random. In Section 3, we construct the efficient estimating equation and establish the asymptotic properties of the proposed estimators. In Section 4, some simulation studies are conducted to evaluate the proposed methods and a real data set is analyzed to illustrate the proposed methods. The proofs of the asymptotic properties are presented in the Appendix.

2. Semiparametric efficiency bound

Model (1.1) concerns the parameter vectors θ and β of interest, the nuisance link function $g(\cdot)$, the unknown conditional variance $\sigma^2(X, Z)$, the conditional distribution of $\varepsilon|X, Z$ and the distribution of (X, Z). For the sake of identifiability, it is assumed that $\|\beta\| = 1$, where $\|\cdot\|$ denotes the Euclidean metric. Because $\|\beta\| = 1$, this means that the true value of β is a boundary point on the unit sphere, and hence $g(X^\top\beta)$ does not have a derivative at the point β (see [41]). For the convenience of computing the score vector, we make the following adjustment. To use the constraint $\|\beta\| = 1$, we use the delete-one-component method proposed by Yu and Ruppert [40]. Without loss of generality, we assume that the parameter vector β has a positive component β_r (otherwise, consider $-\beta$). Let $\beta = (\beta_1, \ldots, \beta_p)^\top$ and $\beta^{(r)} = (\beta_1, \ldots, \beta_{r-1}, \beta_{r+1}, \ldots, \beta_p)^\top$ be a $(p-1)$-dimensional vector after removing the rth component β_r of β. Then, we may write

$$\beta(\beta^{(r)}) = (\beta_1, \ldots, \beta_{r-1}, (1 - \|\beta^{(r)}\|^2)^{1/2}, \beta_{r+1}, \ldots, \beta_p)^\top.$$

The true parameter $\beta_0^{(r)}$ must satisfy the constraint $\|\beta_0^{(r)}\| < 1$. Thus, β is infinitely differential in a neighborhood of $\beta_0^{(r)}$, and the Jacobian matrix is

$$J_{\beta^{(r)}} = \frac{\partial\beta}{\partial\beta^{(r)}} = (b_1, \ldots, b_p)^\top,$$

where $b_s (1 \le s \le p, s \ne r)$ is a $(p-1)$-dimensional unit vector with sth component 1, and $b_r = -(1 - \|\beta^{(r)}\|^2)^{-1/2}\beta^{(r)}$.