Computational Statistics and Data Analysis 68 (2013) 82–97

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation and empirical likelihood for single-index models with missing data in the covariates

Liugen Xue *

College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Article history:
Received 16 March 2012
Received in revised form 24 May 2013
Accepted 19 June 2013
Available online 28 June 2013

Keywords:
Empirical likelihood
Single-index model
Missing data
Confidence region
Imputation method

ABSTRACT

The estimation and empirical likelihood for single-index models with missing covariates are studied. A generalized estimating equations estimator for index coefficients with missing covariates is constructed, and its asymptotic distribution is obtained. The local linear estimator for link function achieves optimal convergence rate. By using the bias-correction and inverse selection probability weighted methods, a class of empirical likelihood ratios is proposed such that each of our class of ratios is asymptotically chi-squared. A simulation study indicates that the proposed methods are comparable in terms of coverage probabilities and average lengths (areas) of confidence intervals (regions). An example of a real data set is illustrated.

1. Introduction

Consider the single-index model

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

where Y is a scalar random variable, $X \in R^p$ is a covariate, $\beta \in R^p$ is an unknown index coefficients, $g(\cdot)$ is an unknown link function, ε is a random error with $E(\varepsilon|X) = 0$ almost surely. For identifiability purposes (Lin and Kulasekera, 2007), we typically assume $\|\beta\| = 1$, with its first non-zero element being positive, where $\|\cdot\|$ denotes the Euclidean norm. For this model, we focus on the case where X value in a sample of size n may be missing and Y is observed completely. That is, we obtain a sample of incomplete observations $\{(X_i, Y_i, \delta_i), 1 \le i \le n\}$ from (X, Y, δ), where all the Y_i's are observed, and $\delta_i = 0$ if X_i are missing and $\delta_i = 1$ otherwise.

The single-index model is one of the most popular semiparametric models in applied quantitative sciences. This model search for a single linear combination of p-covariate X that can capture most information about the relation between response variable Y and covariate X, thereby avoiding the curse of dimensionality. Many authors have investigated the estimation of the index coefficients β with focus on $\sqrt{n}$-estimability and efficiency issues. The proposed methods have the average derivative method (Stoker, 1986; Powell et al., 1989), the sliced inverse regression method (Duan and Li, 1991; Li, 1991), the least squares method (Härdle et al., 1993), the minimum average conditional variance method (Xia et al., 2002), and the estimating function method (Cui et al., 2011). Wang et al. (2010) studied the estimation of parameters of interest for the single-index model with missing response at random. Xue and Zhu (2006) studied the single-index models by the empirical likelihood method. They proposed the estimated and adjusted empirical log-likelihood ratios to construct the confidence region of the regression parameter. The empirical likelihood method has many advantages, for example, it does

* Tel.: +86 13681336028; fax: +86 10 67391738.
E-mail address: lgxue@bjut.edu.cn.

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.06.017

not impose prior constraints on the shape of the region, it does not require the construction of a pivotal quantity and it does not involve a plug-in estimator for the asymptotic variance. For these reasons, this method has found many applications such as in parametric models, nonparametric models and semiparametric models. The relevant papers are, among others, Owen (1988, 1990, 2001), Wang and Rao (2002a,b), Chen and Qin (1993), Chen and Hall (1993), Kolaczyk (1994), Qin and Lawless (1994), Wang et al. (2004), Qin and Zhang (2007), Stute et al. (2007), Zhu and Xue (2006), Xue and Zhu (2007a,b), Xue (2009a,b, 2010), Wang and Xue (2011), Xue and Xue (2011).

In this paper, we are interested in estimating β and the unknown function $g(u)$ as well as constructing confidence region of β in model (1.1) when the covariate X has missing data. A bias-correction method and an inverse selection probability weighted method are used to develop some methods for constructing the estimator and confidence region of β. We construct a class of estimators for β and $g(u)$. It is shown that the estimator of β is asymptotically normal and the estimator of $g(u)$ has optimal convergence rate. A class of empirical log-likelihood ratios for β is constructed. It is shown that each of these ratios is asymptotically chi-squared. We also provide consistent estimators of asymptotic variance. Our results can be used directly to construct the confidence intervals and region of β. The proposed method has the following features: by using the centred covariates and the inverse selection probability weighted imputation for missing X, the resulting empirical log-likelihood ratio has asymptotically chi-squared distribution; this achieves the bias-correction for the empirical likelihood ratio, undersmoothing for estimating nonparametric functions are avoided, and the existing data-driven algorithm can be used to select an optimal bandwidth.

The construction of this paper is as follows. Section 2 gives a methodology for constructing the estimator and the empirical likelihood ratio for β, and gives some results. Section 3 reports the results of a simulation study and an application. Section 4 is the concluding remark. Proofs of theorems are relegated to Appendix.

2. Methodology and main results

2.1. Weighted estimating equations

Throughout this paper, we assume that X is missing at random (MAR). The MAR assumption implies that δ and X is conditionally independent given Y, that is,

$$P(\delta = 1|X, Y) = P(\delta = 1|Y) \equiv \pi(Y), \tag{2.1}$$

where $\pi(Y)$ is termed the selection probability function.

In a general case, the selection probability function $\pi(y)$ is unknown. To construct the estimator and empirical likelihood ratio for β, we need to estimate $\pi(y)$ by using kernel method. Assume that $\{(X_i, Y_i, \delta_i), 1 \le i \le n\}$ is an independent and identically distributed (i.i.d.) sample from model (1.1). For a given bandwidth $h_1 = h_1(n)$ with $0 < h_1 \to 0$ and a kernel function $K^*(y)$ in R, the estimator of $\pi(y)$ is defined as

$$\hat{\pi}(y) = \sum_{i=1}^{n} K^*\left(\frac{Y_i - y}{h_1}\right)\delta_i \Big/ \sum_{j=1}^{n} K^*\left(\frac{Y_j - y}{h_1}\right). \tag{2.2}$$

For given β, we can construct the estimators for $g(u)$ and its derivative $g'(u)$ by using the local linear fitting method proposed by Fan and Gijbels (1996). Let $K(\cdot)$ be a kernel function on the real set R, and $h_2 = h_2(n)$ is a bandwidth sequence tending to 0 with $h_2 > 0$. Then, the estimators of $g(u)$ and $g'(u)$ are defined as

$$\hat{g}(u; \beta) = \sum_{i=1}^{n} W_{ni}(u; \beta) Y_i \tag{2.3}$$

and

$$\hat{g}'(u; \beta) = \sum_{i=1}^{n} \widetilde{W}_{ni}(u; \beta) Y_i, \tag{2.4}$$

where

$$W_{ni}(u; \beta) = \frac{n^{-1}\{\delta_i/\hat{\pi}(Y_i)\} K_{h_2}(X_i^{\mathrm{T}}\beta - u)\{S_{n,2}(u; \beta) - (X_i^{\mathrm{T}}\beta - u) S_{n,1}(u; \beta)\}}{S_{n,0}(u; \beta) S_{n,2}(u; \beta) - S_{n,1}^2(u; \beta)},$$

$$\widetilde{W}_{ni}(u; \beta) = \frac{n^{-1}\{\delta_i/\hat{\pi}(Y_i)\} K_{h_2}(X_i^{\mathrm{T}}\beta - u)\{(X_i^{\mathrm{T}}\beta - u) S_{n,0}(u; \beta) - S_{n,1}(u; \beta)\}}{S_{n,0}(u; \beta) S_{n,2}(u; \beta) - S_{n,1}^2(u; \beta)},$$

$$S_{n,l}(u; \beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(Y_i)} (X_i^{\mathrm{T}}\beta - u)^l K_{h_2}(X_i^{\mathrm{T}}\beta - u), \quad l = 0, 1, 2$$

and $K_{h_2}(\cdot) = h_2^{-1} K(\cdot/h_2)$.