Statistics and Probability Letters 95 (2014) 15–25

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Estimation and test procedures for composite quantile regression with covariates missing at random

Zijun Ning, Linjun Tang *

Department of Statistics, Jiaxing University, Jiaxing, 314001, China

ARTICLE INFO

Article history:
Received 4 April 2014
Received in revised form 29 July 2014
Accepted 2 August 2014
Available online 11 August 2014

Keywords:
Composite quantile regression
General linear model
Missing covariates
Bootstrap

ABSTRACT

In this paper, we study the weighted composite quantile regression (WCQR) for general linear model with missing covariates. We propose the WCQR estimation and bootstrap test procedures for unknown parameters. Simulation studies and a real data analysis are conducted to examine the finite performance of our proposed methods.

1. Introduction

Regression models with missing covariates are applied widely in many scientific areas, especially in biomedical and social sciences (see Little and Rubin, 2002; Tsiatis, 2006 for detailed reviews). The missing at random (MAR) in sense of Robins (1976) is a common assumption for statistical analysis with missing data. When covariates are MAR, the inverse probability weighting approach has received considerable attention. For instance, Robins (1994) proposed the inverse probability weighting estimator for parametric model with missing covariates; moreover, Wang et al. (1998) provided a local version for generalized linear model with missing covariates. In recent years, Liang et al. (2004), Liang (2008), Wang (2009), and Wong et al. (2009) considered the weighted estimation for semiparametric and nonparametric model with missing covariates, based on the inverse probability weighting idea. However, many of these weighted estimation procedures are built on least squares (LS), which can produce an unreliable estimator when the model error has heavy-tailed or skewed distribution. The efficient and stable estimation of regression model with missing covariates is still a challenging problem. Consider a general linear model with missing covariates as follows:

$$Y = \phi^T(X)\beta + \epsilon, \tag{1}$$

where Y is an observable response, $\phi(\cdot)$ is a known $p \times 1$ vector function, β is a $p \times 1$ vector of unknown coefficients. Some values of the covariates, denoted as V with $X = (U, V)$, may be missing at random for some reasons.

To obtain the desirable robust estimator, Sherwood et al. (2014) suggested inverse probability weighted quantile regression with missing covariates. However, a single quantile regression estimation procedure including least absolute deviation (LAD) one can lead to an arbitrary small relative efficiency when compared with LS. Therefore, we do not consider it as a safe alternative to LS. To overcome this drawback, Zou and Yuan (2008) developed the composite quantile regression (CQR) estimation procedure for the linear model. Kai et al. (2011), and Guo et al. (2012) proposed the efficient estimators

* Corresponding author.
E-mail address: tljlqz@163.com (L. Tang).

http://dx.doi.org/10.1016/j.spl.2014.08.003

for semiparametric varying-coefficient partially linear models based on CQR method. Recently, Jiang et al. (2012) and Tang et al. (2012) further extended the CQR method for linear model with censored data. The CQR can significantly improve the relative efficiency of the resulting estimator. Inspired by this nice performance, we consider inverse probability WCQR for model (1), when the selection probabilities are known, estimated nonparametrically and parametrically, respectively.

In statistical modelling, some prior information about the unknown parameters might be available from outside sample sources. To the best of our knowledge, use of reliable prior information can greatly improve upon the efficiency of regression analysis. However, once these prior conditions have been imposed, the first task is to test whether they are true. There is a little literature on the hypothesis test of parametric and nonparametric models with missing response. For example, Sun et al. (2009) investigated the hypothesis test of general linear model with missing response. However, research on parametric test of regression model with missing covariates has been comparatively limited. In this paper, we develop the WCQR-based test procedure for unknown parameters in model (1). Without loss of generality, we consider the following linear hypothesis:

$$H_0 : A\beta = b \quad \text{v.s.} \quad H_1 : A\beta \neq b, \tag{2}$$

where A is a $k \times p$ known matrix with $\text{rank}(A) = k$ $(0 < k \leq p)$ and b is a $k \times 1$ vector of known constant vector. From the idea of Zhao (2004) and Chen et al. (2008), we construct the test statistic for hypothesis (2) based on the weighted residual sums of quantile from WCQR fit. Finally, we propose a wild bootstrap approach to determine the critical value of the test statistic, which avoids estimating asymptotic distribution of test statistic.

The paper is organized as follows. In Section 2, we propose the weighted composite quantile regression estimation and test for general linear model with covariates missing. We conduct extensive simulation studies and a real data analysis to illustrate the proposed methods in Section 3. We conclude the article with a discussion in Section 4. All the technical conditions and proofs are relegated to the Appendix.

2. Methodology

Let $\{(Y_i, \delta_i, U_i, V_i)\}_{i=1}^n$ be a random sample from model (1), such that

$$Y_i = \phi^T(X_i)\beta + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{3}$$

where δ_i is a missing indicator individual, $\delta_i = 0$ if V_i is missing, otherwise $\delta_i = 1$. Suppose that V_i is MAR in sense that

$$\pi(Z_i) = P(\delta_i = 1 | Y_i, V_i, U_i) = P(\delta_i = 1 | Y, U_i), \tag{4}$$

where $Z_i = (Y_i, U_i^T)^T$. Furthermore, we assume that β is the same across different quantile models. Our interest lies in both estimation and test for model (3) based on WCQR approach.

2.1. Estimation

Let $\rho_\tau(r) = \tau I(r \geq 0) + (1 - \tau)I(r < 0)$ be the check loss function of τ quantile regression. Denote $\tau_k = k/(1 + q)$ where q is a composite level. When the selection probability function $\pi(\cdot)$ is known, we define the WCQR estimator $\hat{\beta}$ of β as

$$\left(\hat{\boldsymbol{a}}, \hat{\beta}\right) = \arg\min_{\boldsymbol{a},\beta} \sum_{k=1}^{q} \sum_{i=1}^{n} \frac{\delta_i}{\pi(Z_i)} \rho_{\tau_k}\left(Y_i - a_k - \phi(X_i)^T\beta\right), \tag{5}$$

where $\boldsymbol{a} = (a_1, \ldots, a_q)$, and a_k is the τ_kth quantile of ϵ.

In practice, the selection probability function $\pi(\cdot)$ is usually unknown. It is well known that nonparametric smoothing is a powerful tool to estimate unknown selection probability. Let $L(\cdot)$ be a 2-dimensional symmetric kernel function. The nonparametric smoothing estimator of $\pi(z)$ based on the data $\{(Z_i, \delta_i)\}_{i=1}^n$ is defined as

$$\hat{\pi}(z) = \frac{\sum_{j=1}^{n} \delta_j L_h\left(z - Z_i\right)}{\sum_{j=1}^{n} L_h\left(z - Z_j\right)}, \tag{6}$$

where $L_h(\cdot) = L_h(\cdot/h)/h^2$ is kernel function and h is a bandwidth. Let $\hat{\pi}(z)$ be nonparametric smoothing estimator of $\pi(z)$. Then WCQR estimator with $\hat{\pi}(z)$, denoted by $\hat{\beta}_N(t)$, can be defined as

$$\left(\hat{\boldsymbol{a}}_N, \hat{\beta}_N\right) = \arg\min_{\boldsymbol{a},\beta} \sum_{k=1}^{q} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(Z_i)} \rho_{\tau_k}\left(Y_i - a_k - \phi(X_i)^T\beta\right). \tag{7}$$