Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Estimation and inference of combining quantile and least-square regressions with missing data

Linjun Tang *, Shengchao Zheng, Zhangong Zhou

Department of Statistics, Jiaxing University, Jiaxing 314001, China

A R T I C L E I N F O

Article history:
Received 11 March 2017
Accepted 26 September 2017
Available online xxxx

AMS 2000 subject classifications:
62F03
62F05
62F12

Keywords:
Missing data
Inverse-probability weights
Smoothing techniques
Empirical likelihood

A B S T R A C T

In this paper, we consider how to incorporate quantile information to improve estimator efficiency for regression model with missing covariates. We combine the quantile information with least-squares normal equations and construct an unbiased estimating equations (EEs). The lack of smoothness of the objective EEs is overcome by replacing them with smooth approximations. The maximum smoothed empirical likelihood (MSEL) estimators are established based on inverse probability weighted (IPW) smoothed EEs and their asymptotic properties are studied under some regular conditions. Moreover, we develop two novel testing procedures for the underlying model. The finite-sample performance of the proposed methodology is examined by simulation studies. A real example is used to illustrate our methods.

1. Introduction

The regression analysis with missing covariates is a practical problem in biomedical, epidemiologic and social research. The missing at random (MAR) in sense of Rubin (1976) is a common assumption for statistical analysis with missing data and is often reasonable in many practical situations. A naive approach of handling missing data is a complete-case (C–C) analysis, which uses only data points with complete observation. However, a C-C analysis may lead to a biased estimator when the missing data mechanism is MAR. To overcome this problem, Robins, Rotnitsky, and Zhao (1994) developed an inverse probability weighting (IPW) estimating method for regression model with missing covariates. In the past ten years, the IPW method has received considerable attention as a well-known technique for handling missing data. For example, Little and Rubin (2002), and Tsiatis (2006) proposed the IPW estimators for linear model with missing covariates. Liang (2008) and Wong, Guo, Chen, et al. (2009) further extended this method to semi-parametric and nonparametric models with missing data, respectively. However, the aforementioned estimations are mainly built on least squares (LS), which can cause an unreliable estimator without normality of errors.

To obtain a robust estimator, Sherwood, Wang, and Zhou (2013) developed IPW quantile regression (IPW-QR) method for linear model with missing covariates. Much of the literature views quantile regression (QR) as an alternative to LS (see e.g., Yang and Liu, 2016; Sun and Sun, 2015). A major difficulty of QR is that the asymptotic covariance of estimators is often cumbersome to obtain due to need an estimate of regressor density. Furthermore, QR estimator only considers some quantile moment condition and resultant estimator might suffer efficiency loss. These motivated Zhou, Alan, and Wan (2011) to combine mean and median regression and proposed the coherent estimation framework. Their theoretical and numerical results have shown that LS regression combined with median one not only leads to a more efficient estimators when model

* Corresponding author.
E-mail address: tlj@zjxu.edu.cn (L. Tang).

https://doi.org/10.1016/j.jkss.2017.09.005
1226-3192/

error are symmetric, but also results in a relatively simple calculation of estimators standard error. Based on the same idea, Zhao, Wang, and Zhou (2016) considered the estimator of censored linear model by combining the quantile information with censored least-squares normal equations. However, the interest of this paper is to extend this idea to regression model with missing covariates when some quantile information of model error is available. Consider the model

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

where Y is an observable response variable, β is a $p \times 1$ vector of parameter, and some values of the covariates, denoted as V with $X = (U^T, V^T)^T$, may be missing at random for some reasons. Let $\{(Y_i, U_i, V_i, \delta_i)\}_{i=1}^n$ be a random sample from population of (Y, U, V, δ), where $Z_i = (Y_i, U_i^T)^T$ are fully observed, and $\delta_i = 0$ if V_i is missing, otherwise $\delta_i = 1$. In current paper, we assume that V_i is MAR. That is

$$P(\delta_i = 1|Y_i, U_i, V_i) = P(\delta_i = 1|Y_i, U_i) = \pi(Z_i, \gamma) > 0, \tag{1.2}$$

where $\pi(\cdot, \gamma)$ is a known function and γ is unknown parameter.

Our objective in this paper is to improve the statistical inference on model (1.1) by using some prior quantile information on model error. Firstly, we combine least-squares with quantile regression to develop a coherent estimation framework. We use a kernel-based smoothing technique to overcome the lack of smoothness of the objective EEs, and establish the MSEL estimators based on IPW smoothed EEs. Secondly, we propose a novel statistic based on difference between empirical likelihood ratios of the null hypotheses and alternative hypotheses, and discuss its limiting distribution. Finally, we develop two testing procedures based on adjusted test statistic and bootstrap method. Numerical studies show that two proposed testing procedures are both indeed powerful.

The rest of this paper is arranged as follows. In Section 2, we apply smoothing technique and IPW method to construct an asymptotic unbiased estimating equations for missing covariates. The estimation and test procedures based on empirical likelihood method are established in Sections 3 and 4. Simulation studies and a real data analysis are given in Section 5. In Section 6 we provide our conclusions. The technical conditions and proofs of theorems are provided in Appendix.

2. The IPW smoothing estimating equation

Along lines of Zhou et al. (2011), we can estimate β in model (1.1) without any missing data based on the following EEs:

$$\psi(Y, X, \beta) = \begin{pmatrix} \psi_1(Y, X, \beta) \\ \psi_2(Y, X, \beta) \end{pmatrix} = \begin{pmatrix} X(Y - X^T\beta) \\ X[\tau - I(Y - X^T\beta \le c_\tau)] \end{pmatrix}, \tag{2.1}$$

where the first part of (2.1) is based on the LS normal equations, and the second part of (2.1) is based on τ-quantile conditions of ε, i.e., $P(\varepsilon \le c_\tau) = \tau$. By the MAR assumption and the iterative expectation formula, we have

$$E\Big[\frac{\delta_i}{\pi(Z_i, \gamma)}\psi(Y_i, X_i, \beta)\Big] = E\psi(Y_i, X_i, \beta) = \mathbf{0}. \tag{2.2}$$

Therefore, we can show that

$$\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(Z_i, \gamma)}\psi(Y_i, X_i, \beta) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi(Z_i, \gamma)}\begin{pmatrix} \psi_1(Y_i, X_i, \beta) \\ \psi_2(Y_i, X_i, \beta) \end{pmatrix} = \mathbf{0}, \tag{2.3}$$

are unbiased EEs. It is noted that (2.3) is consisting of $2p$ EEs, but only has p unknown parameters. Hence, it is over-determined for estimating β. An ordinary estimating method is infeasible here, but EL method has achieved respectability. However, a major difficulty here is that functions $\psi_2(Y, X, \beta)$ obtained from quantile information are non-differentiable about β. To solve this problem, we introduce the kernel-based technique developed in Zhou et al. (2011) to smooth $\psi_2(Y, X, \beta)$. Consider a smoothing r-order kernel function $l(\cdot)$ satisfying

$$\int u^j l(u)du = \begin{cases} 1 & if\ \ j = 0 \\ 0 & if\ \ 1 \le j \le r-1 \\ \nu_r \ne 0 & if\ \ j = r \end{cases}. \tag{2.4}$$

for some integer $r \ge 2$. The bandwidth $b_n \to 0$ and $nb_n \to \infty$ as $n \to \infty$. For the sake of convenience we define the condition τ-quantile function of Y as $q_\tau(X, \beta) = X^T\beta + c_\tau$. In current paper, we consider the smoothed version of $\psi_2(Y, X, \beta)$ as follows:

$$\phi_2(Y, X, \beta) = X\Big\{\tau - \int I\big(Y - q_\tau(X, \beta) + b_n\xi \le 0\big)\mathrm{d}L(\xi)\Big\} = X^T\Big[\tau - L\Big(\frac{q_\tau(X, \beta) - Y}{b_n}\Big)\Big], \tag{2.5}$$

where $L(\xi) = \int_0^\xi l(u)\mathrm{d}u$. Let β_0 be the true value of β. Assume that $Q_X(\mu) = E_X\big(X\tau - I(Y - \mu \le 0)\big)$ is a rth continuously differentiable in the neighborhood of β_0. We have

$$E\phi_2(Y, X, \beta_0) = E\Big\{\int Q_X(\mu)\mathrm{d}L\Big(\frac{\mu - q_\tau(X, \beta_0)}{b_n}\Big)\Big\}$$