Statistics and Probability Letters 127 (2017) 120–130

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Imputation in nonparametric quantile regression with complex data

Yanan Hu [a], Yaqi Yang [a], Chunyu Wang [a], Maozai Tian [a,b,c,*]

[a] Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China

[b] School of Statistics and Information, Xinjiang University of Finance and Economics, 830012, Xinjiang, China

[c] School of Statistics, Lanzhou University of Finance and Economics, Lanzhou, 730101, Gansu

ARTICLE INFO

Article history:
Received 14 October 2016
Received in revised form 4 March 2017
Accepted 4 March 2017
Available online 27 March 2017

Keywords:
Complex data
Missing covariates
Multiple imputation
Quantile regression

ABSTRACT

This paper considers nonparametric quantile regression models for complex data of mixed categorical and continuous variables together with missing values at random (MAR). In consideration of the increasingly popular application of multiple imputation for handling missing data and the advantages of nonparametric quantile regression, we propose an effective and accurate multiple imputation method. The estimation procedure not only does well in modeling with mixed categorical and continuous data, but also makes full use of the entire data set to achieve increased efficiency. The proposed estimator is asymptotically normal. In simulation study, we compare the performance of the multiple imputation method with complete case (CC), Regression imputation and nearest-neighbor imputation methods, and outline advantages and drawbacks of the different methods. Simulation studies show that the proposed multiple imputation method performs better.

1. Introduction

Variables used as covariates often have missing values (Yang and Kim, in press). A simple approach, called a complete-case analysis, uses only observations without missing values. The complete case method can lead to biased estimators and is inefficient relative to procedures that use observations with partial information.

Imputation is a widely used method to exploit the full information contained in the data set. In imputation, the missing values are replaced with plausible values derived under a set of assumptions. Well-known imputation methods include mean imputation and nearest-neighbor imputation.

Two types of imputation procedures are single imputation and repeated imputation. In single imputation, each missing value is replaced with a single imputed value. Examples of single imputation methods include Cheng (1994) and Wang et al. (2004). In repeated imputation, each missing value is replaced with multiple imputed values. Two broad categories of repeated imputation procedures are multiple imputation and fractional imputation Kim and Shao (2013).

Repeated imputation overcomes limitations associated with single imputation (Landerman et al., 1997). The use of more than one imputed value can lead to more precise estimators than those based on a single imputed value. The use of multiple imputed values also allows estimation of the variance associated with the imputation procedure. We focus on a frequentist version of the type of multiple imputation discussed in Little and Rubin (1987). Our multiple imputation paradigm is closely

* Corresponding author at: Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China.
E-mail address: mztian@ruc.edu.cn (M. Tian).

http://dx.doi.org/10.1016/j.spl.2017.03.003
0167-7152/

related to that of Wei et al. (2012). Multiple imputation creates $L > 1$ complete data sets. Each complete data set is analyzed by standard complete-data methods, and the results are then combined.

Quantile regression, as first introduced by (Koenker and Bassett, 1978), is gradually developing into an important modeling tool, due to its flexibility in exploring the relationship between the response and the covariates. Compared with mean regression, quantile regression is more explicable and robust, when the distribution of data is typically skewed or the data contains some outliers. However, it is not a well-developed topic to effectively handle missing values especially missing covariates with quantile regression method. Wei et al. (2012) proposed a multiple imputation estimator for parameter estimation in linear quantile regression models with covariates missing at random. Sherwood et al. (2013) studied a weighted quantile regression approach for estimating the conditional quantiles of health care cost data with missing covariates. Admittedly, the parametric modeling approach is not robust to functional form specification and may introduce a large bias when the model is misspecified. Compared with traditional parametric estimation techniques, nonparametric techniques have been applied to a variety of problems and offer a great deal of flexibility. McMillen and McDonald (1997) pointed out that nonparametric estimation procedures for modeling polycentric cities are flexible enough to account for functional form misspecification, and the nonparametric estimates are more accurate than OLS regression. Stone (1985) considered the problem of estimating a multivariate joint distribution, and showed that the parametric approach starts with the assumption of an a priori model for distribution function that contains finitely many unknown parameters, while the nonparametric approach eschews such an assumption. Wang and Chen (2009) proposed a nonparametric imputation of the missing values from a kernel estimator of the conditional distribution of the missing variable given the always observable variable, and used empirical likelihood to construct a profile likelihood for the parameter.

In applied settings, however, one frequently encounters a mixture of discrete and continuous data. Rather than using the conventional frequency approach to deal with the discrete variables, Li and Racine (2008) proposed a method to smooth the discrete and the continuous variables, which have several attractive features and give us some enlightenment.

The method proposed in this paper has several features, which admits both categorical and continuous data, provides flexibility and robustness benefitting from nonparametric modeling, combines quantile regression with missing data, and develops a multiple imputation to improve estimation efficiency.

The remaining part of the paper is organized as follows. In Section 2 we develop an effective technique of multiple imputation in nonparametric quantile regression with missing covariates and investigate main results of the asymptotic properties for the proposed estimator. Section 3 provides the comparison between multiple imputation method and complete case, regression imputation and neighbor-nearest imputation method through simulation study. Section 4 concludes the paper with some discussions.

2. Estimation

We consider that regressors contain a mix of discrete and continuous variables. Define $X = (X^c, X^d)$, where X^c is a $p \times 1$ continuous random vector, and X^d is a $q \times 1$ discrete random vector. We consider a nonparametric quantile regression model given by

$$Q_\tau(Y|X^c, X^d) = m_\tau(X^c, X^d), \tag{1}$$

where τ is the quantile and lie in $(0, 1)$, and $m(\cdot)$ is an unknown continuously differentiable function. Denote $f(x) = f(x^c, x^d)$ as the joint density function of (X^c, X^d). Here X^d may be missing, while X^c is always observed. We use n for the total sample size, n_1 for the complete observations and n_0 for the observations including X^d missing, i.e. $n = n_1 + n_0$. Thus, observations can be expressed as $\{(y_i, x_i^c, x_i^d) : i = 1, \ldots, n_1\}$ and $\{(y_j, x_j^c, \cdot) : j = n_1 + 1, \ldots, n\}$. The MAR assumption in this paper, following from Wei et al. (2012), is defined as follows:

$$P(\delta = 1|Y, X^c, X^d) = P(\delta = 1|X^c),$$

where δ is an indicator variable and $\delta = 1$ means X^d is missing. As Chen et al. (2015) said, the missing function depends on the observed covariate but not on the response even when the response data are fully observed.

It is well known that the selection of smoothing parameters is of crucial importance to estimate the model (1). Li and Racine (2004) proposed a natural extension of Aitchison and Aitken (1976) work to the problem of mixed categorical and continuous data in a nonparametric regression framework, and obtained smoothing parameters from the least squares cross-validation. According to Li and Racine (2004), we obtain the product kernel function for the categorical variables given by

$$L_\lambda(X_i^d, X_j^d) = \left[\prod_{s=1}^{r_1} \lambda_s^{|X_{is}^d - X_{js}^d|}\right]\left[\prod_{s=r_1+1}^{q} \lambda_s^{I(X_{is}^d \neq X_{js}^d)}\right],$$

where $I(\cdot)$ denote an indicator function, X_{is}^d denote the sth component of X_i^d, and $\lambda_s \in [0, 1]$ is a bandwidth. $X_{i1}^d, \ldots, X_{ir_1}$ are ordinal variables and the left are categorical. The product kernel function used for the continuous variables is given by

$$W_h(X_i^c, X_j^c) = \prod_{s=1}^{p} h_s^{-1} w\left(\frac{X_{is}^c - X_{js}^c}{h_s}\right),$$