**Journal of Mathematics and Statistics 10 (1): 30-44, 2014**
ISSN: 1549-3644
© 2014 Science Publications
doi:10.3844/jmssp.2014.30.44 Published Online 10 (1) 2014 (http://www.thescipub.com/jmss.toc)

# AN EFFECTIVE TECHNIQUE OF MULTIPLE IMPUTATION IN NONPARAMETRIC QUANTILE REGRESSION

**Yanan Hu, Qianqian Zhu and Maozai Tian**

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

Received 2013-10-04; Revised 2013-12-03; Accepted 2014-01-04

## ABSTRACT

In this study, we consider the nonparametric quantile regression model with the covariates Missing at Random (MAR). Multiple imputation is becoming an increasingly popular approach for analyzing missing data, which combined with quantile regression is not well-developed. We propose an effective and accurate two-stage multiple imputation method for the model based on the quantile regression, which consists of initial imputation in the first stage and multiple imputation in the second stage. The estimation procedure makes full use of the entire dataset to achieve increased efficiency and we show the proposed two-stage multiple imputation estimator to be asymptotically normal. In simulation study, we compare the performance of the proposed imputation estimator with Complete Case (CC) estimator and other imputation estimators, e.g., the regression imputation estimator and k-Nearest-Neighbor imputation estimator. We conclude that the proposed estimator is robust to the initial imputation and illustrates more desirable performance than other comparative methods. We also apply the proposed multiple imputation method to an AIDS clinical trial data set to show its practical application.

**Keywords:** Bandwidth Selection, Local Linear Fitting, Missing Covariates, Nonparametric Quantile Regression, Two-stage Multiple Imputation

## 1. INTRODUCTION

Quantile regression has been widely used in analyzing the relationship between response and covariates since its first introduction in (Koenker and Bassett, 1978). Compared with mean regression, quantile regression is able to depict the impact of covariates on various quantiles of the response, which provides more information for analysis. Furthermore, quantile regression is robust to outliers in data and distribution-free for error term. Due to its advantages, quantile regression has illustrated its increasingly importance in modeling and has attracted great attention in data analysis and empirical applications, nonparametric quantile regression modeling is such an example. Consider the following nonparametric regression model:

$$Y = m(X) + \in$$

Where:
$m(\cdot)$ = The unknown real function and
$\in$ = The error term

Based on the above model, we consider the following nonparametric quantile regression model Equation (1.1):

$$Q_\tau(Y \mid X = x) = c_\tau + m(x) \tag{1.1}$$

where quantile $\tau \in (0, 1)$, $Q_\tau(Y|X = x)$ is the $\tau$-th conditional quantile of Y given X = x. $c_\tau$ is the $\tau$-th quantile of error term $\in$ and satisfies $c_\tau = F^{-1}(\in)$, where $F(\in)$ is the unknown distribution function of $\in$. Here, without loss of generality, covariate vector x does not contain constant 1, which means there is no intercept term in m(x) and ensures the identification of the model. This model overcomes many disadvantages of usually used parametric models in

**Corresponding Author:** Maozai Tian, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, China

which misspecification could be encountered. Nonparametric regression does not assume that the relationship between response and covariates to be linear or satisfy some specified form, which might be more reasonable for most of data set and thus more flexible than parametric models. Especially when data set does not present some kind of parametric form, nonparametric regression model could be a plausible choice since it avoids the great bias due to the wrong model form assumption and brings increasing accuracy and more reasonable explanations.

The above nonparametric quantile regression model can be widely applied to many empirical data analysis, where the data set is complete. However, it is unavoidable to face with data set with missing data, so it is necessary to extend the above model to deal with missing data. In practice, missing data is very pervasive and the reasons for missing are various. More details Little and Rubin (1978); Robins *et al*. (1994) and Vach (1994). In this study, we pay more attention to the nonparametric quantile regression model (1.1) with the covariates missing at random, which has the following form Equation (1.2):

$$Q_\tau(x,z) \triangleq Q_\tau(Y \mid X = x, Z = z) = c_\tau + m(x,z) \tag{1.2}$$

where, (X, Z) are covariate vectors, X may be missing whereas Z is all observed in sample interval. Denote n as the sample size. For notation simplicity, we suppose that the first $n_1$ observations are complete while the remaining $n_0$ are missing in X. Therefore, rewrite the sample as $\{(Y_i, X_i, Z_i): i = 1,\cdots, n_1\}$ and $\{(Y_j,\cdot, Z_j): j = n_1+1,..., n\}$. Let δ be a missing indicator whose value is 1 when X is observed and else 0 when X is missing. Then, $\delta_i = 1$ for $i = 1,\cdots, n_1$ while $\delta_i = 0$ for $i = n_1+1,..., n$. Here we assume that X is MAR which takes the form of conditional independence, i.e., X and δ are conditionally independent given (Y, Z) Equation (1.3):

$$P(\delta = 1 \mid Y, X, Z) = P(\delta = 1 \mid Y, Z) \tag{1.3}$$

In order to estimate model (1.2), we may just consider the observed data and ignore the observations with missing values, which is called the CC analysis. Although we can obtain consistent estimator for m(x, z) through CC analysis under MAR assumption, it may be misleading and inefficient when missing rate is high. Therefore, it is necessary to construct a more reasonable estimator to make use of the information in data set, e.g., imputation estimator. Particularly, the multiple imputation methods often bring more reliable inference than single imputation methods and perform better in missing data problems. In this study, we focus on the estimation of model (1.2) under MAR assumption based on local linear fitting and propose an effective and easy-to-use two stage multiple imputation estimator, which improves the estimation efficiency to a large extent.

In the context of mean regression, parametric or nonparametric regression models with missing data have been studied in many papers. Anderson (1957) derived the maximum likelihood estimates of parametric models and Cheng (1994) and Chu and Cheng (1995) studied the nonparametric regression estimation with missing response. Wang and Rao (2001; 2002) and Wang *et al.* (2004) studied the estimation of generalized linear models, linear models, semiparametric models with missing response, respectively. Furthermore, quantile regression models with missing data also have been considered in literature. It should be noted that the above research mainly consider the models with missing response rather than missing covariates. Under mean regression, Liang *et al*. (2004) considered the partially linear model with covariate missing depending on other complete covariates and response. Wu and Wu (2001) proposed a multiple imputation method for missing covariates in non-linear mixed-effects models and applied the proposed method to HIV Dynamics. Robins *et al*. (1994) studied the regression coefficients estimation with missing covariates. Wang (2009) also considered the estimation of partial linear models with covariables data missing at random. With respect to quantile regression, Wei *et al*. (2012) studied the multiple imputation for parametric quantile regression model with missing covariates, which provided a new imputation method. However, nonparametric quantile regression with missing covariates has not been considered up to now. Based on the existing research and methods, we propose a two-stage multiple imputation method for nonparametric quantile regression with missing covariates, which greatly enriches the methods to cope with missing data in quantile regression.

The rest of the paper is organized as follows. In Section 2, we develop nonparametric quantile regression with missing covariates based on a two-stage multiple imputation method and present main results of the asymptotic properties for the proposed estimator. Section 3 compares our methods with regression imputation method, k-Nearest-Neighbour and Nearest-Neighbour methods through simulation study. Discussion is available in Section 4.

# 2 ESTIMATION WITH MULTIPLE IMPUTATION

In this section, we present the estimation of model (1.2) under CC case and propose a two stage multiple imputation estimator. For the CC estimator and the proposed estimator, we further study its large sample properties.

## 2.1. CC Estimator

For model (1.2), we first consider the model estimation under CC case, which is the basis of the two-stage multiple imputation estimator.

To estimate $Q_\tau$ (Y|X = x), the conditional quantile of Y given X = x in model (1.2) under CC case, we apply classical local linear fitting and quantile regression method and consequently have the following objective function Equation (2.1):

$$R_{n_1}(\beta)=\sum_{i=1}^{n_1}\rho_\tau\left(Y_i-\beta_0-\left(X_i-x,Z_i-z\right)\beta_1\right) K_H\left(X_i-x,Z_i-z\right) \tag{2.1}$$

where, $\beta_0 = c_\tau + m(x, z)$, $\beta_1 = (\partial m(x, z)/\partial x, \partial m(x, z)/\partial z)^T$ and $\beta = (\beta_0, \beta_1^T)$. $\rho_\tau(u) = \tau u I_{[0,\infty)}(u)-(1-\tau)u I_{(-\infty,0)}(u)$ is the check function used in quantile regression, which is one kind of loss function. $I(\cdot)$ is the usual indicator function. $K_H(\cdot)$ is the kernel function satisfying $K_H(Z)=\frac{1}{\det(H)}K\left(H^{-1}z\right)$ and H represents the bandwidth matrix.

By minimizing $R_{n1}(\beta)$ in (2.1) with respect to β, we can obtain the estimate of β under CC case Equation (2.2):

$$\hat{\beta}_{n_1}(\tau)=\left(\hat{\beta}_{n_1,0}(\tau),\hat{\beta}^T_{n_1,1}(\tau)\right)=\operatorname*{Arg\,min}_{\beta} R_{n_1}(\beta) \tag{2.2}$$

Via (2.2) we can obtain the conditional quantile estimate of Y using the complete data only, i.e., $\hat{Q}_\tau$ (x, z) = $\hat{\beta}_{n_1,0}(\tau)$, which is the so-called CC estimator.

## 2.2. Two-Stage Multiple Imputation Estimator

In this subsection, we propose a two-stage multiple imputation estimator for model (1.2). The basic idea of the two-stage multiple imputation estimator is to impute the missing data via the estimated conditional density $\hat{f}(x\,|\,y,z)$ and then estimate model (1.2) based on the complete data including imputed data.

To obtain this estimator, two stages are performed, where initial imputation values are realized in the first stage while multiple imputation values are obtained based on these initial imputation values. Then we discuss about these two stages in detail.

### 2.2.1. First-Stage Imputation

In the first stage, we can obtain initial imputation values through many imputation methods. Here we consider the following three methods to get initial imputation values:

- Regression Imputation. Based on the MAR assumption in Section 1 and the dependence of x on z, construct linear regression model for x given z with the complete data and obtain the parametric estimates. Then impute the missing x via the prediction values based on the corresponding z
- k-Nearest-Neighbor Imputation. For j = $n_1$ + 1,...,n, find the k nearest data pairs ($y_l$, $z_l$) (l = 1,..., k) of data pair ($y_j$, $z_j$) in the complete data and the corresponding points $x_l$ (l = 1,..., k) are the k nearest points in distance of missing data $x_j$. Then impute $x_j$ by averaging these points, i.e., $\tilde{x}_j=\frac{1}{k}\sum_{l=1}^{k}x_l$
- Nearest-Neighbor Imputation. Different with k-Nearest-Neighbor imputation, the Nearest-Neighbor imputation just considers the nearest one point of missing x as the imputation value, i.e., k = 1. For j = $n_1$ + 1,...,n, find the nearest data pair ($y_l$, $z_l$) of data pair ($y_j$, $z_j$) in the complete data and the corresponding point $x_l$ are the nearest point in distance of missing data $x_j$. Then impute $x_j$ through this point, i.e., $\tilde{x}_j = x_l$

## Remark:

The first imputation method is based on regression imputation while the third method belongs to matching method. The above regression imputation requires the linear relationship between missing covariate and regression variables. Matching is nonparametric imputation method which allows imputation without estimating conditional distribution of missing variable. Further information about regression imputation and matching method, Little and Rubin (1987) and Chen and Shao (2000).

It should be noted that a reasonable two-stage imputation estimator should be insensitive to the above initial imputation methods. In other words, if our proposed two-stage multiple imputation estimator is effective and reasonable, it should be stable under different initial imputation methods. In the simulation study, we will illustrate the robustness of the proposed estimator to initial imputation. Based on these initial imputation values of missing x, we can estimate the conditional density f(y|x, z) and then obtain the estimated conditional density $\hat{f}(x \mid y,z)$. We discuss the above estimation process in the following second-stage imputation.

### 2.2.2. Second-Stage Imputation

In this stage, we realize the multiple imputation based on the estimated conditional density $\hat{f}(x \mid y,z)$ and estimate model (1.2) using the whole data after multiple imputation. This stage can be concluded as the following steps:

Step 1: Estimate conditional density f(x|y, z). According to Bayes formula, f(x|y, z) ∝ f(y|x, z)f(x|z). It is reasonable to estimate f(x|y, z) through estimating f(x|z) and f(y|x, z) respectively, which can be realized via the following steps.

Step 1a: Estimate conditional density f(x|z). Model x given z parametrically as f(x|z, η) and obtain the estimate $\hat{\eta}$ and the estimated conditional density $\hat{f}(x \mid z)$ of x given z can be denoted as $f(x \mid z, \hat{\eta})$.

Step 1b: Estimate conditional density f(y|x, z). The quantile function is the inverse distribution function, so the density function can be expressed as the reciprocal of the first derivative of the quantile function at the corresponding quantile level. Here we choose $K_n$ quantile levels $\tau_k = k/(K_n + 1)$ ($k = 1,..., K_n$), similarly and approximate the conditional density f(y|x, z) as follows Equation (2.3):

$$\hat{f}\left(y \mid x, \hat{Q}_\tau(x,z)\right) = \sum_{k=1}^{K_n} \frac{\tau_{k+1} - \tau_k}{\hat{Q}_{\tau_{k+1}}(x,z) - \hat{Q}_{\tau_k}(x,z)} \\ I\left\{\hat{Q}_{\tau_k}(x,z) \le y < \hat{Q}_{\tau_{k+1}}(x,z)\right\} \tag{2.3}$$

where, $\hat{Q}_{\tau k}(x,z)$ is the estimated $\tau_k$-th conditional quantile of Y in model (1.2) with the whole data set including initial imputed data from first-stage imputation.

At last, normalize $\hat{f}(y \mid x,z)\hat{f}(x \mid z)$ to be a density, then we get the estimated conditional density $\hat{f}(x \mid y,z)$:

Step 2: Multiple imputation based on estimated conditional density $\hat{f}(x \mid y,z)$. First, obtain empirical distribution function $\hat{F}(x \mid y,z)$ via estimated conditional density $\hat{f}(x \mid y,z)$. Then draw random numbers $u_i^*, i = 1,..., n_0$ from uniform distribution U(0, 1). Finally, regard $u_i^*, i = 1,..., n_0$ as the quantile levels and obtain the corresponding quantiles from empirical distribution function $\hat{F}(x \mid y,z)$, which can be seen as the imputation values.

Step 3: Estimation of model (1.2) using the whole data after multiple imputation. Consider a new objective function including the observed data and the l-th imputed data set as follows:

$$\begin{aligned} R_{n(l)}(\beta) &= \sum_{i=1}^{n_1} \rho_\tau\left(Y_i - \beta_0 - \left(X_i - x, Z_i - z\right)\beta_1\right) \\ &K_H\left(X_i - x, Z_i - z\right) \\ &+ \sum_{i=n_1+1}^{n_1} \rho_\tau\left(Y_i - \beta_0 - \left(X_i - x, Z_i - z\right)\beta_1\right) \\ &K_H\left(X_i - x, Z_i - z\right) \\ &= \sum_{i=1}^{n_1} \rho_\tau\left(Y_j - \left(1, \left(X_i - x\right)^T, \left(Z_i - z\right)^T\right)\beta\right) \\ &K_H\left(X_i - x, Z_i - z\right) \\ &+ \sum_{j=n_1+1}^{n} \rho_\tau\left(Y_j - \left(1, \left(X_{j(l)} - x\right)^T \left(Z_j - z\right)^T\right)\beta\right) \\ &K_H\left(X_{j(l)} - x, Z_j - z\right) \end{aligned}$$

Minimize $R_{n(l)}(\beta)$, we have $\hat{\beta}_{n(l)}(\tau) = \left(\hat{\beta}_{n(l),0}(\tau), \hat{\beta}^T_{n(l),1}(\tau)\right) = \operatorname{Arg}\min_\beta R_{n(l)}(\beta)$ as the estimated coefficient under the l-th imputation data. Repeat the imputation estimation step L times and obtain the two-stage multiple imputation estimator Equation (2.4):

$$\hat{\beta}_0^*(\tau) = \frac{1}{L}\sum_{l=1}^{L} \hat{\beta}_{n(l),0}(\tau) \tag{2.4}$$

For the two-stage multiple imputation estimator $\hat{\beta}*(\tau)$ obtained based on the above two-stage imputation, we derive its asymptotic properties in the following subsection.

### 2.3. Large Sample Properties

In this section, we give the asymptotic distribution of the two-stage multiple imputation estimator $\hat{\beta}^*(\tau)$. Let $h(\tau; X,Z) = 1/Q_\tau(X, Z)$ be the density of Y given X and Z at τ-th quantile. Recall that $\hat{\beta}_{n_1}(\tau)$ is the CC estimator in section 2.1 and $\hat{\beta}_{n(l)}(\tau)$ is the estimator obtained from the objective function based on the whole data including the l-th imputed data of missing values in section 2.2.2. Define the following objective function:

$$\tilde{R}_{n_0(l)}(\beta) = \sum_{j=n_1+1}^{n} \rho_\tau \left\{ Y_i - \left(1, \left(\tilde{X}_{j(l)} - x\right)^T, \left(Z_j - z\right)^T\right)\beta \right\}$$
$$K_H\left(\tilde{X}_{j(l)} - x, Z_j - z\right)$$

and denote $\hat{\beta}_{n_0}(\tau)$ as the estimators obtained from $\tilde{R}_{n_0(l)}(\beta)$ based on the imputed data of missing values only, i.e., $\hat{\beta}_{n_0}(\tau) = \underset{\beta}{\mathrm{Argmin}}\tilde{R}_{n_0(l)}(\beta)$.

The above three estimators are the basis for the two-stage multiple imputation estimator $\hat{\beta}^*(\tau)$. Then define the functions as follows:

$$H_0(\beta) = E_{(Y,X,Z)}\begin{bmatrix}\rho_\tau\left\{Y - \left(1, (X-x)^T, (Z-z)^T\right)\beta\right\} \\ K_H(X-x, Z-z)\end{bmatrix},$$

$$H_0(\beta) = E_{(Y,\tilde{X},Z)}\begin{bmatrix}\rho_\tau\left\{Y - \left(1, (\tilde{X}-x)^T, (Z-z)^T\right)\beta\right\} \\ K_H(\tilde{X}-x, Z-z)\end{bmatrix}$$

where, (X,Y,Z) is the observed data set while $\left(\tilde{X}, Y, Z\right)$ is the imputed data set. Given (Y,Z), $\tilde{X}$ follows the conditional distribution $\hat{f}(x \mid y,z)$.

To obtain the asymptotic properties for $\hat{\beta}^*(\tau)$, we list the following assumptions needed in proof.

### Assumption 1:

There exists a $\beta(\tau) \in R^p$ such that β(τ) uniquely minimizes the objective function $H_0(\beta)$, i.e., $\beta(\tau) = \mathrm{Argmin}_\beta H_0(\beta)$.

### Assumption 2:

There exists a compact set $\Omega \in R^p$ and $\beta_{(l)}(\tau) \in \Omega$, such that $\beta_{(l)}(\tau) \mathrm{Arg\,min}_\beta \tilde{H}_0(\beta)$.

### Assumption 3:

The covariate X has bounded support χ. The true conditional density $f(x|z) = f(x|z, \eta = \eta_0)$, where $f(x|z, \eta)$ is a continuous function of η uniformly for (x, z) in a neighbourhood of $\eta_0$ and is bounded away from zero and infinity for all (x, z).

### Assumption 4:

The true coefficient functions $\beta_0(\tau)$ are smooth functions on (0, 1) and for any $X \in \chi$ and Z:

- $0 < h(\tau; x, z) < \infty$ and $\lim_{\tau\to 0} h(\tau; x, z) = \lim_{\tau\to 1} h(\tau; x, z) = 0$
- There exist constants M and $\nu_1, \nu_2 > -1$, such that the first derivative of $h(\cdot)$ satisfies:

$$\sup_x |h'(\tau; x, z)| < M\tau^{\nu_1}(1-\tau)^{\nu_2}$$

### Assumption 5:

The matrix $\Psi_\tau = (\partial/\partial\beta(\tau))E[\phi_\tau(Y_i - (1, (X_i-x)^T, (Z_i-z)^T)\beta(\tau))K_h(X_i-x, Z_i-z)(1, (X_i-x)^T, (Z_i-z)^T)^T]$, is positive definite, where $\phi_\tau(u) = \tau - I(u < 0)$.

### Assumption 6:

The d-dimensional kernel function K(·) is a bounded density function with a compact support $C^d$ within the interior of the support of f(x) such that $\int K(u)du = 1$, $\int uK(u)du = 0_d$, $\int uu^T K(u)du > 0_{d\times d}$.

### Assumption 7:

The bandwidth matrix H of the kernel function satisfies det(H) → 0 and n · det(H) → ∞, as n→∞.

### Remark:

The Assumption 1 and Assumption 2 ensure the existence of solutions for objective functions. Assumption 3 and Assumption 4 focus on the conditional density f(y|x, z). Assumption 6 and Assumption 7 are common in nonparametric estimation, which represent the assumptions for kernel functions and bandwidths, respectively.

Additionally, we also make the following definitions:

$$V_1 = \operatorname{Var}\left[\begin{array}{l} \phi_\tau\left(Y_i\left(1,\left(X_i-x\right)^T,\left(Z_i-z\right)^T\right)\beta(\tau)\right) \\ K_H\left(X_i-x,Z_i-z\right)\times\left(1,\left(X_i-x\right)^T,\left(Z_i-z\right)^T\right)^T \end{array}\right]$$

$$V_0 = \lim_{n\to\infty}\operatorname{Var}\left[\begin{array}{l} \phi_\tau\left(Y_j\left(1,\left(\tilde{X}_{j(l)}-x\right)^T,\left(Z_j-z\right)^T\right)\beta(\tau)\right) \\ K_H\left(\tilde{X}_{j(l)}-x,Z_j-z\right) \\ \times\left(1,\left(\tilde{X}_{j(l)}-x\right)^T,\left(Z_j-z\right)^T\right)^T \end{array}\right]$$

$$U_0 = \underset{n\to\infty}{\operatorname{Cov}}\left[\begin{array}{l} \phi_\tau\left(Y_j-\left(1,\left(\tilde{X}_{j(l)}-x\right)^T,\left(Z_j-z\right)^T\right)\beta(\tau)\right) \\ K_h\left(\tilde{X}_{j(l)}-x,Z_j-z\right) \\ \times\left(1,\left(\tilde{X}_{j(l)}-x\right)^T\left(Z_j-z\right)^T\right)^T, \\ \phi_\tau\left(Y_j-\left(1,\left(\tilde{X}_{j(l')}-x\right)^T,\left(Z_j-z\right)^T\right)\beta(\tau)\right) \\ \times K_h\left(\tilde{X}_{j(l')}-x,Z_j-z\right)\left(1,\left(\tilde{X}_{j(l)}-x\right)^T,\left(Z_j-z\right)^T\right)^T \end{array}\right]$$

Based on the above regularity conditions, the two-stage multiple imputation estimator $\hat{\beta}*(\tau)$ has the asymptotic distribution in the following theorem.

### Theorem 1:

Under (1.3) and the above Assumptions 1-7, for $K_n \to \infty$ and $K_n n^{-1} \to 0$, the multiple estimator $\sqrt{n.\det(H)}\left(\hat{\beta}*(\tau)-\beta(\tau)\right)$ converges in distribution to a multivariate Gaussian vector. Specifically:

$$\sqrt{n.\det(H)}\left(\hat{\beta}^*(\tau)-\beta(\tau)\right) \to N\left(0,\Psi_\tau^{-1}\sum\Psi_\tau\right)$$

Where:

$$\sum = (\lambda+1)^{-1}V_1 + (1+1/\lambda)^{-1}\left[L^{-1}V_0+\left\{(L-1)/L\right\}U_0\right]$$

Based on the Theorem 1, we gives the large sample property of conditional quantile estimator $\hat{Q}_\tau(x,z)$ as follows.

### Theorem 2:

Based on (1.3), the above Assumptions 1-7 and Theorem 1, $\hat{\beta}_0^*(\tau)=\hat{Q}_\tau(y)=e^T\hat{\beta}*(\tau)$ has the following asymptotic distribution:

$$\sqrt{n.\det(H)}\left(\hat{\beta}_0^*(\tau)-\beta_0(\tau)\right) \to N\left(0,e^T\Psi_\tau^{-1}\sum\Psi_\tau e\right)$$

where, e = $(1,0,...,0)^T$.

More details of the proofs are available all request.

### 2.4. Bandwidth Selection

It is well known that the selection of bandwidths in nonparametric regression estimation is of vital significance. The nonparametric estimation results depend on the bandwidth selection to a large extent. Silverman (1986) pointed out that the choice of bandwidth is much more important than the choice of kernel function. Thus, it is necessary to choose reasonable bandwidths to improve the performance of estimation. There are many bandwidth selection methods, such as Plug-in method and cross-validation method. Based on the bandwidth selection in mean regression and quantile regression proposed in Yu and Jones (1998) and Silverman (1986), we discuss about the selection of bandwidths in estimating model (1.2).

According to Yu and Jones (1998), we have the following bandwidth selection formula for quantile regression Equation (2.5):

$$h_\tau = h_{mean}\left\{\frac{\tau(1-\tau)}{\phi\left(\Phi^{-1}(\tau)\right)^2}\right\}^{1/5} \tag{2.5}$$

where, $\tau$ is the quantile level, $h_\tau$ is the optimal bandwidth for the $\tau$-th quantile regression, $h_{mean}$ is the optimal bandwidth for the mean regression estimation, $\varphi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the standard normal distribution respectively.

In terms of the optimal bandwidth for the mean regression estimation $h_{mean}$, we choose the Silverman's rule-of-thumb bandwidth, i.e., $h_{mean} \approx 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ can be the sample estimator of standard deviation $\sigma$. Based on (2.5) and the rule-of-thumb bandwidth, we obtain the optimal bandwidth for model (1.2).

## 3. NUMERICAL SIMULATION

In this section, we implement three simulation examples to illustrate the finite sample performance of the two-stage multiple imputation estimator and compare the performance of the proposed imputation estimator with the CC estimator and other imputation estimators. Specifically, we utilize the two-stage multiple imputation based on the three initial imputation methods in section 2.2.1 and compare these results with that of the first-stage imputation and the CC estimator. Denote the CC estimator, the three first-stage imputation estimators (the regression imputation estimator, the k-Nearest-Neighbor imputation estimator and the Nearest-Neighbor imputation estimator) and the two-stage multiple imputation estimator based on the above three initial imputation methods as CC, RI, kNN, NN, TSMI1, TSMI2 and TSMI3, respectively.

The first example represents the linear case of function m(x, z) to be estimated while the second example is on behalf of nonlinear case of function m(x, z) and the third example stands for the heteroscedastic case. In both the three simulation examples, we consider different sample sizes n = 60, 120 and 200, respectively and distinct missing probability function P(y, z) under different quantile levels τ = 0.25, 0.5 and 0.75. In terms of kernel function in estimation, we choose Gaussian kernel $K(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$ and product kernel $K_h(x, z) = K_h(x)K_h(z)$. For the selection of bandwidths, here we choose bandwidths for the above 7 estimators according to the selection rule in section 2.4.

For the missing probability function P(y, z), we choose the following three functions

### Case 1:

$$\Delta_1 = P(y,z) = P(\delta=1 \mid Y=y, Z=z)$$
$$= \frac{1}{1+\exp\{-\ln(2)-0.2(y-4)-0.1(z-4)\}}$$

### Case 2:

$$\Delta_2 = P(y,z) = P(\delta=1 \mid Y=y, Z=z)$$
$$= \frac{1}{1+\exp\{0.1-0.1(y-3)-0.2(z-2)\}}$$

### Case 3:

$$\Delta_3 = P(y,z) = P(\delta=1 \mid Y=y, Z=z)$$
$$= \frac{1}{1+\exp\{0.5-0.2(y-3)-0.1(z-3)\}}$$

Through the above missing cases, we can study the efficiency of each estimator under different missing rates.

In order to evaluate and compare the performance of the proposed three estimators and other 4 estimators, we calculate the Mean Square Error (MSE) as follows:

$$\mathrm{MSE} = n^{-1}\sum_{i=1}^{n}\left(Q_\tau(X_i, Z_i) - \hat{Q}_\tau(X_i, Z_i)\right)^2$$

and replicate the three simulations k = 100, respectively, to obtain the Average Mean Square Error (AMSE), $\mathrm{AMSE}=\frac{1}{k}\sum_{j=1}^{k}\mathrm{MSE}_j$ . Furthermore, we also calculate the Asymptotic Relative Efficiency (ARE) of our proposed estimators with CC estimator and first-stage imputation estimators. For instance, the ARE of our proposed estimator TMSI1 with the CC estimator is $\mathrm{ARE}_{\mathrm{TMSI}} = \frac{\mathrm{AMSE}_{\mathrm{CC}}}{\mathrm{AMSE}_{\mathrm{TMSI1}}}$, where AMSECC is the AMSE of the CC estimator while AMSETMSI1 is the AMSE of TMSI1.

### Example 1:

Consider the following linear quantile regression model Equation (3.1):

$$Y_i = 1 + X_i + Z_i + \epsilon_i \tag{3.1}$$

where the covariate (X, Z) are jointly normal with mean vector $(4, 4)^T$, variance $(1, 1)^T$ and correlation 0.5 and $\in$ is from standard normal distribution N(0, 1). For this model, we consider the above three missing cases. **Table 1 and 4** illustrate the AMSE values of the 7 estimators and ARE values of our proposed estimators with other estimators for model (3.1), respectively.

### Example 2:

Then we consider a nonlinear function for m(x, z). The model is Equation (3.2):

$$Y_i = 1 + X_i + \sin(Z_i) + \varepsilon_i \tag{3.2}$$

**Table 1.** AMSE of Linear Model (4.1)

| | P(y, z) | n | CC | RI | kNN | NN | TSMI1 | TSMI2 | TSMI3 |
|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | $\Delta_1$ | 60 | 0.4163 | 0.4936 | 0.4609 | 0.4780 | 0.3641 | 0.3633 | 0.3632 |
| | | 120 | 0.0510 | 0.0697 | 0.0623 | 0.0705 | 0.0410 | 0.0413 | 0.0412 |
| | | 200 | 0.2084 | 0.3170 | 0.2924 | 0.3180 | 0.1848 | 0.1849 | 0.1848 |
| | $\Delta_2$ | 60 | 0.4558 | 0.6174 | 0.5595 | 0.6034 | 0.3722 | 0.3701 | 0.3680 |
| | | 120 | 0.3023 | 0.4626 | 0.4228 | 0.4599 | 0.2425 | 0.2417 | 0.2419 |
| | | 200 | 0.2224 | 0.4137 | 0.3774 | 0.4224 | 0.1853 | 0.1852 | 0.1846 |
| | $\Delta_3$ | 60 | 0.4729 | 0.6123 | 0.5488 | 0.5779 | 0.3687 | 0.3697 | 0.3676 |
| | | 120 | 0.3221 | 0.4975 | 0.4515 | 0.4887 | 0.2433 | 0.2435 | 0.2427 |
| | | 200 | 0.2364 | 0.4308 | 0.3949 | 0.4330 | 0.1846 | 0.1846 | 0.1842 |
| $\tau = 0.5$ | $\Delta_1$ | 60 | 0.3243 | 0.3280 | 0.3281 | 0.3463 | 0.3020 | 0.3021 | 0.3012 |
| | | 120 | 0.2241 | 0.2283 | 0.2248 | 0.2321 | 0.2069 | 0.2067 | 0.2068 |
| | | 200 | 0.1712 | 0.1733 | 0.1760 | 0.1827 | 0.1573 | 0.1571 | 0.1570 |
| | $\Delta_2$ | 60 | 0.3568 | 0.3587 | 0.3725 | 0.3778 | 0.3070 | 0.3061 | 0.3052 |
| | | 120 | 0.2584 | 0.2547 | 0.2521 | 0.2677 | 0.2075 | 0.2073 | 0.2073 |
| | | 200 | 0.1824 | 0.1874 | 0.2006 | 0.1919 | 0.1573 | 0.1571 | 0.1572 |
| | $\Delta_3$ | 60 | 0.3530 | 0.5243 | 0.4794 | 0.4996 | 0.3092 | 0.3096 | 0.3092 |
| | | 120 | 0.2545 | 0.4393 | 0.3955 | 0.4295 | 0.2099 | 0.2090 | 0.2093 |
| | | 200 | 0.1935 | 0.3904 | 0.3601 | 0.3986 | 0.1566 | 0.1564 | 0.1563 |
| $\tau = 0.75$ | $\Delta_1$ | 60 | 0.3734 | 0.4609 | 0.4262 | 0.4742 | 0.3434 | 0.3438 | 0.3414 |
| | | 120 | 0.2574 | 0.3681 | 0.3315 | 0.3694 | 0.2408 | 0.2406 | 0.2405 |
| | | 200 | 0.1943 | 0.3093 | 0.2750 | 0.3136 | 0.1770 | 0.1766 | 0.1769 |
| | $\Delta_2$ | 60 | 0.4120 | 0.5785 | 0.5265 | 0.5687 | 0.3415 | 0.3403 | 0.3397 |
| | | 120 | 0.2925 | 0.4645 | 0.4115 | 0.4620 | 0.2412 | 0.2406 | 0.2400 |
| | | 200 | 0.2116 | 0.4103 | 0.3672 | 0.4233 | 0.1762 | 0.1758 | 0.1755 |
| | $\Delta_3$ | 60 | 0.3914 | 0.5740 | 0.5230 | 0.5534 | 0.3476 | 0.3447 | 0.3458 |
| | | 120 | 0.2983 | 0.4848 | 0.4394 | 0.4864 | 0.2417 | 0.2414 | 0.2417 |
| | | 200 | 0.2160 | 0.4186 | 0.3776 | 0.4365 | 0.1760 | 0.1757 | 0.1752 |

where the covariate (X, Z) are jointly normal with mean vector $(4, 4)^T$, variance $(1, 1)^T$ and correlation 0.5 and $\in \sim N(0, 1)$. For model (3.2), we also choose the above three missing probability functions. The AMSE values of the 7 estimators and ARE values of our proposed estimators with other estimators for model (3.2) are given in **Table 2 and 5**, respectively.

## Example 3:

A remarkable advantage of quantile regression is that it does not require strict assumptions on error distribution, which brings us convenience to analyze model with heteroscedasticity. Thus, here we consider the following heteroscedastic model Equation (3.3):

$$Y_i = X_i Z_i + Z_i \in_i \tag{3.3}$$

where the covariate (X, Z) are jointly normal with mean vector $(4, 4)^T$, variance $(1, 1)^T$ and correlation 0.5 and $\in$ is from standard normal distribution. For model (3.3), we still use the above three missing cases. The AMSE values of the 7 estimators and ARE values of our proposed estimators with other estimators for model (3.3) are shown in **Table 3 and 6** respectively.

### 3.1. Simulation Results Analysis

**Table 1-3** illustrate the estimation results AMSE of CC, RI, kNN, NN, TSMI1, TSMI2 and TSMI3 estimators for model (3.1), model (3.2), model (3.3), respectively, with different sample sizes n = 60, 120 and 200, respectively and distinct missing probability function p(y, z) under different quantile levels $\tau = 0.25$, 0.5 and 0.75. From these tables, overall, the estimation effects are the best under $\tau = 0.5$ for all the 7 estimators, which is consistent with the conclusions for quantile regression models. Via the comparison of AMSE values for the 7 estimators under the same sample size, the same missing function and the same quantile level, we conclude that the estimation performance of our proposed estimators TSMI1, TSMI2 and TSMI3 is uniformly better than that of the CC estimator and the initial imputation estimators. Compared with the CC estimator, the initial imputation estimators RI, kNN and NN have similar estimation results and even perform worse than the CC estimator, while our proposed estimators improve a lot than the CC estimator. Apparently, it is necessary to use our proposed estimators to improve estimation performance.

**Table 2.** AMSE of Nonlinear Model (4.2)

| | P(y, z) | n | CC | RI | kNN | NN | TSMI1 | TSMI2 | TSMI3 |
|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | $\Delta_1$ | 60 | 0.4871 | 0.6165 | 0.5865 | 0.6181 | 0.3819 | 0.3809 | 0.3800 |
| | | 120 | 0.3221 | 0.4965 | 0.4462 | 0.4806 | 0.2491 | 0.2496 | 0.2489 |
| | | 200 | 0.2338 | 0.4306 | 0.4004 | 0.4373 | 0.1880 | 0.1877 | 0.1879 |
| | $\Delta_2$ | 60 | 0.5147 | 0.6914 | 0.6299 | 0.6617 | 0.3796 | 0.3773 | 0.3768 |
| | | 120 | 0.3515 | 0.5555 | 0.5460 | 0.5503 | 0.2506 | 0.2500 | 0.2493 |
| | | 200 | 0.2663 | 0.5086 | 0.4749 | 0.5003 | 0.1865 | 0.1864 | 0.1860 |
| | $\Delta_3$ | 60 | 0.5963 | 0.7959 | 0.7451 | 0.7724 | 0.3743 | 0.3756 | 0.3747 |
| | | 120 | 0.4225 | 0.6848 | 0.6454 | 0.6433 | 0.2531 | 0.2530 | 0.2520 |
| | | 200 | 0.1800 | 0.3487 | 0.3393 | 0.3359 | 0.1090 | 0.1087 | 0.1087 |
| $\tau = 0.5$ | $\Delta_1$ | 60 | 0.3706 | 0.5238 | 0.4926 | 0.5177 | 0.3119 | 0.3107 | 0.3116 |
| | | 120 | 0.2690 | 0.4490 | 0.3942 | 0.4328 | 0.2099 | 0.2101 | 0.2102 |
| | | 200 | 0.1886 | 0.3925 | 0.3644 | 0.3956 | 0.1570 | 0.1570 | 0.1569 |
| | $\Delta_2$ | 60 | 0.3946 | 0.5825 | 0.5467 | 0.5759 | 0.3076 | 0.3075 | 0.3066 |
| | | 120 | 0.2795 | 0.5003 | 0.4664 | 0.4848 | 0.2112 | 0.2108 | 0.2111 |
| | | 200 | 0.2098 | 0.4653 | 0.4184 | 0.4598 | 0.1577 | 0.1574 | 0.1572 |
| | $\Delta_3$ | 60 | 0.4650 | 0.6814 | 0.6446 | 0.6813 | 0.3092 | 0.3092 | 0.3087 |
| | | 120 | 0.3316 | 0.5969 | 0.5556 | 0.5729 | 0.2125 | 0.2130 | 0.2128 |
| | | 200 | 0.2649 | 0.5712 | 0.5501 | 0.5640 | 0.1675 | 0.1672 | 0.1669 |
| $\tau = 0.75$ | $\Delta_1$ | 60 | 0.4088 | 0.5579 | 0.5391 | 0.5880 | 0.3418 | 0.3404 | 0.3405 |
| | | 120 | 0.2982 | 0.4920 | 0.4325 | 0.4922 | 0.2404 | 0.2403 | 0.2404 |
| | | 200 | 0.2184 | 0.4252 | 0.3841 | 0.4328 | 0.1754 | 0.1753 | 0.1751 |
| | $\Delta_2$ | 60 | 0.4024 | 0.6049 | 0.5800 | 0.6209 | 0.3418 | 0.3396 | 0.3397 |
| | | 120 | 0.3186 | 0.5571 | 0.5110 | 0.5326 | 0.2402 | 0.2401 | 0.2405 |
| | | 200 | 0.2385 | 0.4916 | 0.4382 | 0.4952 | 0.1753 | 0.1752 | 0.1752 |
| | $\Delta_3$ | 60 | 0.4746 | 0.7244 | 0.7044 | 0.7510 | 0.3378 | 0.3385 | 0.3393 |
| | | 120 | 0.3572 | 0.6397 | 0.6019 | 0.6449 | 0.2427 | 0.2420 | 0.2421 |
| | | 200 | 0.1612 | 0.3413 | 0.3247 | 0.3478 | 0.1030 | 0.1028 | 0.1029 |

**Table 3.** AMSE of Heteroscedastic Model (4.3)

| | P(y, z) | n | CC | RI | kNN | NN | TSMI1 | TSMI2 | TSMI3 |
|---|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | $\Delta_1$ | 60 | 7.4379 | 7.8215 | 7.3706 | 7.4371 | 6.6702 | 6.6908 | 6.6903 |
| | | 120 | 4.7398 | 5.4485 | 4.8840 | 5.0029 | 4.3043 | 4.3098 | 4.3055 |
| | | 200 | 3.8177 | 4.5728 | 4.0704 | 4.1675 | 3.4440 | 3.4482 | 3.4436 |
| | $\Delta_2$ | 60 | 8.1419 | 9.1333 | 8.0609 | 8.3306 | 6.6151 | 6.6442 | 6.6226 |
| | | 120 | 5.2419 | 6.7107 | 5.9150 | 6.1060 | 4.2878 | 4.2907 | 4.2882 |
| | | 200 | 4.1749 | 5.7374 | 5.0413 | 5.2337 | 3.4457 | 3.4407 | 3.4433 |
| | $\Delta_3$ | 60 | 8.1459 | 8.9268 | 8.0955 | 8.0714 | 6.7563 | 6.7274 | 6.7320 |
| | | 120 | 5.3136 | 6.5546 | 5.6952 | 5.7854 | 4.3051 | 4.2933 | 4.2965 |
| | | 200 | 4.2479 | 5.4410 | 4.7731 | 4.8984 | 3.4324 | 3.4298 | 3.4316 |
| $\tau = 0.5$ | $\Delta_1$ | 60 | 6.6802 | 7.0337 | 6.8267 | 7.0758 | 6.1920 | 6.2081 | 6.2041 |
| | | 120 | 4.0949 | 4.8229 | 4.4451 | 4.5662 | 3.8816 | 3.8790 | 3.8857 |
| | | 200 | 3.1105 | 3.9112 | 3.5443 | 3.7113 | 2.9140 | 2.9132 | 2.9120 |
| | $\Delta_2$ | 60 | 6.6055 | 7.7472 | 7.3155 | 7.3147 | 5.6392 | 5.6482 | 5.6207 |
| | | 120 | 4.3871 | 5.8509 | 5.1849 | 5.5298 | 3.8602 | 3.8595 | 3.8555 |
| | | 200 | 3.4791 | 5.0172 | 4.4262 | 4.7402 | 2.9112 | 2.9082 | 2.9088 |
| | $\Delta_3$ | 60 | 6.5070 | 7.4823 | 7.1850 | 6.9774 | 5.7121 | 5.6902 | 5.6897 |
| | | 120 | 4.4310 | 5.6454 | 5.0598 | 5.3750 | 3.9054 | 3.9016 | 3.8931 |
| | | 200 | 3.3833 | 4.6930 | 4.2232 | 4.5003 | 2.8648 | 2.8555 | 2.8583 |
| $\tau = 0.75$ | $\Delta_1$ | 60 | 6.5826 | 7.2507 | 7.0541 | 7.1712 | 6.3269 | 6.3502 | 6.3015 |
| | | 120 | 4.8126 | 5.5770 | 5.2269 | 5.4253 | 4.6149 | 4.6222 | 4.6284 |
| | | 200 | 3.5127 | 4.3220 | 3.9366 | 4.2053 | 3.3541 | 3.3496 | 3.3501 |
| | $\Delta_2$ | 60 | 7.3087 | 8.8290 | 8.2471 | 8.4216 | 6.3694 | 6.3552 | 6.3519 |
| | | 120 | 5.1900 | 6.6381 | 6.1926 | 6.4209 | 4.6202 | 4.6137 | 4.6119 |
| | | 200 | 3.8991 | 5.5529 | 4.9208 | 5.3910 | 3.3481 | 3.3431 | 3.3457 |
| | $\Delta_3$ | 60 | 7.0599 | 8.1882 | 7.6723 | 7.8852 | 6.5087 | 6.4697 | 6.4990 |
| | | 120 | 4.9981 | 6.3682 | 5.8842 | 6.2399 | 4.6580 | 4.6503 | 4.6301 |
| | | 200 | 3.8518 | 5.1829 | 4.6411 | 5.1500 | 3.3463 | 3.3395 | 3.3455 |

**Table 4.** ARE of Linear Model (4.1)

| | P(y, z) | n | $\frac{CC}{TSMI1}$ | $\frac{R1}{TSMI1}$ | $\frac{CC}{TSMI2}$ | $\frac{kNN}{TSMI2}$ | $\frac{CC}{TSMI3}$ | $\frac{NN}{TSMI3}$ |
|---|---|---|---|---|---|---|---|---|
| τ = 0.25 | $\Delta_1$ | 60 | 1.1434 | 1.3556 | 1.1460 | 1.2687 | 1.1463 | 1.3162 |
| | | 120 | 1.2439 | 1.6996 | 1.2339 | 1.5079 | 1.2369 | 1.7090 |
| | | 200 | 1.1276 | 1.7152 | 1.1270 | 1.5814 | 1.1272 | 1.7203 |
| | $\Delta_2$ | 60 | 1.2247 | 1.6587 | 1.2315 | 1.5115 | 1.2388 | 1.6398 |
| | | 120 | 1.2470 | 1.9080 | 1.2509 | 1.7492 | 1.2498 | 1.9011 |
| | | 200 | 1.2002 | 2.2321 | 1.2010 | 2.0378 | 1.2051 | 2.2882 |
| | $\Delta_3$ | 60 | 1.2824 | 1.6605 | 1.2790 | 1.4844 | 1.2864 | 1.5720 |
| | | 120 | 1.3238 | 2.0452 | 1.3228 | 1.8543 | 1.3272 | 2.0138 |
| | | 200 | 1.2809 | 2.3340 | 1.2809 | 2.1396 | 1.2836 | 2.3512 |
| τ = 0.5 | $\Delta_1$ | 60 | 1.0736 | 1.0860 | 1.0733 | 1.0860 | 1.0767 | 1.1498 |
| | | 120 | 1.0835 | 1.1037 | 1.0844 | 1.0874 | 1.0840 | 1.1226 |
| | | 200 | 1.0884 | 1.1015 | 1.0899 | 1.1207 | 1.0901 | 1.1632 |
| | $\Delta_2$ | 60 | 1.1625 | 1.1685 | 1.1659 | 1.2171 | 1.1690 | 1.2376 |
| | | 120 | 1.2456 | 1.2278 | 1.2463 | 1.2156 | 1.2464 | 1.2913 |
| | | 200 | 1.1600 | 1.1919 | 1.1612 | 1.2767 | 1.1606 | 1.2213 |
| | $\Delta_3$ | 60 | 1.1418 | 1.6957 | 1.1405 | 1.5486 | 1.1417 | 1.6157 |
| | | 120 | 1.2121 | 2.0924 | 1.2173 | 1.8920 | 1.2159 | 2.0524 |
| | | 200 | 1.2360 | 2.4936 | 1.2375 | 2.3030 | 1.2382 | 2.5504 |
| τ = 0.75 | $\Delta_1$ | 60 | 1.0873 | 1.3421 | 1.0862 | 1.2397 | 1.0938 | 1.3889 |
| | | 120 | 1.0689 | 1.5286 | 1.0696 | 1.3776 | 1.0702 | 1.5359 |
| | | 200 | 1.0982 | 1.7479 | 1.1004 | 1.5573 | 1.0984 | 1.7725 |
| | $\Delta_2$ | 60 | 1.2065 | 1.6940 | 1.2107 | 1.5471 | 1.2130 | 1.6743 |
| | | 120 | 1.2129 | 1.9259 | 1.2160 | 1.7108 | 1.2190 | 1.9253 |
| | | 200 | 1.2011 | 2.3291 | 1.2035 | 2.0887 | 1.2057 | 2.4118 |
| | $\Delta_3$ | 60 | 1.1261 | 1.6514 | 1.1354 | 1.5170 | 1.1318 | 1.6002 |
| | | 120 | 1.2343 | 2.0060 | 1.2360 | 1.8204 | 1.2344 | 2.0127 |
| | | 200 | 1.2271 | 2.3785 | 1.2292 | 2.1493 | 1.2327 | 2.4914 |

**Table 5.** ARE of Nonlinear Model (4.2)

| | P(y, z) | n | $\frac{CC}{TSMI1}$ | $\frac{R1}{TSMI1}$ | $\frac{CC}{TSMI2}$ | $\frac{kNN}{TSMI2}$ | $\frac{CC}{TSMI3}$ | $\frac{NN}{TSMI3}$ |
|---|---|---|---|---|---|---|---|---|
| τ = 0.25 | $\Delta_1$ | 60 | 1.2753 | 1.6142 | 1.2789 | 1.5398 | 1.2817 | 1.6265 |
| | | 120 | 1.2927 | 1.9931 | 1.2901 | 1.7872 | 1.2940 | 1.9310 |
| | | 200 | 1.2436 | 2.2907 | 1.2451 | 2.1326 | 1.2439 | 2.3268 |
| | $\Delta_2$ | 60 | 1.3559 | 1.8214 | 1.3643 | 1.6696 | 1.3661 | 1.7563 |
| | | 120 | 1.4028 | 2.2167 | 1.4061 | 2.1841 | 1.4103 | 2.2077 |
| | | 200 | 1.4282 | 2.7274 | 1.4287 | 2.5471 | 1.4319 | 2.6897 |
| | $\Delta_3$ | 60 | 1.5931 | 2.1264 | 1.5875 | 1.9837 | 1.5914 | 2.0614 |
| | | 120 | 1.6696 | 2.7057 | 1.6703 | 2.5513 | 1.6767 | 2.5525 |
| | | 200 | 1.6516 | 3.2006 | 1.6561 | 3.1225 | 1.6559 | 3.0907 |
| τ = 0.5 | $\Delta_1$ | 60 | 1.1881 | 1.6793 | 1.1925 | 1.5854 | 1.1891 | 4.0000 |
| | | 120 | 1.2817 | 2.1397 | 1.2801 | 1.8760 | 1.2796 | 2.0592 |
| | | 200 | 1.2009 | 2.4994 | 1.2010 | 2.3208 | 1.2023 | 2.5218 |
| | $\Delta_2$ | 60 | 1.2830 | 1.8937 | 1.2833 | 1.7778 | 1.2871 | 1.8784 |
| | | 120 | 1.3233 | 2.3687 | 1.3257 | 2.2122 | 1.3240 | 2.2965 |
| | | 200 | 1.3304 | 2.9515 | 1.3327 | 2.6584 | 1.3342 | 2.9247 |
| | $\Delta_3$ | 60 | 1.5038 | 2.2039 | 1.5037 | 2.0845 | 1.5063 | 2.2068 |
| | | 120 | 1.5606 | 2.8091 | 1.5572 | 2.6090 | 1.5587 | 2.6925 |
| | | 200 | 1.6104 | 3.4112 | 1.6131 | 3.2907 | 1.6154 | 3.3788 |
| τ = 0.75 | $\Delta_1$ | 60 | 1.1959 | 1.6321 | 1.2010 | 1.5838 | 1.2005 | 1.7268 |
| | | 120 | 1.2402 | 2.0464 | 1.2409 | 1.7999 | 1.2401 | 2.0472 |
| | | 200 | 1.2450 | 2.4233 | 1.2459 | 2.1911 | 1.2475 | 2.4719 |
| | $\Delta_2$ | 60 | 1.1775 | 1.7701 | 1.1849 | 1.7078 | 1.1848 | 1.8280 |
| | | 120 | 1.3262 | 2.3192 | 1.3270 | 2.1286 | 1.3248 | 2.2151 |
| | | 200 | 1.3605 | 2.8044 | 1.3614 | 2.5012 | 1.3612 | 2.8264 |
| | $\Delta_3$ | 60 | 1.4049 | 2.1443 | 1.4022 | 2.0810 | 1.3990 | 2.2135 |
| | | 120 | 1.4715 | 2.6354 | 1.4757 | 2.4866 | 1.4751 | 2.6635 |
| | | 200 | 1.5648 | 3.3129 | 1.5685 | 3.1590 | 1.5669 | 3.3802 |

**Table 6**. ARE of Heteroscedastic Model (4.3)

| | P(y, z) | n | $\frac{CC}{TSMI1}$ | $\frac{R1}{TSMI1}$ | $\frac{CC}{TSMI2}$ | $\frac{kNN}{TSMI2}$ | $\frac{CC}{TSMI3}$ | $\frac{NN}{TSMI3}$ |
|---|---|---|---|---|---|---|---|---|
| $\tau = 0.25$ | $\Delta_1$ | 60 | 1.1274 | 1.1726 | 1.1239 | 1.1016 | 1.1240 | 1.1116 |
| | | 120 | 1.1012 | 1.2658 | 1.0998 | 1.1332 | 1.1009 | 1.1620 |
| | | 200 | 1.1085 | 1.3278 | 1.1072 | 1.1804 | 1.1086 | 1.2102 |
| | $\Delta_2$ | 60 | 1.2308 | 1.3807 | 1.2254 | 1.2132 | 1.2294 | 1.2579 |
| | | 120 | 1.2225 | 1.5651 | 1.2217 | 1.3785 | 1.2224 | 1.4239 |
| | | 200 | 1.2116 | 1.6651 | 1.2134 | 1.4652 | 1.2125 | 1.5200 |
| | $\Delta_3$ | 60 | 1.2057 | 1.3212 | 1.2108 | 1.2034 | 1.2100 | 1.1990 |
| | | 120 | 1.2343 | 1.5225 | 1.2376 | 1.3265 | 1.2367 | 1.3466 |
| | | 200 | 1.2376 | 1.5852 | 1.2385 | 1.3916 | 1.2379 | 1.4274 |
| $\tau = 0.5$ | $\Delta_1$ | 60 | 1.1328 | 1.1359 | 1.1298 | 1.0996 | 1.1306 | 1.1405 |
| | | 120 | 1.0549 | 1.2425 | 1.0557 | 1.1459 | 1.0538 | 1.1751 |
| | | 200 | 1.0674 | 1.3422 | 1.0677 | 1.2166 | 1.0682 | 1.2745 |
| | $\Delta_2$ | 60 | 1.1714 | 1.3738 | 1.1695 | 1.2952 | 1.1752 | 1.3014 |
| | | 120 | 1.1365 | 1.5157 | 1.1367 | 1.3434 | 1.1379 | 1.4342 |
| | | 200 | 1.1951 | 1.7234 | 1.1963 | 1.5220 | 1.1961 | 1.6296 |
| | $\Delta_3$ | 60 | 1.1392 | 1.3099 | 1.1436 | 1.2627 | 1.1436 | 1.2263 |
| | | 120 | 1.1346 | 1.4455 | 1.1357 | 1.2968 | 1.1382 | 1.3807 |
| | | 200 | 1.1810 | 1.6382 | 1.1848 | 1.4790 | 1.1837 | 1.5744 |
| $\tau = 0.75$ | $\Delta_1$ | 60 | 1.0518 | 1.1460 | 1.0480 | 1.1108 | 1.0561 | 1.1380 |
| | | 120 | 1.0428 | 1.2085 | 1.0412 | 1.1308 | 1.0398 | 1.1722 |
| | | 200 | 1.0473 | 1.2886 | 1.0487 | 1.1752 | 1.0485 | 1.2553 |
| | $\Delta_2$ | 60 | 1.1475 | 1.3862 | 1.1500 | 1.2977 | 1.1506 | 1.3258 |
| | | 120 | 1.1233 | 1.4368 | 1.1249 | 1.3422 | 1.1254 | 1.3923 |
| | | 200 | 1.1646 | 1.6585 | 1.1663 | 1.4719 | 1.1654 | 1.6113 |
| | $\Delta_3$ | 60 | 1.0847 | 1.2580 | 1.0912 | 1.1859 | 1.0863 | 1.2133 |
| | | 120 | 1.0730 | 1.3672 | 1.0748 | 1.2653 | 1.0795 | 1.3477 |
| | | 200 | 1.1511 | 1.5489 | 1.1534 | 1.3897 | 1.1514 | 1.5394 |

In terms of the different missing functions, we can see that, on the whole, all the 7 estimators perform worse as the missing rates increase under the same sample size and the same quantile level, which is common for analyzing data sets with missing values. However, for this conclusion some exceptions exist as sample size increases and when it is big enough. These two conclusions reflect the relative importance of imputation when sample size is small and missing rate is high.

Furthermore, **Table 4-6** show ARE values of TSMI1, TSMI2 and TSMI3 with CC and corresponding first-stage imputation estimators for model (3.1), model (3.2), model (3.3), respectively, with different sample sizes and various missing probability functions under distinct quantile levels. According to **Table 4-6**, with all the ARE values are larger than 1, we find that our proposed two-stage multiple imputation estimators are uniformly more effective than the CC estimator and the first-stage multiple imputation estimators for all the models considered. For any one of the above models, under the same quantile level and the same missing function, overall, the relative efficiency of our proposed estimators increases as sample size increases. Additionally, under the same quantile level and the same sample size, overall, the relative efficiency of our proposed estimators increases as missing rate increases. What's more, the advantages of our proposed estimators are more obvious when model is nonlinear or heteroscedastic and at extreme quantile levels.

In addition, we can see that the estimation results of our proposed estimators TSMI1, TSMI2 and TSMI3 are very close, which reflects the robustness of our two-stage multiple imputation estimator to the initial imputation methods. This point is of vital importance for the application of our methods. Based on this good property, we can choose one kind of initial imputation methods to realize our two-stage multiple imputation, which provides great convenience for implementation and applications.

## 4. EMPIRICAL DATA ANALYSIS

In this section, we apply the proposed two-stage multiple imputation method to the ACTG 315 data set, which can be found on the website:

http://www.urmc.rochester.edu/biostat/people/faculty/wusite/datasets/data/ACTG315LongitudinalDataNLMEData3.cfm. Meanwhile we analyze this data set using CC method for comparison. The ACTG 315 data set comes from an AIDS clinical trial group (ACTG 315) study which aimed to investigate the relationship between virologic and immunologic responses in AIDS clinical trials. In this data set, virologic response RNA was measured by viral load while immunologic response was measured by CD4 cell count. The ACTG 315 data set has been analyzed by many papers. Liang *et al*. (2004) analyzed this data set via partially linear models. Wu and Wu (2002) used non-linear mixed effects models for this data set in which more details about the data can be found. Similar research Wu and Wu (2001) and Zeger and Diggle (1994). Recently, Grun and Hornik (2012) used a mixed effects model while accounting for censored longitudinal data. Guo *et al*. (2014) considered the multi-index regression models with missing covariates at random to study the effect of the tumor necrosis factor. However, these papers just constructed mean regression models to analyze this data set, we may want to obtain more information from the analysis. For instance, we may be more interested in the influence of covariates on different quantiles of response variable; we may want to explore the influence pattern without specifying the model form in advance. Such analysis aims can be realized by the nonparametric quantile regression model, which is the interested model in this article.

The data set we used here has 317 observations in total with 20.19% CD4 cell counts missing. Similar to the analysis in Liang *et al*. (2004), we here choose the viral load as the response Y while CD4 cell count as the missing covariate X and time as the complete covariate Z. According to the related research, the missingness of CD4 cell counts is due to the distinct measure times of CD4 cell counts and viral load. Thus, it is reasonable to assume this data as MAR.

Since the missing rate is relatively high, CC analysis may lead to information loss to some extent and hence imputation for the missing data can be necessary to consider. Based on the above mentioned analysis aims and data imputation requirement, we apply model (1.2) to this data set and utilize the proposed two-stage multiple imputation methods to estimate the model. To verify and compare the performance of our proposed two-stage multiple imputation methods, CC method is also implemented. Here we consider quantile levels $\tau$ = 0.25, 0.5 and 0.75 and choose Gaussian kernel and product kernel. In terms of the bandwidths selection, we use the bandwidths obtained according to the selection rule in Section 2.4.

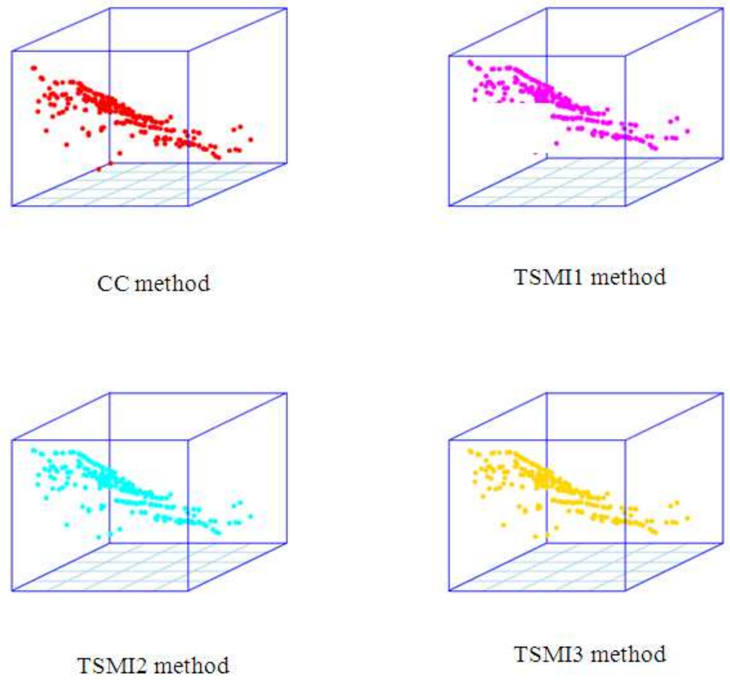
**Table 7** lists the Average Residual Sum of Squares (ARSS), which is calculated as $\text{ARSS} = n^{-1}\sum_{i=1}^{n}\left(Y_i - \hat{Q}_\tau\left(x_i, z_i\right)\right)^2$. CC, TSMI1, TSMI2 and TSMI3 represent the ARSSs of the CC method and the proposed two-stage multiple imputation methods, respectively. **Figure 1-3** show the estimation results of quantile function Q(x, z) based on different methods under $\tau$ = 0.5, 0.25 and 0.75, respectively.

From **Table 7**, we can know that, overall, the smaller values of ARSS of our proposed two-stage multiple imputation methods show that our methods perform better than CC method in terms of data fitting. We also calculate the relative efficiency of our methods compared with the CC method, which is measured via the ratio of ARSSs and we find that our proposed two-stage multiple imputation methods can improve about 5% under $\tau$ = 0.5. In addition, the estimation result under $\tau$ = 0.5 is best, which is common on quantile regression. From **Fig. 1**, we can see, under $\tau$ = 0.5, our three multiple imputation methods show similar results, which reflects the robustness of the proposed imputation method to the initial imputation. Furthermore, our estimation results represent bigger variation of viral load between different time, which shows the distinct influence of cd4 cell count on virologic response under different time. Therefore, our proposed two-stage multiple imputation methods reflect more helpful information to some extent due to their full use of more data information. Similar conclusion can be obtained from **Fig. 2 and 3**.

In addition, from the comparison of these three figures, we can see the different influence patterns among distinct quantile levels of the viral load. In other words, at different virologic response levels, immunologic responses show diversity. Such additional information and conclusion from our analysis can provide more useful signal for relevant research.

**Table 7.** ARSS of model (1.2) based on CC and two-stage multiple imputation methods

| | CC | TSMI1 | TSMI2 | TSMI3 |
|---|---|---|---|---|
| $\tau$ = 0.25 | 0.5196 | 0.5139 | 0.5206 | 0.5107 |
| $\tau$ = 0.5 | 0.3949 | 0.3774 | 0.3797 | 0.3750 |
| $\tau$ = 0.75 | 0.6838 | 0.6693 | 0.6636 | 0.6701 |



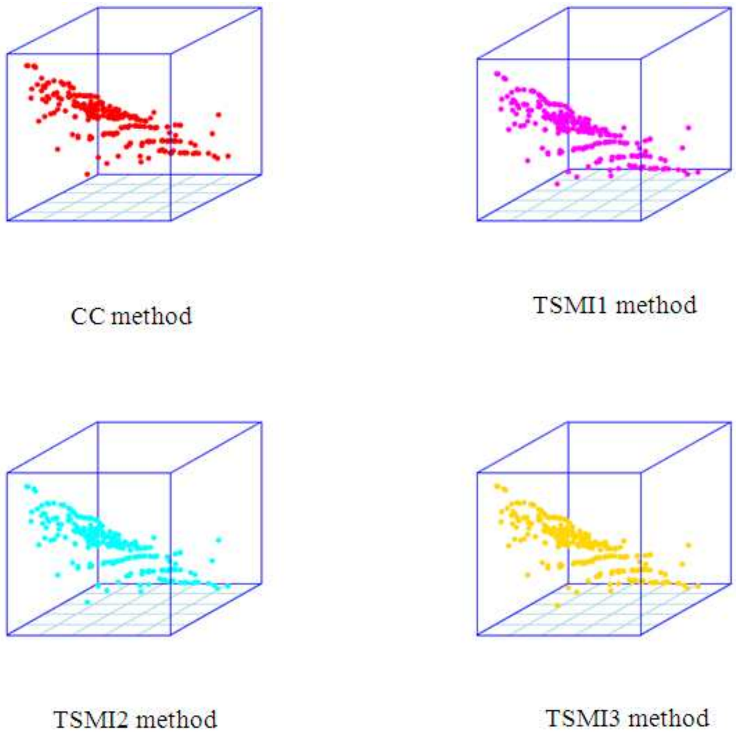
CC method

TSMI1 method

TSMI2 method

TSMI3 method

**Fig. 1.** Estimation result of Q(x, z) at quantile τ = 0.5



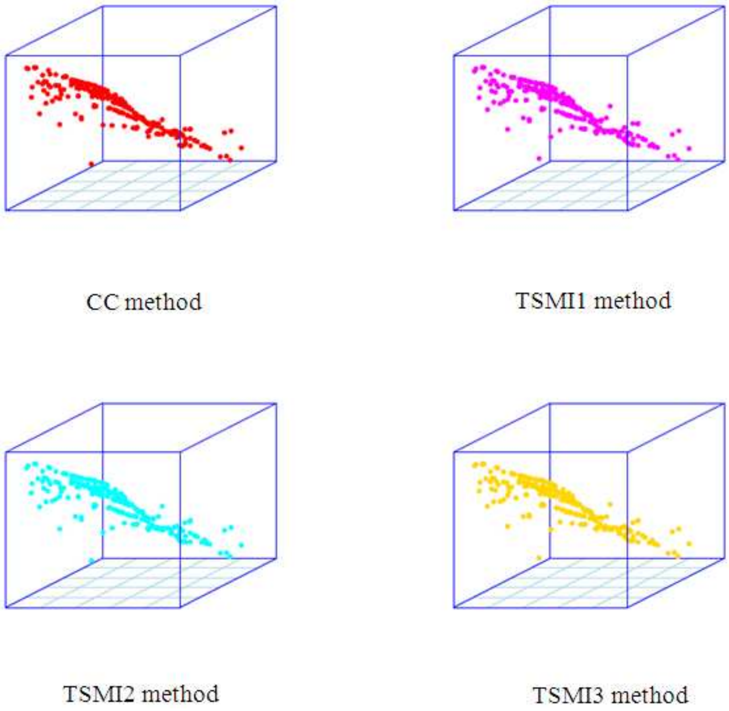
CC method

TSMI1 method

TSMI2 method

TSMI3 method

**Fig. 2.** Estimation result of Q(x, z) at quantile τ = 0.25



CC method

TSMI1 method

TSMI2 method

TSMI3 method

**Fig. 3.** Estimation result of Q(x, z) at quantile $\tau = 0.75$

## 5. CONCLUSION

In this study, we study the nonparametric quantile regression model with the covariates missing at random. We propose an effective and convenient two-stage multiple imputation method for the model and construct the two-stage multiple imputation estimator and give the asymptotic properties of the proposed estimator. Via several simulation examples, we compare the finite sample performance of the proposed estimators under different initial imputation methods with CC estimator, the regression imputation estimator, k-Nearest-Neighbor imputation estimator and the Nearest-Neighbor imputation estimator, which reflects the accuracy and efficiency of the proposed method. In empirical analysis, we construct nonparametric quantile regression model and apply the proposed multiple imputation methods to analyze the ACTG 315 data set and we find that our methods could fit better and give more useful information than CC method.

In addition, the research can be extended to additive quantile regression models with missing covariates, which can avoid the "curse of dimensionality" in nonparametric regression. Furthermore, we can apply the proposed two-stage multiple imputation method to semiparametric models which have more flexibility and interpretation.

## 6. ACKNOWLEDGEMENT

The work was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (Grant No. 10XNL018).

## 7. REFERENCES

Anderson, T.W., 1957. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Am. Stat. Assoc., 52: 200-203. DOI: 10.1080/01621459.1957.10501379

Chen, J. and J. Shao, 2000. Nearest neighbor imputation for survey data. J. Official Stat., 16: 113-132.

Cheng, P.E., 1994. Nonparametric estimation of mean functionals with data missing at random. J. Am. Stat. Assoc., 89: 81-87. DOI: 10.1080/01621459.1994.10476448

Chu, C.K. and P.E. Cheng, 1995. Nonparametric regression estimation with missing data. J. Stat. Plann. Inform., 48: 85-99. DOI: 10.1016/0378-3758(94)00151-K

Grun, B. and K. Hornik, 2012. Modelling human immunodeficiency virus ribonucleic acid levels with finite mixtures for censored longitudinal data. J. Royal Stat. Soc., 61: 201-218. DOI: 10.1111/j.1467-9876.2011.01007.x

Guo, X., W. Xu and L. Zhu, 2014. Multi-index regression models with missing covariates at random. J. Multivariate Anal., 123: 345-363. DOI: 10.1016/j.jmva.2013.10.006

Koenker, R. and G. Bassett, 1978. Regression quantiles. Econometrica, 46: 33-50.

Liang, H, S.J. Wang, J.M. Robins and R.J. Carroll, 2004. Estimation in partially linear models with missing covariates. J. Am. Stat. Assoc., 99: 357-367. DOI: 10.1198/016214504000000421

Little, R.J.A. and D.B. Rubin, 1987. Statistical Analysis with Missing Data. 1st Edn., Wiley, New York, ISBN-10: 0471802549, pp: 278.

Robins, L., A. Rotnizky and L.P. Zhao, 1994. Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc., 89: 846-866. DOI: 10.1080/01621459.1994.10476818

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. 1st Edn., CRC Press, ISBN-10: 0412246201, pp: 176.

Vach, W., 1994. Logistic Regression with Missing Values and Covariates. 1st Edn., Springer-Verlag.

Wang, Q. and J.N.K. Rao, 2001. Empirical likelihood for linear regression models under imputation for missing responses. Canad. J. Stat., 29: 597-608. DOI: 10.2307/3316009

Wang, Q. and J.N.K. Rao, 2002. Empirical likelihood-based inference under imputation for missing response data. Annals Stat., 30: 896-924.

Wang, Q., O. Linton and W. Hardle, 2004. Semiparametric regression analysis with missing response at random. J. Am. Stat. Assoc., 99: 334-345. DOI: 10.1198/016214504000000449

Wang, Q.H., 2009. Statistical estimation in partial linear models with covariate data missing at random. Ann. Inst. Stat. Math., 61: 47-84. DOI: 10.1007/s10463-007-0137-1

Wei, Y., Y. Ma and R.J. Carroll, 2012. Multiple imputation in quantile regression. Biometrika, 99: 423-438. DOI: 10.1093/biomet/ass007

Wu, H. and L. Wu, 2001. A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. Stat. Med., 20: 1755-1769. DOI: 10.1002/sim.816

Wu, L. and H. Wu, 2002. Missing time-dependent covariates in human immunodeficiency virus dynamic models. J. Royal Stat. Soc. Ser. C, 51: 297-318. DOI: 10.1111/1467-9876.00270

Yu, K. and M.C. Jones, 1998. Local linear quantile regression. J. Am. Stat. Assoc., 93: 228-237. DOI: 10.1080/01621459.1998.10474104

Zeger, S.L. and P.J. Diggle, 1994. Semiparametric models for longitudinal data with application to $CD_4$ cell numbers in HIV seroconverters. Biometrics, 50: 689-699.