Generalized Method of Moments Estimator Based On Semiparametric Quantile Regression Imputation

Senniang Chen*and Cindy Yu†

Abstract

In this article, we consider an imputation method to handle missing response values based on semiparametric quantile regression estimation. In the proposed method, the missing response values are generated using the estimated conditional quantile regression function at given values of covariates. We adopt the generalized method of moments for estimation of parameters defined through a general estimation equation. We demonstrate that the proposed estimator, combining both semiparametric quantile regression imputation and generalized method of moments, is an effective alternative to parameter estimation when missing data is present. The consistency and the asymptotic normality of our estimators are established and variance estimation is provided. Results from limited simulation studies and an empirical study are presented to show the adequacy of the proposed method.

Key Words: generalized method of moments, imputation, semiparametric quantile regression.

1 Introduction

Missing data is a frequently encountered issue in many disciplines. Missing data analysis is important because an inference based on ignoring missingness will undermine efficiency and lead to biases and misleading conclusions. There is a large literature handling missing data which can be

*Department of Statistics, Iowa State University, Ames, IA 50011, USA. snchen@iastate.edu
†Department of Statistics, Iowa State University, Ames, IA 50011, USA. cindyyu@iastate.edu

basically categorized into three categories: observed likelihood-based approaches, inverse probability weighting methods, and imputation methods. The main motivation for imputation is to provide a complete data set so that the resulting point estimates are consistent among different users. Once the data set is filled in, one can simply treat it as if there were no missing and apply standard analysis techniques. Due to its intuitive simplicity, imputation becomes particularly popular among practitioners and is the focus of our article.

Many different imputation approaches have been developed in the literature and some prominent examples are included as follows. A pioneer work, multiple imputation (MI) by Rubin (1987), uses Bayesian methods to generate pseudo values from the posterior predictive distribution and imputes multiple data sets. However, the variance estimator of MI, despite its simplified form, requires some special conditions called congeniality and self-sufficiency (Meng 1994) to converge. Fractional imputation was proposed to retain both estimation efficiency of multiple imputation and consistency of the Rao-Shao variance estimator (Rao and Shao 1992). In fractional imputation, multiple missing values are imputed for each missing cell with assigned weights. Kim (2011) proposed parametric fractional imputation (PFI) to reduce computation burden by using the idea of importance sampling and calibration weighting. But both PFI and MI assume a parametric regression model, therefore suffer from model misspecification. Instead of creating artificial values, hot-deck imputation (HDI) replaces missing units with observed data through matching methods. By using covariate information, the matching method could be classifying donors and recipients into similar classes (Brick and Kalton 1996; Kim and Fuller 1996), or creating metric to match donors and recipients (Rubin 1986; Little 1988). More examples are documented in Andridge and Little (2010). In a recent work by Wang and Chen (2009), multiple imputed values are independently drawn from observed respondents with probabilities proportional to kernel distances between missing cells and the donors. Both HDI and Wang and Chen (2009) are purely non-parametric, so the stability and accuracy of their estimators depend on the dimensionality and the sample size of a problem. In fact, finite sample biases are observed in both of these non-parametric methods in our simulation study. It might be due to the fact that a donor with higher probability of being present is more likely selected for imputing than a donor with lower probablity of being present, which possibly results in a distorted conditional density when the covariate is non-uniformaly distributed. For more detailed discussions about this issue, see Section 3.

For a researcher who wants neither fully parametric nor purely non-parametric approach, we

propose an imputation method through semiparametric quantile regression as a solution. Define $f(y|\mathbf{x})$ as the conditional density where y is the response subject to missing and $\mathbf{x}$ is the covariate, and $q(\tau|\mathbf{x})$ as the τ-th conditional quantile function, which is the inverse conditional distribution function $F^{-1}(\tau|\mathbf{x})$. Instead of estimating $f(y|\mathbf{x})$ parametrically or non-parametrically, we estimate $q(\tau|\mathbf{x})$ semiparametrically using observed data under the missing at random (MAR) assumption, a term coined by Rubin (1976). Then multiple imputed values $y_j^*(j = 1, \cdots, J)$ are obtained by $y_j^* = \hat{q}(\tau_j|\mathbf{x})$ where τ_j is independently drawn from Uniform$[0, 1]$. The semiparametric quantile regression imputation (hereafter called SQRI) is expected to possess some attractive features. Firstly, the entire conditional distribution function is used to draw imputed values, hence preserving the conditional density of the filled-in response values. Secondly, because different conditional quantiles instead of conditional means or actual observations are used in imputation, the method is less sensitive to outliers as quantiles are known to be less affected by extremes. Thirdly, it does not require strong model assumptions as in a fully parametric solution, thus is robust against model violation. Lastly, imputed values can be easily created through random numbers generated from Uniform$[0, 1]$.

In this article, we are interested in estimating parameters defined through a general estimation equation. After imputation, the data set is regarded as complete and the generalized method of moments (GMM) is used for parameter estimation. However, combining GMM estimation with SQRI (hereafter called SQRI-GMM) has not been studied, to our best knowledge. So it is not clear, despite its aforementioned theoretical appeals, whether the proposed method can be advocated as an effective alternative in imputation. There are two main goals in this article. The first goal is to establish rigorously large sample theories of our GMM estimator based on SQRI, and the second goal is to evaluate its finite sample performance through numerical simulation. We study our first goal carefully in Section 2 and investigate our second goal in Section 3 through addressing the following three questions: (1) Can our SQRI-GMM method really reduce biases significantly caused by model misspecification, compared to MI and PFI? Our simulations are contrived to cover different kinds of misspecified mean structures, and performances of the estimators are compared; (2) Can our SQRI-GMM method have competative finite sample performance compared to some other non-parametric imputations? This question is interesting since both hot-deck imputation and Wang and Chen (2009) are also robust against model violations. (3) Can our SQRI-GMM method provide a credible inference? The coverage probability of the confidence interval based on our

SQRI-GMM estimator is studied in the simulation. Through our analyses of these three important questions, our article demonstrates the numerical advantages of our GMM estimator through SQRI and confirms its adequacy of being an attractive alternative for imputing.

We are not the first ones who use quantile regression for imputation. There are a few papers pertaining to quantile regression imputation in the literature, see e.g. Munoz and Rueda (2009), Wei et al (2012) and Yoon (2013). Our article is distinctive from these papers in terms of objective, type of imputation and theory. (i) For objective, Wei et al (2012) and Yoon (2013) were only interested in estimating quantile regression coefficients. However our method is designed for a general framework and can be used for estimating parameters defined through any general estimation equation. The article by Munoz and Rueda (2009) focuses on imputation strategy only, and parameter estimation is not an objective of the paper. It is worth noting that the setting in Wei et al (2012) is also different since they dealt with missing covariate, not missing response. (ii) For type, Wei et al (2012) imputed multiple data sets, while Munoz and Rueda (2009) proposed a single and deterministic imputation. However our method utilizes fractional imputation. (iii) For theory, both Wei et al (2012) and Yoon (2013) showed the asymptotic theories for their quantile regression coefficient estimators when a linear quantile regression model is assumed. However, our method is fully semiparametric and permits a penalty to penalize the model complexity. The key idea of our proof which is substantially different from Wei et al (2012) and Yoon (2013) is used to arrive at the consistency and normality, and the asymptotic theory of our semiparametric approach gets more involved than a linear parametric approach. Because the primary interest of Munoz and Rueda (2009) was computation strategy, no theory was offered in their article.

The rest of article is organized as follows. In Section 2, we introduce our imputation method through semiparametric quantile regression with penalty and present large sample theories of our SQRI-GMM estimator. Section 3 compares our method with other competing methods through simulation studies and reports the statistical inference results of our SQRI-GMM estimator. Section 4 analyzes an income data set from Canadian Census Public Use Tape. The Appendix outlines proofs of the theorems appearing in the main text. Details of the proofs are collected in the supplemental file Chen and Yu (2014).

2 Proposed GMM Estimator Based On Semiparametric Quantile Regression Imputation (SQRI)

In this section, we introduce our GMM estimator based on SQRI. Section 2.1 builds the framework and discusses SQRI using penalized B-splines, Section 2.2 contains theoretical results for asymptotic consistency and normality of the unweighted SQRI-GMM estimator, and Section 2.3 extends the large sample theories for the weighted SQRI-GMM estimator.

2.1 SQRI using penalized B-splines

We consider $(\mathbf{x}_i, y_i)^T, i = 1, \cdots, n$, to be a set of i.i.d. observations of random variables $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d_x-dimension variable always observed and Y is the response variable subject to missing. Let $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is missing. We assume that δ and Y are conditionally independent given $\mathbf{X}$, i.e.

$$P(\delta = 1|Y = y, \mathbf{X} = \mathbf{x}) = P(\delta|\mathbf{X} = \mathbf{x}) := p(\mathbf{x}),$$

a condition termed as missing at random in Rubin (1976). The primary interest of this article is to estimate a d_θ-dimensional parameter $\boldsymbol{\theta}_0$ which is the unique solution to $E\{\mathbf{g}(Y, \mathbf{X}; \boldsymbol{\theta})\} = 0$, and make inference on $\boldsymbol{\theta}_0$. Here $\mathbf{g}(Y, \mathbf{X}; \boldsymbol{\theta}) = (g_1(Y, \mathbf{X}; \boldsymbol{\theta}), \cdots, g_r(Y, \mathbf{X}; \boldsymbol{\theta}))^T$ is a vector of r estimating functions for $r \geq d_\theta$. Let $q_\tau(\mathbf{x})$ be the unknown conditional $100\tau\%$ quantile of response Y given $\mathbf{X} = \mathbf{x}$ and be defined as $P(Y < q_\tau(x)|\mathbf{X} = \mathbf{x}) = \tau$ for a given $\tau \in (0, 1)$. When $\tau = 0.5$, $q_\tau(\mathbf{x})$ is the conditional median of Y. It is easy to show that $q_\tau(\mathbf{x})$ satisfies

$$q_\tau(\mathbf{x}) = \arg\min_{h(\mathbf{x})} E\{\rho_\tau(Y - h(\mathbf{x}))|\mathbf{X} = \mathbf{x}\},$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$, the check function proposed in Koenker and Bassett (1978). Many papers have studied the estimation of $q_\tau(\mathbf{x})$ based on parametric methods, and a summary of those work can be found in Koenker (2005). Nevertheless, it is inevitable that parametric model assumptions fail in some scenarios. Nonparametric quantile regression, including the kernel quantile regression in Yu and Jones (1994) and the smoothing spline method in Koenker et al (1994), has also been intensively studied. One one hand, the smoothing spline method demands lots of computing

time, but on the other hand, the unpenalized spline, despite of its cheap computing cost, tends to give a wiggle curve. In this article, we employ a semiparametric quantile regression method based on penalized B-splines, as suggested in Yoshida (2013). This penalized spline method not only provides a relatively smoothed quantile function but also reduces computation burden.

To simplify notations, we assume X is an univariate variable with a distribution function $F_x(x)$ on $[0,1]$. We discuss how to deal with multivariate $\mathbf{X}$ in Section 3 and how to rescale $\mathbf{X}$ on any compact set into $[0,1]$ in Section 4. Let $K_n - 1$ be the number of knots within the range $(0,1)$, and p be the degree of B-splines. In order to construct the p-th degree B-spline basis, we define equidistantly located knots as $\kappa_k = K_n^{-1}k, (k = -p+1, \cdots, K_n + p)$. Note there are $K_n - 1$ knots located in $(0,1)$. The p-th B-spline basis is

$$\mathbf{B}(x) = (B_{-p+1}^{[p]}(x), B_{-p}^{[p]}(x), \cdots, B_{K_n}^{[p]}(x))^T,$$

where $B_k^{[p]}(x)(k = -p+1, \cdots, K_n)$ are defined recursively as

$$B_k^{[0]}(x) = \begin{cases} 1, \kappa_{k-1} < x \leq \kappa_k, \\ 0, \text{ otherwise,} \end{cases} \quad \text{for } k = -p+1, \cdots, K_n + p,$$

$$B_k^{[s]}(x) = \frac{x - \kappa_{k-1}}{\kappa_{k+s-1} - \kappa_{k-1}} B_k^{[s-1]}(x) + \frac{\kappa_{k+s} - x}{\kappa_{k+s} - \kappa_k} B_{k+1}^{[s-1]}(x), \text{ for } k = -p+1, ..., K_n+p-s, \text{ and } s = 1, \cdots, p.$$

Readers can refer to de Boor (2001) for more details and properties of the B-spline functions. The estimated conditional quantile regression function is $\hat{q}_\tau(x) = \mathbf{B}^T(x)\hat{\mathbf{b}}(\tau)$, where $\hat{\mathbf{b}}(\tau)$ is a $(K_n+p)\times 1$ vector and is obtained by

$$\hat{\mathbf{b}}(\tau) = \arg\min_{\mathbf{b}(\tau)} \sum_{i=1}^{n} \delta_i \rho_\tau[y_i - \mathbf{B}^T(x_i)\mathbf{b}(\tau)] + \frac{\lambda_n}{2}\mathbf{b}^T(\tau)\mathbf{D}_m^T\mathbf{D}_m\mathbf{b}(\tau). \tag{1}$$

Here $\lambda_n(>0)$ is the smoothing parameter, and $\mathbf{D}_m$ is the m-th difference matrix and is $(K_n + p - m) \times (K_n + p)$ dimensional with its element defined as

$$d_{ij} = \begin{cases} (-1)^{|i-j|} \begin{pmatrix} m \\ |i-j| \end{pmatrix} & 0 \leq j - i \leq m \\ 0 & o.w. \end{cases}.$$

where the notation $\begin{pmatrix} m \\ k \end{pmatrix}$ is the choose function given by $(k!(m-k)!)^{-1}m!$ and m is the order of penalty. For example when $m = 2$, the smooth fit is shrunk toward a straight line. As discussed in Yoshida (2013), the difference penalty $\mathbf{b}^T(\tau)\mathbf{D}_m^T\mathbf{D}_m\mathbf{b}(\tau)$ is used to remove computational difficulty occuring when the penalty term is defined through an integral, and it controls the smoothness of the estimated quantile regression function. Section 3 discusses how we choose the numbers $(\lambda_\tau, m, K_n, p)$ in practice. To control the variability of the estimating functions with imputed values, we generate J independent imputed values $\{y_{ij}^*\}_{j=1}^J$ when y_i is missing as follows.

1. Simulate $\tau_j \sim$ Uniform(0,1) independently for $j = 1, 2, \cdots, J$;

2. For each $j = 1, 2, ..., J$, $\hat{\mathbf{b}}(\tau_j)$ is calculated as

$$\hat{\mathbf{b}}(\tau_j) = \arg\min_{\mathbf{b}(\tau)} \sum_{i=1}^{n} \delta_i \rho_{\tau_j}[y_i - \mathbf{B}^T(x_i)\mathbf{b}(\tau)] + \frac{\lambda_n}{2}\mathbf{b}^T(\tau)\mathbf{D}_m^T\mathbf{D}_m\mathbf{b}(\tau);$$

3. For the missing unit i, J independent values are generated as

$$y_{ij}^*|x_i = \hat{q}_{\tau_j}(x_i) = \mathbf{B}^T(x_i)\hat{\mathbf{b}}(\tau_j), j = 1, 2, \cdots, J;$$

Repeat step 3 for every missing unit in the data set. Then we use $\delta_i\mathbf{g}(x_i, y_i; \boldsymbol{\theta}) + (1 - \delta_i)J^{-1}\sum_{j=1}^J \mathbf{g}(x_i, y_{ij}^*; \boldsymbol{\theta})$ as the estimating function for the i-th observation.

Some imputation methods used the conditional mean of Y given $X = x$ for imputing, such as in Cheng (1994) and Wang and Rao (2002), but they do not work for a general parameter estimation. In some parametric imputation method, imputation and estimation steps are integrally correlated, meaning that updating parameters and re-imputing based on most updated parameters are iteratively done. This might require heavey computing time. In the SQRI described above, imputation and estimation steps are totally separate, therefore can handle any general parameter estimation. Also standard analysis tools can be directly applied to imputed data without re-imputation in the SQRI. The PFI in Kim (2011) avoids re-imputation by adjusting weights of imputed values based on iteratively updated parameters. However, any parametric imputation method, including PFI and MI, might suffer from model misspecification. Non-parametric imputation, such as HDI or the work proposed in Wang and Chen (2009) using kernel distance, assumes no parametric model. But

the stability and accuracy of non-parametric estimators depend on sample size and dimensionality of the problem. The SQRI provides a useful comprise between a fully parametric approach and a purely non-parametric approach because of its advantages metnioned in Section 1.

Assuming the number of knots $K_n - 1$ and the smoothing parameter λ_τ depend on n, by Barrow and Smith (1978), there exists $\mathbf{b}^*(\tau)$ that satisfies

$$\sup_{x\in(0,1)} |q_\tau(x) + b^a_\tau(x) - \mathbf{B}^T(x)\mathbf{b}^*(\tau)| = o(K_n^{-(p+1)}), \tag{2}$$

where $b^a_\tau(x) = \frac{q^{(p+1)}_\tau(x)}{(p+1)!K_n^{p+1}} Br_p(\frac{x-\kappa_{k-1}}{K_n^{-1}})$ if $\kappa_{k-1} \le x < \kappa_k$, and $q^{(p+1)}_\tau(x)$ is the $(p+1)$-th derivative of $q_\tau(x)$ with respect to x. Here $Br_p(\cdot)$ is the p-th Bernoulli polynomial inductively defined as $Br_0(x) = 1$, and $Br_p(x) = \int_0^x pB_{p-1}(z)dz + b_p$, where $b_p = -p\int_0^1\int_0^x Br_{p-1}(z)dzdx$ is the p-th Bernoulli number (Barrow and Smith (1978) and Yoshida (2013)). The following Lemma gives the asymptotic property of $\hat{q}_\tau(x) = \mathbf{B}^T(x)\hat{\mathbf{b}}(\tau)$ where $\hat{\mathbf{b}}(\tau)$ is defined in (1).

Lemma 1: Under condition 1 given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_n = O(n^v)$ for $v \le (2p+3)^{-1}(p+m+1)$, we have

(i)

$$\sqrt{\frac{n}{K_n}}[\hat{q}_\tau(x) - \mathbf{B}^T(x)\mathbf{b}^*(\tau) + b^\lambda_\tau(x)] \to_d N(0, V_\tau), \tag{3}$$

(ii)

$$\sqrt{\frac{n}{K_n}}[\hat{q}_\tau(x) - q_\tau(x) + b^a_\tau(x) + b^\lambda_\tau(x)] \to_d N(0, V_\tau), \tag{4}$$

for a given $x \in (0,1)$ and $\tau \in (0,1)$, where

$$\begin{aligned}
b^\lambda_\tau(x) &= \tfrac{\lambda_n}{n}\mathbf{B}^T(x)\left(\mathbf{\Phi}(\tau) + \tfrac{\lambda_n}{n}\mathbf{D}_m^T\mathbf{D}_m\right)^{-1}\mathbf{D}_m^T\mathbf{D}_m\mathbf{b}^*(\tau),\\
V_\tau(x) &= \lim_{n\to\infty}\tfrac{\tau(1-\tau)}{K_n}\mathbf{B}^T(x)\left(\mathbf{\Phi}(\tau) + \tfrac{\lambda_n}{n}\mathbf{D}_m^T\mathbf{D}_m\right)^{-1}\mathbf{\Phi}\left(\mathbf{\Phi}(\tau) + \tfrac{\lambda_n}{n}\mathbf{D}_m^T\mathbf{D}_m\right)^{-1}\mathbf{B}(x),\\
\mathbf{\Phi} &= \int_0^1 p(x)\mathbf{B}(x)\mathbf{B}^T(x)dF_x(x),\\
\mathbf{\Phi}(\tau) &= \int_0^1 p(x)f_{y|x}(q_\tau(x))\mathbf{B}(x)\mathbf{B}^T(x)dF_x(x).
\end{aligned} \tag{5}$$

Here $f_{y|x}(\cdot)$ is the conditional density of Y given $X = x$. We can see that there are two sources of asymptotic biases in $\hat{q}_\tau(x)$. One is $b^a_\tau(x)$ which is the model bias between the true function $q_\tau(x)$ and the spline model used, see equation (2). Another source of bias $b^\lambda_\tau(x)$ is introduced by adding penalty term into the quantile regression. When there is no penalty term ($\lambda_n = 0$), this

bias vanishes. Both of these two bias terms have an order $O_p(n^{-\frac{p+1}{2p+3}})$. The nature of the proof corresponds to Theorem 1 of Yoshida (2013) except that we are dealing with missing data while there is no missingness in Yoshida (2013). The detailed proof of this order and Lemma 1 can be found in the supplemental file Chen and Yu (2014).

We define $G(\boldsymbol{\theta}) = E\{\mathbf{g}(Y, X; \boldsymbol{\theta})\}$ and our estimating function as

$$\mathbf{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i \mathbf{g}(y_i, x_i; \boldsymbol{\theta}) + (1-\delta_i)\frac{1}{J}\sum_{j=1}^{J}\mathbf{g}(y_{ij}^*, x_i; \boldsymbol{\theta})\}. \tag{6}$$

We consider the generalized method of moments (GMM), a usual estimation equation approach, to make inference on $\boldsymbol{\theta}$. Our proposal of combining SQRI with GMM is attractive, since it works for a general parameter estimation and has aforementioned appeals of SQRI. In Sec 2.2 and 2.3, we establish rigorously large sample theories of our SQRI-GMM estimators and also provide variance estimation for inference. Section 2.2 presents basic theories for unweighted SQRI-GMM estimator, while Section 2.3 extends theoreies to deal with weighted SQRI-GMM estimator.

2.2 Unweighted GMM estimator based on SQRI

The unweighted GMM-SQRI is obtained as

$$\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathbf{G}_n^T(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}). \tag{7}$$

We first present Lemma 2, regarding the asymptotic normality of $\mathbf{G}_n(\boldsymbol{\theta}_0)$.

Lemma 2: Under conditions 1 and 2 (a) $\sim$ (b) given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_\tau = O(n^v)$ for $v \leq (2p+3)^{-1}(p+m+1)$, as $n \to \infty$ and $J \to \infty$ we have

$$\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) \to_d N(0, V_G(\boldsymbol{\theta}_0)), \tag{8}$$

where

$$V_G(\boldsymbol{\theta}) = Var(\xi_i(\boldsymbol{\theta})), \tag{9}$$

$$\xi_i(\boldsymbol{\theta}) = \mathbf{g}(y_i, x_i; \boldsymbol{\theta}) + (1-\delta_i)\left[\mu_{g|x}(x_i; \boldsymbol{\theta}) - \mathbf{g}(y_i, x_i; \boldsymbol{\theta})\right] + \delta_i C_p h_n(y_i, x_i; \boldsymbol{\theta})\mathbf{B}(x_i), \tag{10}$$

$$h_n(y_i, x_i; \boldsymbol{\theta}) = \int_{-\infty}^{+\infty}\int_0^1 \dot{\mathbf{g}}_y(q_\tau(x), x; \boldsymbol{\theta})\mathbf{B}^T(x)H_n^{-1}(\tau)\psi_\tau(e_i(\tau))d\tau dF_X(x) \tag{11}$$

$$H_n(\tau) = \Phi(\tau) + \frac{\lambda_n}{n}\mathbf{D}_m^T\mathbf{D}_m,\ e_i(\tau) = y_i - \mathbf{B}^T(x_i)\mathbf{b}^*(\tau),\ \psi_\tau(u) = \tau - 1_{u<0}, \tag{12}$$

$$\mu_{g|x}(x;\boldsymbol{\theta}) = E\{\mathbf{g}(y,x;\boldsymbol{\theta})|X=x\},\ C_p = E\{1-p(x)\} \text{ and } \dot{\mathbf{g}}_y(y,x;\boldsymbol{\theta}) = \frac{\partial \mathbf{g}(y,x;\boldsymbol{\theta})}{\partial y}. \tag{13}$$

Justification of Lemma 2 is crucial to show consistency and asymptotic normality of our SQRI-GMM estimator (Pakes and Pollard 1989). We decompose $\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0)$ into three terms

$$\begin{aligned}\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) \;=\; & \underbrace{\frac{1}{\sqrt{n}}\sum_i^n \mathbf{g}(y_i,x_i;\theta_0)}_{:=B_1} + \underbrace{\frac{1}{\sqrt{n}}\sum_i^n[(1-\delta_i)(\mu_{g|x}(x_i;\theta_0) - \mathbf{g}(y_i,x_i;\theta_0)]}_{:=B_2} \\ & + \underbrace{\frac{1}{\sqrt{n}}\sum_i^n[(1-\delta_i)(\hat{\mu}_{g|x}(x_i;\theta_0) - \mu_{g|x}(x_i;\theta_0)]}_{:=B_3},\end{aligned} \tag{14}$$

where $\hat{\mu}_{g|x}(x_i;\boldsymbol{\theta}) = J^{-1}\sum_{j=1}^J \mathbf{g}(y_{ij}^*, x_i, \boldsymbol{\theta})$ and $y_{ij}^* = \mathbf{B}^T(x_i)\hat{\mathbf{b}}(\tau_j)$. Terms B_1 and B_2 are simple since they are sums of i.i.d. random variables. However term B_3 is much more complicated because it involves additional randomness from uniformly distributed random variable τ_j, and it also depends on the estimated coefficients $\hat{\mathbf{b}}(\tau_j)$ calculated using all respondents. Therefore the summands in B_3 are not independent. The key idea in the proof is to replace B_3 by $\tilde{B}_3 = E(B_3|A_R)$ where $A_R = \{(y_i,x_i)|\delta_i = 1, i = 1,\cdots,n\}$, and to show the following two results: (1) $\tilde{B}_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \delta_i C_p h_n(y_i,x_i;\boldsymbol{\theta}_0)\mathbf{B}(x_i) + o_p(1)$, and (2) $\tilde{B}_3 - B_3 = o_p(1)$. Combing these two results with equation (14) gives the asymptotic normality in Lemma 2.

Remark 1: When there is no missing, $\xi_i(\boldsymbol{\theta}_0)$ in equation (10) coincides with $\mathbf{g}(y_i,x_i;\boldsymbol{\theta}_0)$.

Theorem 1: Under conditions 1 and 2 given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_\tau = O(n^v)$ for $v \le (2p+3)^{-1}(p+m+1)$, as $n\to\infty$ and $J\to\infty$ we have

(i)

$$\widehat{\boldsymbol{\theta}}_n \to_p \boldsymbol{\theta}_0;$$

(ii)

$$\sqrt{n}\Sigma^{-1/2}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_d N(\mathbf{0}, \mathbf{I}_{\mathbf{d}_\theta\times\mathbf{d}_\theta}),$$

where

$$\Sigma(\boldsymbol{\theta}) = \left\{\Gamma^T(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})\right\}^{-1}\Gamma^T(\boldsymbol{\theta})V_G(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})\left\{\Gamma^T(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})\right\}^{-1} \text{ and } \Gamma(\boldsymbol{\theta}) = E\{\frac{\partial \mathbf{g}(Y,X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\}. \tag{15}$$

Theorem 1 shows that $\widehat{\boldsymbol{\theta}}_n$ is consistent and asymptoticly normal. With Lemma 2 and justifications of the following 2 conditions: (1) $\sup_{\boldsymbol{\theta}}(1+|\mathbf{G}(\boldsymbol{\theta})|)^{-1}|\mathbf{G}_n(\boldsymbol{\theta})-\mathbf{G}(\boldsymbol{\theta})| = o_p(1)$ and (2) $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|<\zeta_n}(n^{-1/2}+|\mathbf{G}(\boldsymbol{\theta})|+|\mathbf{G}(\boldsymbol{\theta})|)^{-1}|\mathbf{G}_n(\boldsymbol{\theta})-\mathbf{G}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}_0)| = o_p(1)$ for every positive sequence ζ_n converging to zero, Theorem 1 can be proved following Corollary 3.2 and Theorem 3.3 of Pakes and Pollard (1989). Here the notation of $|\cdot|$ represents the norm of a matrix, defined as $|A| = \sqrt{trace(A'A)}$. To consider variance estimation for $\widehat{\boldsymbol{\theta}}_n$, let an estimator of $\xi(\boldsymbol{\theta})$ be

$$\hat{\xi}(\boldsymbol{\theta}) = \mathbf{g}(y_i,x_i;\boldsymbol{\theta}) + (1-\delta_i)\left\{\hat{\mu}_{g|x}(x_i;\boldsymbol{\theta}) - \mathbf{g}(y_i,x_i;\boldsymbol{\theta})\right\} + \delta_i\hat{C}_p\hat{h}_n(y_i,x_i;\boldsymbol{\theta})\mathbf{B}(x_i), \tag{16}$$

where

$$\begin{aligned}
\hat{h}_n(x_i,y_i;\boldsymbol{\theta}) &= \frac{1}{n}\frac{1}{J}\sum_{k=1}^{n}\sum_{j=J}^{L}\dot{\mathbf{g}}_y(\hat{q}_{\tau_j}(x_k),x_k;\boldsymbol{\theta})\mathbf{B}^T(x_k)\hat{H}_n^{-1}(\tau_j)\psi_{\tau_j}(\hat{e}_i(\tau_j)),\\
\hat{e}_i(\tau_j) &= y_i - \mathbf{B}^T(x_i)\hat{\mathbf{b}}(\tau_j),\\
\hat{H}_n(\tau_j) &= \hat{\Phi}(\tau_j) + \frac{\lambda_n}{n}D_m^TD_m,\\
\hat{\Phi}(\tau_j) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\hat{f}_{Y|X}(x_i,\hat{q}_{\tau_j}(x_i))\mathbf{B}(x_i)\mathbf{B}^T(x_i) \text{ with } \hat{q}_{\tau_j}(x_i) = \mathbf{B}^T(x_i)\hat{\mathbf{b}}(\tau_j),\\
\hat{f}_{Y|X}(x,y) &= \frac{\frac{1}{nab}\sum_{i=1}^{n}\delta_i\kappa(\frac{y-y_i}{a})\kappa(\frac{x-x_i}{b})}{\frac{1}{na}\sum_{i=1}^{n}\delta_i\kappa(\frac{x-x_i}{a})}\\
\text{and } \hat{C}_p &= n^{-1}\sum_{i=1}^{n}(1-\delta_i).
\end{aligned}$$

Here the estimation of $\hat{f}_{Y|X}(x,y)$ uses a Normal kernel $\kappa(\cdot)$ and bandwidths a or b for x (or y). The estimator of $\Gamma(\boldsymbol{\theta}_0)$ is obtained by

$$\hat{\Gamma}(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n}\left\{\sum_{i=1}^{n}[\delta_i\frac{\partial\mathbf{g}(y_i,x_i;\widehat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}} + (1-\delta_i)\frac{1}{J}\sum_{j=1}^{J}\sum_{j=1}^{J}\frac{\partial\mathbf{g}(y_{ij}^*,x_i;\widehat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}]\right\}.$$

Then the variance estimator of $\widehat{\boldsymbol{\theta}}_n$ is $\hat{V}(\widehat{\boldsymbol{\theta}}_n) = n^{-1}\widehat{\Sigma}(\widehat{\boldsymbol{\theta}}_n)$ where $\widehat{\Sigma}(\widehat{\boldsymbol{\theta}}_n))$ is calculated as

$$\widehat{\Sigma}(\widehat{\boldsymbol{\theta}}_n) = \left\{\hat{\Gamma}^T(\widehat{\boldsymbol{\theta}}_n)\hat{\Gamma}(\widehat{\boldsymbol{\theta}}_n)\right\}^{-1}\hat{\Gamma}^T(\widehat{\boldsymbol{\theta}}_n)\hat{V}_G(\widehat{\boldsymbol{\theta}}_n)\hat{\Gamma}(\widehat{\boldsymbol{\theta}}_n)\left\{\hat{\Gamma}^T(\widehat{\boldsymbol{\theta}}_n)\hat{\Gamma}(\widehat{\boldsymbol{\theta}}_n)\right\}^{-1}, \text{ and} \tag{17}$$

$$\hat{V}_G(\boldsymbol{\theta}) = \frac{1}{n-1}\sum_{i=1}^{n}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i(\boldsymbol{\theta})\right\}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_i(\boldsymbol{\theta})\right\}^T. \tag{18}$$

Corollary 1: Under conditions $1 \sim 3$ given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_\tau = O(n^v)$ for $v \leq (2p+3)^{-1}(p+m+1)$, as $n \to \infty$ and $J \to \infty$ we have

$$\sqrt{n}\widehat{\Sigma}^{-1/2}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_d N(\mathbf{0}, \mathbf{I}_{\mathbf{d}_\theta \times \mathbf{d}_\theta}).$$

Corollary 1 allows us to construct confidence intervals based on asymptotic normality using the estimated variance estimator.

2.3 Weighted GMM estimator based on SQRI

A weighted GMM estimator is calculated by minimizing $\mathbf{G}_n^T(\boldsymbol{\theta})\mathbf{W}\mathbf{G}_n(\boldsymbol{\theta})$ for a $r \times r$ positive definite weight matrix $\mathbf{W}$. It can be shown that taking $\mathbf{W} \propto V_G^{-1}(\boldsymbol{\theta}_0)$ will result in the most efficient estimator in the class of all asymptotic normal estimators using arbitrary weight matrices. In practice, $\mathbf{W}$ is replaced by the inverse of the random matrix $\hat{V}_G(\boldsymbol{\theta})$ defined in (18) and the weighted GMM estimator is obtained as

$$\widehat{\boldsymbol{\theta}}_n^w = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathbf{G}_n^T(\boldsymbol{\theta})\hat{V}_G^{-1}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}). \tag{19}$$

The following Lemma proves that $\hat{V}_G^{-1}(\boldsymbol{\theta})$ is close to the fixed non-singular matrix $V_G^{-1}(\boldsymbol{\theta}_0)$ uniformly over a sequence of shrinking neighborhoods, an important condition for $\widehat{\boldsymbol{\theta}}_n^w$ to be consistent and asymptotically normal.

Lemma 3: Under conditions $1 \sim 3$ given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_\tau = O(n^v)$ for $v \leq (2p+3)^{-1}(p+m+1)$, as $n \to \infty$ and $J \to \infty$ we have

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\zeta_n} \left\| \hat{V}_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta}_0) \right\| = o_p(1),$$

for a sequence of positive numbers ζ_n that converges to zero.

The following theorem presents the large sample properties of the weighted GMM estimator $\widehat{\boldsymbol{\theta}}_n^w$.

Theorem 2: Under conditions $1 \sim 3$ given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_\tau = O(n^v)$ for $v \leq (2p+3)^{-1}(p+m+1)$, as $n \to \infty$ and $J \to \infty$ we have

(i)

$$\widehat{\boldsymbol{\theta}}_n^w \to_p \boldsymbol{\theta}_0;$$

(ii)

$$\sqrt{n}\Sigma_w^{-1/2}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n^w - \boldsymbol{\theta}_0) \to_d N(\mathbf{0}, \mathbf{I}_{\mathbf{d}_\theta \times \mathbf{d}_\theta}),$$

where $\Sigma_w(\boldsymbol{\theta}) = \left\{\Gamma^T(\boldsymbol{\theta})V_G^{-1}(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})\right\}^{-1}$.

When Lemma 3 holds, the results in Theorem 2 follow immediately from Lemma 3.4 and Lemma 3.5 of Pakes and Pollard (1989).

Remark 2: The asymptotic variance of the most efficient GMM estimator based on full observations is $n^{-1}\left\{\Gamma^T(\boldsymbol{\theta}_0)Var^{-1}(\mathbf{g}(y_i, x_i; \boldsymbol{\theta}_0))\Gamma(\boldsymbol{\theta}_0)\right\}^{-1}$. It can be shown that $V_G(\boldsymbol{\theta}_0)$ in equation (9) can also be expressed as

$$\begin{aligned} V_G(\boldsymbol{\theta}_0) &= Var(\mathbf{g}(y, x; \boldsymbol{\theta}_0)) - E[(1 - p(x))\sigma^2_{g|x}(x; \boldsymbol{\theta}_0)] \\ &\quad + C_p^2 E\left\{\delta_i h_n(x_i, y_i; \boldsymbol{\theta})\mathbf{B}(x_i)\mathbf{B}^T(x_i)h_n^T(x_i, y_i; \boldsymbol{\theta})\right\} \\ &\quad + 2C_p E\left\{\delta_i h_n(x_i, y_i)\mathbf{B}(x_i; \boldsymbol{\theta})\mathbf{g}^T(y_i, x_i; \boldsymbol{\theta}_0)\right\}, \end{aligned} \tag{20}$$

where $\sigma^2_{g|x}(x; \boldsymbol{\theta}_0) = Var(\mathbf{g}(y, x; \boldsymbol{\theta}_0)|X = x)$. So when missing is low, i.e. $p(x)$ is large and C_p is close to zero, the efficiency of $\widehat{\boldsymbol{\theta}}_n^w$ is close to the asymptotic efficiency of the best GMM estimator based on full observations.

Remark 3: When $r = d_\theta$, the semiparametric efficiency bound defined in Chen, Hong and Tarozzi (2007) is $\Sigma_{speb}(\boldsymbol{\theta}_0) = \left\{\Gamma^T(\boldsymbol{\theta}_0)E^{-1}[\sigma^2_{g|x}(x; \boldsymbol{\theta}_0)/p(x) + \mu_{g|x}(x; \boldsymbol{\theta}_0)\mu^T_{g|x}(x; \boldsymbol{\theta}_0)]\Gamma(\boldsymbol{\theta}_0)\right\}^{-1}$. Rewrite $V_G(\boldsymbol{\theta}_0) = E\{p(x)\sigma^2_{g|x}(x; \boldsymbol{\theta}_0)\} + V\{\mu_{g|x}(x, \boldsymbol{\theta}_0)\} + C_p^2 E\{\delta_i h_n(x_i, y_i; \boldsymbol{\theta})\mathbf{B}(x_i)\mathbf{B}^T(x_i)h_n^T(x_i, y_i; \boldsymbol{\theta})\} + 2C_p E\{\delta_i h_n(x_i, y_i)\mathbf{B}(x_i; \boldsymbol{\theta})\mathbf{g}^T(y_i, x_i; \boldsymbol{\theta}_0)\}$. Our esitmator will achieve the semiparametric efficiency bound if $V_G(\boldsymbol{\theta}_0) \leq E[\sigma^2_{g|x}(x; \boldsymbol{\theta}_0)/p(x) + \mu_{g|x}(x; \boldsymbol{\theta}_0)\mu^T_{g|x}(x; \boldsymbol{\theta}_0)]$, i.e.

$$\begin{aligned} E\{(\tfrac{1}{p(x)} - p(x))\sigma^2_{g|x}(x; \boldsymbol{\theta}_0)\} &\geq C_p^2 E\{\delta_i h_n(x_i, y_i; \boldsymbol{\theta})\mathbf{B}(x_i)\mathbf{B}^T(x_i)h_n^T(x_i, y_i; \boldsymbol{\theta})\} \\ &\quad + 2C_p E\{\delta_i h_n(x_i, y_i)\mathbf{B}(x_i; \boldsymbol{\theta})\mathbf{g}^T(y_i, x_i; \boldsymbol{\theta}_0)\}. \end{aligned} \tag{21}$$

It can be shown that when the conditions in Theorem 2 hold, the right hand side of equation (21) has order $O(K_n^{-1})$ (see derivation of this order in Chen and Yu (2014)). However the left side is $O(1)$. So when $K_n \to \infty$, equation (21) will likely happen. This might explain why in our simulation studies our esimator has slightly smaller Monte Carlo variances than the non-parametric imputation estimator by Wang and Chen (2009) which is claimed to have the semiparametric efficiency bound when $r = d_\theta$.

The variance estimator for $\widehat{\boldsymbol{\theta}}_n^w$ can be simply computed as $\hat{V}(\widehat{\boldsymbol{\theta}}_n^w) = n^{-1}\widehat{\Sigma}_w(\widehat{\boldsymbol{\theta}}_n^w)$ where $\widehat{\Sigma}_w(\widehat{\boldsymbol{\theta}}_n^w) = \{\hat{\Gamma}(\widehat{\boldsymbol{\theta}}_n^w)^T\hat{V}_G^{-1}(\widehat{\boldsymbol{\theta}}_n^w)\hat{\Gamma}(\widehat{\boldsymbol{\theta}}_n^w)\}^{-1}$. The following Corollary shows that the central limit theory still holds after replacing $\Sigma_w(\boldsymbol{\theta})$ by its estimator, thus an inference can be legitimately made based on the weighted SQRI-GMM estimator and its variance estimator.

Corollary 2: Under conditions $1 \sim 3$ given in the Appendix, and assuming $q_\tau(x) \in C^{p+1}$, $K_n = O(n^{\frac{1}{2p+3}})$, and $\lambda_\tau = O(n^v)$ for $v \leq (2p+3)^{-1}(p+m+1)$, as $n \to \infty$ and $J \to \infty$ we have

$$\sqrt{n}\widehat{\Sigma}_w^{-1/2}(\widehat{\boldsymbol{\theta}}_n^w)(\widehat{\boldsymbol{\theta}}_n^w - \boldsymbol{\theta}_0) \to_d N(\mathbf{0}, \mathbf{I}_{\mathbf{d}_\theta \times \mathbf{d}_\theta}).$$

3 Simulation Studies

The second goal of our article is to evaluate the finite sample performances of our proposed estimator through simulation studies. For this purpose, we investigate the following 3 questions: (i) Can our proposed method significantly reduce biases caused by model misspecification, compared to parametric imputation methods such as MI and PFI? (ii) Is our proposed method better or competitive, compared to non-parametric imputation methods such as hot-deck imputation and the method proposed in Wang and Chen (2009)? (iii) Can a credible inference be made based on our proposed method?

We specify the simulation set-up as follows. The response y_i is generated from a model $y_i = m(\mathbf{x}_i) + \epsilon_i$, where $m(\mathbf{x}_i)$ is the mean function and ϵ_i are iid $N(0, 0.1^2)$. We consider four different mean functions listed below following the design of simulation studies in Breidt et al (2005) to cover a range of correct and incorrect model specification.

$$\begin{aligned}
\textit{linear}: &\quad m(x_i) = 1 + 2(x_i - 0.5), \\
\textit{bump}: &\quad m(x_i) = 1 + 2(x_i - 0.5) + \exp\{-30(x_i - 0.5)^2\}, \\
\textit{cycle}: &\quad m(x_i) = 0.5 + 2x_i + sin(3\pi x_i), \\
\textit{bivariate}: &\quad m(x_{1i}, x_{2i}) = 1 + 2(x_{1i} - 0.5) + 2\exp\{-10(x_{2i} - 0.4)^2\}.
\end{aligned}$$

The covariate x_i for the first three univariate models (or x_{1i} and x_{2i} for the last bivariate model) are all independently and identically simulated from a truncated normal distribution $N(0.5, 0.3^2)$

on interval $[0, 1]$. The missing mechanism considered is a logistic regression model

$$\begin{aligned} p(x_i) &= \frac{\exp(1+0.5x_i)}{1+\exp(1+0.5x_i)} && \text{for models } \textit{linear, bump, cycle}, \\ \text{or} \quad p(x_{1i}, x_{2i}) &= \frac{\exp(0.2+x_1+0.5x_2)}{1+\exp(0.2+x_1+0.5x_2)} && \text{for model } \textit{bivariate}. \end{aligned}$$

The missing rates in all situations are about 20% . We are interested in estimating three parameters, the marginal mean of response variable $\mu_y = E(Y)$, the marginal standard deviation of response variable $\sigma_y = \sqrt{Var(Y)}$ and the correlation between the response and covariate variables $\rho = corr(X, Y)$. So $\boldsymbol{\theta} = (\mu_y, \sigma_y, \rho)$ and the corresponding estimating function is defined as

$$g(x_i, y_i, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = \begin{pmatrix} x_i - \mu_x \\ y_i - \mu_y \\ (x_i - \mu_x)^2 - \sigma_x^2 \\ (y_i - \mu_y)^2 - \sigma_y^2 \\ (x_i - \mu_x)(y_i - \mu_y) - \rho\sigma_x\sigma_y \end{pmatrix}. \tag{22}$$

For model *bivariate*, $\boldsymbol{\theta} = (\mu_y, \sigma_y, \rho_1, \rho_2)$ where $\rho_1 = corr(X_1, Y)$ and $\rho_2 = corr(X_2, Y)$ and the estimating function is defined in an analogous way. Note that μ_x and σ_x^2 are the mean and variance of covariate and are treated as nuisance parameters. If there is no missing, parameter $\boldsymbol{\theta}$ can be estimated as

$$\begin{aligned} &\hat{\mu}_y = \tfrac{1}{n}\textstyle\sum_{i=1}^n y_i, \quad \hat{\sigma}_y^2 = \tfrac{1}{n-1}\textstyle\sum_{i=1}^n (y_i - \hat{\mu}_y)^2, \\ &\hat{\mu}_x = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i, \quad \hat{\sigma}_x^2 = \tfrac{1}{n-1}\textstyle\sum_{i=1}^n (x_i - \hat{\mu}_y)^2, \quad \hat{\rho} = \frac{n^{-1}\sum_{i=1}^n (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)}{\hat{\sigma}_x \hat{\sigma}_y}. \end{aligned} \tag{23}$$

For each model, 1000 replicate samples of size $n = 200$ are created and the following seven estimators are calculated to compare our semiparametric imputation method to several parametric and non-parametric imputation methods.

- **Full**: An estimator based on the full observations. $\hat{\boldsymbol{\theta}}$ is calculated using equation (23).
- **Resp**: A naive estimator based on the respondents only. $\hat{\boldsymbol{\theta}}$ is calculated using equation (23) after ignoring missing.
- **SQRI-GMM**: Our proposed estimator defined in (19), which combines the semiparametric quantile regression imputation and weighted GMM estimation.

- **MI**: The multiple imputation estimator proposed in Rubin (1987). The R package 'mi' by Gelman (2013) is employed to obtain J multiple imputed data sets. Estimators in (23) are calculated for each imputed data set, and the MI estimators are the average of them across multiple imputed data sets.

- **PFI**: The parametric fractional imputation estimator proposed in Kim (2011). Under PFI, multiple imputed values $y_{ij}^*(j = 1, \cdots, J)$ are generated from a proposed conditional density $\tilde{f}(y|x)$ and their associated fractional weights w_{ij}^* are computed using $\tilde{f}(y|x)$ and the assumed conditional density $f(y|x; \hat{\eta}^0)$ where $\hat{\eta}^0$ is the parameter associated with the conditional density and is initially given. $\hat{\eta}$ is updated by maximizing the score function of the density $f(y_i; \eta)$ using the imputed values and their weights, then the fractional weights w_{ij}^* are re-calculated. This is done iteratively until $\hat{\eta}$ converges. The PFI estimators are calculated using equation (23), with the missing y_i values replaced by $\sum_{j=1}^{J} w_{ij}^* y_{ij}^*$.

- **NPI-EL**: The non-parametric imputation estimator proposed in Wang and Chen (2009). In NPI-EL, multiple imputed values $y_{ij}^*(j = 1, \cdots, J)$ are independently drawn from the respondent group ($\delta_i = 1$) with the probability of selecting y_s with $\delta_s = 1$

$$P(y_{ij}^* = y_s) = \frac{K\{(x_s - x_i)/h\}}{\sum_{m=1}^{n} \delta_m K\{(x_m - x_i)/h\}},$$

  where $K(\cdot)$ is a d_x-dimensional kernel function and h is a smoothing bandwidth. In our simulations, Gaussian kernel is used and h is prescribed by the cross-validation method. NPI-EL estimator is obtained using the empirical likelihood method for a general estimation problem where the estimating function for missing unit i is replaced by $J^{-1} \sum_{j=1}^{J} \mathbf{g}(y_{ij}^*, x_i; \boldsymbol{\theta})$.

- **HDFI**: A hot-deck fractional imputation estimator. Under HDFI, multiple imputed values $y_{ij}^*(j = 1, \cdots, J)$ are independently drawn from a donor pool which in our study consists of 20 nearest neighbors identified through the Euclidean distance. HDFI estimators are calculated using (23) with the missing y_i replaced by $J^{-1} \sum_{j=1}^{J} y_{ij}^*$.

Estimator Full (or Resp) is included in order to help us gauge how far away our proposed estimator is from the ideal case (or the case of simply ignoring missing). Estimator NPI-EL and HDFI are non-parametric imputation methods, while estimator MI and PFI are parametric imputation methods in both of which y_i is assumed to satisfy $Y|X = x \sim N(\boldsymbol{\beta}^T \mathbf{x}, \sigma^2)$ for some $\sigma > 0$.

Our SQRI-GMM is semiparametric as we use penalized B-spline to estimate conditional quantile regression. For penalized B-spline quantile estimators, typically the degree of B-spline p and the degree of the difference matrix m are kept fixed and low, for example $p \leq 3$ and $m \leq 2$. We set $p = 3$ and $m = 2$, a popular choice in practice as suggested in Yoshida (2013). For a given K_n (where $K_n = \#$ of knots $+ 1$), the smoothing parameter λ_n is prescribed via the generalized approximation cross-validation (GACV) method discussed by Yuan (2006). We obtain results for a variety of choices of K_n and conclude $K_n = 5$ suffices in our examples. In the *bivariate* model, the same specifications are used to obtain bases $\mathbf{B}(x_1)$ and $\mathbf{B}(x_2)$ on x_1 and x_2 separately, then $\mathbf{B}(\mathbf{x})$ is their augmentation, $\mathbf{B}(\mathbf{x}) = (\mathbf{B}^T(x_1), \mathbf{B}^T(x_2))^T$. For all the five imputation methods described above, we use both $J = 10$ and $J = 100$. Our simulation studies show that $J = 10$ is sufficient for our proposed estimator to accurately estimate parameters. We summarize the numerical findings for $J = 10$ below. The stories for $J = 100$ remain the same.

Table 1-2 present the Monte Carlo relative biases and variances of the seven estimators for the fourn models. To offer easy bias comparisons, we plot absolute values of the ratios of relative biases for various estimators to the relative biases of our estimator in Figure 1 where ratios bigger than 1 indicate our estimator has smaller relative biases. According to Table 1 and 2, the relative biases in our estimator SQRI-GMM are less than 1% in all cases. And compared to all other estimators, the relative biases of our estimator are much closer to those of Full estimator in nearly all cases, which confirms its good theoretical properties. Compared to Resp estimator which has selection biases due to ignoring missing, our estimator has much smaller biases and also gives relatively smaller variances because of using additonal covariate information in the missing units.

The following findings are summarized to answer Question (i). Compared to the two parametric methods MI and PFI, our estimator has similar relative biases when the model is correctly specified (*linear*). When the model assumption is wrong (*bump, clcle, bivariate*), our estimator significantly reduces biases in MI and PFI in most of the cases, with exceptions arising in the last *bivariate* model when estimating μ_y and σ_y where SQRI-GMM, MI and PFI all have relative bias less than 1%. In Figure 1, the curves with square and triangle symbols represent the relative bias ratios from MI and PFI respectively. It can be seen in Figure 1(b) $\sim$ (d) that most of ratios are bigger than 1, ranging from about 2 to 50, except for parameters μ_y and σ_y in Figure 1(d) where our estimator is not much worse. From Table 1 and 2, it is not surprising to see that when the model is correct, our semiparametric estimator has larger variance than the two parametric ones. However, when the

model is incorrect, our estimator has slightly better or close efficiency compared to MI and PFI.

The following findings are summarized to answer Question (ii). Compared to the two non-parametric estimators NPI-EL and HDFI, our estimator has considerably smaller biases, with one exception arising in the last *bivariate* model when estimating ρ_2. This superior performance is shown in Figure 1(a) $\sim$ (d) where the curves with circle and star symbols represent the relative bias ratios from NPI-EL and HDFI respectively. Most of the ratios are bigger than 1, ranging from about 1 to 80, except for parameter ρ_2 in Figure 1(d) where our estimator and NPI-EL estimator have close relative biases (-0.0070 for ours and 0.0056 for NPI-EL). According to Table 1 and 2, generally our estimator has slightly smaller variances than NPI-EL estimator, and sometimes is better and never much worse when compared to HDFI.

The biases observed in the two non-parametric methods can be possibly explained by the fact that respondents with higher probability of being present are more likely selected for imputing than respondents with lower probability of being present when x is non-uniformly distributed. A made-up example is plotted in Figure 2 to help illustration. This example mimics the *linear* model in the simulation where y has a linear relationship with x, x is a truncated normal centered at $x = 0.45$, and the units with higher x value have higher probabilities of being present. Suppose we want to impute y value at $x = 0.25$ using HDFI and assume there is not any observation between $x \in (0.15, 0.35)$, an exaggerating situation to facilitate explanation. The donor group consists of 10 nearest neighbors (highlighted bigger dots) that have about same distances from $x = 0.25$. There are 9 respondents around $x = 0.40$ and only 1 respondent at $x = 0.10$ due to non-uniform distribution of x. The location of $J^{-1}\sum_{j=1}^{J} y_{ij}^{*}$ at $x = 0.25$ calculated from the 10 donors is marked by symbol $*$ in Figure 3. These imputed values will pull the true conditional mean up, resulting in overestimation of μ_y. It is consistent with the findings in Table 1 and 2 that both NPI-EL and HDFI overestimate the marginal mean μ_y in all cases. Similar overestimating effect will occur if there are observations between $x \in (0.15, 0.35)$ because there are more donors on the right side of $x = 0.25$ than the left side of $x = 0.25$ for the same reason. This argument can also explain biases found in NPI-EL. Under NPI-EL, the 10 highlighted dots have same chances of being drawn as imputed value because they have same kernel distances from $x = 0.25$. Therefore, more imputed values will be from those 9 respondents at $x = 0.40$, resulting in a bigger $J^{-1}\sum_{j=1}^{J} y_{ij}^{*}$ value. In Table 1 and 2, NPI-EL and HDFI have the same directions of over or under estimation in all situations. Another possible reason is that both NPI-EL and HDFI are kind of local methods which might occasionally

suffer from unstable estimates in regions with high missing rates. However our estimator is based on a global quantile regression method, and it is less sensitive to the presence of such regions relative to purely non-parametric methods.

The following findings are summarized to answer Question (iii). Table 3 contains the coverage probabilities of our SQRI-GMM estimator based on the asymptotic normality in Corollary 2 and a Boostrap method for both $J = 10$ and $J = 100$. In most of cases with $J = 10$, the coverage probabilities based on normality are close to the nominal level 0.95 except that under-coverage is found for μ_y and σ_y in model *linear*. This is a common issue when a confidence interval is constructed based on normal approximation of a GMM estimator. After increasing the number of imputation from $J = 10$ to $J = 100$, all coverage probabilities based on normality improve in general, though the coverage for ρ in model *linear* still low (about 0.86). A Boostrap method then is conducted to calculate the confidence intervals. The Boostrap algorithm is described as follows.

1. Draw a simple random sample χ_n^* with replacement from the original sample $\chi_n = (X_i, Y_i, \delta_i)_{i=1}^n$;
2. Implement the semiparametric quantile regression to impute values for the missing cells in χ_n^*;
3. Estimate $\hat{\boldsymbol{\theta}}$ using our SQRI-GMM estimator.
4. Repeat step 1 $\sim$ 3 for B times, then we have $\hat{\boldsymbol{\theta}}^1, \hat{\boldsymbol{\theta}}^2, \cdots, \hat{\boldsymbol{\theta}}^B$.

The 2.5% and 97.5% percentiles of $\{\hat{\boldsymbol{\theta}}^b\}_{b=1}^B$ give the lower and upper bounds of the 95% confidence interval. We use $B = 400$ in our simulation. In general, the Boostrap method has a slightly better performance over normality method, offering satisfactory coverage probabilities close to 0.95 even when $J = 10$.

In summary, our simulation studies confirm the validity of our proposed estimator in finite sample estimation.

4 Empirical Study

In this section, our proposed SQRI-GMM estimator is applied to a real data consisting of $n = 205$ Canadian workers, all of whom were educated to grade 13. A description of this data set can

be found in Ruppert et al (2003) and Ullah(1985), by whom the source was identified as a 1971 Canadian Census Public Use Tape. A copy of the data can be obtained in the R package 'SemiPar' by Wand (2013). The study variable (y) is log(annual income) and the covariate variable (x) is the age, which is rescaled into $[0, 1]$ by $x = (age - min(age))/(max(age) - min(age))$. The sample estimates of (μ_y, σ_y, ρ) when there is no missing are $(13.49, 0.636, 0.231)$. Missingness is created artificially by deliberately deleting some of the y values according to the missing mechanism $p(x) = exp(1-0.5x)/\{1+exp(1-0.5x)\}$, which results in 30% missing rate. All the five imputation estimators described in the simulation are computed using the real data with artificial missing.

The variance estimator for MI is a function of the point estimator and the variance estimator based on each imputed data set. We use GMM to obtain both point and varaince estimators for each imputed data set. The variance estimators for PFI and HDFI are computed using a boostrap method which is similar to the boostrap method described in Section 3 execetp that different impuation methods are employed in step 2. The confidence interval for NPI-EL is obtained via the boostrap method introduced in Wang and Chen (2009). Table 4 reports the relative biases (relative to the sample estimates of (μ_y, σ_y, ρ) based on full observations) and 95% confidence interval widths for five estimators. Figure 4 draws the scatterplot of log(income) versus age. When estimating μ_y, all estimators give relative biases less than 1%. However when estimating σ_y and ρ, the relative biases of our estimator are smaller in magnitude than those of other estimators. This might be due to some features of the data: no obvious mean structure since age 22, but likely exisetnce of heteroscedasticity in variance. In general, our estimator has slightly narrower confidence intervals compared to others except for the MI estimator when estimating ρ. Overall, this case study demonstrates the empirical effectiveness of our SQRI-GMM estimator.

Acknowledgments

The authors thank Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

Breidt, F. J., Opsomer, J. D., Johnson, A. A., and Ranalli, M. G. (2007), "Semiparametric Model-Assisted Estimation for Natural Resource Surveys," *Statistics Canada* 33(1), 35-44.
Andridge, R. R., and Little, R. J. A. (2010). "A Review of Hot Deck Imputation for Survey Non-

response," *International Statistical Review*,78(1), 40-64.
Barrow,D.L., and Smith, P. W.(1978),"Asymptotic Properties of the Best $L_2[0,1]$ Approximation by Splines with variable knots," *Quaterly of Applies Mathmatics* 33, 293-304
Brick, J. M., and Kalton, G., (1996). "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research 5*(3),215-238
Dempster, A. P.,Laird, N. M., and Rubin, D. B.(1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, *39*(1), 1–38
Cheng, P. E.,(1994). Nonparametric Estimation of Mean Functionals with Data Missing at Random. *Biometrika*. *89*(435), 81–87
Horvitz, D. G. and Thompson, D. J., (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663–685
Kim, J. K.,(2011). Parametric fractional imputation for missing data analysis. *Biometrika*. *98*(1), 119–132
Koenker, R., Bassett, G. (1978), "Regression quantiles," *Econometrica 46*, 33 50
Koenker, R., Ng, P., Portnoy, S. (1994)" Quantile smoothing splines," *Biometrika 81*, 673 680 de Boor, C.(2001), "A Practical Guide to Splines," *Springer-Verlag*
Koenker, R. (2005) "Quantile Regression" *Econometric Society Monograph Series, Cambridge University Press*.
Little, R.J.A. (1988). "Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values," *Applied Statistics 37*, 2338.
Meng, X. L.(1994). Multiple-Imputation Inferences with Uncongenial Sources of Input *Statistical Science*. *9*(4), 538–558
Pakes, A. and Pollard D.(1989). Simulation and the Asymptotic of Optimization Estimators *Econometrica.57*(4),1027–1057
Rao, J. N. K., and Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot-Deck Imputation. *Biometrika*. *79*(4), 811–822
Rubin, D. B. (1976). Inference and Missing Data *Biomerika*, *63*(3), 581-592.
Rubin, D. B.,(1987). Multiple Imputation for Nonresponse in Surveys. *New York: Wiley*.
Seaman, S. R. and White, I. R.(2011). Review of Inverse Probability Weighting for Dealing with Missing data. *Statistical Methods in Medical Research*, *22*(3), 278–295
Wang, D.,and Chen, S. X.,(2009). Empirical Likelihood for Estimating Equations with Missing Values. *The Annals of Statistics*. *37*(1), 490–517
Wei, Y. and Yang, Y. J. (2012), Multiple Imputation in Quantile Regression. *Biometrika 99*(2), 423438
Munoz, J. F., and Rudeda, M. (2009) "New imputation methods for missing data using quantiles,"*Journal of Computational and Applied Mathematics 232*(2)305-317
Yoon, J.(2013)"Quantile Regression Analysis with Missing Response, with Application to Inequality Measure and Data Combination," to be appear
Yoshida,T.,(2013),"Asymptotics for penalized spline estimators in quantile regression,"*Communications in Statistics - Theory and Methods*
Yu,K., and Jones, M.C.,(1994),"Local Linear Quantile Regression," *Journal of the American Statistical Assocaiation*. *93*(44), 228–237
Chen and Yu (2014), Supplement to "Generalized Method of Moments Estimator Based On Semiparametric Quantile Regression Imputation".

Appendix

The notation of $|\cdot|$ represents the norm of a matrix, defined as $|A| = \sqrt{trace(A'A)}$ and the notation of $\|\cdot\|$ denotes the sup-norm in all arguments for functions. We first discuss some technical assumptions.

1. Assumptions for penalized semiparametric quantile regression: (a) There exists $\gamma > 0$ such that $E[|g(y,x;\boldsymbol{\theta})|^{2+\gamma}] < \infty$. (b) The explanatory variable X has distribution function $F_x(x)$ on a compact set $[0,1]$. (c) The knots for the B-spline basis are equidistantly located as $\kappa_k = k/K_n$ for $k = -p+1, \cdots, K_n + p$. (d) The order of the difference matrix is $m < p$. (e) $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n p(x_i)B(x_i)B^T(x_i)$ exists and converges to Φ where Φ is defined as $\Phi = \int_0^1 \mathbf{B}(x)\mathbf{B}^T(x)p(x)dF_x(x)$. (f) $\lim_{n\to\infty} n^{-1}\sum_{i=1}^n p(x_i)f_{Y|X}(q_\tau(x))B(x_i)B^T(x_i)$ exists and converges to $\Phi(\tau)$, where $\Phi(\tau) = \int_0^1 \mathbf{B}(x)\mathbf{B}^T(x)p(x)f_{Y|X}(q_\tau(x))dF_x(x)$. (g) The smoothing parameters λ_n is a positive sequence such that λ_n^{-1} is larger than the maximum eigenvalue of $\Phi(\tau)^{-1/2}D_m^T D_m \Phi(\tau)^{-1/2}$.

2. Assumptions for GMM method: (a) $\boldsymbol{\theta}_0$ is the unique solution to the general estimating equation $E[\mathbf{g}(x,y,\boldsymbol{\theta})] = 0$ and $\boldsymbol{\theta}_0$ is an interior point of Θ. (b) $\mathbf{g}(y,x,\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$ and twice differentiable with respect to y. $\dot{\mathbf{g}}_\theta(y,x,\boldsymbol{\theta}) = \frac{\partial \mathbf{g}(y,x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, $\ddot{\mathbf{g}}_{\theta,y}(y,x,\boldsymbol{\theta}) = \frac{\partial^2 \mathbf{g}(y,x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial y}$ and $\ddot{\mathbf{g}}_y(y,x,\theta) = \frac{\partial^2 \mathbf{g}(y,x,\boldsymbol{\theta}_0)}{\partial y^2}$ are bounded for all $\boldsymbol{\theta} \in \Theta$, x, and y. (c) $\mathbf{G}(\boldsymbol{\theta}) = E[\mathbf{g}(y,x;\boldsymbol{\theta})]$ is differentiable at $\boldsymbol{\theta}_0$ with a derivative matrix Γ of full rank. (d) $||\mathbf{G}_n(\hat{\boldsymbol{\theta}}_n)|| \leq o_p(n^{-1/2}) + \inf_{\boldsymbol{\theta}} |\mathbf{G}_n(\boldsymbol{\theta})|$. (e) $E\left[|\dot{\mathbf{g}}_\theta(y,x;\boldsymbol{\theta})||\dot{\mathbf{g}}_\theta(y,x;\boldsymbol{\theta})|^T\right]$ is bounded.

3. General assumptions for variance estimator: (a) The bandwidth of kernel density estimator, a and b satisfy $a \to 0$, $b \to 0$, $na \to \infty$ and $nb \to \infty$. (b) $f_{Y|X}(x,y)$ is differentiable with respect to y. (c) $\|\Gamma(\boldsymbol{\theta})\|$ and $\|V_G(\boldsymbol{\theta})\|$ is bounded away from 0, where $V_G(\boldsymbol{\theta}) = Var[\xi_i(\boldsymbol{\theta})]$ and $\xi_i(\boldsymbol{\theta}) = \mathbf{g}(y_i,x_i;\boldsymbol{\theta}) + (1-\delta_i)\left[\mu_{g|x}(x_i,\boldsymbol{\theta}) - \mathbf{g}(y_i,x_i;\boldsymbol{\theta})\right] + \delta_i C_p h_n(x_i,y_i,\boldsymbol{\theta})\mathbf{B}(x_i)$.

The proof for Lemma 1 and Corollary 2 are skipped here because the proof for Lemma 1 is very similar to Theorem 1 of Yoshida (2013) except that we are dealing with missing data and the proof of Corollary 2 is straitforward and similar to the proof for Corollary 1. We give outlines of the proofs for rest of the theories stated in the text in this Appendix. More detailed proofs (including the skipped ones), the facts referred hereafter and their justifications can be found in the supplemental file Chen and Yu (2014).

A: Proof of Lemma 2

We can decompose $\mathbf{G}_n(\boldsymbol{\theta}_0)$ as in (14). The key idea in our proof is to replace B_3 by $\tilde{B}_3 = E(B_3|A_R)$ where $A_R = \{\delta_i, (y_i,x_i)|\delta_i = 1, i = 1,\cdots,n\}$, and to show the following two results: (1)$\tilde{B}_3 = n^{-1/2}\sum_{i=1}^n \delta_i C_p h_n(x_i,y_i,\boldsymbol{\theta}_0)\mathbf{B}(x_i) + o_p(1)$, and (2) $\tilde{B}_3 - B_3 = o_p(1)$.

(1) To show $\tilde{B}_3 = n^{-1/2}\sum_{i=1}^n \delta_i C_p h_n(x_i,y_i,\boldsymbol{\theta}_0)\mathbf{B}(x_i) + o_p(1)$: We can further decompose $\tilde{B}_3$ into

two terms,

$$\begin{aligned}\tilde{B}_3 &= \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left\{(1-\delta_i)\frac{1}{J}\sum_{j=1}^{J}[\mathbf{g}(\hat{q}_{\tau_j}(x_i),x_i;\boldsymbol{\theta}_0)-\mathbf{g}(q_{\tau_j}(x_i),x_i;\boldsymbol{\theta}_0)]|A_R\right\}}_{\tilde{B}_{31}} \\ &\quad + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E\left\{(1-\delta_i)\frac{1}{J}\sum_{j=1}^{J}[\mathbf{g}(q_{\tau_j}(x_i),x_i;\boldsymbol{\theta}_0)-\mu_{g|x}(x_i;\boldsymbol{\theta}_0)]|A_R\right\}}_{\tilde{B}_{32}}.\end{aligned} \tag{A.1}$$

It is obvious that $\tilde{B}_{32}=0$ because for any x, $E_{\tau|x}\left[\mathbf{g}(q_\tau(x),x,\boldsymbol{\theta}_0)\right]=E_{y|x}\left[\mathbf{g}(y,x,\boldsymbol{\theta}_0)\right]=\mu_{g|x}(x;\boldsymbol{\theta}_0)$. For $\tilde{B}_{31}$, assuming that $\mathbf{g}(x_i,y_i,\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$, then we have

$$\mathbf{g}(\hat{q}_{\tau_j}(x_i),x_i,\boldsymbol{\theta}_0)-\mathbf{g}(q_{\tau_j}(x_i),x_i,\boldsymbol{\theta}_0)=\dot{\mathbf{g}}_y(q_{\tau_j}(x_i),x_i;\boldsymbol{\theta}_0)[\hat{q}_{\tau_j}(x_i)-q_{\tau_j}(x_i)]+\ddot{\mathbf{g}}_y(\tilde{q}_{\tau_j}(x_i),x_i:\boldsymbol{\theta}_0)[\hat{q}_{\tau_j}(x_i)-q_{\tau_j}(x_i)]^2, \tag{A.2}$$

for $\tilde{q}_{\tau_j}(x_i)$ lying between $q_{\tau_j}(x_i)$ and $\hat{q}_{\tau_j}(x_i)$. By equation (A.2), we have

$$\begin{aligned}\tilde{B}_{31} &= \frac{n_m}{\sqrt{n}}E\left\{\frac{1}{J}\sum_{j=1}^{J}\dot{\mathbf{g}}_y(q_{\tau_j}(x),x,\boldsymbol{\theta}_0)[\hat{q}_{\tau_j}(x)-q_{\tau_j}(x)]|A_R\right\} \\ &\quad +\frac{n_m}{\sqrt{n}}E\left\{\frac{1}{J}\sum_{j=1}^{J}\ddot{\mathbf{g}}_y(\tilde{q}_{\tau_j}(x),x;\boldsymbol{\theta}_0)[\hat{q}_{\tau_j}(x)-q_{\tau_j}(x)]^2|A_R\right\},\end{aligned}$$

where $n_m=n-\sum_{i=1}^{n}\delta_i$ and $x \perp\!\!\!\perp A_R$. By Fact 3 in Chen and Yu (2014), we have

$$E\left\{\frac{1}{J}\sum_{j=1}^{J}\ddot{\mathbf{g}}_y(\tilde{q}_{\tau_j}(x),x,\boldsymbol{\theta}_0)[\hat{q}_{\tau_j}(x)-q_{\tau_j}(x)]^2|A_R\right\}=O(\frac{K_n}{n}). \tag{A.3}$$

By Lemma 1, we have

$$\sqrt{n}\left(\hat{q}_\tau(x)-q_\tau(x)\right)=\frac{1}{\sqrt{n}}\mathbf{B}^T(x)H_n^{-1}(\tau)\sum_{i=1}^{n}\delta_i\mathbf{B}(x_i)\psi_\tau(e_i(\tau))-\frac{\lambda_n}{\sqrt{n}}\mathbf{B}^(x)H_n^{-1}(\tau)C_n(\tau)-\sqrt{n}b_\tau^a(x)+o_p(1), \tag{A.4}$$

where $C_n(\tau)=\mathbf{D}_m^T\mathbf{D}_m\mathbf{b}^*(\tau)$. Then $\tilde{B}_{31}$ can be written as

$$\tilde{B}_{31}=\frac{C_p}{\sqrt{n}}\sum_{i=1}^{n}\delta_i h_n(x_i,y_i,\boldsymbol{\theta}_0)\mathbf{B}(x_i)-\sqrt{n}C_pC_{1n}-\sqrt{n}C_pC_{2n}+o_p(1), \tag{A.5}$$

where $h_n(x_i,y_i,\boldsymbol{\theta}_0)$, C_{1n}, C_{2n} and C_p are defined in Lemma 2. The asymptotic order of C_{1n} and C_{2n} are $C_{1n}=\frac{\lambda_n}{n}E_{x,\tau}\left\{\dot{\mathbf{g}}_y(q_\tau(x),x,\boldsymbol{\theta})\mathbf{B}^T(x)H_n^{-1}(\tau)C_n(\tau)\right\}\overset{\text{by Fact 2}}{=}O(\frac{\lambda_n}{n}K_n^{-m})=O(K_n^{-(p+2)})$, and $C_{2n}=E[\dot{\mathbf{g}}_y(q_\tau(x),x,\boldsymbol{\theta}_0)\mathbf{b}_\tau^a(x)]=O(K_n^{-(p+2)})$. Thus we have $\tilde{B}_{31}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_iC_ph_n(x_i,y_i,\boldsymbol{\theta}_0)\mathbf{B}(x_i)+o_p(1)$.
(2) To show $\tilde{B}_3-B_3=o_p(1)$: By Chebychev's inequality, we only need to show that $E[\tilde{B}_3-B_3]^{\otimes 2}\leq E\{[\hat{\mu}_{g|x}(x_i;\boldsymbol{\theta}_0)-\mu_{g|x}(x_i;\boldsymbol{\theta}_0)]^{\otimes 2}\}=o(1)$. First of all, we can decompose $\hat{\mu}_{g|x}(x_i;\boldsymbol{\theta}_0)-\mu_{g|x}(x_i;\boldsymbol{\theta}_0)$

into two terms,

$$\hat{\mu}_{g|x}(x_i;\boldsymbol{\theta}_0)-\mu_{g|x}(x_i;\boldsymbol{\theta}_0) = \underbrace{\frac{1}{J}\sum_{j=1}^{J}[\mathbf{g}(\hat{q}_{\tau_j}(x),x;\boldsymbol{\theta}_0)-\mathbf{g}(q_{\tau_j}(x),x;\boldsymbol{\theta}_0)]}_{S_n} + \underbrace{\frac{1}{J}\sum_{j=1}^{J}[\mathbf{g}(q_{\tau_j}(x),x;\boldsymbol{\theta}_0)-\mu_{g|x}(x_i;\boldsymbol{\theta}_0)]}_{Q_n}$$

It is equivalent to show that $E[Q_n^{\otimes 2}] = o(1)$, $E[Q_n S_n^T] = o(1)$ and $E[S_n^{\otimes 2}] = o(1)$. Details to show these orders can be found in Chen and Yu (2014). Combing step (1) & 2, together with equation (14), we can write $\sqrt{n}\mathbf{G}_n(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_i^n \xi_i(\boldsymbol{\theta}) + o_p(\frac{1}{\sqrt{n}})$ where $\xi_i(\boldsymbol{\theta})$ is defined in (10). Then by central limit theorem, we have $V^{-1/2}(\xi_i(\boldsymbol{\theta}_0))n^{-1/2}\sum_{i=1}^n \xi_i(\boldsymbol{\theta}_0) \sim_d N(\mathbf{0}, \mathbf{I}_{\mathbf{r}\times\mathbf{r}})$, where $V(\xi_i(\boldsymbol{\theta}_0)) \doteq \sigma_g^2(\boldsymbol{\theta}_0) - E[(1-p(x))\sigma_{g|x}^2(x;\boldsymbol{\theta}_0)] + C_p^2 E\{p(x_i)h_n(x_i,y_i,\boldsymbol{\theta}_0)\mathbf{B}(x_i)\mathbf{B}^T(x_i)h_n^T(x_i,y_i,\boldsymbol{\theta}_0)\} + 2C_p E\{\delta_i h_n(x_i,y_i)\mathbf{B}(x_i)\mathbf{g}^T(y_i,x_i;\boldsymbol{\theta}_0)\}$.

B: Proof of Theorem 1

We verify the two condistions stated after Theorem 1. Let

$$\begin{aligned}
&\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) \\
=\ & \frac{1}{n}\sum_i^n \{\mathbf{g}(y_i,x_i;\boldsymbol{\theta}) - E[(\mathbf{g}(y_i,x_i;\boldsymbol{\theta})]\} + \frac{1}{\sqrt{n}}B_2(\boldsymbol{\theta}) + \frac{1}{\sqrt{n}}\tilde{B}_3(\boldsymbol{\theta}) + o_p(\frac{1}{\sqrt{n}}) \\
=\ & \frac{1}{n}\sum_i^n \left\{\mathbf{g}(y_i,x_i;\boldsymbol{\theta}) - E[(\mathbf{g}(y_i,x_i;\boldsymbol{\theta})] + (1-\delta_i)(\mu_{g|x}(x_i;\boldsymbol{\theta}) - \mathbf{g}(y_i,x_i;\boldsymbol{\theta}) + \delta_i C_p h_n(x_i,y_i;\boldsymbol{\theta})\mathbf{B}(x_i)\right\},
\end{aligned}$$

where $B_2(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^n \left[(1-\delta_i)(\mu_{g|x}(x_i,\boldsymbol{\theta}) - g(y_i,x_i;\boldsymbol{\theta})\right]$, $B_3(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^n \left[(1-\delta_i)(\hat{\mu}_{g|x}(x_i,\boldsymbol{\theta}) - \mu_{g|x}(x_i,\boldsymbol{\theta})\right]$ and $\tilde{B}_3(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^n \delta_i C_p h_n(x_i,y_i,\boldsymbol{\theta})\mathbf{B}(x_i)$. By the law of large numbers, we have $\frac{1}{n}\sum_i^n\{\mathbf{g}(y_i,x_i;\boldsymbol{\theta}) - E[(\mathbf{g}(y_i,x_i;\boldsymbol{\theta})]\} = o_p(1)$, $\frac{1}{n}\sum_i^n\{(1-\delta_i)(\mu_{g|x}(x_i;\theta) - \mathbf{g}(y_i,x_i;\boldsymbol{\theta})\} = o_p(1)$ and $\frac{1}{n}\sum_i^n\{\delta_i C_p h_n(x_i,y_i;\boldsymbol{\theta})\mathbf{B}(x_i)\} = o_p(1)$. Thus we have $\|\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})\| = o_p(1)$ and $\sup_\theta \frac{|\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})|}{1+|\mathbf{G}_n(\boldsymbol{\theta})|+|\mathbf{G}(\boldsymbol{\theta})|} \le \|\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})\| = o_p(1)$. So condition (1) holds.

To prove condition 2, it is sufficient to show that for every sequence $\{\zeta_n\}$ of positive numbers converging to zero, $\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) - \mathbf{G}_n(\boldsymbol{\theta}_0) = o_p(n^{-1/2})$ for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \zeta_n$. Since $\mathbf{G}(\boldsymbol{\theta}_0) = 0$, then

$$\begin{aligned}
\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) - \mathbf{G}_n(\boldsymbol{\theta}_0) &= \frac{1}{n}\sum_i^n[\mathbf{g}(y_i,x_i;\boldsymbol{\theta}) - E(\mathbf{g}(y,x;\boldsymbol{\theta})] + B_2(\boldsymbol{\theta}) + B_3(\boldsymbol{\theta}) \\
&- \left\{\frac{1}{n}\sum_i^n[\mathbf{g}(y_i,x_i;\boldsymbol{\theta}_0) - E(\mathbf{g}(y,x;\boldsymbol{\theta}_0)] + B_2(\boldsymbol{\theta}_0) + B_3(\boldsymbol{\theta}_0)\right\} + o_p(\frac{1}{\sqrt{n}}).
\end{aligned} \tag{B.1}$$

Because $|\boldsymbol{\theta} - \boldsymbol{\theta}_0| < \zeta_n$, we can show (details in Chen and Yu (2014)) that

$$\frac{1}{n}\sum_i^n[\mathbf{g}(y_i,x_i;\boldsymbol{\theta}) - E(\mathbf{g}(y,x;\boldsymbol{\theta})] - \frac{1}{n}\sum_i^n[\mathbf{g}(y_i,x_i;\boldsymbol{\theta}_0) - E(\mathbf{g}(y,x;\boldsymbol{\theta}_0)] = O_p(\frac{1}{\sqrt{n}})o_p(1) = o_p(\frac{1}{\sqrt{n}}),$$

$\frac{1}{\sqrt{n}}B_2(\boldsymbol{\theta}) - \frac{1}{\sqrt{n}}B_2(\boldsymbol{\theta}_0) = o_p(\frac{1}{\sqrt{n}})$, and $\frac{1}{\sqrt{n}}\tilde{B}_3(\boldsymbol{\theta}) - \frac{1}{\sqrt{n}}\tilde{B}_3(\boldsymbol{\theta}_0) = o_p(\frac{1}{\sqrt{n}})$. Thus, we have $||\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) - \mathbf{G}_n(\boldsymbol{\theta}_0)|| = o_p(\frac{1}{\sqrt{n}})$, for $\{\zeta_n\} \to 0$.

C: Proof of Corollary 1

It is sufficient to show $\hat{V}_G(\hat{\boldsymbol{\theta}}_n) = \hat{V}ar(\hat{\xi}_i(\hat{\boldsymbol{\theta}}_n))$ converges to $V_G(\boldsymbol{\theta}_0) = Var(\xi_i(\boldsymbol{\theta}_0))$, $\hat{\Gamma}(\hat{\boldsymbol{\theta}}_n) \to^p \Gamma(\boldsymbol{\theta}_0)$ and $\left[\hat{\Gamma}^T(\hat{\boldsymbol{\theta}}_n)\hat{\Gamma}(\hat{\boldsymbol{\theta}}_n)\right]^{-1} \to^p \left[\Gamma^T(\boldsymbol{\theta}_0)\Gamma(\boldsymbol{\theta}_0)\right]^{-1}$. Here $\hat{V}ar\left(\hat{\xi}(\hat{\boldsymbol{\theta}}_n)\right) = \frac{1}{n-1}\sum_{i=1}^n \left(\hat{\xi}_i(\hat{\boldsymbol{\theta}}_n) - \frac{1}{n}\sum_{i=1}^n \hat{\xi}_i(\hat{\boldsymbol{\theta}}_n)\right)^{\otimes 2}$, which is a consistent estimator of $Var\left(\xi_i(\boldsymbol{\theta}_0)\right)$, if $\hat{\xi}_i(\hat{\boldsymbol{\theta}}_n) \to^p \xi_i(\boldsymbol{\theta}_0)$. The rest part of proof is to show that $\hat{\xi}_i(\hat{\boldsymbol{\theta}}_n) \to^p \xi_i(\boldsymbol{\theta}_0)$ by showing that $\mathbf{g}(y_i, x_i, \hat{\boldsymbol{\theta}}_n) \to^p \mathbf{g}(y_i, x_i, \boldsymbol{\theta}_0)$, $\hat{\mu}_{g|x}(x_i, \hat{\boldsymbol{\theta}}_n) \to^p \mu_{g|x}(x_i, \boldsymbol{\theta}_0)$ and $\hat{h}_n(x_i, y_i; \hat{\boldsymbol{\theta}}_n) \to^p h_n(x_i, y_i; \boldsymbol{\theta}_0)$. Similarly, because of $\hat{q}_\tau(x) \to^p q_\tau(x)$ and $\hat{\boldsymbol{\theta}}_n \to^p \boldsymbol{\theta}_0$, as $n \to \infty$, and $J \to \infty$, we have $\hat{\Gamma}(\hat{\boldsymbol{\theta}}_n) \to^p \Gamma(\boldsymbol{\theta}_0)$, and when $\Gamma(\boldsymbol{\theta}_0)$ is bounded away from $\mathbf{0}_{d_\theta \times d_\theta}$ for all $\boldsymbol{\theta} \in \Theta$, we have $\left[\hat{\Gamma}^T(\hat{\boldsymbol{\theta}}_n)\hat{\Gamma}(\hat{\boldsymbol{\theta}}_n)\right]^{-1} \to^p \left[\Gamma^T(\boldsymbol{\theta}_0)\Gamma(\boldsymbol{\theta}_0)\right]^{-1}$. Details can be found in Chen and Yu (2014).

D: Proof of Lemma 3

Since $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |\hat{V}_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta}_0)| \le \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |\hat{V}_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta})| + \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |V_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta}_0)|$, we only need to show that both $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |\hat{V}_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta})| = o_p(1)$, and $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |V_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta}_0)| = o_p(1)$. Similar to the proof in Corollary 1, we will have $\hat{V}_G(\boldsymbol{\theta}) = V_G(\boldsymbol{\theta}) + o_p(1)$ for all $\boldsymbol{\theta}$. Assuming $\|V_G(\boldsymbol{\theta})\|$ is bounded away from 0, so $\hat{V}_G^{-1}(\boldsymbol{\theta}) = \left[\mathbf{I} + V_G^{-1}(\boldsymbol{\theta})\left(\hat{V}_G(\boldsymbol{\theta}) - V_G(\boldsymbol{\theta})\right)\right]^{-1} V_G^{-1}(\boldsymbol{\theta}) = O_p(1)$ and $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |\hat{V}_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta})| = \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |\hat{V}_G^{-1}(\boldsymbol{\theta})\left[\hat{V}_G(\boldsymbol{\theta}) - V_G(\boldsymbol{\theta})\right] V_G^{-1}(\boldsymbol{\theta})| = o_p(1)$. By Taylor expansion,

$$\begin{aligned} V_G(\boldsymbol{\theta}) - V_G(\boldsymbol{\theta}_0) &= E\left[\xi_i^{\otimes 2}(\boldsymbol{\theta}) - \xi_i^{\otimes 2}(\boldsymbol{\theta}_0)\right] - \left\{E^{\otimes 2}\left[\xi_i(\boldsymbol{\theta})\right] - E^{\otimes 2}\left[\xi_i(\boldsymbol{\theta}_0)\right]\right\} \\ &= E\left[2\xi_i(\tilde{\boldsymbol{\theta}})\dot{\xi}_{i,\theta}(\tilde{\boldsymbol{\theta}})\right](\boldsymbol{\theta} - \boldsymbol{\theta}_0) - 2E\left[\xi_i(\tilde{\boldsymbol{\theta}})\right] E\left[\dot{\xi}_{i,\theta}(\tilde{\boldsymbol{\theta}})\right](\boldsymbol{\theta} - \boldsymbol{\theta}_0) = o_p(1) \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, and $\dot{\xi}_{i,\theta}(\boldsymbol{\theta}) = \frac{\partial \xi_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \dot{\mathbf{g}}_\theta(y_i, x_i; \boldsymbol{\theta}) + (1 - \delta_i)\left\{E\left[\dot{\mathbf{g}}_\theta(y_i, x_i; \boldsymbol{\theta})\right] - \dot{\mathbf{g}}_\theta(y_i, x_i; \boldsymbol{\theta})\right\} + \delta_i C_p \frac{\partial h_n(x_i, y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{B}(x_i)$. Thus we have $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |V_G^{-1}(\boldsymbol{\theta}) - V_G^{-1}(\boldsymbol{\theta}_0)| = \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\zeta_n} |V_G(\boldsymbol{\theta})\left[V_G^{-1}(\boldsymbol{\theta}) - V_G(\boldsymbol{\theta}_0)\right] V_G^{-1}(\boldsymbol{\theta}_0)| = o_p(1)$.

Table 1: The Monte Carlo relative biases and variances of the seven estimators for the mean models *linear* and *bump*. The number of replicates in the Monte Carlo is 1000 and the sample size is 200. The number of imputed values is J.

(a). Model *linear*: $m(x) = 1 + 2(x - 0.5)$

		μ_y		σ_y		ρ	
		RBias (×100)	Var (×100)	RBias (×100)	Var (×100)	RBias (×100)	Var (×100)
	Full	0.251	0.116	-0.200	0.041	-0.138	0.135
	Resp	3.325	0.163	-0.598	0.058	-0.151	0.146
J=10	SQRI-GMM	0.286	0.128	-0.407	0.047	-0.204	0.156
	MI	0.254	0.119	-0.197	0.044	-0.134	0.148
	PFI	0.251	0.120	-0.463	0.044	-0.332	0.149
	NPI-EL	1.110	0.133	-1.338	0.048	-2.486	0.160
	HDFI	0.365	0.121	-1.142	0.046	-1.538	0.144
J=100	SQRI-GMM	0.256	0.121	-0.404	0.045	-0.196	0.152
	MI	0.241	0.119	-0.186	0.044	-0.132	0.144
	PFI	0.241	0.119	-0.433	0.043	-0.372	0.147
	NPI-EL	1.114	0.14	-1.305	0.047	-2.286	0.151
	HDFI	0.364	0.121	-1.140	0.046	-0.585	0.146

(b). Model *bump*: $m(x) = 1 + 2(x - 0.5) + \exp\{-30(x - 0.5)^2\}$

		μ_y		σ_y		ρ	
		RBias (×100)	Var (×100)	RBias (×100)	Var (×100)	RBias (×100)	Var (×100)
	Full	0.027	0.056	-0.527	0.079	-0.635	0.213
	Resp	0.906	0.071	-3.934	0.106	-1.449	0.312
J=10	SQRI-GMM	0.033	0.064	-0.790	0.083	-0.604	0.224
	MI	0.072	0.071	-3.814	0.100	-1.417	0.293
	PFI	0.084	0.072	-4.176	0.099	-1.424	0.284
	NPI-EL	0.314	0.061	-3.542	0.093	-4.712	0.27
	HDFI	0.244	0.062	-3.555	0.097	-2.317	0.254
J=100	SQRI-GMM	0.018	0.059	-0.768	0.082	-0.619	0.224
	MI	0.077	0.070	-3.833	0.099	-1.355	0.281
	PFI	0.084	0.070	-4.150	0.099	-1.412	0.280
	NPI-EL	0.316	0.061	-3.492	0.091	-4.689	0.265
	HDFI	0.239	0.061	-3.528	0.096	-2.358	0.254

Table 2: The Monte Carlo relative biases and variances of the seven estimators for the mean models *cycle* and *bivariate*. The number of replicates in the Monte Carlo is 1000 and the sample size is 200. The number of imputed values is J.

(c). Model *cycle*: $m(x) = 0.5 + 2x + sin(3\pi x)$

		μ_y		σ_y		ρ	
		RBias (×100)	Var (×100)	RBias (×100)	Var (×100)	RBias (×100)	Var (×100)
	Full	0.037	0.182	0.092	0.065	-0.025	0.047
	Resp	1.942	0.266	1.973	0.082	1.177	0.057
J=10	SQRI-GMM	0.057	0.197	-0.193	0.066	-0.024	0.050
	MI	-0.200	0.211	2.500	0.086	1.373	0.058
	PFI	-0.211	0.210	2.207	0.086	1.250	0.058
	NPI-EL	0.115	0.198	-0.592	0.073	-1.773	0.060
	HDFI	0.219	0.187	-0.604	0.070	-1.054	0.056
J=100	SQRI-GMM	0.058	0.185	-0.172	0.066	-0.037	0.050
	MI	-0.220	0.208	2.508	0.083	1.387	0.056
	PFI	-0.215	0.209	2.190	0.083	1.218	0.056
	NPI-EL	0.111	0.194	-0.612	0.073	-1.779	0.059
	HDFI	0.222	0.187	-0.608	0.070	-1.096	0.056

(d). Model *bivariate*: $m(x) = 1 + 2(x_1 - 0.5) + 2\exp\{-10(x_2 - 0.4)^2\}$

		μ_y		σ_y		ρ_1		ρ_2	
		RBias (×100)	Var (×100)	RBias (×100)	Var (×100)	RBias (×100)	Var (×100)	RBias (×100)	Var (×100)
	Full	-0.084	0.305	-0.132	0.100	-0.265	0.152	-0.645	0.453
	Resp	0.676	0.401	0.186	0.136	-0.724	0.204	4.101	0.574
J=10	SQRI-GMM	-0.094	0.308	-0.369	0.104	-0.295	0.154	-0.704	0.459
	MI	0.027	0.361	0.346	0.128	-0.925	0.196	3.642	0.548
	PFI	0.027	0.364	0.074	0.129	-0.996	0.196	3.543	0.541
	NPI-EL	0.306	0.326	-1.090	0.113	-2.177	0.173	0.562	0.484
	HDFI	0.562	0.332	-1.684	0.117	-2.712	0.184	2.602	0.496
J=100	SQRI-GMM	-0.088	0.308	-0.361	0.103	-0.300	0.154	-0.691	0.458
	MI	0.024	0.359	0.346	0.124	-0.898	0.194	3.615	0.540
	PFI	0.030	0.358	-0.038	0.125	-0.994	0.194	3.545	0.538
	NPI-EL	0.299	0.322	-1.073	0.112	-2.171	0.172	0.562	0.480
	HDFI	0.562	0.330	-1.684	0.117	-2.749	0.182	2.557	0.494

Table 3: The coverage probabilities of the 95% C.I. of our SQRI-GMM estimator for the 4 models.

(a). Model *linear*: $m(x) = 1 + 2(x - 0.5)$

	J=10			J=100		
	μ_y	σ_y	ρ	μ_y	σ_y	ρ
Normality	0.934	0.937	0.817	0.931	0.938	0.856
Boostrap	0.928	0.953	0.933	0.932	0.953	0.967

(b). Model *bump*: $m(x) = 1 + 2(x - 0.5) + \exp\{-30(x - 0.5)^2\}$

	J=10			J=100		
	μ_y	σ_y	ρ	μ_y	σ_y	ρ
Normality	0.930	0.940	0.940	0.947	0.942	0.942
Boostrap	0.944	0.949	0.950	0.937	0.946	0.949

(c). Model *cycle*: $m(x) = 0.5 + 2x + sin(3\pi x)$

	J=10			J=100		
	μ_y	σ_y	ρ	μ_y	σ_y	ρ
Normality	0.944	0.939	0.913	0.943	0.941	0.915
Boostrap	0.943	0.947	0.941	0.932	0.947	0.943

(d). Model *bivariate*: $m(x) = 1 + 2(x_1 - 0.5) + 2\exp\{-10(x_2 - 0.4)^2\}$

	J=10				J=100			
	μ_y	σ_y	ρ_1	ρ_2	μ_y	σ_y	ρ_1	ρ_2
Normality	0.953	0.923	0.972	0.939	0.953	0.928	0.977	0.942
Boostrap	0.953	0.945	0.963	0.950	0.948	0.944	0.958	0.947

Table 4: Relative biases and 95% C.I. widths for the 5 imputation estimators in the case study. The relative biases is defined as $(\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_0)/\hat{\boldsymbol{\theta}}_0$, where $\hat{\boldsymbol{\theta}}_0$ is the estimate based on full observations. The number of imputed values is $J = 100$.

	μ_y			σ_y			ρ		
	Est	RBias (×100)	Width	Est	RBias (×100)	Width	Est	RBias (×100)	Width
Full $\hat{\boldsymbol{\theta}}_0$	13.49			0.636			0.231		
SQRI-GMM	13.46	0.22	0.15	0.630	0.95	0.153	0.242	4.75	0.362
MI	13.48	0.07	0.20	0.623	2.01	0.181	0.345	49.24	0.301
PFI	13.48	0.07	0.20	0.614	3.55	0.201	0.331	42.83	0.371
NPI-EL	13.49	0	0.19	0.594	6.59	0.179	0.296	27.85	0.396
HDFI	13.49	0	0.31	0.595	8.66	0.292	0.306	32.11	0.503

Figure 1: The comparisons of relative biases for the four models. The y-axis is for absolute values of the ratios of relative biases of various estimators to the relative biases of our SQRI-GMM estimator, and the x-axis is for different parameters. Ratios with values greater than 1 indicate our estimator has smaller relative biases.

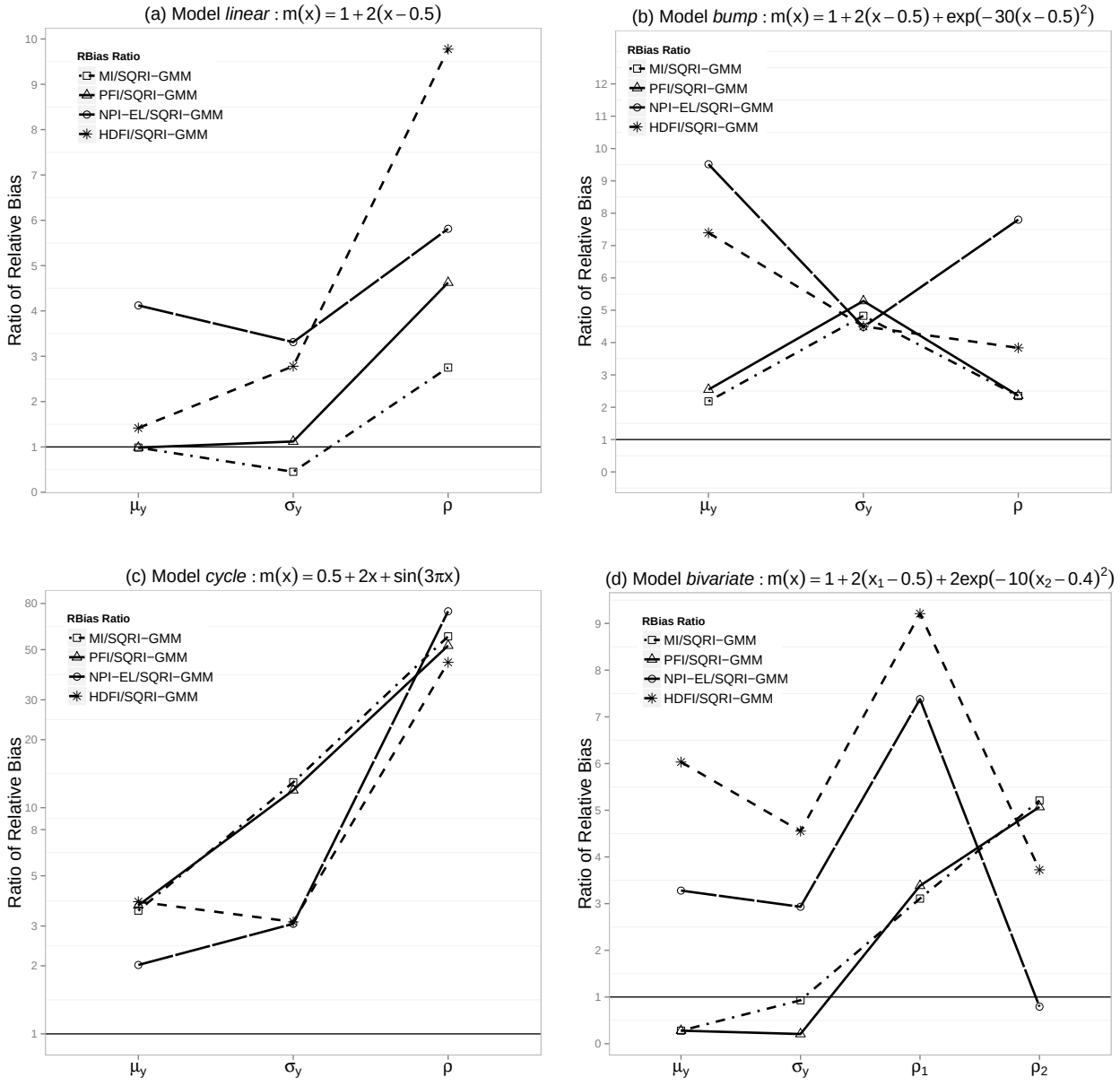

Figure 2: A made-up example to explain the finite sample biases observed in the two non-parametric impuation methods.

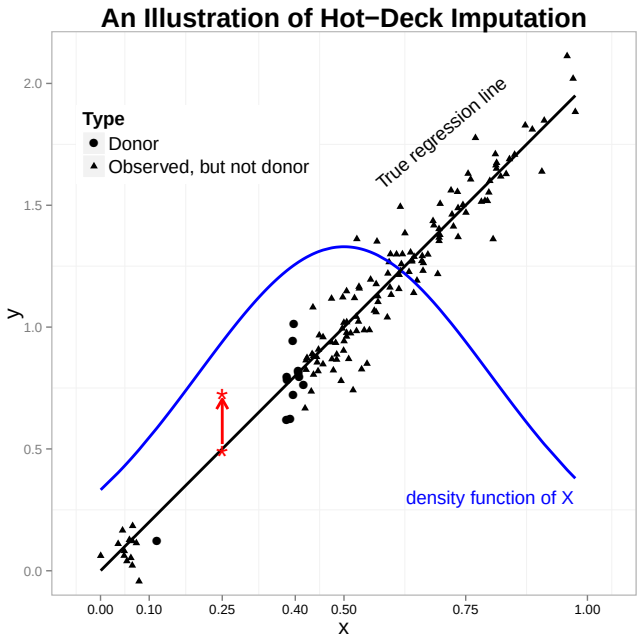

Figure 3: The scatterplot of log(income) versus age in the case study.