QUANTILE REGRESSION ANALYSIS WITH MISSING RESPONSE, WITH APPLICATIONS TO INEQUALITY MEASURES AND DATA COMBINATION

JUNGMO YOON
THE ROBERT DAY SCHOOL OF ECONOMICS AND FINANCE
CLAREMONT MCKENNA COLLEGE

Abstract. We propose a quantile regression method which effectively handles missing values due to non-response. We illustrate the usefulness of our method by two examples. First example is the estimation of income inequality measures when a significant proportion of earnings are missing in survey data. Second example is when we need to combine more than two samples because no single data contains all the relevant variables. We propose a flexible imputation method where missing values in response are drawn from the conditional quantile function of the response at given values of regressors. Once missing values are imputed, a second-step quantile regression is performed, if needed, by using the filled-in, completed sample. We establish the consistency and the asymptotic normality of this two-step procedure and compare its performance with matching estimators.

1. Introduction

In many survey data commonly used in economics, missing values due to non-response are common and often become a source of errors. Econometric analysis that analyze non-reponse in surveys is therefore an important problem for empirical research. Quantile regression has emerged as one of indispensable tool-kits for researchers because it highlights possible sources of heterogeneity among individuals. With survey data such as the Current Population Survey (CPS) or the U.S. Census, it has been successfully exploited for this purpose (Chamberlain (1994), Buchinsky (1994), Abadie, Angrist, and Imbens (2002) and Angrist, Chernozhukov, and Fernández-Val (2006)). But the literature on quantile regression has mostly focused on the case where missing values are not present or to be easily ignored. The so-called complete-case analysis using only the pair of observations with complete information can be misleading when it causes bias and inefficient when the missing rate is too high. So our first goal in this paper is to propose an effective and easy-to-use method to handle missing values in quantile regression analysis.

This Version: July, 2010. First Draft: November, 2009. Please send comments to jyoon@cmc.edu.

We propose a flexible imputation method where imputed values are drawn from its conditional distribution function. Quantile regression is used to estimate the entire family of conditional quantile functions, the inverse of conditional distribution functions. Once missing values are imputed from their conditional quantile functions, a second-step quantile regression is performed, if needed, by using the filled-in, completed sample. We study the large sample properties and the numerical performance of this quantile regression based imputation method.

Imputation methods have a long history in econometrics and statistics in both missing data problem and treatment effect estimation. Examples include Dagenais (1973), Gouriéroux and Monfort (1981), Rubin (1987), Cheng (1994), Robins, Rotnitzky, and Zhao (1995), Wang, Linton, and Härdle (2004), Hahn (1998), to name a few. In most cases, the conditional mean of the response at given values of covariates is estimated and used for the imputed values. The virtue of our method is that we use the entire conditional distribution function of response to draw imputed values, hence, it preserves the distribution of the filled-in response variable, whereas imputing conditional mean does not. This can be a merit when applied to moment condition based methods (See further discussions in Zhou, Wan, and Wang (2008)).

One can view our imputation method as an alternative to matching procedures. Matching is a nonparametric imputation method, therefore, robust to the model specification. Its intuitive appeal and robustness make it popular in literature (Heckman, Ichimura, and Todd (1998), Heckman, Ichimura, Smith, and Todd (1998), Rosenbaum and Rubin (1983), Dehejia and Wahba (1999), Imbens (2004), Abadie and Imbens (2006)). But because it is a purely nonparametric method, the stability and accuracy of a matching estimator depend on the dimensionality of a problem and the sample size. It has been noted (Wang, Linton, and Härdle (2004)) that to justify the conditional independence assumption, which is essential to most imputation methods, one needs a rich set of covariates. It necessarily make the missing data problem a high dimensional one.

So our second goal is to develop an alternative of matching procedure which can be readily used when a researcher does not always want to utilize a fully nonparametric imputation method possibly because of the sample size or the number of regressors. Because many practical methods in non-parametric, semi-parametric, or parametric quantile regression models are available, one can impose necessary amount of structures through quantile regression in his model according to his needs. We will further discuss details on several possible choices in Section 2. We compare the performance of several competing imputation

methods in diverse experimental designs. Quantile regression based methods prove to be preferable when the data generating structure has some degree of structures or the dimension in covariate space is large. Another nonparametric method in imputation is Cheng and Chu (1996) which share the same advantages and disadvantages of a matching method.

Other methods such as the inverse probability weighting can possibly be adapted to the problems we consider. Based on these ideas, a general framework of missing data problem has been formulated in the context of moment conditions (Wooldridge (2007) and Chen, Hong, and Tarozzi (2008)). But imputation has certain intuitive appeal. Once missing values are filled-in, one can use standard techniques developed for complete data, which are readily available in most cases (Little and Rubin (2002)). This simplicity and intuitive appeal makes imputation methods easy to use and particularly popular among practitioner.

We have two examples to illustrate the usefulness of our method. Both involve the case where responses are missing at random. First example is the estimation of various income inequality measures when a significant proportion of earnings are missing. Estimates of the trend of income inequality in the U.S. are mostly based on monthly earning files in CPS. But CPS income information is plagued with high non-response rates. The complete-case analysis, using only the observed individuals, causes bias in the estimate of the marginal distribution of income, therefore, bias in income inequality estimates. One would hope that the practice of the Census Bureau allocating earnings for some of non-respondents would solve the problem. During the 1980s, the earning imputation rates were stable around 15%, but it has steadily increased in recent years, and in 2001, 31% of weekly wage data in CPS earnings file were imputed by the Census (Hirsch and Schumacher (2004)). But many recent studies show that allocated earnings are often unreliable and can be a source of bias (Lillard, Smith, and Welch (1986), Heckman and LaFontaine (2006), Bollinger and Hirsch (2006)).

As a remedy we show that the quantile regression based imputation method can effectively deal with the problem. Using the CPS Outgoing Rotation Group files, we find that using the CPS allocated earnings may results in misleading results for the trend of income inequality, and dropping all the allocated earnings does not always fix the problem. Although income inequality has been increased steadily over time in the U.S., we find that not every sub-groups share the same experience; in some sub-groups, the within-group inequality has risen much more than the entire population, but in others it has not been changed much.

Second example concerns when some crucial information in the main sample is only partially observed but there is a possibly small but independent data source where the

missing variable in the first sample is fully observed. In such a case researchers want to combine two samples. For example, in South African Census data, respondents are only asked to indicate the income category that best describes the personal or household incomes. The response, personal or household earnings, is not a continuous but a categorical variable. Although quantile regression with categorical response variable is perhaps conceivable, to the best of author's knowledge, there is no established inference procedure yet.

In the spirit of imputation methods, it would be desirable if we can fill the gap in discrete variable and make it continuous. In South African case, it is possible because the Labor Force Survey, an independent source of data, reports wages without censoring. Under the conditional independence assumption we will elaborate later, one can combine two samples. Regarding the discrete earnings in Census as incomplete observations, one can allocate reasonable values for the Census data based on the conditional distribution function estimated from the Labor Force Survey. Since the Census earning files provide interval information, the imputation procedure needs to take the constraint into account. With the combined sample, we estimate the income equation and find that both ethnicity and schooling variables affect the distribution of income significantly. But their effects are quite heterogenous; for example, a white worker gets more preferential treatment as he moves to the high end of income distribution, but a black worker face a hardship as he moves to the low end of the income distribution.

The rest of the paper is organized as follows. In Section 2, we develop quantile regression based imputation methods present main results. Section 3 compares our methods with various matching estimators through simulation exercises. Section 4 apply our methodology to the monthly CPS earnings files and present the trend of income inequalities in the U.S. from 1984 to 2005. We highlight some notable features concerning the trend of within group inequality. Section 5 illustrate how to use our methodology to merge more than two samples. Section 6 concludes. Proofs for Theorems and Lemmas appeared in the main text are collected in the Appendix.

2. Methods and Main Results

2.1. **Quantile Regression-based Imputation Method.** We observe (Y_i, X_i, D_i) where Y_i is a one-dimensional response, X_i is a k-dimensional covariates, and D_i is a missing indicator taking $D_i = 1$ when Y_i is observed, $D_i = 0$ when Y_i is missing. Denote $F_{Y|X}(y|x)$ the conditional distribution of Y given $X = x$, and its inverse, $Q_{Y|X}(u|x)$ the conditional quantile function. We will use notation Y_x for the conditional random variable $Y|X = x$.

Consider a case where Y values are missing at random (MAR), but not missing completely at random (MCAR); we allow the missing probability to vary depending on characteristics of individuals. The MAR assumption in this paper takes the form of a conditional independence meaning that Y and D are conditionally independent given X,

$$\Pr(D = 1|Y, X) = \Pr(D = 1|X). \tag{2.1}$$

To illustrate the meaning of the assumption, let us consider an income equation where Y is a person's log-income and X include her individual characteristics. The notion that the missing probability is not directly affected by her income does not mean that they are un-correlated. The MAR assumption allows the missing probability to be related to her income, but it is only because both missing probability and the income are determined by the same factors, her education level and years of job experience.

Denote $p(x) = \Pr(D = 1|X = x)$ the conditional probability of the response being observed, often called, the propensity score, and $p = \int p(x) f_X(x) dx$ the overall non-missing rate. The equation (2.1), through Bayes law, implies the following distributional restriction,

$$F_{Y|X,D}(Y|X, D = 1) = F_{Y|X,D}(Y|X, D = 0). \tag{2.2}$$

It is crucial for identification. By using observations with $D = 1$, one can estimate the conditional distribution function of respondents, and by virtue of equation (2.2), infer the shape of the conditional distribution functions of non-respondents. So the first step in our method is to use observations with $D_i = 1$ to estimate the conditional quantile function,

$$Q_{Y|X}(u|X = x, D = 1) = h(x, u, 1) \ , \quad u \in (0, 1).$$

In view of (2.2), for a fixed x, $h(x, u, 1) = h(x, u, 0)$, hence, for an individual with missing values in response, one can draw an imputed values Y_x^* from $\hat{h}(x, u, 0)$. To be specific, draw a uniform random variables $u \sim \text{Uniform}(0, 1)$ and evaluate the conditional quantile function at $U = u$,

$$Y_{x_i}^* = \hat{Q}_{Y|X}(u|X = x_i, D_i = 0). \tag{2.3}$$

This imputation step gives us a completed data $\{(D_i Y_i + (1 - D_i) Y_i^*, X_i), 1 \leq i \leq n\}$ for later use. One can actually draw independently m times from the uniform distribution and

obtain m sets of completed data $\{(D_iY_i + (1-D_i)Y^*_{x_ij}, X_i), 1 \le i \le n\}_{j=1}^m$. This multiple imputation improves the efficiency of the resulting estimator. We have two examples to illustrate how to use this augmented data set; first, estimating the marginal distribution of responses, second, merging more than two samples.

Application 1 : Income Inequality Measures in the presence of Missing Data

In this example, the main concern is the marginal quantile function of response, $Q_Y(\tau)$ for $\tau \in (0,1)$. It has an important role to understand inequality measures because most inequality measures, the inter-quantile range, the variance, the Lorenz curve, and the Gini coefficient, are defined as functionals of $Q_Y(\tau)$. We propose the following estimator for a marginal quantile function with possibly missing responses,

$$\hat{Q}_Y(\tau) = \inf_y \{y : n^{-1} \sum_{i=1}^n [D_i\, I(Y_i \le y) + (1-D_i)\, m^{-1} \sum_{j=1}^m I(Y^*_{x_ij} \le y)] \ge \tau\}. \tag{2.4}$$

A quantile function is defined as the inverse of an empirical distribution function as usual, but when Y_i is missing, it is replaced by its imputed values $Y^*_{x_ij}$, $j = 1, \ldots, m$. A nonparametric distribution function estimator of Cheng and Chu (1996) can be viewed as a special case of (2.4) where one imputes infinitely many times. To see this, increase the number of imputation m, bearing in mind that $Y^*_{x_ij}$'s are independently drawn from $\hat{F}_{Y|X}(y|x_i)$, then by invoking the law of large numbers,

$$\hat{Q}_Y(\tau) = \inf_y \{y : n^{-1} \sum_{i=1}^n [D_i\, I(Y_i \le y) + (1-D_i)\, \hat{F}_{Y|X}(y|x_i)] \ge \tau\}. \tag{2.5}$$

What Cheng and Chu proposed is to use a non-parametric kernel method to obtain $\hat{F}_{Y|X}(\cdot|x_i)$ and plug it into the definition of marginal distribution function. We analyze the case where one imputes infinitely many times but also where one imputes a few times, or even once. Also our quantile regression method provides flexibility for the first step conditional distribution/quantile estimates. Depending on the sample size and the dimension of the problem (the number of covariates), one can use a purely non-parametric method (local polynomial quantile regression or series estimator), a semi-parametric method (additive or partially linear quantile regression model), or parametrically specified model for each quantile (linear and non-linear quantile regression). Details on some of those first

step estimation methods will be discussed shortly in Section 2.2. Estimates of inequality measures can be easily obtained by plugging-in $\hat{Q}_Y(\tau)$ to the definitions, the interquartile range $R_Y = Q_Y(.75) - Q_Y(.25)$, the variance $\sigma_Y^2 = \int_0^1 Q_Y(u)^2 du$, the Lorenz curve $L_Y(t) = \int_0^t Q_Y(u)du / \int_0^1 Q_Y(u)du$, and the Gini coefficient $G_Y = 2\int_0^1 (t - L_Y(t))dt$.

Remark The expressions in (2.5) suggests that when the object of interest is the marginal distribution function of the response, one needs to impute in the place of missing response either the whole conditional distribution function or random draws from the conditional distribution function. Allocating conditional mean would not solve the problem.

Application 2 : Data Combination

In the second example, we consider quantile regression analysis when one needs to combine more than two samples. Suppose sample A has variables $(\tilde{Y}, X)$ and sample B has (Y, X). Suppose that it would be better to use sample A because of its larger sample size, but its response variable, $\tilde{Y}$, has only partial information regarding the variable of interest Y. On the other hand, the sample B has all relevant variables but sample size is too small, so it is desirable to use information in both A and B. For example, in the South African Census data we use in Section 5, the respondent's income is only recorded as a categorical variable, so for an individual in sample A, the information available is $\tilde{Y}_i \in (c_j, d_j)$ where c_j and d_j are end-points of j-th category.

Let $n = n_a + n_b$ the sample size for the combined sample, and n_a and n_b sample sizes for A and B. For the combined sample, we introduce a new notation δ for the indicator. Let $\delta_i = 1$ when i-th observation in the combined sample comes from sample B, and $\delta_i = 0$ when it comes from sample A. Note that δ is not a random variable, it is a simple indicator with designated binary value depending on where the observation originates.

Under the condition (2.2), we use sample B to fit the conditional quantile function and draw missing values from the conditional quantiles as described earlier. For example, when $\tilde{Y}$ is a categorical variable, use the fitted conditional quantile to draw observations,

$$Y_{x_i}^* \sim \hat{Q}_{Y|X}(u|x_i) \text{ where } u \sim \text{Uniform}(\hat{F}_{Y|X}(c_j|x_i), \hat{F}_{Y|X}(d_j|x_i)).$$

where $\hat{Q}_{Y|X}(u|x)$ is the first step conditional quantile function estimate, and $\hat{F}_{Y|X}(y|x)$ is its inverse. This imputation procedure, when repeated m times, provides us the completed data $\{(\delta_i Y_i + (1-\delta_i)Y_{x_i j}^*, X_i), 1 \le i \le n\}_{j=1}^m$, then we run a second step linear quantile regression,

$$\tilde{\beta}(\tau) = \arg\min_{\beta}\{\sum_{i=1}^{n}\delta_i\,\rho_\tau(Y_i - X_i^T\beta) + \sum_{i=1}^{n}(1-\delta_i)m^{-1}\sum_{j=1}^{m}\rho_\tau(Y^*_{x_i,j} - X_i^T\beta)\}. \tag{2.6}$$

The large sample properties of the resulting two-step quantile regression estimator will be studied in Section 2.3.

2.2. **First-Step Conditional Quantile Estimation.** In order to implement our quantile regression based imputation method, we should be able to consistently estimate conditional quantile functions in the first place. Since we are interested in distributional features of Y at a given $X = x$ at this stage, we can only use observations with $D_i = 1$. In general, our first step quantile regression estimator is obtained by,

$$\hat{\beta}(u) = \arg\min_{\beta}\ \sum_{i=1}^{n} D_i\,\rho_u(Y_i - h(X_i,\beta)) \tag{2.7}$$

where $\rho_u(\cdot)$ is the check function $\rho_u(z) = (u - I(z \le 0))z$. Here $\beta(u)$ is used as a generic notation for parameters that characterizes the u-th conditional quantile. The above procedure yields a consistent estimates thanks to the MAR assumption. To see this, consider the moment equation that describe quantile regression problem, $\psi(W,b) = \dot{h}(X)(I(Y < h(X,b)) - u)$ where $W = (Y,X)$ and $\dot{h}(x) = \partial h(x,b)/\partial b$. Without any missing values, the identification condition guarantees that the moment condition is satisfied at $b = \beta(u)$, that is, $E[\psi(W,\beta(u))] = 0$. With missing values, under the MAR assumption, the same value becomes the unique solution of the equation $E[D\psi(W,b)]$ because

$$\begin{aligned} E[D\psi(W,\beta(u))] &= E[D\dot{h}(X)(I(Y < h(X,\beta(u))) - u)] \\ &= E[E[D\dot{h}(X)|X]\cdot E[I(Y < h(X,\beta(u))) - u|X]] \\ &= E[P(X)\dot{h}(X)\cdot\{F_{Y|X}(h(X,\beta(u))|X) - u\}] = E[P(X)\dot{h}(X)\cdot\{u-u\}] = 0. \end{aligned}$$

The third equality is due to the MAR assumption and the last equality is because of the definition of the u-th conditional quantile. This consideration guarantees the consistency of the estimator (Wooldridge (2007)).

We now discuss estimation. The choice of the conditional quantile function $h(x,\beta(u))$ involves a trade-off. One can leave the functional form of $h(x,\beta(u))$ completely unspecified except some smoothness conditions and turn to purely nonparametric estimation methods, such as, local polynomial regression (Chaudhuri (1991), Chaudhuri, Doksum, and Samarov

(1997), Yu and Jones (1998)), spline (Koenker, Ng, and Portnoy (1994)), or method of sieve (Ai and Chen (2003)). A flexible modeling in conditional quantile is certainly desirable due to concerns for mis-specification.

But at the same time, we need to be cautious of the dimensionality of a problem and its possible impact on nonparametric function estimation. As Wang, Linton, and Härdle (2004) forcefully argued, the same assumption that motivates an imputation method often requires a high dimensional covariate space. It was the MAR assumption that imposes a strong distributional restriction through (2.2), and therefore, enables identification. To justify the assumption that the missing probability is not directly affected by the outcome, it is very likely that we need a rich set of conditioning variables. When it is the case, a fully nonparametric method, whether it is used for the conditional mean, conditional distribution, or propensity score, will suffer from the curse of dimensionality. This dimensionality issue often manifests itself via poor finite sample performance of the resulting estimator. A parametric quantile regression where a finite-dimensional parametric function specify each quantile can be a reasonable choice.

One can compromise between the need of flexible model specification and the curse of dimensionality by using semi-parametric models. The additive model (Koenker, Ng, and Portnoy (1994), Horowitz and Lee (2005)), the partially linear model (Lee (2003)), and the single index model (Wu, Yu, and Yu (2008)) can be helpful in this regard.

A researcher should evaluate her own research problem and choose the right degree of complexity. She may choose different estimation methods, a non-parametric, a semi-parametric, or a parametric model, for the conditional quantile estimation. We present with some details one possible choice for each category. Further details can be found in Section 3.

Method 1: Local Polynomial Quantile Regression

As an illustration of nonparametric implementation, we consider a local polynomial quantile regression model. The unknown u-th conditional quantile $h(x,u)$ is approximated by a linear function $h(z,u) = h(x,u) + h'(x,u)(z-x) = a + b(z-x)$ for z which is close to x. The constant a estimate conditional quantiles, so we define

$$\hat{Q}_{Y|X}(u|x) = \hat{h}(x,u) = \hat{\beta}_0$$

where $\hat{\beta}_0$ is the solution of the following minimization problem,

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} D_i \, \rho_u(y_i - \beta_0 - \beta_1(x_i - x)) K\left(\frac{x_i - x}{d_n}\right).$$

Here $K(\cdot)$ is an appropriately chosen kernel function and d_n is the bandwidth that controls the smoothness of the estimated quantile function. Yu and Jones (1998) discuss a quantile-specific bandwidth selection rule (see also Fan, Hu, and Truong (1994) and Ruppert, Sheather, and Wand (1995)). Because nonparametric methods require a large sample size and small dimension of a problem, it would be wise to use a purely nonparametric approach only when the imputation is based on a small number of covariates.

Method 2: Additive Quantile Regression

As an example of semi-parametric models, we consider the additive model (Hastie and Tibshirani (1990)). This approach reduces the burden of high dimensionality of a nonparametric method by imposing additive structure on the estimated function. Let $x_i = (w_i, z_{1i}, \ldots, z_{Ji})$, then the u-th conditional quantile is modeled as

$$Q_{Y|X}(u|x_i) = w_i^T \beta(u) + \sum_{j=1}^{J} g_j(z_{ij}, u)$$

Following Koenker, Ng, and Portnoy (1994), parameters of the model, β and $g = (g_1, \ldots, g_J)$, are estimated by the following penalized minimization problem,

$$\min_{\beta,g} \sum_{i=1}^{n} D_i \, \rho_u(y_i - w_i^T \beta - \sum_{j=1}^{J} g_j(z_{ij})) + \sum_{j=1}^{J} \lambda_j \int |g_j''(z)| dz.$$

The choice of tuning parameters $\lambda = (\lambda_1, \ldots, \lambda_J)$ determines the smoothness of the estimated functions. Koenker (2010) discuss selection rules based on the information criteria. We follow one of his suggestions which choose λ that minimize

$$\text{SIC}(\lambda) = n \log \hat{\sigma}(\lambda) + p(\lambda) \frac{\log(n)}{2},$$

where $\hat{\sigma}(\lambda) = n^{-1} \sum_{i=1}^{n} D_i \, \rho_u(y_i - \hat{h}(w_i, z_i))$, and $p(\lambda)$ is the effective dimension of the fitted model $\hat{h}(w_i, z_i) = w_i^T \beta + \sum_{j=1}^{J} g_j(z_{ij})$.

Method 3: Linear Quantile Regression

Consider a case where conditional quantile function is linear in parameters, then $\hat{\beta}(u)$ is estimated by

$$\hat{\beta}(u) = \arg\min_{\beta} \sum_{i=1}^{n} D_i \, \rho_u(Y_i - x_i^T \beta), \tag{2.8}$$

and the u-th conditional quantile is obtained by $\hat{Q}_{Y|X}(u|x) = x^T \hat{\beta}(u)$, $u \in (0,1)$. Some obvious advantages of this approach include the ease of implementation and the parametric rate of convergence for the resulting conditional quantile function estimates. Since we use a parametric model for each quantile, one can expect that it would produce reliable outcomes even in a high dimensional problem. One needs to be careful about the possible mis-specification, though. Several goodness-of-fit tests are available in Koenker and Xiao (2002) and He and Zhu (2003).

2.3. **Large Sample Properties.** This section studies large sample properties of two estimators proposed in Section 2.1. The limiting distribution of two-step estimators are most straightforward when the conditional quantile function is estimated by linear quantile regression. We focus on this case. We need the following conditions.

Condition Q. *Suppose that (i) the data $\{(Y_i, X_i, 1 \le i \le n)\}$ are independent and identically distributed across i, (ii) the MAR assumption (2.1) holds, (iii) the conditional density $f_{Y|X}(y|x)$ exists, and is bounded and uniformly continuous in y uniformly in x over the support of X, (iv) the covariates have a bounded moment, i.e., $E||X||^{2+\epsilon} < \infty$ for some $\epsilon > 0$, (v) for any fixed $u \in (0,1)$, there exist positive definite matrices $A_0(p)$ and $A_1(u,p)$ such that $A_0(p) = E[p(x)\, xx^T]$ and $A_1(u,p) = E[p(x) f(Q_Y(u|x)|x)\, xx^T]$.*

Condition (i) allows $F_{Y|X}(y|x)$ to have different shapes across different values of x, so for instance it allows the conditional heteroskedasticity. Condition (ii) is the MAR assumption, which among other things makes the identification possible. Conditions (iii)-(v) are conventional in quantile regression literature. We now summarize the results of the procedure (2.8) in the following Lemma.

Lemma 1. *Assume Condition Q, then the quantile regression estimate $\hat{\beta}(u)$ from (2.8) is*

(1) consistent, $\hat{\beta}(u) = \beta(u) + o_p(1)$, and

(2) $\sqrt{n}(\hat{\beta}(u) - \beta(u))$ converges in distribution to a k-dimensional Gaussian random variables with mean zero and covariance

$$\Sigma(u,u') = (\min(u,u') - uu') A_1^{-1}(u,p) A_0(p) A_1^{-1}(u',p).$$

With the above results in hand, we obtain the estimate for the conditional quantile function by $\hat{Q}_{Y|X}(u|x) = x^T\hat{\beta}(u),\ u \in (0,1)$, and draw imputing values from it. Finally, with the augmented, completed samples, we estimate the marginal quantile function according to (2.4). In the next theorem, we prove the consistency and the asymptotic normality of the marginal quantile estimate defined in (2.5). A previous version of this paper has results for the estimates defined in equation (2.4) but at the cost of longer proofs.

Theorem 1. *Assume Condition Q and further assume that the support of X, $\mathcal{X}$ is a compact subset of R^k. Then*

(1) The marginal quantile estimate is consistent, $\hat{Q}_Y(\tau) = Q_Y(\tau) + o_p(1)$

(2) $\sqrt{n}(\hat{Q}_Y(\tau) - Q_Y(\tau))$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2(\tau)/f(Q_Y(\tau))^2$ where

$$\begin{aligned}\sigma^2(\tau) &= \tau(1-\tau) - E[(1-p(x))F(Q_Y(\tau)|x)(1-F(Q_Y(\tau)|x))] \\ &+ E[(1-p(x))(1-p(x'))f(Q_Y(\tau)|x)f(Q_Y(\tau)|x')\Sigma(F(Q_Y(\tau)|x), F(Q_Y(\tau)|x'))].\end{aligned}$$

Imputing the missing values multiple times increases the efficiency of the quantile function estimator. When we increase the number of imputation m by one, its variance is reduced by the order of $(1/m^2)E[(1-p(x))F(Q_Y(\tau)|x)(1-F(Q_Y(\tau)|x))]$.

Once we have an estimate for the marginal quantile function, estimates for measures for the inequality can be simply obtained by plugging-in the empirical marginal quantile function; the inter-quartile range $\hat{R}_Y = \hat{Q}_Y(.75) - \hat{Q}_Y(.25)$, variance $\hat{\sigma}_Y^2 = \int_0^1 \hat{Q}_Y(u)^2 du$, Lorenz curve $\hat{L}_Y(t) = \int_0^t \hat{Q}_Y(u)du / \int_0^1 \hat{Q}_Y(u)du$, and Gini coefficient $\hat{G}_Y = 2\int_0^1 (t - \hat{L}_Y(t))dt$. To make exposition simple, we present results of the inter-quantile range here, but other measures can be treated similarly.

Lemma 2. *Assume conditions in Theorem 1. Then (1) the estimate for the Inter-quantile range is consistent $\hat{R}_Y - R_Y = o_p(1)$, and (2) $\sqrt{n}(\hat{R}_Y - R_Y)$ converges in distribution to a Gaussian distribution with mean zero and variance*

$$\sigma^2(\tau)/f(Q_Y(\tau))^2 + \sigma^2(\tau')/f(Q_Y(\tau'))^2 - 2\sigma(\tau,\tau')/f(Q_Y(\tau))f(Q_Y(\tau'))$$

where

$$\begin{aligned}&\sigma(\tau,\tau') = \{\min(\tau,\tau') - \tau\tau'\} \\ -\ &E[(1-p(x))\{\min(F(Q_Y(\tau)|x), F(Q_Y(\tau')|x)) - F(Q_Y(\tau)|x)F(Q_Y(\tau')|x)\}] \\ +\ &E[(1-p(x))(1-p(x'))f(Q_Y(\tau)|x)f(Q_Y(\tau')|x')\Sigma(F(Q_Y(\tau)|x), F(Q_Y(\tau')|x'))].\end{aligned}$$

We now present results for data combination. Recall that in sample B we observe full information about responses but in sample A we only observe none or a part of information. We fully observes realized values of covariates in both samples. The first step fit of the conditional quantile function is done by using sample B,

$$\hat{\beta}(u) = \arg\min_{\beta} \sum_{i=1}^{n} \delta_i \, \rho_u(Y_i - X_i^T \beta).$$

In view of Lemma 1, we have the following result.

Lemma 3. *Assume Condition Q but replace the condition (v) with the followings; (v') there exist positive definite matrices $A_0 = E[xx^T]$ and $A_1(u) = E[f(Q_Y(u|x)|x)\, xx^T]$, and as sample size increases, $n_b/n \to p$ and,*

$$n^{-1}\sum_{i=1}^{n} \delta_i \, x_i x_i^T \to p\, A_0, \qquad n^{-1}\sum_{i=1}^{n} \delta_i \, f(Q_Y(u|x_i)|x_i)\, x_i x_i^T \to p\, A_1(u)$$

$$n^{-1}\sum_{i=1}^{n} (1-\delta_i) \, x_i x_i^T \to (1-p)\, A_0, \qquad n^{-1}\sum_{i=1}^{n} (1-\delta_i) \, f(Q_Y(u|x_i)|x_i)\, x_i x_i^T \to (1-p)\, A_1(u).$$

Then (1) the quantile regression estimate is consistent, $\hat{\beta}(u) = \beta(u) + o_p(1)$, and

(2) $\sqrt{n}(\hat{\beta}(u) - \beta(u))$ converges in distribution to a k dimensional Gaussian random variables with mean zero and covariance

$$\Sigma(u, u') = 1/p(\min(u, u') - uu') A_1^{-1}(u) A_0 A_1^{-1}(u').$$

The first step provides the conditional quantile function estimate $\hat{Q}_{Y|X}(u|x)$ for any $x \in \mathcal{X}$, and its inverse $\hat{F}_{Y|X}(y|x)$. Suppose that following procedures around the equation (2.6), we have the completed observations for combined samples $\{(1-\delta_i)Y^*_{x_i,j} + \delta_i Y_i, X_i, 1 \leq i \leq n\}_{j=1}^{m}$ and run a second step quantile regression. The crucial step toward the final result is the following Bahadur representation.

Lemma 4. *Given conditions in Lemma 3, the following asymptotic linear representation for the quantile regression estimates holds,*

$$\begin{aligned}
\sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau)) &= -n^{-1/2} A_1^{-1}(\tau) \sum_{i=1}^{n} \{\delta_i x_i (I(Y_i \leq x_i \beta(\tau)) - \tau)\} \\
&- n^{-1/2} A_1^{-1}(\tau) \sum_{i=1}^{n} \{(1-\delta_i) x_i \, m^{-1} \sum_{j=1}^{m} (I(Y^*_{x_i} \leq x_i \beta(\tau)) - \tau)\} + o_p(1).
\end{aligned}$$

We are ready to state the asymptotic distribution of the quantile regression based on the combined samples.

Theorem 2. *Assume conditions in Lemma 3. Then the quantile regression estimate defined in equation (2.6) is (1) consistent* $\tilde{\beta}(u) = \beta(u) + o_p(1)$*, and (2)* $\sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau))$ *converges in distribution to a mean zero Gaussian random variables with variance matrix*

$$\begin{aligned} & \tau(1-\tau)(p + \frac{1-p}{m})A_1^{-1}(\tau)A_0A_1^{-1}(\tau) \\ + \quad & \tau(1-\tau)\frac{(1-p)^2}{p}E[f(Q(\tau|x)|x)f(Q(\tau|x')|x')\{x^TA_1^{-1}(\tau)A_0A_1^{-1}(\tau)x'\}]A_1^{-1}(\tau)A_0A_1^{-1}(\tau). \end{aligned}$$

3. Comparison with Matching Estimators

3.1. **Matching: Alternative Imputation Method.** Matching is an alternative procedure to draw Y_x^* for missing responses. Matching methods also impute missing responses, but by assigning outcomes of respondents with observably similar attributes to non-respondents. Matching is a nonparametric imputation method. Matching allows imputation without estimating conditional distribution of Y, therefore, it is robust to the model specification. In this sense, matching is comparable to our method 1, the nonparametric first step estimation of conditional quantiles. One needs to bear in mind that just as any other nonparametric methods, the stability and accuracy of final estimates crucially depend on the dimensionality of the problem and the sample size.

There are two conditions that allow matching estimators to work. The first assumption is the equation (2.2), which is often called in literature the 'unconfoundedness' or 'selection on observable' assumption. It requires that conditional on the observable attributes, the distribution of outcomes for respondents is the same as the distribution for non-respondents. But unlike regression based methods, if some of covariates are continuous variables, one cannot condition on the exactly same $X = x$, therefore, in practice, we define neighbors of a missing response based on distances between covariates. Following notations in Imbens (2004), for i-th person with $D_i = 0$, we define $l_m(i)$ the index l that satisfies

$$\sum_{j:D_j=1} I(||x_j - x_i|| \leq ||x_l - x_i||) = m.$$

So $l_m(i)$ is the index of m-th closest unit among respondents in terms of distance in covariates, and $l_1(i)$ is the index of the nearest neighbor. Let $s_m(i) = \{l_1(i), \ldots, l_m(i)\}$ be

the set of index for the first m closest neighbor for the individual i, then the imputed values are determined by

$$\{Y^*_{x_i1}, \ldots, Y^*_{x_im}\} = \{Y_{x_ij} : j \in s_m(i)\}.$$

When the matching is performed based on x values, we will call it a matching on covariates. One alternative is the matching procedure based on the propensity score. If the propensity score $p(x)$ is known, the index set $s_m(i)$ can be constructed by distance $||p(x_j) - p(x_i)||$. This seems to reduces dimensionality in matching procedure, but since the propensity scores is unknown in most cases and need to be estimated, the burden of dimensionality is simply moved to the estimation of propensity score. When matching is based on propensity score, we will call it a matching on propensity score. The second condition for a matching estimator is

$$0 < \Pr(D = d|X) < 1\,, \quad d \in \{0, 1\}.$$

This assumption, often called the 'overlap' or 'common support' condition, requires that for every non-respondent a match can be found among respondents. The missing probability should be a smooth function of x. For example, if missing mechanism is a deterministic process, that is, for certain subgroups with particular $X = x$, people either report or do not report their income with probability one, the common support condition is violated.

This seemingly innocuous condition is in fact crucial for matching estimators. Heckman, Ichimura, and Todd (1998) and Heckman, Ichimura, Smith, and Todd (1998) emphasized the comparability or the balance between two groups as a condition for the success of matching. They argued that in the National Supported Work Demonstration (NSW) data set, famously explored by LaLonde (1986), the treatment group, people who are eligible for job-training programs, are not the same kind of individuals commonly found in CPS or PSID sample, the control group, and this discrepancy caused troubles in many matching methods when applied to NSW data set. One can check the balance of covariates comparing distribution functions of control variables between two groups. In a high dimensional problem, one alternative is to compare distributions of the propensity scores between two groups. Following Dehejia and Wahba (1999) we take the second approach in our simulation exercise.

3.2. **Relative Performances of Estimators.** We examine relative performances of five estimators we discussed so far. For quantile regression based imputation methods, we consider two cases where the conditional quantile function is estimated by a parametric linear

model (QRL) and a semiparametric additive model (QRA). For matching based imputation methods, we consider a matching on covariates (MX) and a matching on propensity score (MP). We also consider the complete case estimator (CC). The object of estimation is the marginal quantile function when some responses are missing at random.

Exact evaluation of relative efficiency is of course difficult. It depends on many factors including the specification of missing probability, the distribution of regressors, and the error distribution of regression models. We do not try to provide a definite answer, instead we choose explicit assumptions on various factors and examine relative performances of five estimators in such circumstances.

Our first simulation exercise follows a particular real data analysis, estimating income equation, as close as possible. The response is defined by the logarithm of weekly wages and regressors include education, gender, race, age, marital status, and union membership. We draw values of covariates from the empirical distributions of regressors we calculate from 2005 CPS Outgoing Rotation Group file. The response is generated

$$y_i = \alpha_{0i} + \sum_{j=1}^{10} \alpha_j w_{ij} + \beta_1 z_i + \varepsilon_i$$

where y represents log weekly wages, dummy variables $w_1, \ldots, w_{10}$ are for female, three ethnic groups (white, black, hispanic), marital status, four levels of schooling, and union status. One continuous variable z is the age of a person and the error ε is drawn from the standard normal distribution. We calculate least squares estimates of the above equation with CPS ORG 2005 file (with respondents only) and obtain the values of the regression coefficients. To be specific $(\alpha_0, \ldots, \alpha_{10}) = (2, -0.21, 0.05, -0.06, -0.11, 0.16, 0.22, 0.34, 0.64, 0.96, 0.23)$ and $\beta_1 = 0.01$.

In order to evaluate the effects of the common support condition on relative performances, we consider two cases, differed by the distribution of propensity score. The Case I for the propensity score is where two groups are well-balanced in terms of their observable attributes. In this case, the propensity scores for most individuals are strictly away from 0 and 1, meaning that everyone in the sample has a decent chance of being observed. This ensures that it is easier to find neighbors not far away from a missing observation. The Case II is where the missing mechanism is close to a deterministic process; the propensity scores among non-respondents and respondents are close to 0 and 1 respectively. This is the case where we basically compare two distinct groups with narrow range for overlap. We expect that nonparametric methods including matching estimators may be less effective in such circumstances. But note that the Case II is rather a common situation in empirical

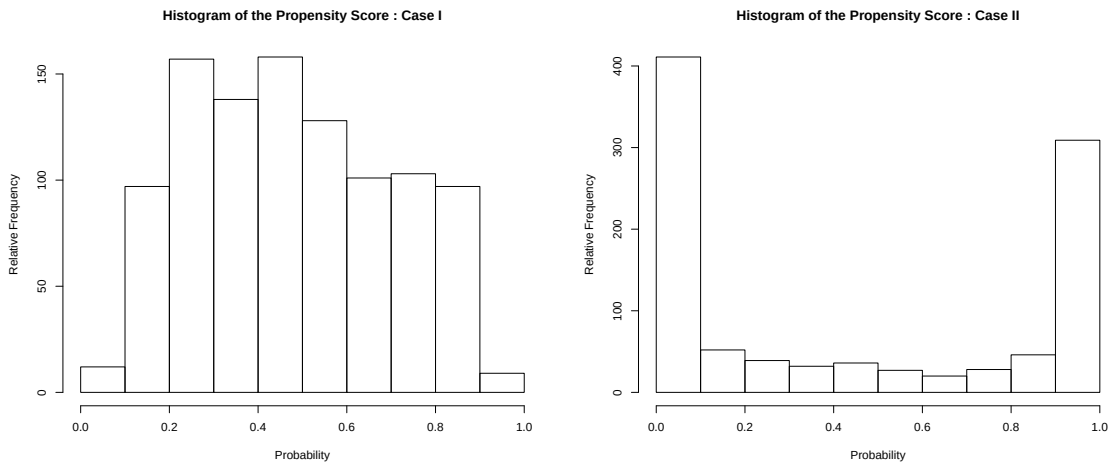

FIGURE 1. Histograms of Propensity Scores. The left-hand side plot (Case I) is when two groups are well-balanced in terms of their observable attributes. The right-hand side plot (Case II) is when the missing mechanism is close to a deterministic process, that is, when we try to match two distinct groups.

analysis. For example, in the LaLonde data set, the distribution of propensity scores is as extreme as the Case II (See Figures 1 & 2 in Dehejia and Wahba (1999)). The propensity score is generated by

$$p(x_i) = \Phi_\sigma(\gamma_{0i} + \sum_{j=1}^{10} \gamma_j w_{ij} + \zeta_1 z_i)$$

where $\Phi_\sigma(\cdot)$ is the normal distribution function with mean 0 and variance σ^2. The values for the parameters are $(\gamma_0, \ldots, \gamma_{10}) = (1.5, .1, -0.05, 0.15, 0.12, -0.4, -0.2, 0.3, -0.5, -0.52, -0.3)$ and $\zeta_1 = -0.03$. Propensity scores for Cases I and II are generated by setting $\sigma = 1$ and $\sigma = 0.2$. Under the MAR assumption, the missing probability and the responses can be correlated because they share the same control variables. In our simulation the correlation coefficient between y_i and $p(x_i)$ ranges from -0.15 to -0.3.

We want to emphasize certain bias issue in the complete case (CC) analysis (See also Cheng and Chu (1996)). Under the MAR assumption, the CC estimator causes bias when we estimate marginal quantiles. The marginal quantile estimator under CC analysis

$$\hat{Q}_c(\tau) = \inf_y \{y : [\sum_{i=1}^{n} D_i \, I(Y_i \leq y) / \sum_{i=1}^{n} D_i] \geq \tau\}$$

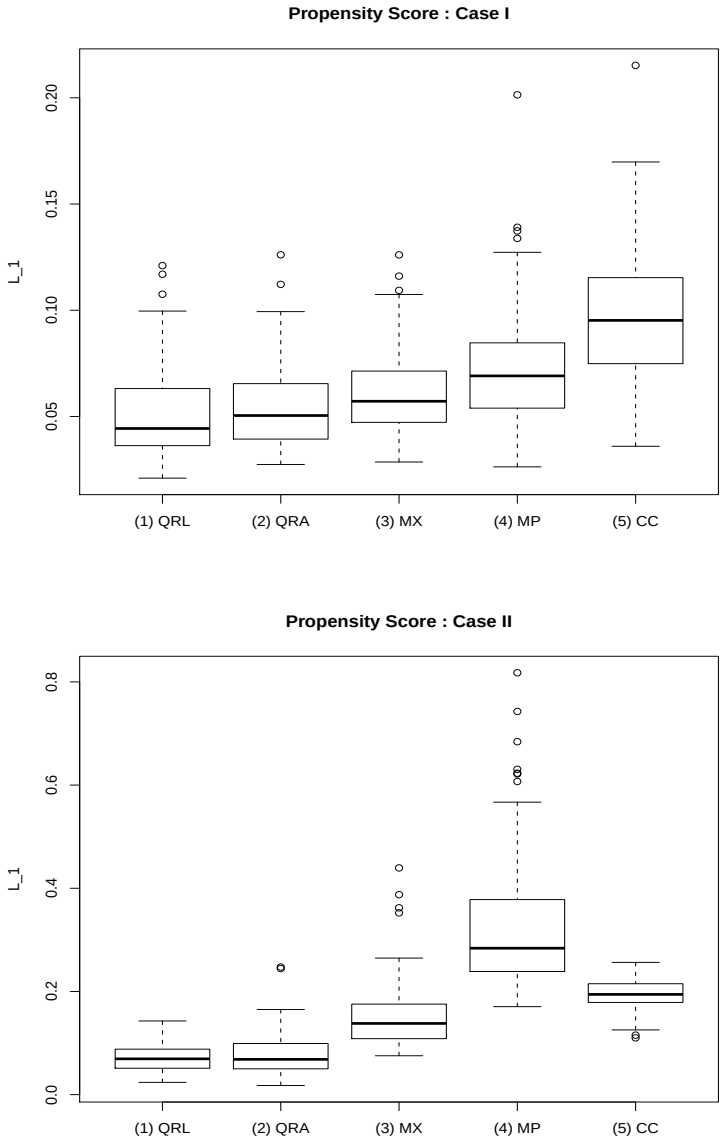

FIGURE 2. Box plots of L_1 distance measures for five estimators. 'QLR' and 'QRA' are imputation methods by linear and additive quantile regression models. 'MX' and 'MP' are matching methods based on covariates and propensity score. Finally 'CC' is the complete-case analysis. Data generating process is described in main text. The sample size is $1,000$, and $p = 0.5$, only a half of responses are observed. Results are based on 100 simulation runs.

would be a consistent estimator for the τ-th quantile of a function $E[p(X)F(y|X)]/p$, which is different from the τ-th quantile of $F_Y(\cdot)$. Unless $p(x) = p$ for all x, that is, unless we are willing to assume a stronger assumption of MCAR, the two quantiles would not be equal.

Non-missing Rate	**Case I**					**Case II**				
	QRL	QRA	MX	MP	CC	QRL	QRA	MX	MP	CC
$n = 500$										
0.2	0.125	0.129	0.155	0.165	0.147	0.196	0.210	0.226	0.500	0.248
0.3	0.107	0.107	0.129	0.143	0.125	0.161	0.186	0.204	0.456	0.224
0.4	0.088	0.085	0.105	0.104	0.112	0.115	0.128	0.180	0.419	0.216
0.5	0.073	0.074	0.078	0.089	0.105	0.111	0.112	0.163	0.361	0.214
0.6	0.065	0.065	0.069	0.075	0.084	0.082	0.088	0.147	0.335	0.172
0.7	0.051	0.047	0.056	0.066	0.072	0.059	0.063	0.109	0.260	0.129
$n = 1,000$										
0.2	0.085	0.086	0.107	0.128	0.118	0.137	0.160	0.204	0.433	0.229
0.3	0.068	0.072	0.089	0.100	0.104	0.111	0.130	0.185	0.438	0.216
0.4	0.062	0.061	0.072	0.084	0.099	0.087	0.102	0.160	0.380	0.200
0.5	0.048	0.053	0.058	0.070	0.090	0.069	0.083	0.155	0.346	0.194
0.6	0.042	0.046	0.051	0.060	0.076	0.055	0.066	0.142	0.306	0.169
0.7	0.040	0.036	0.044	0.051	0.066	0.044	0.049	0.101	0.259	0.127
$n = 5,000$										
0.2	0.041	0.045	0.058	0.081	0.113	0.055	0.068	0.154	0.375	0.232
0.3	0.034	0.038	0.049	0.071	0.102	0.050	0.061	0.147	0.363	0.226
0.4	0.027	0.031	0.039	0.060	0.094	0.040	0.053	0.141	0.340	0.215
0.5	0.023	0.027	0.033	0.052	0.086	0.030	0.038	0.128	0.306	0.200
0.6	0.020	0.023	0.027	0.041	0.077	0.027	0.032	0.114	0.264	0.171
0.7	0.016	0.018	0.022	0.034	0.066	0.020	0.024	0.086	0.257	0.131

TABLE 1. Averages of L_1 distance measures based on 100 simulation runs. 'QLR' and 'QRA' are imputation methods by linear and additive quantile regression models. 'MX' and 'MP' are matching methods based on covariates and propensity score. Finally 'CC' is the complete-case analysis. Data generating process is described in main text. Three sample sizes 500, $1,000$, $5,000$, and six different values for non-missing rates ranges from $p = 0.2$ to $p = 0.7$.

Throughout simulations, we consistently use $m = 1$ for the number of imputation, that is, for the quantile based methods, we draw and impute missing responses only once, for the matching method we use the nearest neighbor method. For the additive quantile regression step, we regard all dummy variables as linear covariates and the age variable as a nonparametric component. The propensity score matching is performed assuming that we know the true propensity score, so we ignored the estimation errors in propensity scores.

The Box plots in Figure 2 shows results of 100 simulation runs for one experimental design. We obtain five marginal quantile estimates $\hat{Q}^s_Y(\tau)$ where the index s takes values in $\text{QRL}, \text{QRA}, \text{MX}, \text{MA}, \text{CC}$. We compared the estimated marginal quantile functions with an infeasible estimate $\hat{Q}^t_Y(\tau)$ which is the empirical quantile function with full, both observed

and missing, observations. Two measures of performance, L_1 and L_2 measures, are defined by

$$L_l = \left(\int_{(0,1)} |\hat{Q}_Y^t(\tau) - \hat{Q}_Y^s(\tau)|^l d\mu(\tau) \right)^l, \quad l = 1, 2,$$

where $d\mu(\tau)$ is an empirical density of equally distanced 200 points in the unit interval with equal weights.

Non-missing Rate	**Case I**					**Case II**				
	QRL	QRA	MX	MP	CC	QRL	QRA	MX	MP	CC
$n = 500$										
0.2	0.157	0.160	0.198	0.211	0.180	0.235	0.244	0.287	0.584	0.275
0.3	0.135	0.135	0.167	0.187	0.148	0.197	0.224	0.253	0.535	0.247
0.4	0.115	0.109	0.136	0.138	0.133	0.146	0.161	0.231	0.491	0.234
0.5	0.096	0.096	0.109	0.121	0.121	0.137	0.139	0.206	0.427	0.228
0.6	0.085	0.083	0.094	0.101	0.099	0.106	0.110	0.185	0.397	0.184
0.7	0.069	0.063	0.077	0.088	0.084	0.081	0.082	0.144	0.314	0.141
$n = 1,000$										
0.2	0.107	0.108	0.138	0.166	0.140	0.166	0.189	0.250	0.500	0.247
0.3	0.087	0.090	0.118	0.131	0.119	0.137	0.155	0.235	0.515	0.230
0.4	0.079	0.079	0.095	0.111	0.112	0.107	0.124	0.201	0.445	0.212
0.5	0.063	0.067	0.079	0.092	0.101	0.087	0.104	0.193	0.406	0.203
0.6	0.054	0.059	0.069	0.079	0.085	0.072	0.082	0.178	0.362	0.177
0.7	0.052	0.048	0.058	0.067	0.074	0.059	0.063	0.129	0.305	0.135
$n = 5,000$										
0.2	0.051	0.062	0.075	0.105	0.120	0.070	0.087	0.188	0.435	0.237
0.3	0.043	0.055	0.062	0.089	0.106	0.064	0.079	0.184	0.423	0.230
0.4	0.036	0.046	0.050	0.077	0.098	0.052	0.067	0.176	0.395	0.218
0.5	0.031	0.040	0.044	0.067	0.089	0.039	0.050	0.158	0.367	0.203
0.6	0.027	0.033	0.035	0.054	0.079	0.035	0.043	0.142	0.314	0.174
0.7	0.022	0.027	0.029	0.045	0.068	0.025	0.033	0.108	0.301	0.134

TABLE 2. Averages of L_2 distance measures based on 100 simulation runs. 'QLR' and 'QRA' are imputation methods by linear and additive quantile regression models. 'MX' and 'MP' are matching methods based on covariates and propensity score. Finally 'CC' is the complete-case analysis. Data generating process is described in main text. Three sample sizes 500, 1,000, 5,000, and six different values for non-missing rates ranges from $p = 0.2$ to $p = 0.7$.

The overall non-missing rate, one minus the overall missing rate, ranges from 0.2 to 0.7. In an extreme case where $p = 0.2$, only 20% of responses are assumed to be observed, which seems extreme. But especially in the data combination exercise this figure is fairly

common. Often, the incomplete data set with many observations is matched or combined with a complete data set with fewer observations. For example, our analysis in Section 5, the Census including the entire population of a country do not have complete information about income, so we use a much smaller survey, the Labor Force Survey, to obtain the imputed values. In those cases the non-missing rate p can be quite small.

Figure 2 shows box-plots of L_1 distance measures based on 100 simulation runs for the case where $p = 0.5$ and sample size $1,000$. Further details are shown in Table 1 for L_1 measures and Table 2 for L_2 measures. A few observations are in order. First, in Case I, when two groups are well-balanced, two quantile regression methods (QRL, QRA) and two matching estimators (MX, MP) are quite comparable in terms of efficiency. Quantile regression based imputation methods outperform matching estimators by $10\% \sim 25\%$ depending on circumstances. In general, matching estimators are less efficient, but it is a cost that a purely nonparametric method needs to pay when experimental designs are more or less parametric. Second, in Case II, with unbalanced groups, the matching estimators tend to underperform significantly. This is an instance where a matching proves to be unreliable and we may get benefits from parametric specifications via regression models. Third, between two matching methods, the propensity score matching (MP) is more problematic than the covariate matching (MX). It seems to be a by-product of our experimental design because we manipulate the distribution of propensity scores to reproduce extreme cases.

Fourth, even when sample size increases, the bias in complete-case analysis does not go away. Unlike other estimators, the distance measures in CC estimator fail to be decreased along with the sample size. It shows that the bias issue in CC estimator is a serious concern. Fifth, the performance of additive quantile regression model (QRA) is more or less the same as that of linear quantile regression model (QRL). The data generating process is indeed linear in parameters, so QRL is supposed to be efficient but one can suspect that this might not be the case when we introduce non-linearity in the model.

To further explore this possibility, we have a second experimental design where the log-wages are generated by

$$y_i = \alpha_{0i} + \sum_{j=1}^{10} \alpha_j w_{ij} + \beta_1 z_{1i} + h(z_{2i}) + \varepsilon_i$$

where the newly added regressor z_2 representing job experience has a nonlinear effect, $h(a) = 1/(1 + e^{-a})$ and $a = (z_2 - \bar{z}_2)/5$. Other factors in the experiment, including the distribution of regressors and propensity scores remain the same. In QRA, we regard all dummy variables as linear covariates and age (z_1) and job experience (z_2) as nonparametric

Non-missing Rate	**Case I**					**Case II**				
	QRL	QRA	MX	MP	CC	QRL	QRA	MX	MP	CC
$n = 500$										
0.2	0.127	0.134	0.149	0.167	0.134	0.185	0.216	0.217	0.507	0.223
0.3	0.101	0.112	0.130	0.141	0.117	0.137	0.152	0.174	0.447	0.191
0.4	0.087	0.089	0.102	0.120	0.102	0.112	0.134	0.155	0.461	0.180
0.5	0.076	0.074	0.088	0.100	0.087	0.097	0.121	0.133	0.393	0.158
0.6	0.058	0.061	0.067	0.084	0.079	0.082	0.100	0.113	0.356	0.140
0.7	0.048	0.048	0.054	0.066	0.056	0.058	0.075	0.090	0.248	0.115
$n = 1,000$										
0.2	0.092	0.092	0.117	0.130	0.135	0.135	0.137	0.178	0.480	0.219
0.3	0.072	0.075	0.091	0.119	0.114	0.096	0.107	0.147	0.467	0.223
0.4	0.059	0.061	0.071	0.085	0.104	0.094	0.104	0.140	0.392	0.213
0.5	0.051	0.050	0.060	0.077	0.093	0.067	0.082	0.116	0.341	0.198
0.6	0.044	0.043	0.050	0.059	0.075	0.058	0.062	0.105	0.325	0.168
0.7	0.036	0.035	0.039	0.050	0.062	0.043	0.047	0.071	0.258	0.128
$n = 5,000$										
0.2	0.040	0.041	0.052	0.085	0.118	0.062	0.073	0.116	0.423	0.225
0.3	0.033	0.035	0.041	0.076	0.108	0.045	0.050	0.103	0.370	0.224
0.4	0.030	0.032	0.033	0.059	0.099	0.045	0.051	0.093	0.315	0.216
0.5	0.023	0.027	0.030	0.053	0.088	0.033	0.040	0.092	0.303	0.204
0.6	0.020	0.022	0.021	0.044	0.079	0.025	0.030	0.071	0.274	0.173
0.7	0.017	0.017	0.018	0.035	0.069	0.020	0.022	0.051	0.261	0.135

TABLE 3. Averages of L_1 distance measures based on 100 simulation runs. Results for the second experimental designs.

components. Results for this variation are in Table 3. To save space, we report summaries of L_1 measures only, the L_2 measure shows the same tendency. The results are quite comparable to Table 1 although the distance measures are increased slightly reflecting the added complexity in the model. In the second experimental design, QRL is mis-specified, but its performance is almost the same as QRA. It shows that QRL is robust at least up to some degrees of model mis-specification.

4. TREND OF THE U.S. INCOME INEQUALITY

The monthly CPS income file is the basis of official reports on the U.S. unemployment rate and income inequality. But CPS like most other household surveys conducted by the U.S. Census Bureau suffer from the increasing trend of non-response rates concerning questions on earnings. During the periods 1998-2003, only 50% of white and Hispanic males

in monthly CPS Outgoing Rotation Group reports their income. For black males, earnings data is available only for 38% of the sample (Heckman and LaFontaine (2006)).

In response to this worrying trend, the census allocates or imputes earnings for some of those with missing income values. During the 1980s, the earning imputation rates were stable around 15%, but it has steadily increased in recent years, and in 2001 31% of weekly wage data in CPS earnings file were imputed by the census (Hirsch and Schumacher (2004)).

The census use a 'hot deck' procedure to match each non-respondent to a respondent whose background characteristics are identical. The main categories for matching are gender, age, race, occupation, and education. There are several important studies which analyze the potential bias issue due to the allocated earnings in the CPS. Hirsch and Schumacher derive an expression for 'match bias', which occurs when certain individual attributes, such as union status, are included in covariates of a wage regression but excluded from the matching criteria. It can be understood as a form of the omitted variable bias where non-match characteristics such as union status fail to be included in the first step matching procedure.

Another source of bias occurs when the match criteria uses too coarse categorical information (Bollinger and Hirsch (2006)). For example, Census use three education categories; less than high school, high school graduate and some colleges, and bachelor's degree and above. Heckman and LaFontaine found that the census matching procedure causes an upward bias in returns of education for those who obtains a high school certification through an equivalence exam. For non-respondents who obtain high school degree through equivalence certificate, their wages are frequently matched to wages of high school graduates or those with some college education, therefore, potentially inflated.

When the main research question is to know the stochastic relationship between the response and covariates, in other words, when we want to know certain features of conditional distribution of wage given individual attributes, one can simply discard individuals with missing or allocated values and fit a model only with those observations with complete information. This standard treatment, the complete-case analysis, was taken in all of the above-mentioned research, but unfortunately can not be used when we are interested in the marginal distribution of response variable. It introduces bias unless we are willing to assume that the missing mechanism is completely random. Since income inequality measures are in essence dispersion variables of the marginal distribution of income, this complete-case analysis would bring bias in inequality measures. A coherent imputation method is needed.

In Table 4 we show both missing and allocation rates in CPS Outgoing Rotation Group (CPS-ORG) from the period of 1979 to 2005. Not all of missing earnings are imputed,

Year	Private Sector Sample			Ethnic Minority Sample		
	Missing	Allocated	Total	Missing	Allocated	Total
1979	0.221	0.131	0.352	0.264	0.125	0.389
1980	0.231	0.125	0.355	0.274	0.118	0.392
1981	0.233	0.119	0.352	0.286	0.112	0.398
1982	0.252	0.105	0.356	0.307	0.104	0.411
1983	0.248	0.104	0.352	0.308	0.104	0.412
1984	0.222	0.115	0.337	0.268	0.122	0.390
1985	0.215	0.112	0.327	0.258	0.122	0.380
1986	0.209	0.085	0.294	0.251	0.086	0.337
1987	0.200	0.109	0.310	0.239	0.120	0.359
1988	0.193	0.118	0.311	0.231	0.127	0.359
1989	0.191	0.111	0.302	0.229	0.125	0.354
1990	0.192	0.113	0.305	0.234	0.136	0.370
1991	0.203	0.109	0.312	0.246	0.132	0.377
1992	0.206	0.111	0.317	0.252	0.135	0.387
1993	0.200	0.121	0.321	0.241	0.154	0.394
1994	0.176	NA	NA	0.210	NA	NA
1995	0.172	NA	NA	0.201	NA	NA
1996	0.177	0.180	0.356	0.208	0.224	0.432
1997	0.169	0.183	0.351	0.204	0.228	0.431
1998	0.163	0.195	0.358	0.190	0.241	0.431
1999	0.158	0.230	0.388	0.180	0.277	0.457
2000	0.155	0.250	0.405	0.173	0.294	0.468
2001	0.165	0.256	0.421	0.188	0.314	0.502
2002	0.171	0.250	0.421	0.200	0.301	0.500
2003	0.173	0.262	0.435	0.202	0.303	0.505
2004	0.171	0.258	0.429	0.202	0.307	0.509
2005	0.166	0.255	0.421	0.193	0.310	0.503

TABLE 4. Missing and Allocation Rates in CPS Outgoing Rotation Group earnings files from the period of 1979 to 2005.

though. A significant portions of non-responses are not allocated by the Census and remained to be missing in CPS-ORG earning files. We use all employed wage workers in private sectors ages 16 or older. When respondent's weekly wages are too small, meaning that they are less than the first percentile of all reported earnings at a given year, we exclude them from our samples. We report missing and allocation rates for both all private sectors employees and ethnic minorities.

From 1979, the Census started to include imputed wages in the edited earning fields as well as allocation flags indicating which individuals reported their income and which had their wages imputed. Both or one of weekly wages and hours of works per week were

imputed. The 'Missing' in Table 1 is the proportion of missing weekly wages or hours of work in edited fields. The 'Allocated' is the proportion of imputed weekly wages or hours of work. The 'Total' is simply the sum of these two rates, it captures the proportion of respondents whose earnings have not been reported in the original surveys.

Beginning in 1989, Hirsch and Schumacher reports that the earning flags are unreliable. The flags identify only a quarter of those who had their wages assigned. To fix this inconsistency, they recommend to compare unedited and edited variables. Those who with missing unedited earnings or hours, but have proper edited earnings or hours will be treated as those whose wages are allocated. From 1994 to 1995, the unedited weekly wages are not available, nor does the imputation flags. So allocation rates for those two years are left unspecified. Accurate flags variables were included again from 1996 ORG files.

The increasing trend of allocation and missing rates are obvious from the table. The overall missing rate have been steadily increased and reached above 42% for the whole population in our sample and 50% for ethnic minorities which includes black, hispanic, and asian ethnic groups. If we simply discard all missing or imputed earnings as suggested by previous studies, we end up throwing too many individuals away especially in ethnic minority groups and that would hurt precision of inequality measures.

We now describe the quantile regression based imputation method proposed in this paper. Hourly wages are defined by the weekly earnings divided by the hours of work per week. Nominal wages are adjusted by Consumer Price Index, so they are comparable in 2000 dollars. Regressors include age, gender, marital status, ethnicity (white, black, hispanic, asian), schooling (high school drop-outs or less, high school graduates, some college, associate or bachelor's degree, postgraduate degrees), and union status. The information on union status is available from 1984, so we restrict our sample periods from 1984 to 2005.

The imputation is performed as we describe in the earlier section. For an equally spaced grid $u = \{0.01, 0.02, \ldots 0.99\}$, we run quantile regression of logarithm of earnings on the control variables each year. Based on the fitted conditional quantile functions, a random imputation described in (2.3) is performed, which provides the filled-in data set for the entire sample. Based on these completed hourly wages, we calculate the inter-quartile range of the marginal distribution of wages in each year. Figures 3-5 report this inequality trends for some selected sub-groups. When the inequality measures are estimated with the filled-in values by our imputation procedures, they are drawn in black, solid lines. For the matter of comparison, we show two other possible estimates for the trends; inequality measures by simply dropping both imputed and missing wages, which are in blue, dotted lines, and

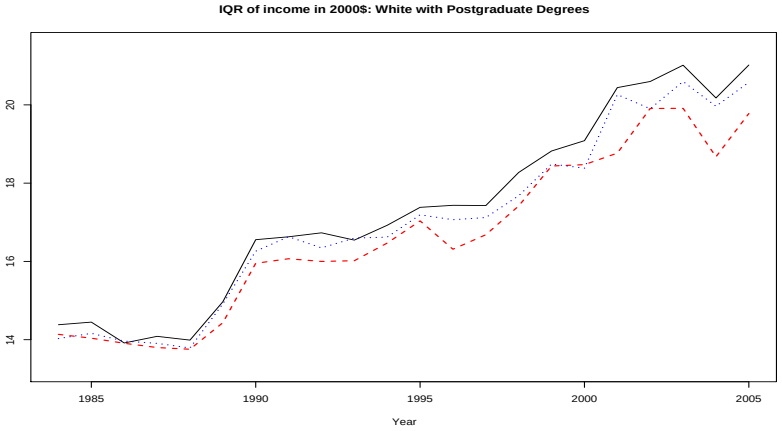

FIGURE 3. Inter-quartile Range. White with Postgraduate degrees.

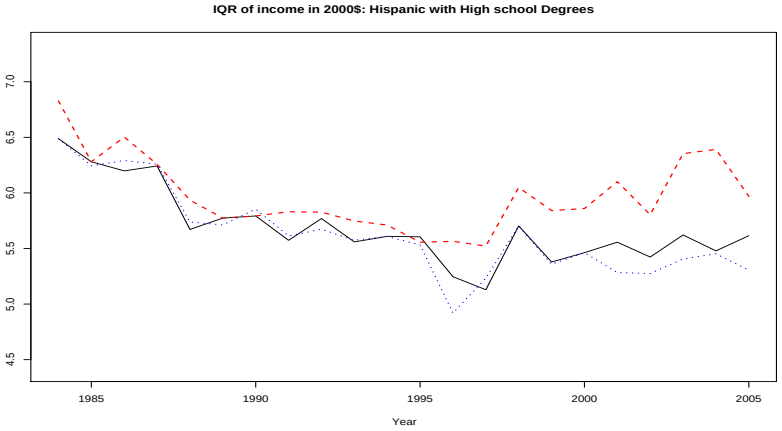

FIGURE 4. Inter-quartile Range. Hispanic, High school graduates.

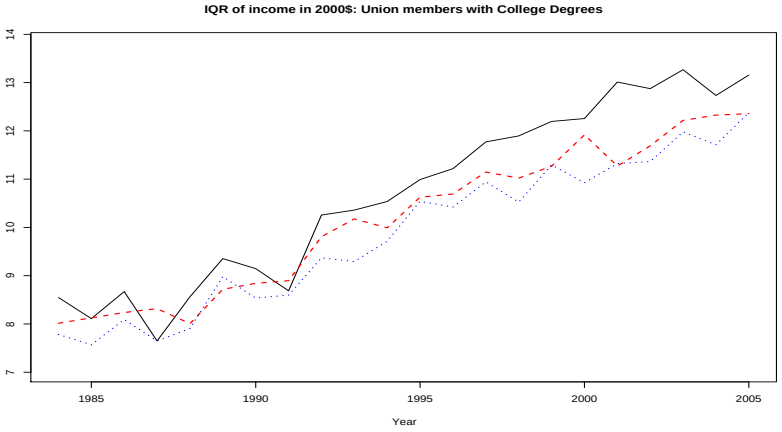

FIGURE 5. Inter-quartile Range. Union members with Bachelor degrees.

those by using the Census allocated earnings but dropping missing values are in red, dashed lines.

Over the sample periods, the overall inequality has risen by 26.5%; the inter-qartile range of hourly wages for the whole sample measured in 2000$ has been increased from $8.39 in the year 1984 to $10.61 in 2005. But this trend has not been shared equally by all sub-populations, some groups has experienced sharper growth in inequality among themselves, but others hardly have had any increase.

Figure 3 shows the trend for white with postgraduate degrees, i.e., masters, doctoral, or other professional degrees. It is well known that the higher education has contributed to the increasing trends of inequality over time, but it shows that even among people with postgraduate degrees, wages has become increasingly dispersed. The inter-quartile range for hourly wages has been increased by 46.5%, from $14.38 in 1984 to $21.01 in 2005. When we use the Census imputed earnings, it systematically under-estimate the dispersion of the overall income in this subgroup. This discrepancy or bias is less severe in the beginning of the periods, which reflects the fact that the missing or allocation rates are small in the earlier periods, but it gets bigger, and in 2000s the gap becomes 5-8%. When we entirely drop all missing and the Census allocated earnings, it helps to correct the bias, but not entirely.

Fig 4 shows the trends among hispanic group with high school degrees. It shows that the inequality in this subgroup has stagnated or actually decreased over time. The inter-quartile range for hourly wages in this group is $6.48 in 1984 and barely moved to $5.61 in 2005, measured in 2000$. Unlike the previous subgroup, the use of the Census allocated income over-estimate the inequality throughout the periods. In 2000s, it inflate the actual inequality by 6-14%. We again note that dropping all incomplete observations helps to correct the Census allocation induced bias. But we emphasize that it is not always the case. Figure 5 shows the trend of income inequality among union members with bachelor degrees. The within group inequality has been increased by 53.9%, from $8.54 in 1984 to $13.15 in 2005, this rate is higher than the overall trend. We note that dropping incomplete observations all together do not correct the bias for this sub-group, which is present in the trend with the Census allocated income. Both trends under-estimate the actual inequality by 4-15%.

When features in the marginal distribution of the response, earnings in our exercise, is the variable of interest, a special care is warranted for the non-response. The use of the Census provided imputed earnings should be carefully handled. Dropping all the non-respondents

would be either misleading in the worst case or inefficient even in the best case. The quantile regression based imputation provides a promising way to correct the high non-response rate in CPS.

5. Merging Two Surveys: Income Equation Estimation in South Africa

The Census in South Africa is conducted by the Statistics South Africa in every ten years. The people who were present in the country on the night before the Census day, regardless of the nationality and types residents, are all visited. The Census data covers 10% of the South African households, including more than 3.7 million individuals in the latest survey conducted in October 2001. The sampling unit is the household but respondents in each household report information on behalf of other members in the household. For example the Census income files report both individual and household earnings. The large sample size and its rich contents of information regarding demographic changes and social mobility make the Census desirable to study social and economic changes, but unfortunately, one of important variables, the earning, is only available as a categorical variable. Respondents are asked to indicate the income category that best describes the personal or household incomes. In 2001 Census, there were 12 categories of earnings to choose from. Earnings as a categorical variable makes it difficult to apply the standard quantile regression techniques which typically requires a continuous response variable.

This difficulty can be resolved if we have an independent sample and it has both continuous earning variable and all relevant determinants of income. In South African, the Labor Force Survey (LFS) also conducted by the Statistics South Africa, fits this criteria. It was conducted twice per year since 2000, but went through a major revisions in 2005 and has been available in every quarter. It has information on economic activity and employment and other relevant demographic variables. In the survey conducted in February 2001, there are about 28,000 households with 72,000 individuals. The important distinction for our purpose is that it reports income in its original form. By combining the Census with the Labor Force Survey, we can overcome the short-comings in earnings variable in the Census, but still utilize its large sample size and its representativeness of the population.

Our LFS sample is a subsample of the survey conducted in February 2001, it has all employed individuals with paying jobs, age from 18 to 65. People who runs their own business or work for their own or family farms are not included. This leaves us 18,128 workers, among which about 25% refuse to answer their earnings or respondents do not know other family member's income. We exclude workers whose earning are too low, when they are below the first percentile of all earnings. The covariates that used to explain

the income are age, gender, marital status, ethnicity (4), schooling (6), industry (11). Although the union membership is available in LFS sample, it is not included in the Census questionnaire, so we do not use the union status as a control variable. The size of the LFS sample we used in the first step quantile regression is 13,486.

We now describe the 2001 Census data. Because the huge sample size, we randomly select 10% individuals among the original Census sample, which represent 1% of the South African population. Our sample include workers who reported to have jobs during the seven days before the Census date and who worked for more than 5 hours per week. Workers who report to make no income or those who work for their own farms are excluded. We have 69,312 workers in our Census sample. Around 6% of the Census earnings files were imputed by the Statistics South Africa by a hot deck procedure. Just as the LFS sample, we have age, gender, marital status, ethnicity, schooling, industry as covariates.

Variable	LFS	Census
Female	0.372	0.430
Age	37.280	38.450
Married	0.594	0.612
Single	0.333	0.315
White	0.138	0.194
Black	0.654	0.633
Indian or Asian	0.033	0.046
Other Ethnic groups	0.175	0.127
No Schooling	0.067	0.087
Some Primary	0.159	0.123
Complete Primary	0.068	0.057
Some Secondary	0.309	0.293
Complete Secondary	0.324	0.374
Bachelor or Higher	0.060	0.065
sample size	18,128	69,312

TABLE 5. Summary Statistics of the LFS and the Census Samples. Proportions of each variable are reported except age. Compositions of variables are very similar across two samples, with possible exceptions of female and white dummy variables. Earnings are not reported here because the Census only reports categorical earning variable, which makes wages in two samples incomparable.

Table 5 gives the summary statistics. Throughout the variables, two samples we use, the LFS and the Census, are quite homogeneous. The Census sample has higher proportions of women and white compared to the LFS, but the compositions of other control variables are

very similar. We do not provide summary statistics for annual income because two samples are non comparable.

With the LFS sample, we run quantile regression of logarithm of yearly earnings on control variables for each percentile $u = \{0.01, 0.02, \ldots 0.99\}$. This gives us a conditional quantile function estimate for any given values of covariates, with which we generate imputed earnings for the Census sample. For each individual in the Census, the values of covariates are given, then the conditional quantile function of income given covariates can be fixed. We randomly draw annual wages from this conditional quantile function, under the constraint that the person's income should be inside of the interval that is provided by the Census earnings files. When earnings were imputed, we disregard the interval wage information provided by the Census, and treat them as completely missing observations.

With the imputed earnings in the Census sample, we have the continuous income variable in both LFS and Census samples. The sample size of the merged sample is 82,798. We run quantile regressions of logarithm of wages on the set of covariates at various values of τ, $\tau = \{0.05, 0.1, 0.2, \ldots, 0.9, 0.95\}$. Figure 4 shows this second stage quantile regression results for some selected set of covariates. It has the point estimates and the confidence bands of both quantile regression and the least squares. We will discuss effects of race and education on the conditional distribution of earnings.

On the average, the least squares fits tell us that white population earn more than other ethnic groups (indians, asians, and other racial groups) by 55%, while black population earns less than other groups by 40%. There is racial difference in terms of earnings, but an interesting fact is that the direction of this bias appears differently among white and black workers. For a white worker, the wage gap is 32% at the lower 5-th percentile in wage distribution, but it becomes as much as 80% at the upper 95-th percentile compared to the wages of other ethnic groups. For a black worker, the gap is greater at the lower end of wage distribution; it is -48% at the 10-th percentile, but the gap keeps shrinking as he moves to the higher end to become -35% at the 90-th percentile of income distribution.

One possible explanation is that different ethnic groups tend to have different types of occupations in terms of earning power. In an extreme case, suppose that white workers hold the most lucrative positions, low paying positions are mainly filled by black workers, and those in the middle are claimed by other ethnic groups. Then when we compare income distributions of white and other ethnic groups, the biggest gap will be found in the upper 95-th percentile but the difference will get smaller in the low 5-th percentile. It reflects

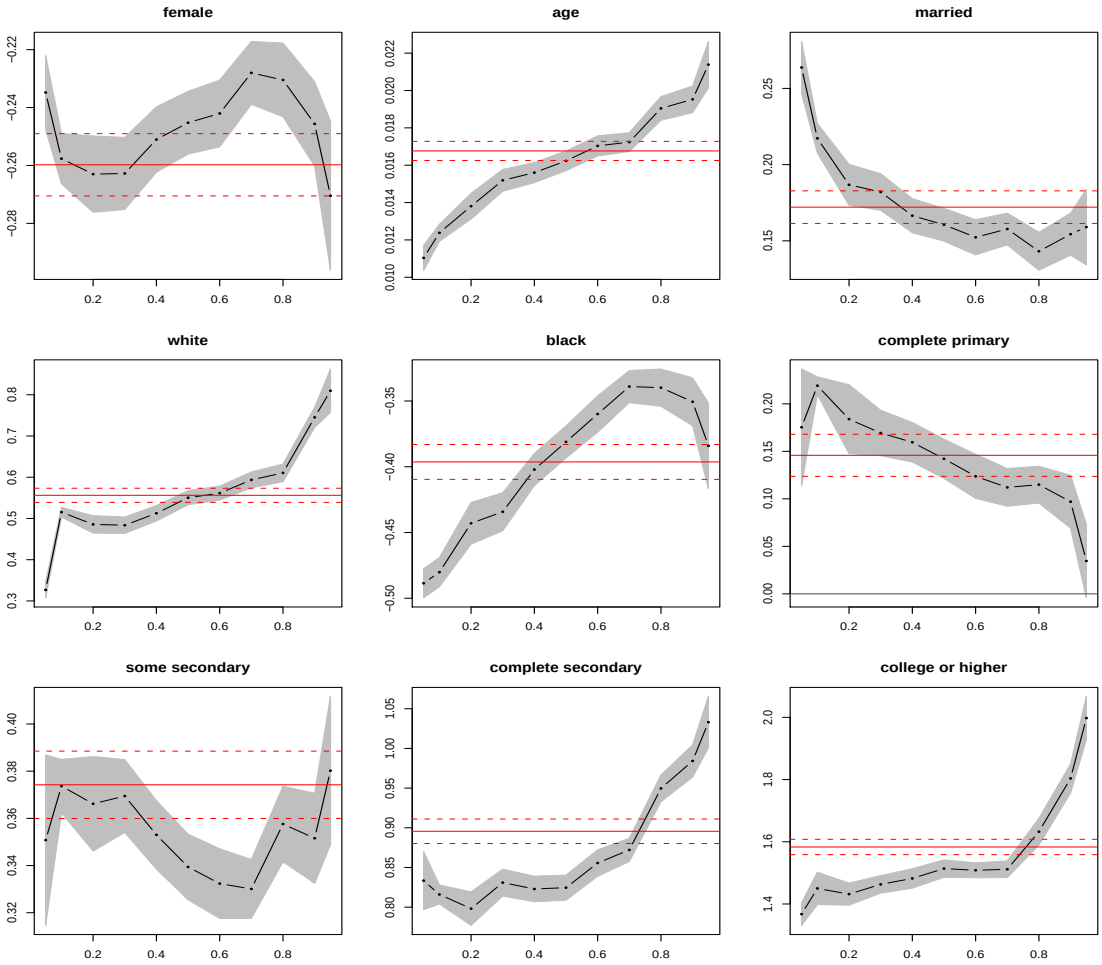

FIGURE 6. Quantile regression estimates for the income equation with combined samples. The South African Census data which has categorical income variable is combined with the Labor Force Survey which has continuous income variable.

differing levels of overlaps. When we compare income distributions of black workers and other ethnic groups, we expect to observe the opposite.

Since we control for education level and the industry, this tendency is not easily explained without considering possible discrimination in labor markets. In this regard, white workers at higher end jobs seem to be the main recipients of favoritism. Black workers at high paying jobs tend to get more equal treatments. It is perhaps because after long history of successful civil right movement in South Africa, social awareness against the discrimination starts to make impacts on more visible part of society, high paying positions.

Four levels of years of schooling, 'complete primary', 'some secondary', 'complete secondary', and 'bachelor or higher degrees', are compared to the combined groups of 'no schooling' and 'some primary'. Although additional education attainment always helps workers earn more, their relative impacts on income distribution are different. Having completed primary education helps workers on the average, but its effect on the high end of the income distribution almost disappears, 3.5% at the 95-th percentile. For the higher end of income distribution, the relative merit of completing primary schools compared to not finishing primary schools is insignificant. It is perhaps because both educational levels are not adequate preparation for well-paying jobs, therefore, those who tend to earn more with less schooling may attain their economic status through other factors such as fortune in labor markets.

The effect of having some years in secondary schools appears to be positive and stable. But the merit of having completed secondary schools (grade 12 with diploma or certificate) or having bachelor or postgraduate degrees appears significantly and gets better as one moves to the higher end of the income distribution. Earnings of those with college education or more are 140% higher than those who do not complete primary schools at the 5-th percentile, but the gap consistently gets larger and becomes 200% at the 95-th percentile in income distribution. The benefits of higher education is evident in South African labor markets.

6. Conclusion

This paper achieves two goals. One is to develop a quantile regression analysis with possibly missing responses. We apply this methodology to data combination problem. The other is to propose a flexible imputation method which can be used as an alternative to matching methods. We apply it to the trend of income inequality in the U. S. where the high non-response rates poses a threat to the validity in conventional analysis. We establish large sample properties of the proposed method. Quantile regression based imputation methods show superior performance compared to matching estimators when the data generating structure has some degree of structures or when there are many control variables. If one does not always want to use a fully nonparametric imputation method like matching, we argue that the method in this paper can be an effective alternative.

Appendix A. Proof of Theorems

This appendix collects proofs of Theorems and Lemmas appeared in the main text. For notational convenience, we will drop the subscripts in distribution and quantile functions.

$||x||$ means the Euclidean norm of a vector x. We first prove Lemma 1 and present its Corollaries. These results will be used later to prove Theorems 1 and 2.

A.1. **Proofs of Lemma 1 and its Corollaries.** Define

$S_n(u,\beta) = n^{-1}\sum_{i=1}^n D_i x_i\{I(y_i \le x_i^T\beta) - u\}$ and $S_\infty(u,\beta) = E[Dx\{I(y \le x^T\beta) - u\}]$.

(a) Note that $\beta(u)$ is the unique solution of $S_\infty(u,\beta) = 0$. It is because $S_\infty(u,\beta(u)) = E[Dx\{I(y \le x^T\beta(u)) - u\}] = E[p(x)xE[I(y \le x^T\beta(u)) - u|x]] = E[p(x)x\{F(x^T\beta(u)|x) - u\}] = E[p(x)x\{u-u\}] = 0$.

(b) We show the uniform convergence, that is, for any compact set $\mathcal{B}$, $S_n(u,\beta) = S_\infty(u,\beta) + o_p(1)$ uniformly in $\beta \in \mathcal{B}$ for any fixed u. The pointwise convergence holds by Khintchine's law of large numbers because $-||x_i|| \le D_i x_i\{I(y - x^T\beta \le 0) - u\} \le ||x_i||$ and the finite first moment under the given conditions. Next we claim that the empirical process $\beta \to S_n(u,\beta)$ is stochastically equicontinuous. To show this, we follow arguments in Qu (2008, Lemma A.1). For a given u, $S_n(u,\beta(u)) = n^{-1}\sum_{i=1}^n D_i x_i\{I(y_i \le x_i^T\beta(u)) - u\} = n^{-1}\sum_{i=1}^n D_i x_i\{I(F(y_i|x_i) \le u) - u\}$. Let v_i be a uniform random variable, then $S_n(u,\beta(u)) = n^{-1}\sum_{i=1}^n D_i x_i\{I(v_i \le u) - u\}$. Since $E[I(v_i \le u)] = u$, the vector $D_i x_i\{I(v_i \le u) - u\}$ is a sequence of a martingale difference, then the stochastically equicontinuity holds. The stochastic equicontinuity coupled with the pointwise convergence implies the uniform convergence.

(c) Given the identification (a) and the uniform convergence (b), the consistency follows by the usual argument of Z-estimator, for example, by Theorem 5.9 in van der Vaart (1998). To establish the asymptotic normality, we show the following steps.

(d) We claim that $\sqrt{n}S_n(u,\hat\beta(u)) = o_p(1)$. We have

$$|S_n(u,\hat\beta(u))| \le |\sum_{i=1}^n D_i I(y_i = x_i^T\hat\beta(u))| \cdot \sup_i ||x_i||/n \le k \cdot o_p(n^{-1/2}).$$

The first inequality is by the monotonicity of the sub-gradient condition (Ruppert and Carroll (1980)) and the second inequality is due to the following two facts. First, in a k-dimensinoal quantile regression problem, the solution $\hat\beta(u)$ in equation (2.7) is characterized by the k exact fits (Theorem 3.3 in Koenker and Bassett (1978)). Second, the assumption $E||x_i||^{2+\epsilon} < \infty$ implies $\sup_i ||x_i||/n = o_p(n^{-1/2})$. It is because $\Pr(\sup_i ||x_i|| \ge n^{1/2}) \le n\Pr(||x_i|| > n^{1/2}) \le nE||x_i||^{2+\epsilon}/n^{(2+\epsilon)/2} = o_p(1)$, which establish the result.

(e) We next claim that the empirical process $\beta \to \sqrt{n}(S_n(u,\beta) - E[S_n(u,\beta)])$ is stochastically equicontinuous over $\mathcal{B}$ for a fixed u. Following arguments in Chernozhukov and Hansen

(2006), Lemma B.2, the function space $\{DX(I(Y \leq X^T\beta) - u), \beta \in \mathcal{B}\}$ is Donsker with a squre-integrable envelope $2\max_{j\in 1,\ldots l}|X_j|$, which in turn implies the claim.

(f) Since we have established the stochastically equicontinuity, for any $\hat{\beta}(u) \xrightarrow{p} \beta(u)$,

$$\sqrt{n}\left[\{S_n(u,\beta) - E[S_n(u,\beta)]\}|_{\beta=\hat{\beta}(u)} - \{S_n(u,\beta(u)) - E[S_n(u,\beta(u))\}\right] \xrightarrow{p} 0. \tag{A.1}$$

Plugging-in $E[S_n(u,\beta(u))] = 0$ and $\sqrt{n}S_n(u,\hat{\beta}(u)) = o_p(1)$ into equation (A.1), we have,

$$\sqrt{n}\left[E[S_n(u,\beta)]|_{\beta=\hat{\beta}(u)} + S_n(u,\beta(u))\right] = o_p(1)$$

By Taylor expansion of $E[S_n(u,\beta)]|_{\beta=\hat{\beta}(u)}$ around $\hat{\beta}(u)$, we obtain the expansion,

$$\sqrt{n}(\hat{\beta}(u) - \beta(u)) = A_1^{-1}(u,p)\sqrt{n}S_n(u,\beta(u)) + o_p(1) \tag{A.2}$$

where $A_1(u,p) = (\partial/\partial\beta)E[S_n(u,\beta)]|_{\beta=\beta(u)} = E[p(x)f(x^T\beta(u)|x)xx^T]$.

Finally, $\sqrt{n}S_n(u,\beta(u)) = 1/\sqrt{n}\sum_{i=1}^n D_i x_i\{I(y_i \leq x_i^T\beta) - u\} = 1/\sqrt{n}\sum_{i=1}^n D_i x_i\{I(F(y_i|x_i) \leq u) - u\}$ converges in distribution to a normal distribution with mean zero and covariance $(\min(u,u') - uu')A_0(p)$ where $A_0(p) = E[p(x)xx^T]$. Therefore, from equation (A.2), we conclude

$$\sqrt{n}(\hat{\beta}(u) - \beta(u)) \xrightarrow{d} N(0, \Sigma(u,u')),$$

where $\Sigma(u,u') = (\min(u,u') - uu')A_1^{-1}(u,p)A_0(p)A_1^{-1}(u,p)$.

From Lemma 1, we have a series of Corollaries. For the conditional quantile, we have the following.

Corollary 1. *Assume Conditions in Lemma 1, then for any $x \in \mathcal{X}$, $\hat{Q}_{Y|X}(u|x) \xrightarrow{p} Q_{Y|X}(u|x)$, and $\sqrt{n}(\hat{Q}_{Y|X}(u|x) - Q_{Y|X}(u|x))$ converges in distribution to $T(u,x)$, a Gaussian random variable with mean zero and covariance*

$$E[T(u,x)T(u',x)] = x^T\Sigma(u,u')x = \{\min(u,u') - uu'\}x^TA_1^{-1}(u,p)A_0(p)A_1^{-1}(u',p)x.$$

Proof: This follows from Lemma 1 and the model specification $Q_{Y|X}(u|x) = x^T\beta(u)$. Note $\sqrt{n}(\hat{Q}_{Y|X}(u|x) - Q_{Y|X}(u|x)) = x^T\sqrt{n}(\hat{\beta}(u) - \beta(u))$ then apply results in Lemma 1 for any given x in a compact set $\mathcal{X}$. ∎

For the conditional distribution function, defined by $F(y|x) = \sup\{u \in (0,1)|Q(u|x) \leq y\}$, we have the following result.

Corollary 2. *Assume Conditions in Lemma 1, then* $\hat{F}(y|x) \xrightarrow{p} F(y|x)$*, and* $\sqrt{n}(\hat{F}(y|x) - F(y|x))$ *converges in distribution to* $Z(y,x) = -f(y|x)T(F(y|x),x)$*, a Gaussian random variable with mean zero and covariance*

$$\begin{aligned} E[Z(y,x)Z(y',x)] &= f(y|x)f(y'|x)x^T\Sigma(F(y|x),F(y'|x))x \\ &= f(y|x)f(y'|x)\{\min(F(y|x),F(y'|x)) - F(y|x)F(y'|x)\} \\ &\quad \cdot x^T A_1^{-1}(F(y|x),p)A_0(p)A_1^{-1}(F(y'|x),p)x. \end{aligned}$$

Proof: It is due to Corollary 1 and an application of the functional delta method (Chapter 21 in van der Vaart (1998)). ■

A.2. **Proof of Theorem 1.** Let us denote $\hat{F}(y)$ the estimated distribution function inside the bracket in equation (2.5). We will study the estimate of the distribution function in the first part of the proof, and then study the properties of quantile function estimate. The distribution function estimate is decomposed into three elements.

$$\begin{aligned} \hat{F}(y) - F(y) &= n^{-1}\sum_{i=1}^{n}\{I(Y_i \leq y) - F(y)\} \\ &+ n^{-1}\sum_{i=1}^{n}\{(1-D_i)\,(F(y|x_i) - I(Y_i \leq y))\} \\ &+ n^{-1}\sum_{i=1}^{n}\{(1-D_i)\,(\hat{F}(y|x_i) - F(y|x_i))\} \\ &= B_1(y) + B_2(y) + B_3(y) \end{aligned}$$

For the consistency, we will show $B_j(y) = o_p(1)$ for $j = 1,2,3$. It is obvious to see $B_1(y) = o_p(1)$ by the weak law of large numbers. The second part, when conditioned on all x_i, is a weighted sum of non-identical but independent Bernoulli random variables centered at zero. By a theorem in Singh (1975) and Lemma 2.1 in Cheng and Chu (1996), the summand uniformly converges in y, which means $B_2(y) = o_p(1)$. The last part, $B_3(y) = o_p(1)$, is the consequence of the consistency in conditional quantile function estimates $\hat{F}(y|x_i) - F(y|x_i) = o_p(1)$ in Corollary 2 and the fact that D is a bounded random variable.

For the asymptotic normality, we will apply the central limit theorem to each component and calculate their covariances. (a) The first part is the sum of mean zero Bernoulli random variables, therefore, $\sqrt{n}B_1(y) \xrightarrow{d} N(0, F(y)(1-F(y)))$. (b) The second part converges in distribution to a mean zero random variable with variance,

$$\begin{aligned} E[(1-D)^2(F(y|x)-I(Y\le y))^2] &= E[E[(1-D)^2(F(y|x)-I(Y\le y))^2|X]] \\ &= E[E[(1-D)^2|X]\,E[(F(y|x)-I(Y\le y))^2|X]] \\ &= E[(1-p(x))F(y|x)(1-F(y|x))]. \end{aligned}$$

The second equality is due to the MAR assumption. We conclude that $\sqrt{n}B_2(y) \xrightarrow{d} N(0, E[(1-p(x))F(y|x)(1-F(y|x))])$.

(c) The covariance between $\sqrt{n}B_1(y)$ and $\sqrt{n}B_2(y)$ is,

$$\begin{aligned} & 2Cov(\sqrt{n}B_1(y), \sqrt{n}B_2(y)) = 2E[\sqrt{n}B_1(y)\cdot\sqrt{n}B_2(y)] \\ = \;& 2E[n^{-1}\sum_{i,j=1}^{n}(I(Y_i\le y)-F(y))(1-D_j)(F(y|x_j)-I(Y_j\le y))] \\ = \;& 2E[n^{-1}\sum_{i=1}^{n}(I(Y_i\le y)-F(y))(1-D_i)(F(y|x_i)-I(Y_i\le y))] \\ & +2E[n^{-1}\sum_{i\neq j}(I(Y_i\le y)-F(y))(1-D_j)(F(y|x_j)-I(Y_j\le y))] \\ = \;& 2E[(1-D)(I(Y\le y)-F(y))(F(y|x)-I(Y\le y))]+0 \\ = \;& -2E[(1-p(x))F(y|x)(1-F(y|x))]. \end{aligned}$$

(d) The last part can be represented as a normalized sum of n random variables

$$\begin{aligned} \sqrt{n}B_3(y) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)(\hat{F}(y|x_i)-F(y|x_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n}(1-D_i)\sqrt{n}(\hat{F}(y|x_i)-F(y|x_i)) \\ &\overset{d}{\sim} \frac{1}{n}\sum_{i=1}^{n}(1-D_i)Z(y,x_i) \end{aligned}$$

This is a normalized sum of mean zero Gaussian random variables, therefore, $E(\sqrt{n}B_3(y)) = 0$. To obtain its variance,

$$\begin{aligned} V(\sqrt{n}B_3(y)) &= V(\frac{1}{n}\sum_{i=1}^{n}(1-D_i)Z(y,x_i)) \\ &= \frac{1}{n}V((1-D_1)Z(y|x_1)) + \frac{n(n-1)}{n^2}Cov((1-D)Z(y,x),(1-D')Z(y,x')) \end{aligned}$$

The first part is $O(1/n)$, so let us focus on the second part. It becomes

$$\begin{aligned} & E[(1-D)(1-D')Z(y,x)Z(y,x')] \\ = \; & E[E[(1-D)(1-D')|x,x']\,E[Z(y,x)Z(y,x')|x,x']] \\ = \; & E[(1-p(x))(1-p(x'))f(y|x)f(y|x')\Sigma(F(y|x),F(y|x'))]. \end{aligned}$$

We conclude that

$$V(\sqrt{n}B_3(y)) = E[(1-p(x))(1-p(x'))f(y|x)f(y|x')\Sigma(F(y|x),F(y|x'))].$$

We further calculate a series of covariances.
(e)The covariance between $\sqrt{n}B_1(y)$ and $\sqrt{n}B_3(y)$ is

$$\begin{aligned} 2Cov(\sqrt{n}B_1(y),\sqrt{n}B_3(y)) &= 2E[1/n\sum_{i=1}^{n}(I(Y_i\le y)-F(y))(1-D_i)(\hat{F}(y|x_i)-F(y|x_i))] \\ &= 2E_XE_{Y|X}[(I(Y\le y)-F(y))(\hat{F}(y|x)-F(y|x))|X] \\ &= 2E_X[(F(y)-F(y))o_p(1)] = 0. \end{aligned}$$

And by the same calculations, one can show that
(f) $2Cov(\sqrt{n}B_2(y),\sqrt{n}B_3(y)) = 0$. To conclude, we have

$$\sqrt{n}(\hat{F}(y)-F(y)) \xrightarrow{d} N(0,\sigma^2(y))$$

where

$$\begin{aligned} \sigma^2(y) &= F(y)(1-F(y)) - E[(1-p(x))F(y|x)(1-F(y|x))] \\ &+ E[(1-p(x))(1-p(x'))f(y|x)f(y|x')\Sigma(F(y|x),F(y|x'))]. \end{aligned}$$

The asymptotic normality of an estimator of a distribution function automatically leads to the asymptotic normality of the corresponding quantile estimator. The conclusion of Theorem 1 is the consequence of the following Lemma.

Lemma 5. *Under conditions in Theorem 1, then* $\sqrt{n}(\hat{Q}(\tau) - Q(\tau))$ *converges to a normal distribution with mean zero and variance* $\sigma(Q(\tau))^2/f(Q(\tau))^2$.

Proof: By a standard argument, we have

$$\begin{aligned} & P(\sqrt{n}(\hat{Q}(\tau) - Q(\tau))k^{-1} \leq y) \\ = \; & P(\hat{Q}(\tau) \leq Q(\tau) + ky/\sqrt{n}) \\ = \; & P(\tau \leq \hat{F}(z_n)) \qquad\qquad \text{for} \quad z_n = Q(\tau) + ky/\sqrt{n} \\ = \; & P(\sqrt{n}(\hat{F}(z_n) - F(z_n)) \geq \sqrt{n}(\tau - F(z_n)) \end{aligned}$$

When we expand $F(z_n)$ around $Q(\tau)$, we have the right hand side of the last equation converging to $-kyf(Q(\tau))$, hence, with the choice of $k = \sigma(Q(\tau))/f(Q(\tau))$ and $k_n = \sigma(z_n)/f(Q(\tau))$, divide both sides by $\sigma(z_n)$

$$P(\frac{\sqrt{n}(\hat{F}(z_n) - F(z_n))}{\sigma(z_n)} \geq -\frac{k}{k_n}y) \to 1 - \Phi(-y) = \Phi(y),$$

since $k/k_n \to 1$, where $\Phi(\cdot)$ is a normal cdf. We have shown that $\sqrt{n}(\hat{Q}(\tau) - Q(\tau))$ converges to a normal distribution with mean zero and variance $k^2 = \sigma^2(Q(\tau))/f^2(Q(\tau))$. This completes the proof of Theorem 1. ∎

We now outline proof for Lemma 2. The covariance can be calculated similarly. By calculating $E[\sqrt{n}B_i(y), \sqrt{n}B_j(y')]$ for $i, j = 1, 2, 3$, one can obtain the following formula.

$$\begin{aligned} \sigma(y, y') \; = \; & \{\min(F(y), F(y')) - F(y)F(y')\} \\ - \; & E[(1 - p(x))\{\min(F(y|x), F(y'|x)) - F(y|x)F(y'|x)\}] \\ + \; & E[(1 - p(x))(1 - p(x'))f(y|x)f(y'|x')\Sigma(F(y|x), F(y'|x'))]. \end{aligned}$$

The inter-quantile range is defined $R_Y = Q_Y(\tau) - Q_Y(\tau')$. The consistency and the asymptotic normality follows from those of the marginal quantiles. The variance of $\sqrt{n}(\hat{R}_Y - R_Y)$ is, by standard calculation, $\sigma^2(\tau)/f(Q_Y(\tau))^2 + \sigma^2(\tau')/f(Q_Y(\tau'))^2 - 2\sigma(\tau, \tau')/f(Q_Y(\tau))f(Q_Y(\tau'))$ where $\sigma^2(\tau) = \sigma^2(Q_Y(\tau))$ and $\sigma(\tau, \tau') = \sigma(Q_Y(\tau), Q_Y(\tau'))$.

A.3. **Proofs of Lemma 4 and Theorem 2.** We present the proof of Lemma 4 first. For the time being suppose that the number of imputation is $m = 1$. Following Gutenbrunner and Jurečková (1992) and Qu (2008), we define

$$
\begin{aligned}
R_n(\tau,\beta) &= n^{-1/2}\sum_{i=1}^{n}\left[\delta_i x_i\{I(y_i \le x_i^T\beta)-\tau\}+(1-\delta_i)x_i\{I(y_i^* \le x_i^T\beta)-\tau\}\right] \\
R_n^d(\tau,\beta) &= n^{-1/2}\sum_{i=1}^{n}\left[\delta_i x_i\{I(y_i \le x_i^T\beta)-F_i(x_i^T\beta)\}+(1-\delta_i)x_i\{I(y_i^* \le x_i^T\beta)-\hat{F}_i(x_i^T\beta)\}\right]
\end{aligned}
$$

We prove the following steps. (a) For a k dimensional vector γ satisfying $||\gamma|| \le M$ for some $M > 0$,

$$
\sup_{||\gamma||\le M} ||R_n(\tau,\beta(\tau)+\gamma/\sqrt{n}) - R_n(\tau,\beta(\tau)) - A_1(\tau)\gamma|| = o_p(1). \tag{A.3}
$$

Observe

$$
\begin{aligned}
R_n(\tau,\beta(\tau)+\gamma/\sqrt{n}) &= R_n^d(\tau,\beta(\tau)+\gamma/\sqrt{n}) - R_n^d(\tau,\beta(\tau)) \\
&+ 1/\sqrt{n}\sum_{i=1}^{n}\delta_i x_i\{F_i(x_i^T\beta(\tau)+x_i^T\gamma/\sqrt{n}) - F_i(x_i^T\beta(\tau))\} \\
&+ 1/\sqrt{n}\sum_{i=1}^{n}(1-\delta_i)x_i\{\hat{F}_i(x_i^T\beta(\tau)+x_i^T\gamma/\sqrt{n}) - \hat{F}_i(x_i^T\beta(\tau))\} \\
&+ R_n(\tau,\beta(\tau)) \\
&= e_1+e_2+e_3+R_n(\tau,\beta(\tau)).
\end{aligned}
$$

First, we claim that $e_1 = o_p(1)$ uniformly in $||\gamma|| \le M$. To this end, let $e_1 = g_1 + g_2$

$$
\begin{aligned}
g_1 &= n^{-1/2}\sum_{i=1}^{n}\delta_i x_i\{I(y_i \le x_i^T\beta(\tau)+x_i^T\gamma/\sqrt{n}) - I(y_i \le x_i^T\gamma\beta(\tau))\}, \\
g_2 &= n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)x_i\{I(y_i^* \le x_i^T\beta(\tau)+x_i^T\gamma/\sqrt{n}) - I(y_i^* \le x_i^T\gamma\beta(\tau))\}.
\end{aligned}
$$

By using results in Qu (2008), we obtain $g_1 = o_p(1)$ and $g_2 = o_p(1)$ uniformly in $||\gamma|| \le M$. Second, we want to establish that $e_2 + e_3 = A_1(\tau)\gamma + o_p(1)$, uniformly in $||\gamma|| \le M$. To this end, observe

$$\begin{aligned} e_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i x_i\{\int_0^1 f_i(x_i^T\beta(\tau)+\frac{x_i^T\gamma}{\sqrt{n}}\,s)ds\cdot\frac{x_i^T\gamma}{\sqrt{n}}\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\delta_i x_i x_i^T\gamma\{f_i(x_i^T\beta(\tau)+\frac{x_i^T\gamma}{\sqrt{n}}\,s^*)(1-0)\}, \end{aligned}$$

for a $s^* \in (0,1)$. The second equality is due to the mean value theorem. Since f_i is uniformly continuous, there exists a $c>0$ so that

$$f_i(x_i^T\beta(\tau)+s^*x_i^T\gamma/\sqrt{n}) = f_i(x_i^T\beta(\tau)) + c\,s^*x_i^T\gamma/\sqrt{n}$$

where the second term on the right hand side is $o_p(1)$ because $\max_i ||x_i||/\sqrt{n} = o_p(1)$. Therefore we have

$$e_2 = \frac{1}{n}\sum_{i=1}^{n}\delta_i f_i(x_i^T\beta(\tau))x_i x_i^T\gamma + o_p(1).$$

By a similar method, one can establish that $e_3 = n^{-1}\sum_{i=1}^{n}\delta_i x_i x_i^T\gamma\{\hat{f}_i(x_i^T\beta(\tau)+\frac{x_i^T\gamma}{\sqrt{n}}\,s^*)$. Because of the uniform convergence of $\hat{f}(\cdot)$ and the uniform continuity of $f(\cdot)$, we have

$$e_3 = n^{-1}\sum_{i=1}^{n}(1-\delta_i)f_i(x_i^T\beta(\tau))x_i x_i^T\gamma + o_p(1).$$

Together with the result for e_2, it shows the desired result, which in turn prove the equation (A.3).

(b) We claim that if γ_n is a vector such that $||R_n(\tau,\beta(\tau)+\gamma_n/\sqrt{n})|| = o_p(1)$, then $||\gamma_n|| = O_p(1)$. This argument follows from Lemma 5.2 in Jurečková (1977) and Lemma A.4 in Koenker and Zhao (1996). For n sufficiently large,

$$\begin{aligned} P(||\gamma_n|| \geq M) &\leq P(||\gamma_n|| \geq M, ||R_n(\tau,\beta(\tau)+\gamma_n/\sqrt{n})|| < \eta) + P(||R_n(\tau,\beta(\tau)+\gamma_n/\sqrt{n})|| \geq \eta) \\ &\leq P(\inf_{||\gamma||\geq M}||R_n(\tau,\beta(\tau)+\gamma/\sqrt{n})|| < \eta) + \epsilon. \end{aligned}$$

So we want to show that for any $\epsilon>0$, we can choose $\eta>0$, $M>0$, and $\tilde{N}>0$ such that $n>\tilde{N}$ means

$$P(\inf_{||\gamma||\geq M}||R_n(\tau,\beta(\tau)+\gamma/\sqrt{n})|| < \eta) < \epsilon.$$

For any γ satisfying $||\gamma|| \geq M$, denote $\gamma = M\lambda e$, where $\lambda = ||\gamma||/M$, $e = \gamma/||\gamma||$. Note that $\lambda \geq 1$, $||e|| = 1$. By Cauchy-Schwarz,

$$\begin{aligned} & P(\inf_{||\gamma||\geq M} ||R_n(\tau, \beta(\tau) + \gamma/\sqrt{n})|| < \eta) \\ = \; & P(\inf_{||\lambda||\geq 1} ||R_n(\tau, \beta(\tau) + M\lambda e/\sqrt{n})|| \times ||e|| < \eta) \\ \leq \; & P(\inf_{||\lambda||\geq 1} |e^T R_n(\tau, \beta(\tau) + M\lambda e/\sqrt{n})| < \eta). \end{aligned}$$

Since $e^T R_n(\tau, \beta(\tau) + M\lambda e/\sqrt{n})$ is a non-decreasing function of λ, the infimum is achieved at $||\lambda|| = 1$ or $||\gamma|| = M$. Thus, we will get the result when we show

$$P(\inf_{||\gamma||=M} |e^T R_n(\tau, \beta(\tau) + Me/\sqrt{n})| < \eta) < \epsilon.$$

The left hand side is less than or equal to

$$\begin{aligned} & P(\inf_{||\gamma||=M} |e^T R_n(\tau, \beta(\tau) + Me/\sqrt{n})| < \eta, \inf_{||\gamma||=M} |e^T (R_n(\tau, \beta(\tau)) + A_1(\tau)\gamma)| \geq 2\eta) \\ + \; & P(\inf_{||\gamma||=M} |e^T (R_n(\tau, \beta(\tau)) + A_1(\tau)\gamma)| < 2\eta) = h_1 + h_2. \end{aligned}$$

The first probability h_1 is less than or equal to

$$P(\sup_{||\gamma||=M} ||R_n(\tau, \beta(\tau) + \gamma/\sqrt{n}) - R_n(\tau, \beta(\tau)) - A_1(\tau)\gamma|| \geq 2\eta).$$

In view of (A.3), it is obvious that we can choose $\tilde{N}_1$ and M so that whenever $n > \tilde{N}_1$ the above probability is less than $\epsilon/2$. The second probability h_2 is less than or equal to

$$P(||e^T R_n(\tau, \beta(\tau))|| \geq Me^T \lambda_1(A_1(\tau))e - 2\eta).$$

where $\lambda_1(A_1(\tau))$ is the smallest eigen-value of $A_1(\tau)$. By following the same argument in part (b) of the proof of Lemma 1, we can establish that $e^T R_n(\tau, \beta(\tau))$ is relatively tight. Along with the assumption (v'), it implies that for a given M we can always choose η and $\tilde{N}_2$ so that $n > \tilde{N}_2$ means the above probability being less than $\epsilon/2$. Finally let $\tilde{N} = \max(\tilde{N}_1, \tilde{N}_2)$ and it establishes the desired result.

(c) Use $\sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau))$ in the place of γ_n. Then $||R_n(\tau, \beta(\tau) + \sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau))/\sqrt{n})|| = ||R_n(\tau, \tilde{\beta}(\tau))|| = o_p(1)$ by the same argument used in Lemma 1. Now by the step (b), $\sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau)) = O_p(1)$ and we can use equation (A.3). We obtain the Bahadur representation in Lemma 4 when we plug-in $\sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau))$ in the place of γ.

We now prove Theorem 2. The Bahadur representation says, when $m = 1$,

$$\begin{aligned}\sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau)) &= -1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}\delta_i x_i\{I(Y_i \le x_i\beta(\tau)) - \tau\} \\ &\quad -1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\{I(Y^*_{x_i} \le x_i\beta(\tau)) - \tau\} + o_p(1)\end{aligned}$$

The first part becomes

$$\begin{aligned}&-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}\delta_i x_i\{I(Y_i \le x_i\beta(\tau)) - \tau\} \\ =\ &-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}\delta_i x_i\{I(F_i(Y_i) \le F_i(x_i\beta(\tau))) - \tau\} \\ =\ &-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}\delta_i x_i\{I(u_i \le \tau) - \tau\} = C_1(\tau)\end{aligned}$$

By the central limit theorem, $C_1(\tau)$ converges to a normal distribution with mean zero and variance $\tau(1-\tau)\,p\,A_1^{-1}(\tau)A_0A_1^{-1}(\tau)$. The second part is decomposed to two components

$$\begin{aligned}&-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\{I(Y^*_{x_i} \le x_i^T\beta(\tau)) - \tau\} \\ =\ &-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\{I(\hat{F}_i(Y^*_{x_i}) \le \hat{F}_i(x_i^T\beta(\tau))) - F_i(x_i^T\beta(\tau))\} \\ =\ &-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\{I(u_i \le \hat{F}_i(x_i^T\beta(\tau))) - F_i(x_i^T\beta(\tau))\} \\ =\ &-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\{I(u_i \le \hat{F}_i(x_i^T\beta(\tau))) - \hat{F}_i(x_i^T\beta(\tau))\} \\ &-1/\sqrt{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\{\hat{F}_i(x_i^T\beta(\tau)) - F_i(x_i^T\beta(\tau))\} = C_2(\tau) + C_3(\tau).\end{aligned}$$

$C_2(\tau)$ is a sum of weighted Bernoulli random variables. Together with a fact $\hat{F}_i(x_i^T\beta(\tau)) \xrightarrow{p} F_i(x_i^T\beta(\tau)) = \tau$, it converges to a mean zero Gaussian distribution with variance

$$\tau(1-\tau)\,(1-p)\,A_1^{-1}(\tau)A_0A_1^{-1}(\tau).$$

Next,

$$C_3(\tau) = -\frac{1}{n}A_1^{-1}(\tau)\sum_{i=1}^{n}(1-\delta_i)x_i\sqrt{n}\{\hat{F}_i(x_i^T\beta(\tau)) - F_i(x_i^T\beta(\tau))\}$$

which is a sum of n asymptotically normal random variables. Following Corollary 2, we may write it as,

$$C_3(\tau) = -A_1^{-1}(\tau)\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)x_i Z(Q(\tau|x_i), x_i)$$

By the same method used in the proof of Theorem 1 (see step (d)), $C_3(\tau)$ converges to a normal distribution with mean zero and variance

$$\tau(1-\tau)\frac{(1-p)^2}{p}E[f(Q(\tau|x)|x)f(Q(\tau|x')|x')\{x^TA_1^{-1}(\tau)A_0A_1^{-1}(\tau)x'\}]A_1^{-1}(\tau)A_0A_1^{-1}(\tau).$$

When we consider a case where $m > 1$, the only difference is in $C_2(\tau)$. Now it becomes

$$C_2(\tau) = -\frac{1}{\sqrt{n}}A_1^{-1}(\tau)\sum_{i=1}^{n}\{(1-\delta_i)x_i\frac{1}{m}\sum_{i=1}^{n}(I(u_{ij} \le \hat{F}_i(x_i^T\beta(\tau))) - \hat{F}_i(x_i^T\beta(\tau)))\}$$

After some calculation, it can be shown that $C_2(\tau)$ converges to normal distribution with mean zero and variance

$$\tau(1-\tau)\,\frac{1-p}{m}\,A_1^{-1}(\tau)A_0A_1^{-1}(\tau).$$

Finally, by combining all three components, we conclude that $\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))$ converges in distribution to a Gaussian distribution with mean zero and variance

$$\begin{aligned} & \tau(1-\tau)(p + \frac{1-p}{m})A_1^{-1}(\tau)A_0A_1^{-1}(\tau) \\ + \quad & \tau(1-\tau)\frac{(1-p)^2}{p}E[f(Q(\tau|x)|x)f(Q(\tau|x')|x')\{x^TA_1^{-1}(\tau)A_0A_1^{-1}(\tau)x'\}]A_1^{-1}(\tau)A_0A_1^{-1}(\tau). \end{aligned}$$

REFERENCES

ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70(1), 91–117.

ABADIE, A., AND G. W. IMBENS (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74(1), 235–267.

AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): "Quantile regression under misspecification, with an application to the U.S. wage structure," *Econometrica*, 74(2), 539–563.

BOLLINGER, C. R., AND B. T. HIRSCH (2006): "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching," *Journal of Labor Economics*, 24(3), 483–519.

BUCHINSKY, M. (1994): "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 62(2), 405–458.

CHAMBERLAIN, G. (1994): "Quantile regression, censoring, and the structure of wages," in *Advances in econometrics, Sixth World Congress, Vol. I (Barcelona, 1990)*, vol. 23 of *Econom. Soc. Monogr.*, pp. 171–209. Cambridge Univ. Press, Cambridge.

CHAUDHURI, P. (1991): "Nonparametric estimates of regression quantiles and their local Bahadur representation," *Ann. Statist.*, 19(2), 760–777.

CHAUDHURI, P., K. DOKSUM, AND A. SAMAROV (1997): "On average derivative quantile regression," *Ann. Statist.*, 25(2), 715–744.

CHEN, X., H. HONG, AND A. TAROZZI (2008): "Semiparametric efficiency in GMM models with auxiliary data," *Ann. Statist.*, 36(2), 808–843.

CHENG, P. E. (1994): "Nonparametric Estimation of Mean Functionals with Data Missing at Random," *Journal of the American Statistical Association*, 89(425), 81–87.

CHENG, P. E., AND C. K. CHU (1996): "Kernel estimation of distribution functions and quantiles with missing data," *Statist. Sinica*, 6(1), 63–78.

CHERNOZHUKOV, V., AND C. HANSEN (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132(2), 491 – 525.

DAGENAIS, M. G. (1973): "The use of incomplete observations in multiple regression analysis : A generalized least squares approach," *Journal of Econometrics*, 1(4), 317 – 328.

DEHEJIA, R. H., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

FAN, J., T. C. HU, AND Y. K. TRUONG (1994): "Robust non-parametric function estimation," *Scand. J. Statist.*, 21(4), 433–446.

GOURIÉROUX, C., AND A. MONFORT (1981): "On the problem of missing data in linear models," *Rev. Econom. Stud.*, 48(4), 579–586.

GUTENBRUNNER, C., AND J. JUREČKOVÁ (1992): "Regression rank scores and regression quantiles," *Ann. Statist.*, 20(1), 305–330.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

HASTIE, T. J., AND R. J. TIBSHIRANI (1990): *Generalized additive models*, vol. 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

HE, X., AND L.-X. ZHU (2003): "A lack-of-fit test for quantile regression," *J. Amer. Statist. Assoc.*, 98(464), 1013–1022.

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1998): "Characterizing selection bias using experimental data," *Econometrica*, 66(5), 1017–1098.

HECKMAN, J. J., H. ICHIMURA, AND P. TODD (1998): "Matching as an econometric evaluation estimator," *Rev. Econom. Stud.*, 65(2), 261–294.

HECKMAN, J. J., AND P. A. LAFONTAINE (2006): "Bias Corrected Estimates of GED Returns," *Journal of Labor Economics*, 24(3), 661–700.

HIRSCH, B. T., AND E. J. SCHUMACHER (2004): "Match Bias in Wage Gap Estimates Due to Earnings Imputation," *Journal of Labor Economics*, 22(3), 689–722.

HOROWITZ, J. L., AND S. LEE (2005): "Nonparametric estimation of an additive quantile regression model," *J. Amer. Statist. Assoc.*, 100(472), 1238–1249.

IMBENS, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86(1), 4–29.

JUREČKOVÁ, J. (1977): "Asymptotic relations of M-estimates and R-estimates in linear regression model," *Ann. Statist.*, 5(3), 464–472.

KOENKER, R. (2010): "Additive Models for Quantile Regression: Model Selection and Confidence Bandaids," Working paper.

KOENKER, R., AND G. BASSETT, JR. (1978): "Regression quantiles," *Econometrica*, 46(1), 33–50.

KOENKER, R., P. NG, AND S. PORTNOY (1994): "Quantile smoothing splines," *Biometrika*, 81(4), 673–680.

KOENKER, R., AND Z. XIAO (2002): "Inference on the quantile regression process," *Econometrica*, 70(4), 1583–1612.

KOENKER, R., AND Q. ZHAO (1996): "Conditional quantile estimation and inference for ARCH models," *Econometric Theory*, 12(5), 793–813.

LALONDE, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–20.

LEE, S. (2003): "Efficient semiparametric estimation of a partially linear quantile regression model," *Econometric Theory*, 19(1), 1–31.

LILLARD, L., J. P. SMITH, AND F. WELCH (1986): "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, 94(3), 489.

LITTLE, R. J. A., AND D. B. RUBIN (2002): *Statistical analysis with missing data*, Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edn.

ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1995): "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *J. Amer. Statist. Assoc.*, 90(429), 106–121.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.

RUBIN, D. B. (1987): *Multiple imputation for nonresponse in surveys*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

RUPPERT, D., AND R. J. CARROLL (1980): "Trimmed least squares estimation in the linear model," *J. Amer. Statist. Assoc.*, 75(372), 828–838.

RUPPERT, D., S. J. SHEATHER, AND M. P. WAND (1995): "An effective bandwidth selector for local least squares regression," *J. Amer. Statist. Assoc.*, 90(432), 1257–1270.

SINGH, R. S. (1975): "On the Glivenko-Cantelli theorem for weighted empiricals based on independent random variables," *Ann. Probability*, 3, 371–374.

VAN DER VAART, A. W. (1998): *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

WANG, Q., O. LINTON, AND W. HÄRDLE (2004): "Semiparametric regression analysis with missing response at random," *J. Amer. Statist. Assoc.*, 99(466), 334–345.

WOOLDRIDGE, J. M. (2007): "Inverse probability weighted estimation for general missing data problems," *J. Econometrics*, 141(2), 1281–1301.

WU, T., K. YU, AND Y. YU (2008): "Single Index Quantile Regression," Working paper.

YU, K., AND M. C. JONES (1998): "Local linear quantile regression," *J. Amer. Statist. Assoc.*, 93(441), 228–237.

ZHOU, Y., A. T. K. WAN, AND X. WANG (2008): "Estimating equations inference with missing data," *J. Amer. Statist. Assoc.*, 103(483), 1187–1199.