Statistics and Its Interface Volume 6 (2013) 369–377

Imputation methods for quantile estimation under missing at random

Shu Yang*, Jae-Kwang Kim and Dong Wan Shin

Imputation is frequently used to handle missing data for which multiple imputation is a popular technique. We propose a fractional hot deck imputation which produces a valid variance estimator for quantiles. In the proposed method, the imputed values are chosen from the set of respondents and are assigned with proper fractional weights that use a density function for the working model. In addition, we consider a nonparametric fractional imputation method based on nonparametric kernel regression, avoiding a parametric distribution assumption and thus giving more robustness. The resulting estimator can be called nonparametric fractionally imputation estimator. Valid variance estimation is also discussed. A limited simulation study compares the proposed methods favorably with other existing methods.

Keywords and phrases: Bahadur representation, Estimating equation, Fractional hot deck imputation, Jackknife variance estimator, Linearization method, Nonparametric imputation, Woodruff variance.

1. INTRODUCTION

Quantile estimation is frequently used in many disciplines. In industry, a device manufacturer may wish to know what are the 10% and 90% quantiles for some features of the production processes to tailor the process to produce 80% of the devices. In finance, for risk management, a bank may need to estimate a lower bound on the changes in the values of its portfolio which will hold with high probability.

We consider imputation methods for quantile estimation, where the missing mechanism is assumed to be missing at random in the sense of Rubin (1987). Under existence of missing data, imputation is often used for missing data analysis to facilitate the parameter estimation, which is a process of replacing missing values with pseudo values so that analysis from different users will be consistent.

There are various ways to impute missing values which lead to different imputation methods. Multiple imputation (MI in the sequel), proposed by Rubin (1987), uses a Bayesian method to generate multiple imputed values which represent the uncertainty about the right value to impute. Parametric fractional imputation (PFI in the sequel), proposed by Kim (2011), is a frequentist version of MI, where fractional weights are assigned to the imputed values to properly represent the point mass of the imputed values.

Imputation has been widely used for handling missing data because a single imputed data can be used to estimate several parameters. However, many papers on imputation focused only on estimating the population mean. Estimating the population quantiles with imputed data is also an important practical problem but is rarely addressed in literature. We discuss MI and PFI in terms of quantile estimation. Moreover, we propose a new imputation method which can be called fractional hot deck imputation (FHDI in the sequel). Instead of generating imputed values, FHDI chooses the imputed values from the set of respondents and assigns fractional weights to imputed values so that the conditional expectation of the estimating function is approximated by the imputed estimating function.

The proposed FHDI method has a nonparametric feature which can easily be modified to a fully nonparametric version called nonparametric fractional imputation (NPFI in the sequel). In NPFI, the whole estimation procedure is fully nonparametric. On the other hand, MI and PFI rely on the model assumptions. Therefore, the proposed methods, FHDI and NPFI, are more robust than the existing methods, MI and PFI, producing less-biased estimators in the cases of failure of the assumed model.

A more important advantage of the fractional imputation methods over the MI method is that the former allow valid variance estimators for quantile estimates while for the latter, a variance estimator using Rubin's formula is not valid. Valid variance estimation for fractional imputation is possible because, unlike MI, the effect of estimated nuisance parameter in imputation is correctly reflected in the replication variance estimation. It is well known that direct application of delete-1 observation jackknife variance estimation is not valid for medians or quantiles. On the other hand, FHDI and NPFI allow a valid two-step variance estimator of quantiles by combining the linearization method (or test inversion method) and the delete-1 observation jackknife variance estimation to the empirical distribution function together.

The rest of the paper is organized as follows. In Section 2, we introduce multiple imputation and parametric fractional imputation in quantile estimation with ignorable missing

*Corresponding author.

data. In Section 3, we develop FHDI and NPFI for quantile estimation. Delete-1 observation jackknife variance estimation is discussed in Section 4. Section 5 presents a limited simulation study and a real data set analysis. Some concluding remarks follow in Section 6.

2. EXISTING METHODS

When the study variable y is fully observed in the sample, an empirical distribution function can be computed from the sample by

$$\hat{F}(y) = n^{-1}\sum_{i=1}^{n} I(y_i < y),$$

and the sample quantile is then computed by the inverse of empirical distribution

$$\hat{\xi}_p = \hat{F}^{-1}(p) = \inf\{y : \hat{F}(y) \geq p\},$$

which can also be viewed as a solution to the estimating equation

$$U(\xi_p) = \sum_{i=1}^{n} U(\xi_p, y_i) = n^{-1}\sum_{i=1}^{n}\{I(y_i < \xi_p) - p\} = 0.$$

The above estimation procedure is robust because it does not require any distribution assumption.

For variance estimation, we consider two types of variance estimators. The first type is the linearization method estimator which is based on the Bahadur representation (Bahadur, 1966)

$$\hat{V}(\hat{\xi}_p) \cong \frac{1}{[\hat{f}(\hat{\xi}_p)]^2}\hat{V}(\hat{F}(\hat{\xi}_{p,FI})), \tag{1}$$

where f is the marginal density function of Y. The second type is the Woodruff variance estimator (Woodruff, 1952) which is based on the so-called test-inversion method. To compute the Woodruff variance estimator, we first construct a normal-based 95% asymptotic confidence interval for p by $\hat{p} \pm 2\sqrt{\hat{V}(\hat{p})} \equiv (\hat{p}_L, \hat{p}_U)$, where $\hat{p} \equiv \hat{F}(\hat{\xi}_p)$. Since $\hat{F}$ is monotone, a normal-based 95% asymptotic confidence interval for ξ_p can be obtained by $(\hat{F}^{-1}(\hat{p}_L), \hat{F}^{-1}(\hat{p}_U)) \equiv (\hat{\xi}_{p,L}, \hat{\xi}_{p,U})$. Thus the Woodruff variance estimator is given by

$$\hat{V}(\hat{\xi}_p) = \left(\frac{\hat{\xi}_{p,U} - \hat{\xi}_{p,L}}{4}\right)^2. \tag{2}$$

We now consider missing cases. Several imputation methods will be adopted to estimate the quantiles. We assume that the study variable y is subject to missing and an auxiliary variable x is observed throughout the sample. Properly incorporating x into the estimation of the quantiles of y can lead to bias correction as well as variance reduction. For simplicity, we assume that the first r elements have both x and y observed and the remaining $n-r$ elements have only x observed.

2.1 Multiple imputation (MI)

Multiple imputation (MI) is a popular technique of imputation proposed by Rubin (1987). In the MI, instead of generating one single value, a set of plausible values are generated to represent the uncertainty about the right value to impute. The complete sample estimator is then applied to each of the multiply imputed data sets. Finally the results are combined from these analysis for inference.

In MI, Bayesian method is used to generate imputed values. Multiple imputation procedure for bivariate normal (x, y) is described in Schenker and Welsh (1988). At each repetition of the imputation $(k = 1, \ldots, m)$, we can calculate the imputed version of the quantile estimator $\hat{\zeta}_{\gamma,I(k)}$ and its variance estimator $\hat{V}_{I(k)}$. The final quantile estimator is computed by

$$\hat{\zeta}_{\gamma,MI} = m^{-1}\sum_{k=1}^{m}\hat{\zeta}_{\gamma,I(k)}.$$

Rubin proposed using the following estimator for the variance of $\hat{\zeta}_{\gamma,MI}$:

$$\hat{V}_{MI} = W_{m,n} + \left(1 + m^{-1}\right) B_{m,n}, \tag{3}$$

where

$$W_{m,n} = m^{-1}\sum_{k=1}^{m}\hat{V}_{I(k)}, \tag{4}$$

and

$$B_{m,n} = (m-1)^{-1}\sum_{k=1}^{m}\left(\hat{\zeta}_{\gamma,I(k)} - \hat{\zeta}_{\gamma,MI}\right)^2. \tag{5}$$

In (4), $\hat{V}_{I(k)}$ is the variance estimator (1) or (2) applied to the k^{th} imputed data set.

Validity of the MI variance estimator requires that the congeniality condition of Meng (1994) holds. Roughly speaking, the congeniality condition means that

$$V(\hat{\theta}_{MI}) = V(\hat{\theta}_n) + V(\hat{\theta}_{MI} - \hat{\theta}_n).$$

Kim (2011) argues that the congeniality condition does not hold when the parameter of interest is $\theta = Pr(Y < c)$ and $\hat{\theta}_n = n^{-1}\sum_{i=1}^{n} I(y_i < c)$ is used to estimate θ under complete response. Because the quantile estimator is also obtained from the sample distribution function, the congeniality condition does not hold for quantiles, which is confirmed numerically in the simulation study in Section 5.

2.2 Parametric fractional imputation (PFI)

Parametric fractional imputation (PFI) was proposed by Kim (2011) for general purpose estimation. In Kim (2011), the PFI method was developed for estimating population mean and proportion under ignorable non-response. The

PFI method can be developed for quantile estimation as well. One advantage of the PFI method is that, if the imputed data is applied to the score function or the estimating function, the resulting estimator is very close to the maximum likelihood estimator. In the PFI method, m imputed values are generated for y_i, $i = r+1, \ldots, n$ and m fractional weights are assigned to the imputed values so that the mean score function or the mean estimating function can be approximated by a weighted sum of the imputed score functions or estimating functions. Let y_{ij}^* be the j^{th} imputed value of missing y_i and w_{ij}^* be the fractional weight assigned to y_{ij}^*. The fractional weights are constructed to satisfy

$$\sum_{j=1}^{m} w_{ij}^* = 1, \tag{6}$$

for each $i = 1, 2, \ldots, n$ and

$$w_{ij}^* \propto \frac{f(y_{ij}^* | x_i; \hat{\theta})}{h(y_{ij}^*)}, \tag{7}$$

where $f(y \mid x; \theta)$ is the conditional density of y given x, $h(y)$ is the density function of the distribution from which y_{ij}^* are generated, $\hat{\theta}$ is the MLE of θ which is obtained by solving

$$\sum_{i=1}^{r} S(\theta; x_i, y_i) = 0,$$

and $S(\theta, x_i, y_i) = \partial \log f(y_i | x_i, \theta)/\partial\theta$ is the score function for the i-th observation, $f(y_i|x_i, \theta)$. Once the fractional weights are constructed, the PFI estimator of ξ_p is given by

$$\hat{\xi}_{p,PFI}^* = \hat{F}_{PFI}^{*-1}(p) = \inf\{y : \hat{F}_{PFI}^*(y) \geq p\}, \tag{8}$$

where

$$\hat{F}_{PFI}^*(y) = n^{-1} \sum_{i=1}^{n} \{\delta_i I(y_i < y) + (1-\delta_i) \sum_{j=1}^{m} w_{ij}^* I(y_{ij}^* < y)\}. \tag{9}$$

and δ_i is the response indicator such that $\delta_i = 1$ for observed y_i and $\delta_i = 0$ for missing y_i.

Variance estimation can be obtained by the linearization method as described in Appendix A.1.

3. PROPOSED METHODS

In MI and PFI, the imputed values are generated from a parametric distribution. Instead of generating imputed values, in the FHDI, the imputed values are taken from the set of respondents. The record providing the value is called the donor and the record with the missing value is called the recipient. Hot deck imputation is initially proposed by Brick and Kalton (1996) to reduce imputation variance by random selection of one imputed value among donors. Kalton and Kish (1984) and Fay (1996) used more than one donor for a recipient to reduce the imputation variance.

3.1 Fractional hot deck imputation (FHDI)

In FHDI, for each missing y_i, a set of m imputed values $\{y_{i1}^*, \ldots, y_{im}^*\}$ are obtained from the set of respondents $\{y_1, \ldots, y_r\}, i = r+1, \ldots, n$. Let w_{ij}^* be the fractional weights assigned to y_{ij}^*, $j = 1, 2, \ldots, m$. In FHDI, we use $m = r$, that is, the j-th imputed value of missing y_i is $y_{ij}^* = y_j$, an observed value, $j = 1, \ldots, r$. In this case, the fractional weights $w_{i1}^*, \ldots, w_{ir}^*$ are computed to satisfy $\sum_{j=1}^{r} w_{ij}^* = 1$ and

$$\sum_{j=1}^{r} w_{ij}^* I(y_j < y) \cong Pr(y_i < y | x_i).$$

Since we can treat $\{y_1, \ldots, y_r\}$ as a set of realizations from $f(y \mid \delta = 1)$, the desired fractional weight assigned to $y_{ij}^* = y_j$ for $\delta_i = 0$ is given by

$$w_{ij}^* \propto \frac{f(y_j | x_i; \hat{\theta})}{f(y_j | \delta_j = 1)}, \tag{10}$$

and $\sum_{j=1}^{m} w_{ij}^* = 1$, for $i = r+1, \ldots, n$. Since

$$\begin{aligned} f(y \mid \delta = 1) &= \int f(y \mid x, \delta = 1) f(x \mid \delta = 1) dx \\ &= \int f(y \mid x) f(x \mid \delta = 1) dx, \end{aligned}$$

where the second equality follows from MAR, a consistent estimator of $f(y_j|\delta_j = 1)$ is given by

$$\hat{f}(y_j | \delta_j = 1) = \frac{\sum_{k=1}^{n} \delta_k f(y_j | x_k, \hat{\theta})}{\sum_{k=1}^{n} \delta_k},$$

which uses the empirical distribution for $f(x \mid \delta = 1)$. That is, it uses

$$\hat{f}(x \mid \delta = 1) = \frac{\sum_{\delta_i = 1} I(x = x_i)}{\sum_{i=1}^{n} \delta_i}. \tag{11}$$

Thus, the fractional weight in (10) is computed by

$$w_{ij}^* = \frac{f(y_j | x_i; \hat{\theta}) / \{\sum_{k=1}^{n} \delta_k f(y_j | x_k, \hat{\theta})\}}{\sum_{l=1}^{r} [f(y_l | x_i; \hat{\theta}) / \{\sum_{k=1}^{n} \delta_k f(y_l | x_k, \hat{\theta})\}]}.$$

Once the weight set $\{w_{ij}^*\}$ is created, the FHDI estimator $\hat{\xi}_{p,FHDI}$ of ξ_p is computed from (8)–(9) with these $\{w_{ij}^*\}$ replacing that in (8)–(9). Variance estimation for $\hat{\xi}_{p,FHDI}$ will be discussed in Section 4.

3.2 Nonparametric fractional imputation (NPFI)

Fractional imputation can be implemented nonparametrically. Cheng (1994) used kernel regression estimators to

estimate mean functionals through empirical estimation of the missing pattern. Chen (2001) and Kim and Yu (2011) used a semi-parametric logistic regression model of mean functionals with non-ignorable missing data. We adopt the kernel regression idea to obtain fractional weights in FHDI, and the resulting imputation method will be called nonparametric fractional imputation, NPFI.

Let $K(\cdot)$ be a symmetric density function on the real line and $h = h_n$ be a smoothing bandwidth such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. A nonparametric regression estimator of $m(x) = E(y|x)$ can be obtained by finding $\hat{m}(x)$ that minimizes

$$\sum_{i=1}^{n} K_h(x_i, x)\delta_i\{y_i - m(x)\}^2,$$

where $K_h(u, x) = h^{-1}K\{(u-x)/h\}$. The minimizer is the well-known Nadaraya-Watson (1964) kernel regression estimator (NW estimator)

$$\hat{m}(x) = \sum_{j=1}^{n} w_{j1}(x)y_j,$$

where

$$w_{i1}(x) = \frac{K_h(x, x_i)\delta_i}{\sum_{j=1}^{n} K_h(x, x_j)\delta_j}, \tag{12}$$

which represents the point mass assigned to y_i when $m(x)$ is approximated by $\sum_{i=1}^{m} w_{i1}(x)y_i$. Consider $\hat{m}(x_i)$ to be a prediction for missing unit i, then

$$\begin{aligned}\hat{\mu}_y &= n^{-1}\sum_{i=1}^{n}\{\delta_i y_i + (1-\delta_i)\hat{m}(x_i)\} \\ &= n^{-1}\sum_{i=1}^{n}\left\{\delta_i y_i + (1-\delta_i)\sum_{j=1}^{n} w_{j1}(x_i)y_j\right\}.\end{aligned}$$

The weight $w_{ij}^* = w_{j1}(x_i)$ is essentially the fractional weight assigned to j^{th} imputed value for missing unit i. Consider implementing fractional hot deck imputation in a nonparametric fashion. Using (10) where $f(y_j|x_i)$ is nonparametrically estimated by a kernel-based method, the final fractional weights can be given by

$$w_{ij}^* = \frac{K_h(x_i, x_j)/C(x_j)}{\sum_{k=1}^{n} K_h(x_i, x_k)\delta_k/C(x_k)},$$

where

$$C(x_j) = \sum_{i=1}^{n} \delta_i K_h(x_i, x_j).$$

Given the weight set $\{w_{ij}^*\}$, the NPFI estimator $\hat{\xi}_{p,NPFI}$ is computed from (8)–(9) with these $\{w_{ij}^*\}$ replacing that in (8)–(9). Variance estimation for $\hat{\xi}_{p,NPFI}$ will be discussed in Section 4.

4. VARIANCE ESTIMATION

One advantage of FHDI and NPFI is that all imputed values are realized values, which enables us to use the replication method for variance estimation. Delete-1 observation jackknife variance estimator is considered. It has been shown that delete-1 observation jackknife variance estimator is valid with smooth differentiable statistics, such as totals, means, proportions and etc; but not with medians or quantiles.

In order to get a valid variance estimator in FHDI and NPI, we consider a two-step procedure using the linearization method or the test inversion method in the first step and the delete-1 observation jackknife variance estimation with the empirical distribution function in the second step. This two-step approach makes the delete-1 observation jackknife variance estimator valid for median or quantile estimators.

Denote HD as either FHDI or NPFI. Based on the linearizaiton method of (1) applied to $\hat{\xi}_{p,HD} = \hat{F}_{HD}^{-1}(p)$, we get

$$V(\hat{\xi}_{p,HD}) \cong \frac{1}{[\hat{f}(\hat{\xi}_{p,HD})]^2} V\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\},$$

or the test inversion method of (2), we get

$$V(\hat{\xi}_{p,HD}) = \left(\frac{\hat{\xi}_{p,U} - \hat{\xi}_{p,L}}{4}\right)^2,$$

where $(\hat{p}_L, \hat{p}_U) = \hat{F}_{HD}(\hat{\xi}_{p,HD}) \pm 2\sqrt{V(\hat{F}_{HD}(\hat{\xi}_{p,HD}))}$, and $(\hat{\xi}_{p,L}, \hat{\xi}_{p,U}) = (\hat{F}^{-1}(\hat{p}_L), \hat{F}^{-1}(\hat{p}_U))$. In either method, we need a consistent estimate of $V(\hat{F}_{HD}(\hat{\xi}_{p,HD}))$.

Notice that $\hat{F}_{HD}(y) = n^{-1}\sum_{i=1}^{n}\{\delta_i I(y_i < y) + (1-\delta_i)\sum_{j=1}^{m} w_{ij}^* I(y_{ij}^* < y)\}$ is a proportion. Create $Z_i = I(y_i < \hat{\xi}_{p,HD})$ and $Z_{ij}^* = I(y_{ij}^* < \hat{\xi}_{p,HD})$. Then

$$\hat{F}_{HD}(\hat{\xi}_{p,HD}) = \bar{Z}_{HD} = n^{-1}\sum_{i=1}^{n}\{\delta_i Z_i + (1-\delta_i)\sum_{j=1}^{m} w_{ij}^* Z_{ij}^*\}.$$

So, Jackknife method can be applied to obtain a consistent estimator for the variance of the average $\hat{F}_{HD}(\hat{\xi}_{p,HD}) = \bar{Z}_{HD}$. Specifically, $V\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\} = V(\bar{Z}_{HD})$ is estimated by the following delete-1 observation jackknife variance estimator

$$\hat{V}_{rep}\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\} = \frac{n-1}{n}\sum_{k=1}^{n}(\bar{Z}_{HD}^{(k)} - \bar{Z}_{HD})^2,$$

where

$$\bar{Z}_{HD}^{(k)} = n^{-1}\sum_{i=1}^{n}\left\{\delta_i w_i^{(k)} Z_i + (1-\delta_i)\sum_{j=1}^{m} w_j^{(k)} w_{ij}^{*(k)} Z_{ij}^*\right\},$$

with

$$w_i^{(k)} = \begin{cases}(n-1)^{-1} & \text{if } i \neq k \\ 0 & \text{if } i = k\end{cases}.$$

For FHDI, $w_{ij}^{*(k)}$ are the replicate fractional hot deck weights computed by

$$w_{ij}^{*(k)} = \frac{w_j^{(k)} f(y_j|x_i;\hat{\theta}^{(k)})/\{\sum_{l=1}^n w_l^{(k)}\delta_l f(y_j|x_l;\hat{\theta}^{(k)})\}}{\sum_{s=1}^m [w_s^{(k)} f(y_s|x_i;\hat{\theta}^{(k)})/\{\sum_{l=1}^n w_l^{(k)}\delta_l f(y_s|x_l;\hat{\theta}^{(k)})\}]}.$$

The k-th replicate of $\hat{\theta}$, denoted by $\hat{\theta}^{(k)}$, satisfies

$$\sum_{i=1}^r w_i^{(k)} S(\hat{\theta}^{(k)}; x_i, y_i) = 0.$$

For NPFI, $w_{ij}^{*(k)}$ are the replicate nonparametric fractional weights computed by

$$w_{ij}^{*(k)} = \frac{w_j^{(k)} K_h(x_i, x_j)/C^{(k)}(x_j)}{\sum_{l=1}^n \{w_l^{(k)} K_h(x_i, x_l)/C^{(k)}(x_l)\}},$$

where $C^{(k)}(x_j) = \sum_{i=1}^n \delta_i w_i^{(k)} K_h(x_i, x_j)$.

Once we obtain the delete-1 observation jackknife variance estimator $\hat{V}_{rep}\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\}$ of $V\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\}$, we can get a consistent variance estimator of $\hat{\xi}_{p,HD}$. If the linearization method is used, we get the variance estimator of $\hat{\xi}_{p,HD}$ as

$$\hat{V}(\hat{\xi}_{p,HD}) \cong \frac{1}{[\hat{f}(\hat{\xi}_{p,HD})]^2} \hat{V}_{rep}\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\}. \tag{13}$$

If the test inversion method is used, we get the variance estimator of $\hat{\xi}_{p,HD}$ as

$$\hat{V}(\hat{\xi}_{p,HD}) = \left(\frac{\hat{\xi}_{p,U} - \hat{\xi}_{p,L}}{4}\right)^2, \tag{14}$$

where $(\hat{\xi}_{p,L}, \hat{\xi}_{p,U}) = (\hat{F}^{-1}(\hat{p}_L), \hat{F}^{-1}(\hat{p}_U))$, $(\hat{p}_L, \hat{p}_U) = \hat{F}_{HD}(\hat{\xi}_{p,HD}) \pm 2\sqrt{\hat{V}_{rep}\{\hat{F}_{HD}(\hat{\xi}_{p,HD})\}}$.

5. SIMULATION STUDY

We performed a limited simulation study and a real data analysis. In Section 5.1, we compared the performance of the proposed method with some other imputation methods in a correctly specified model and a misspecified model. In Section 5.2, we applied FHDI to a real data analysis from the Korea Labor and Income Panel Survey (KLIPS).

5.1 Simulation

Two sets of models were considered to generate the observations. In Model A, we used $y_i = 1 + x_i + e_i$, where $x_i \sim N(0,1)$, $e_i \sim N(0,1)$, x_i and e_i are independent. In Model B, we used $y_i = 1 + x_i + e_i$, where $x_i \sim N(0,1)$, $e_i \sim Exp(1) - 1$, x_i and e_i are independent. Random samples of size $n = 200$ were separately generated from the two models. In addition to (x_i, y_i), we also generated δ_i, the response indicator variable, from Bernoulli distributions with response rate 0.6. Variable x_i is always observed but variable y_i is observed if and only if $\delta_i = 1$. We used $B = 2,000$ Monte Carlo samples in the simulation. In each of the samples, we computed the following five estimators:

1. Full sample (Full) estimator that is computed using the complete observations.
2. Multiple imputation (MI) estimator with imputation size m, where the imputed values are generated from the normal-theory regression model, as considered in Schenker and Welsh (1988).
3. Parametric fractional imputation (PFI) estimator with imputation size m.
4. Fractional hot deck imputation (FHDI) estimator using the full set of respondents as imputation values ($m = n_r$) where n_r is the size of respondents.
5. Nonparametric fractional imputation (NPFI) estimator.

In MI and PFI, we set the imputation size $m = n_r$, the same as that in FHDI, for fair comparison. In both Models A and B, we used a working model which is the normal density with mean $\beta_0 + \beta_1 x$ and variance σ^2 as the imputation model. Thus, the working model is the true model in model A but not true in model B. In NPFI, the nonparametric kernel regression estimator was computed using a Gaussian kernel function with bandwidth $h = an^{-2/5}$, suggested by Cheng (1994), where $a = 0.2$.

We considered four parameters: the mean of $y(\mu_y)$, the 25% quantile ($\xi_{0.25}$), the median ($\xi_{0.5}$) and the 75% quantile ($\xi_{0.75}$) of y. The full sample estimator was used as a benchmark which is unbiased for the parameters considered.

Tables 1 and 3 present Monte Carlo mean, variance and standardized variance of the point estimators, based on 2,000 Monte Carlo samples for Model A and Model B, respectively. The standardized variance is calculated as the ratio of variance and the variance of the full sample estimator multiplied by 100, which measures the increased variance due to imputation relative to the full sample estimator. Comparing the Monte Carlo means in the third column, the imputation estimators are essentially unbiased in estimating the parameters considered in Model A, which is expected since the imputed estimating equations are consistent under the correctly specified model. Comparing the standardized variance in the fourth column, PFI and MI have smaller standardized variances than FHDI and NPFI, which suggests that PFI and MI are more efficient than FHDI and NPFI. The reason is that in PFI and MI, the imputed values are generated according to the conditional distribution $f(y|x)$ directly, whereas in FHDI, the imputed values are taken from the respondents in which case some of the fractional weights can be large and thus dominate other weights resulting lose of efficiency. FHDI and NPFI lose efficiency in order to gain robustness which is shown in Model B. In Model B, both MI and PFI show no robustness against model misspecification. MI and PFI are unbiased

Table 1. Mean, variance and standardized variance of the point estimators, based on 2,000 Monte Carlo samples in Model A

Parameter	Method	Mean	Var	Std Var
μ_y	Full	0.998	0.0102	100
	MI $(m = n_r)$	0.997	0.0132	129
	PFI $(m = n_r)$	0.997	0.0132	129
	FHDI $(m = n_r)$	0.996	0.0135	132
	NPFI	0.996	0.0134	131
$\xi_{0.25}$	Full	0.047	0.0178	100
	MI$(m = n_r)$	0.046	0.0203	114
	PFI $(m = n_r)$	0.046	0.0208	117
	FHDI $(m = n_r)$	0.046	0.0266	149
	NPFI	0.047	0.0265	149
$\xi_{0.5}$	Full	0.997	0.0157	100
	MI $(m = n_r)$	0.998	0.0166	106
	PFI $(m = n_r)$	0.998	0.0173	110
	FHDI $(m = n_r)$	0.999	0.0231	147
	NPFI	0.998	0.0231	147
$\xi_{0.75}$	Full	1.951	0.0182	100
	MI $(m = n_r)$	1.949	0.0199	109
	PFI $(m = n_r)$	1.948	0.0207	113
	FHDI $(m = n_r)$	1.949	0.0272	149
	NPFI	1.945	0.0281	154

Table 3. Mean, variance and standardized variance of the point estimators, based on 2,000 Monte Carlo samples in Model B

Parameter	Method	Mean	Var	Std Var
μ_y	Full	0.995	0.0095	100
	MI $(m = n_r)$	1.003	0.0136	145
	PFI $(m = n_r)$	1.003	0.0136	145
	FHDI $(m = n_r)$	0.996	0.0134	142
	NPFI	0.995	0.0133	141
$\xi_{0.25}$	Full	0.040	0.0125	100
	MI$(m = n_r)$	0.051	0.0137	110
	PFI $(m = n_r)$	0.052	0.0143	115
	FHDI $(m = n_r)$	0.041	0.0175	141
	NPFI	0.042	0.0167	134
$\xi_{0.5}$	Full	0.872	0.0125	100
	MI $(m = n_r)$	0.929	0.0145	117
	PFI $(m = n_r)$	0.929	0.0149	119
	FHDI $(m = n_r)$	0.879	0.0172	138
	NPFI	0.872	0.0175	140
$\xi_{0.75}$	Full	1.800	0.0196	100
	MI $(m = n_r)$	1.872	0.0245	125
	PFI $(m = n_r)$	1.870	0.0251	128
	FHDI $(m = n_r)$	1.803	0.0298	152
	NPFI	1.795	0.0299	153

Table 2. Monte Carlo relative biases and t-statistics of the variance estimators, based on 2,000 Monte Carlo samples in Model A

Parameter	Method	R.B. (%)	t-statistics
μ_y	MI $(m = n_r)$	1.2	0.38
	PFI $(m = n_r)$	0.6	0.19
	FHDI $(m = n_r)$	−4.4	−1.33
	NPFI	−4.0	−1.25
$\xi_{0.25}$	MI$(m = n_r)$	29.8	9.22
	PFI $(m = n_r)$	−0.01	−0.01
	FHDI $(m = n_r)$	−1.5	−0.45
	NPFI	−1.4	−0.45
$\xi_{0.5}$	MI $(m = n_r)$	30.1	9.41
	PFI $(m = n_r)$	−1.5	−0.49
	FHDI $(m = n_r)$	−1.6	−0.51
	NPFI	−1.7	−0.54
$\xi_{0.75}$	MI $(m = n_r)$	30.7	9.66
	PFI $(m = n_r)$	1.5	0.48
	FHDI $(m = n_r)$	−1.6	−0.49
	NPFI	−3.2	−1.02

for estimating population mean but are biased for estimating quantiles due to the misspecified imputation model. On the other hand, the FHDI estimator and the NPFI estimator are essentially unbiased in estimating population mean and quantile. The robustness of the NPFI estimator against this misspecification is due to the fact that nonparametric models avoid restrictive assumptions on the functional form of the regression function. Even though FHDI method is a parametric imputation, it turns out that FHDI estimator is also robust to model misspecification in this case due to the special structure of fractional weights. Further discussion of the robustness feature of FHDI is discussed by Yang and Kim.

For variance estimation, in MI, we can use either linearization variance estimator (1) or Woodruff variance estimator (2) in Rubin's variance function (4). It turns out that the results from these two types of variance estimator are comparable, so we will only present the results from (1). In PFI, we used the variance estimator as described in Appendix A.1. In FHDI and NPFI, we used two-step Jackknife variance estimator (14).

Tables 2 and 4 present Monte Carlo relative biases and t-statistics of the variance estimators to test the significance of the bias of the variance estimators for Model A and Model B, respectively. The relative bias is calculated as $[E_{MC}\{\hat{V}\} - V_{MC}\{\hat{\theta}\}]/V_{MC}\{\hat{\theta}\}$, where $E_{MC}\{\hat{V}\}$ is the Monte Carlo mean of variance estimates $\hat{V}$, and $V_{MC}\{\hat{\theta}\}$ is the Monte Carlo variance of the point estimates $\hat{\theta}$. A justification of the t-statistics can be found in Appendix D of Kim (2004), which claims that the bias is not statistically significant if the t-statistics is less than 2. The relative bias of variance estimator in MI is small (1.2%) for μ_y, but is quite large (29.8%, 30.1%, and 30.7%) for quantiles even when the working model is true (Model A), which is also confirmed by the t-statistics. The t-statistics is small (0.38) for μ_y and is quite large (9.22, 9.41, and 9.66) for quantiles, which exceed 2 by a large amount, indicating that the MI variance estimator is biased for quantile. Rubin's formula is based on the following decomposition,

Table 4. Monte Carlo relative biases and t-statistics of the variance estimators, based on 2,000 Monte Carlo samples in Model B

Parameter	Method	R.B. (%)	t-statistics
μ_y	MI $(m = n_r)$	−1.9	−0.59
	PFI $(m = n_r)$	9.3	2.87
	FHDI $(m = n_r)$	−4.7	−1.46
	NPFI	−3.2	−0.99
$\xi_{0.25}$	MI$(m = n_r)$	81.2	26.26
	PFI $(m = n_r)$	65.5	20.89
	FHDI $(m = n_r)$	3.0	0.99
	NPFI	4.7	1.48
$\xi_{0.5}$	MI $(m = n_r)$	45.7	13.95
	PFI $(m = n_r)$	31.4	9.66
	FHDI $(m = n_r)$	5.2	1.65
	NPFI	3.5	1.09
$\xi_{0.75}$	MI $(m = n_r)$	4.7	1.36
	PFI $(m = n_r)$	−19.9	−6.17
	FHDI $(m = n_r)$	−3.8	−1.19
	NPFI	−0.1	−0.03

$$V(\hat{\theta}_{MI}) = V(\hat{\theta}_n) + V(\hat{\theta}_{MI} - \hat{\theta}_n) \tag{15}$$

where $\hat{\theta}_n$ is the full sample estimator of θ. Basically, W_m term in (3) estimates $V(\hat{\theta}_n)$ and $(1 + m^{-1})B_m$ term in (3) estimates $V(\hat{\theta}_{MI} - \hat{\theta}_n)$. The decomposition (15) holds when $\hat{\theta}_n$ is the MLE of θ, which is the congeniality condition of $\hat{\theta}_n$ (Meng, 1994). For a general case, we have

$$V(\hat{\theta}_{MI}) = V(\hat{\theta}_n) + V(\hat{\theta}_{MI} - \hat{\theta}_n) + 2Cov(\hat{\theta}_{MI} - \hat{\theta}_n, \hat{\theta}_n) \tag{16}$$

and Rubin's variance estimator can be biased. The congeniality condition holds true for $\hat{\mu}_y$; however, it does not hold for the method of moments estimator of quantiles. In PFI, the t-statistics of the variance estimator (Appendix A.1) for all parameters considered here are less than 2 in Model A, however exceed 2 in Model B, suggesting that the variance estimator in PFI is valid if and only if it is under the true model. The reason is that the variance estimator is essentially derived from the observed fisher information which is model dependent. On the other hand, in FHDI and NPFI, the t-statistics of the variance estimators are less than 2, which show that the two-step variance estimator is unbiased.

5.2 Real data analysis

In this section, the proposed FHDI method was applied to real data. The data set used is obtained from the Korea Labor and Income Panel Survey (KLIPS). We used the data set of size (n = 2,506) which consists of the regular wage earners in the sample of year 2008. A brief description of the panel survey can be found at http://www.kli.re.kr/klips/en/about/introduce.jsp. The study variable (y) is the average monthly income for the current year and the auxiliary variable (x) is the average monthly income for the previous year. Figure 1 reports the histogram plot of y in the full sample. The sample distribution of y is skewed to the left and the sample 25% quantile, the median, and 75% quantile of y is $(1.03,\ 1.6,\ 2.3) \times 10^6$ Korean Won. The sample mean of (x, y) is $(1.6643,\ 1.8504) \times 10^6$ Korean Won, the sample correlation between x and y is 0.8144. Figure 2 reports the scatter plot of y versus x. From the figure, the functional relationship for y in terms of x can be treated as linear.

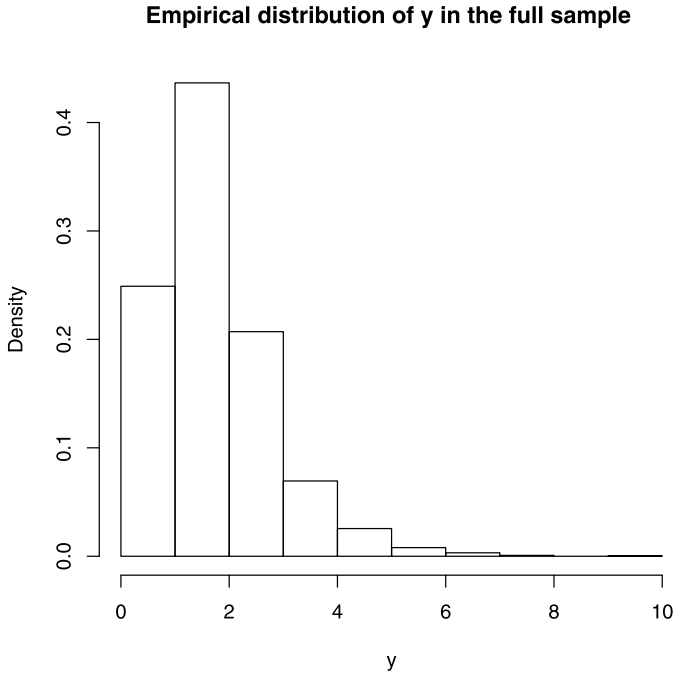

Figure 1. Histogram Plot of Current Year Monthly Income in the Full Sample. Unit (10^6 Korean Won).

From the sample described above, we created artificial missing data by deliberately deleting some of the y values according to the response mechanism Bernoulli(π), where $\pi(x) = \{1 + \exp(-\phi_0 - \phi_1 x)\}^{-1}$, with $(\phi_0, \phi_1) = (-1.1, 1.0)$. From this response mechanism, the response rate is roughly 60%. Figure 3 reports the histogram plot of y in the respondents. Compared to the histogram plot of y in the full sample, the sample distribution of y pertaining to the respondents has shifted to the right. From the respondents, the sample 25% quantile, the median, and the sample 75% quantile of y is $(1.28,\ 1.94,\ 2.7) \times 10^6$ Korean Won, and the sample mean of (x, y) is $(1.9309,\ 2.1117) \times 10^6$ Korean Won.

In FHDI, the sample is partitioned into $4 \times 2 \times 2$ cells by age (<30, $[30, 40)$, $[40, 50)$, $\geqslant 50$), gender (1 = male and 2 = female) and level of education (1 = high school or lower, 2 = college or higher). The sixteen cells consist of sample of size 99, 130, 91, 193, 250, 406, 107, 141, 241, 202, 179, 49, 207, 85, 115, 11, respectively. In each cell, we applied FHDI using the imputation model $y_i = \beta_0 + \beta_1 x_i + e_i$, where $e_i \sim$

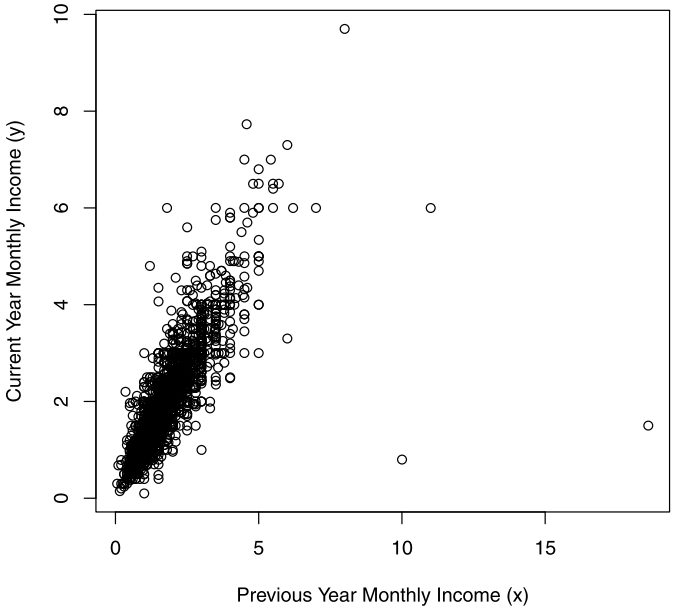

Figure 2. Scatter Plot of Current Year Monthly Income versus Previous Year Monthly Income in 2008 Korea Labor and Income Panel Survey. Unit (10^6 Korean Won).

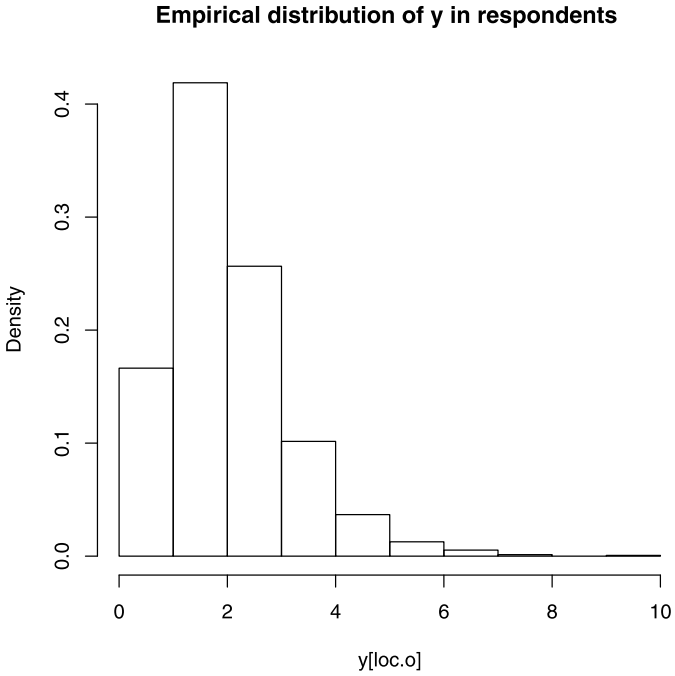

Figure 3. Histogram Plot of Current Year Monthly Income in the respondents. Unit (10^6 Korean Won).

$N(0, \sigma^2)$ to create imputed values for missing units. Once the imputed data was created, the usual complete sample estimators can be applied. For variance estimation, we used the usual delete-1 observation Jackknife variance estimator for μ_y and two-step Jackknife variance estimator (14) for quantiles.

Table 5. Point estimates, estimated variance, 95% confidence intervals from FHDI

	Est	Var Est	95% C.I.
μ_y	1.883	0.0004987	(1.838, 1.928)
$\xi_{0.25}$	1.12	0.04730625	(0.68, 1.55)
$\xi_{0.5}$	1.70	0.00160000	(1.59, 1.75)
$\xi_{0.75}$	2.38	0.00122500	(2.30, 2.44)

Table 5 presents the estimates for μ_y, $\xi_{0.25}$, $\xi_{0.5}$ and $\xi_{0.75}$, their estimated variances, and 95% confidence intervals under missing. The full sample mean $\hat{\mu}_n$, sample quantiles $\hat{\xi}_{0.25,n}$, $\hat{\xi}_{0.5,n}$, and $\hat{\xi}_{0.75,n}$ are successfully captured by the 95% confidence intervals. In conclusion, this case study demonstrates the empirical effectiveness of the FHDI estimator.

6. CONCLUDING REMARKS

In this paper, four imputation methods were considered for quantile estimation with missing data. MI, applied to quantile estimation, does not satisfy the congeniality condition of Meng (1994) and can lead to biased variance estimation. Fractional imputation methods, on the other hand, do not require the congeniality condition and provide consistent variance estimators.

In the correctly specified model, among the four methods, MI and PFI turn out to be more efficient in point estimation than FHDI and NPFI. In the misspecified model, the FHDI and NPFI are shown to be much better than MI and PFI in terms of bias. The PFI is not robust to misspecification of models in variance estimation. In FHDI, there is no random imputation and thus fractional weights are deterministically computed, which enables simplified replication variance estimation. A revised Jackknife variance estimation method produces an essentially unbiased estimator. Properties of FHDI carry over to NPFI. Furthermore, for NPFI method, no parametric model assumption is required and hence the resulting estimator is robust. As with the usual nonparametric methods, the NPFI method may be subject to the curse of dimensionality associated with nonparametric estimation, if the dimension is high. More rigorous theoretical investigation of the NPFI method in the high dimension cases will be a good topic of future study.

ACKNOWLEDGMENT

We thank two anonymous referees for very helpful comments. This work was supported by Basic Research Program (2012-001361) and SRC Program (2011-0030811) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education Science and Technology.

APPENDIX A. APPENDIX SECTION

A.1 Variance estimation in PFI

Since β and θ are information orthogonal, we can use Louis's formula to construct the confidence intervals for β.

(17)
$$I_{obs}(\beta) = -\sum_{i=1}^{n} E\big\{\dot{S}(\beta; y_i)|y_{i,obs}\big\} - \sum_{i=1}^{n} V\big\{S(\beta; y_i)|y_{i,obs}\big\},$$

which can be approximated by

(18)
$$-\sum_{i=1}^{n}\sum_{k=1}^{M} w_i^{*(k)} \dot{S}(\hat{\beta}; y_i^{*(k)}) - \sum_{i=1}^{n}\sum_{k=1}^{M} w_i^{*(k)} \{S(\hat{\beta}; y_i^{*(k)}) - \bar{S}_i(\hat{\beta})\}^{\otimes 2},$$

where $S(\beta; y) = \partial \log f(y; \beta)/\partial\beta$, $\dot{S}(\beta; y) = \partial S(\beta; y)/\partial\beta$ and $\bar{S}_i(\beta) = \sum_{k=1}^{M} w_i^{*(k)} S(\beta; y_i^{*(k)})$.

For variance estimation of $\hat{\eta}$, based on Taylor linearization obtained from $\bar{U}^*(\eta) = 0$, we can write $\bar{U}(\eta|\hat{\gamma}) \approx \bar{U}(\eta_0|\gamma_0) + K'\bar{S}(\gamma_0)$, where K is defined as

$$K = -[E\{\partial\bar{S}(\gamma_0)/\partial\gamma\}]^{-1} E\{S_{mis}(\gamma_0)U(\eta_0)\}.$$

If we write

$$\bar{U}(\eta|\gamma) + K'\bar{S}(\gamma) = n^{-1}\sum_{i=1}^{n}\{\bar{\mathbf{u}}_i(\eta|\gamma) + K'\bar{\mathbf{s}}_i(\gamma)\} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{u}}_i,$$

the plug-in estimator of $\mathrm{Var}(\sum_{i=1}^{n}\tilde{\mathbf{u}}_i)$ is $\sum_{i=1}^{n}(\hat{\mathbf{u}}_i - \bar{\hat{\mathbf{u}}})(\hat{\mathbf{u}}_i - \bar{\hat{\mathbf{u}}})'$, where $\hat{\mathbf{u}}_i = \bar{\mathbf{u}}_i(\hat{\eta}; \hat{\gamma}) + \hat{K}'\bar{\mathbf{s}}_i(\hat{\gamma})$. The terms $\bar{\mathbf{u}}_i(\hat{\eta}; \hat{\gamma})$ and $\bar{\mathbf{s}}_i(\hat{\gamma})$ can be computed from fractional imputation with fractional weights.

Received 29 November 2012

REFERENCES

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics*, **37**, 577–580. MR0189095

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, **5**, 215–238.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, **89**, 81–87.

Fay R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, **91**, 490–498.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley-Interscience. MR1925014

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, **9**, 538–573.

Milliken, G. A. and Johnson, D. E. (2002). *Analysis of Messy Data: Analysis of Covariance*. Chapman and Hall/CRC.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**(1), 141–142.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, **13**, 1919–1939.

Kim, J. K. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics*, **32**, 766–783. MR2060177

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, **98**, 119–132. MR2804214

Kim, J. K. and Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, **91**, 559–578. MR2090622

Kim, J. K. and Yu, C. Y. (2011). A semi-parametric estimation of mean functional with non-ignorable missing data. *Journal of the American Statistical Association*, **106**, 157–165. MR2816710

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. MR0899519

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Annals of Statistics*, **16**, 1550–1566. MR0964938

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, **26**(4), 359–372. MR0185765

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, **47**, 635–646. MR0050845

Shu Yang
Department of Statistics, Iowa State University
Ames, IA 50011
USA
E-mail address: shuyang@iastate.edu

Jae-Kwang Kim
Department of Statistics, Iowa State University
Ames, IA 50011
USA
E-mail address: jkim@iastate.edu

Dong Wan Shin
Department of Statistics, Ewha University
Seoul, 120-750
Republic of Korea
E-mail address: shindw@ewha.ac.kr