*Journal of the Korean Statistical Society (2007), **36**: 3, pp 423–434*

REGRESSION FRACTIONAL HOT DECK IMPUTATION[†]

JAE KWANG KIM[1]

ABSTRACT

Imputation using a regression model is a method to preserve the correlation among variables and to provide imputed point estimators. We discuss the implementation of regression imputation using fractional imputation. By a suitable choice of fractional weights, the fractional regression imputation can take the form of hot deck fractional imputation, thus no artificial values are constructed after the imputation. A variance estimator, which extends the method of Kim and Fuller (2004), is also proposed. Results from a limited simulation study are presented.

AMS 2000 subject classifications. Primary 62D05; Secondary 62J99.
Keywords. Missing data, nonresponse, variance estimation.

1. INTRODUCTION

Consider a finite population of N elements identified by a set of indices $U = \{1, 2, \ldots, N\}$ with N known. Associated with each unit i in the population there are two study variables, x_i and y_i, where x_i is always observed and y_i is subject to nonresponse. Let A denote the set of indices for the elements in a sample selected by a set of probability rules called the *sampling mechanism.* Under complete response, an unbiased estimator of $\theta_1 = N^{-1}\sum_{i=1}^{N} y_i$, the population mean of y, is

$$\widehat{\theta}_1 = \sum_{i \in A} w_i y_i, \tag{1.1}$$

where $w_i = N^{-1}\left[\Pr\left(i \in A\right)\right]^{-1}$ is the inverse of the inclusion probability of unit i divided by N. In addition to the population mean, suppose that we are also

Received October 2006; accepted April 2007.
[†]This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. R01-2005-001-11057-0).
[1]Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea (e-mail: kimj@yonsei.ac.kr)

interested in the population correlation between x_i and y_i. For simplicity, we assume that another parameter of interest is $\theta_2 = N^{-1}\sum_{i=1}^{N} x_i y_i$ and the complete sample estimator of θ_2 is

$$\widehat{\theta}_2 = \sum_{i \in A} w_i x_i y_i. \tag{1.2}$$

To deal with item nonresponse, we define A_R and A_M as the set of indices of the sample respondents and sample nonrespondents, respectively. In many practical cases, the imputed value y_i^* is written as a predicted value plus a residual term

$$y_i^* = \widehat{y}_i + \widehat{e}_i^*, \tag{1.3}$$

where $\widehat{y}_i$ is the predicted value of y_i and $\widehat{e}_i^*$ is an imputed residual selected from $\{\widehat{e}_i = y_i - \widehat{y}_i;\ i \in A_R\}$. If we assume a linear regression model, the predicted value for unit i is of the form $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ for some $(\widehat{\beta}_0, \widehat{\beta}_1)$ computed from the respondents. The residual term is added to preserve the marginal variability of the original data after imputation.

Regression imputation of the form (1.3) has two main motivation. First, under the regression model

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i \quad \text{and} \quad V(y_i \mid x_i) = \sigma^2, \tag{1.4}$$

the resulting imputed estimator for θ_1 is conditionally unbiased and often quite efficient. Secondly, the imputed data set preserves the correlation structure between the two variables and the resulting imputed estimator of θ_2 is also unbiased. Deville and Särndal (1994) and Shao and Wang (2002) also discussed regression imputation in survey sampling.

However, the regression imputation method is not a hot deck imputation in the sense that artificial values can be constructed after the imputation. A desirable property of hot deck imputation is that all imputed values are observed values. For example, imputed values for categorical variables will also be categorical with the same number of categories as observed for the respondents. For this reason, hot deck imputation is the most popular imputation method, especially in the household surveys.

In this article, we propose a new imputation method that combines the advantages of the hot deck imputation and the regression imputation. The way we combine the two imputation methods takes the form of fractional imputation. Fractional imputation was originally motivated to reduce the imputation variance (Kalton and Kish, 1984), but later it also provides a useful tool for variance

estimation (Kim and Fuller, 2004). Thus, a third advantage of using fractional imputation is that it can be made to preserve the correlation structure of the important variables but still take the form of a hot deck imputation.

The proposed method is called the fractional regression hot deck imputation (FRHDI). The FRHDI method shares the advantages of the fractional hot deck imputation discussed in Kim and Fuller (2004) and still preserves the correlation structure of the variables used in the regression model.

The proposed imputation method is introduced in Section 2. Variance estimation is discussed in Section 3. Results from a limited simulation study are reported in Section 4. Concluding remarks are made in Section 5.

2. PROPOSED METHOD

Fractional imputation is a procedure in which more than one donor is used per recipient. Let d_{ij} take the value one if y_i is used as donor for the missing y_j and take the value zero otherwise. Let w_{ij}^* be the factor applied to the original weight for element j when y_i is used as a donor for element j. The factor w_{ij}^* is called the fractional weight. It is the fraction that donor i donates for the missing j. The fractional weights satisfy

$$\sum_{i \in A_R} w_{ij}^* = 1 \quad \text{for} \quad j \in A_M. \tag{2.1}$$

When the fractional imputation is applied to the regression imputation (1.3), the imputed estimator of θ_1 can be constructed as

$$\widehat{\theta}_{1I} = \sum_{i \in A_R} w_i y_i + \sum_{j \in A_M} \sum_{i \in A_R} w_j w_{ij}^* \left(\widehat{y}_j + \widehat{e}_i\right). \tag{2.2}$$

Note that when $\widehat{y}_j$ is linear function of x_j, the weighted mean of the imputed values can be written

$$\sum_{i \in A_R} w_{ij}^* \left(\widehat{y}_j + \widehat{e}_i\right) = \sum_{i \in A_R} w_{ij}^* y_i$$

if

$$\sum_{i \in A_R} w_{ij}^* \left(1, x_i\right) = \left(1, x_j\right) \quad \text{for} \quad j \in A_M. \tag{2.3}$$

Thus, under (2.3), the resulting imputed estimator can be written

$$\widehat{\theta}_{1I} = \sum_{i \in A_R} w_i y_i + \sum_{j \in A_M} \sum_{i \in A_R} w_j w_{ij}^* y_i, \tag{2.4}$$

which is an expression for a hot deck imputation. Condition (2.3) enables the regression fractionally imputed estimator to use only the reported values for imputation. Also, the FRHDI estimator for θ_2 can be written

$$\widehat{\theta}_{2I} = \sum_{i \in A_R} w_i x_i y_i + \sum_{j \in A_M} \sum_{i \in A_R} w_j w_{ij}^* x_j y_i. \tag{2.5}$$

Fractional weights with constraint (2.3) can be constructed using the regression weighting technique. Let w_{ij0}^* be any initial fractional weights satisfying $\sum_{i \in A_R} w_{ij0}^* = 1$. A common choice is $w_{ij0}^* = M^{-1} d_{ij}$ for $j \in A_M$ where M is the number of donors used for fractional imputation. Under the fractional imputation, the regression weighting method can be used to get

$$w_{ij}^* = w_{ij0}^* + (x_j - \bar{x}_{Ij0}) S_{xx,j}^{-1} w_{ij0}^* (x_i - \bar{x}_{Ij0}), \tag{2.6}$$

where

$$S_{xx,j} = \sum_{i \in A_R} w_{ij0}^* (x_i - \bar{x}_{Ij0})^2,$$
$$\bar{x}_{Ij0} = \sum_{i \in A_R} w_{ij0}^* x_i.$$

If some of the final fractional weight w_{ij}^* are negative, then initial fractional weights w_{ij0}^* can be modified to produce nonnegative fractional weights.

The fractional weights w_{ij}^* in (2.6) are obtained by minimizing

$$\sum_{i \in A_R} d_{ij} \left(\frac{w_{ij}^*}{w_{ij0}^*} - 1 \right)^2 \tag{2.7}$$

subject to the constraints in (2.3). If different objective function is used in this optimization problem, then the fractional weights may be different from (2.6) but the unbiasedness of $\widehat{\theta}_{1I}$ and $\widehat{\theta}_{2I}$ still follows as long as they satisfy (2.3). Main advantage of using the objective function (2.7) is that we have a closed-form solution.

To show the unbiasedness of the FRHDI estimator, we assume a model more general than the simple regression model in (1.4). To do this, assume that the sample A is made up of G imputation cells and the imputed residuals are selected in the same cell. The set of elements in cell g is A_g and let $A_{Rg} = A_R \cap A_g$. We assume the cell regression model

$$E(y_i \mid x_i, i \in A_g) = \beta_{0g} + \beta_1 x_i \quad \text{and} \quad V(y_i \mid x_i, i \in A_g) = \sigma_g^2, \tag{2.8}$$

which is an extension of the linear regression model (1.4). Thus, we allow for unequal intercepts and unequal variances at each cell. Also, we assume that the response mechanism is ignorable under the model (2.8) or missing at random using the terminology of Rubin (1976).

Under the cell regression model and ignorable response mechanism, the conditional expectation of the FRHDI estimator of θ_1 is

$$E\left(\widehat{\theta}_{1I} - \widehat{\theta}_1 | A, A_R, \mathcal{X}\right) = E\left[\sum_{j\in A_M} w_j \left\{\left(\sum_{i\in A_R} w_{ij}^* y_i\right) - y_j\right\} \middle| A, A_R, \mathcal{X}\right]$$
$$= \sum_{g=1}^{G} \sum_{j\in A_{Mg}} w_j \left\{\sum_{i\in A_R} w_{ij}^* \left(\beta_{0g} + \beta_1 x_i\right) - \left(\beta_{0g} + \beta_1 x_j\right)\right\},$$

which is equal to zero by (2.3), where $\mathcal{X} = \{(i, x_i) : i \in A\}$. Since the above equality also holds if we replace w_i by $w_i x_i$, the conditional unbiasedness also holds for $\widehat{\theta}_{2I}$. Therefore, the FRHDI estimators are unbiased for both parameters since the complete sample estimators are unbiased under the sampling mechanism.

3. VARIANCE ESTIMATION

We now consider variance estimation for the FRHDI estimators. Under the cell regression model (2.8), the imputed estimator $\widehat{\theta}_{I1}$ in (2.4) is unbiased and the variance can be written

$$\mathrm{Var}\left(\widehat{\theta}_{I1}\right) = \mathrm{Var}\left(\sum_{i\in A} w_i \mu_i\right) + E\left\{\sum_{g=1}^{G} \sum_{i\in A_{Rg}} \alpha_i^2 \sigma_g^2\right\}, \tag{3.1}$$

where $\mu_i = \beta_{0g} + \beta_1 x_i$ for $i \in A_g$ and $\alpha_i = \sum_{j\in A} w_j w_{ij}^*$.

Assume replicates are to be used to estimate the variance and let the replication variance estimator for the complete sample be

$$\widehat{V}\left(\widehat{\theta}\right) = \sum_{k=1}^{L} c_k \left(\widehat{\theta}^{(k)} - \widehat{\theta}\right)^2, \tag{3.2}$$

where $\widehat{\theta}$ is the full sample estimator, $\widehat{\theta}^{(k)}$ is the k^{th} estimate of θ_1 based on the observations included in the k^{th} replicate, L is the number of replicates and c_k is a factor associated with replicate k determined by the replication method. If we treat the imputed values as if observed and apply the complete sample variance

estimator (3.2), the resulting variance estimator has the expectation

$$E\left\{\widehat{V}_1\left(\widehat{\theta}\right)\right\} = V\left(\sum_{i\in A} w_i\mu_i\right) + E\left\{\sum_{k=1}^{L}\sum_{g=1}^{G}\sum_{i\in A_{Rg}} c_k\left(\alpha_{i1}^{(k)} - \alpha_i\right)^2\sigma_g^2\right\}, \quad (3.3)$$

where $\alpha_{i1}^{(k)} = \sum_j w_j^{(k)} w_{ij}^*$ and $w_j^{(k)}$ is the weight for element j in replicate k. The first term in (3.1) can be safely estimated by the naive variance estimator, but the second term of (3.1) is not unbiasedly estimated. To estimate the second term, we use a version of fractional imputation discussed in Kim and Fuller (2004).

We outline a replication variance estimator closely related to that of Kim and Fuller (2004) that changes the fractional replicate weights of the naive variance estimator to produce a consistent estimator of the variance. Let superscript k denote the replicate where element k is in the deleted set. The replicated fractional weights $w_{ij}^{*(k)}$ are to be created in such a way that the resulting variance estimator is unbiased for the variance in (3.1). To achieve unbiasedness, two conditions are needed for the replicated fractional weights. The two conditions are

$$\sum_{i\in A_R} w_{ij}^{*(k)}(1, x_i) = (1, x_j) \quad (3.4)$$

and

$$\sum_{k=1}^{L} c_k\left\{\sum_{i\in A_{Rg}}\left(\alpha_i^{(k)} - \alpha_i\right)^2\right\} = \sum_{i\in A_{Rg}}\alpha_i^2, \quad (3.5)$$

where $\alpha_i^{(k)} = \sum_j w_j^{(k)} w_{ij}^{*(k)}$. Note that, under cell mean model with $x_i \equiv 0$, conditions (3.4) and (3.5) are the same conditions discussed in Kim and Fuller (2004). A sufficient condition for (3.5) is

$$\sum_{k=1}^{L} c_k\left\{\left(\alpha_i^{(k)} - \alpha_i\right)^2 + \sum_{t\in D_{Ri}}\left(\alpha_t^{(k)} - \alpha_t\right)^2\right\} = \alpha_i^2 + \sum_{t\in D_{Ri}}\alpha_t^2, \quad (3.6)$$

where $D_{Ri} = \{t\,;\sum_{j\in A_M} d_{ij}d_{tj} = 1\}$ is the set of donors, other than i, to recipients from donor i.

The fractional weights assigned to donor k are changed so that the expected value of the sum of squares is changed by the proper amount. Consider the class

of the fractional weight in replicate k for the value donated by i to j be

$$w_{ijb}^{*(k)} = \begin{cases} w_{ij}^* - \left(1 - w_{ij0}^*\right) b_k + h_{ik,j} b_k, & \text{if } i = k \text{ and } d_{ij} = 1, \\ w_{ij}^* + w_{ij0}^* b_k + h_{ik,j} b_k, & \text{if } i \neq k \text{ and } d_{kj} = 1 \text{ and } d_{ij} = 1, \\ w_{ij}^*, & \text{otherwise,} \end{cases} \tag{3.7}$$

where M is the number of donors for each recipient,

$$h_{ik,j} = w_{ij0}^* \left(x_i - \bar{x}_{Ij0}\right) S_{xx,j}^{-1} \left(x_k - \bar{x}_{Ij0}\right)$$

and b_k is to be determined. Note that the replicated fractional weights in (3.7) satisfy (3.4) for any value of b_k, since $\sum_{i \in A_R} h_{ik,j} = 0$ and $\sum_{i \in A_R} h_{ik,j} x_i = x_k - \bar{x}_{Ij}$ for any $j \in A_M$. Suitable choice of b_k can satisfy condition (3.6) and the resulting variance estimator will be unbiased.

Using (3.7), the b_k that satisfies (3.6) can be computed as the solution to the quadratic equation

$$\begin{aligned} & c_k \left\{ \alpha_{k0}^{(k)} - \alpha_k - b_k \sum_{j \in A_M} w_j^{(k)} \left(1 - w_{kj0}^* - h_{kk,j}\right) \right\}^2 - c_k \left(\alpha_{k0}^{(k)} - \alpha_k\right)^2 \\ & + \sum_{t \in D_{Rk}} c_k \left\{ \alpha_{t0}^{(k)} - \alpha_t + b_k \sum_{j \in A_M} w_j^{(k)} \left(w_{tj0}^* + h_{tk,j}\right) \right\}^2 \\ & - \sum_{t \in D_{Rk}} c_k \left(\alpha_{t0}^{(k)} - \alpha_t\right)^2 = \alpha_k^2 - \phi_k, \end{aligned} \tag{3.8}$$

where D_{Rk} is the set of donors, other than k, to recipients from donor k, $\alpha_{i0}^{(k)} = \sum_{j \in A} w_j^{(k)} w_{ij}^*$, $\alpha_i = \sum_{j \in A} w_j w_{ij}^*$ and

$$\phi_i = \sum_{k=1}^{L} c_k \left(\alpha_{i0}^{(k)} - \alpha_i\right)^2. \tag{3.9}$$

The difference $\alpha_k^2 - \phi_k$ is the difference between the desired sum of squares for observation k and the sum of squares for the naive estimator. The procedure provides unbiased estimate of the conditional variance of the imputed estimator under the assumption of common variance within each cell.

4. SIMULATION STUDY

To test our theory, we performed two limited simulation studies. In the first simulation study, $B = 5,000$ Monte Carlo random samples of size $n = 100$ are

generated from the linear regression model

$$y_i = 2 + x_i + e_i, \tag{4.1}$$

where x_i and e_i are independently generated from the standard normal distribution with zero correlation. In addition to (x_i, y_i), z_i are generated from the uniform $(0, 1)$ distribution, independent of (x_i, y_i).

From each sample, we also generated a response indicator variable R_i from a Bernoulli distribution with the response rate $p = 0.65$. The Y_i is observed if and only if $R_i = 1$. The x_i and z_i are observed throughout the sample.

For the imputation mechanism, we used two imputation methods. The first one is the fractional imputation (FI) with M fractions, where the M donors for each missing unit are selected with M closest x-values. In this simulation, we used two values of M, $M = 5$ and $M = 10$. The initial fractional weights are set to $w_{ij0}^* = M^{-1} d_{ij}$ and modified to satisfy (2.3). The second imputation method used is the multiple imputation (MI), where the imputed values are generated using the same linear regression model (4.1) with the method described in Schenker and Welsh (1988).

Four parameters are estimated. The parameters are

$$\begin{aligned}
\theta_1 &= \text{mean of } Y, \\
\theta_2 &= \text{mean of } Y \text{ where } Z < 0.25, \\
\theta_3 &= \text{proportion of } Y \leq 1.0 \text{ and} \\
\theta_4 &= \text{slope for the regression of } Y \text{ on } X.
\end{aligned}$$

For fractional imputation, the variance estimation method proposed in Section 3 was applied. The variance estimator for multiple imputation was adopted from Rubin (1987).

The mean and variance of the imputed estimator are calculated based on the Monte Carlo sample generated by the linear regression model in (4.1). Both imputation methods are unbiased for the four parameters and the simulation results of the point estimators are not listed here. For variance estimation, the relative bias and the t-statistics are calculated. The t-statistic is the statistic used to test the significance of the bias of the variance estimator.

Table 4.1 shows the Monte Carlo relative biases and the Monte Carlo t-statistics of the variance estimators. The variance estimator for fractional imputation has negligible relative biases except for the domain mean estimator. For the variance estimation of domain mean estimator, the relative bias for fractional

TABLE 4.1 *Monte Carlo relative biases and t-statistics of the variance estimators under the linear regression model, based on 5,000 samples*

Parameter	*Method*	*Relative bias (%)*	*t-statistic*
Mean	*FI* ($M = 5$)	0.2	0.11
	MI ($M = 5$)	4.4	2.20
	FI ($M = 10$)	0.9	0.45
	MI ($M = 10$)	4.6	2.25
Domain	*FI* ($M = 5$)	8.1	3.93
	MI ($M = 5$)	34.9	16.62
	FI ($M = 10$)	4.4	2.02
	MI ($M = 10$)	37.2	17.98
Proportion	*FI* ($M = 5$)	2.4	1.19
	MI ($M = 5$)	22.1	11.08
	FI ($M = 10$)	−1.4	−0.07
	MI ($M = 10$)	23.8	11.80
Slope	*FI* ($M = 5$)	4.8	2.00
	MI ($M = 5$)	1.1	0.53
	FI ($M = 10$)	4.3	2.00
	MI ($M = 10$)	1.8	0.84

imputation is much smaller than the multiple imputation and decreases as M increase, which was already discussed in Kim and Fuller (2004). A more surprising result is that multiple imputation does not perform well for variance estimation of the estimated proportion, even when the imputation model is correct. Note that multiple imputation shows negligible biases in the variance estimators for θ_1 and θ_4, which were explicitly included in the imputation model. Multiple imputation variance estimators are approximately unbiased only for the estimates of the parameters explicitly included in the imputation model. The variance estimators under fractional imputation are approximately unbiased for all parameter estimators in this setup.

In the second simulation study, the samples are generated from the following quadratic regression model

$$y_i = 2 + \sqrt{0.5}\left(x_i^2 - 1\right) + e_i, \tag{4.2}$$

where x_i and e_i are the same as in the first simulation. We also used the same z_i and R_i variables that were generated in the first simulation. The parameters and the imputation methods we consider are the same as in experiment one. Thus, in the second simulation, the imputation model still uses the simple linear regression model and the true model (4.2) is not in the class of the model used

TABLE 4.2 *Monte Carlo relative biases and t-statistics of the variance estimators under the quadratic regression model, based on 5,000 samples*

Parameter	*Method*	*Relative bias (%)*	*t-statistic*
Mean	*FI* ($M = 5$)	1.2	0.57
	MI ($M = 5$)	−0.26	−0.13
	FI ($M = 10$)	0.1	0.08
	MI ($M = 10$)	1.3	0.65
Domain	*FI* ($M = 5$)	9.2	4.39
	MI ($M = 5$)	61.4	27.45
	FI ($M = 10$)	5.83	2.75
	MI ($M = 10$)	66.4	30.15
Proportion	*FI* ($M = 5$)	2.8	1.40
	MI ($M = 5$)	27.32	13.28
	FI ($M = 10$)	1.83	0.90
	MI ($M = 10$)	28.71	14.08
Slope	*FI* ($M = 5$)	0.8	0.37
	MI ($M = 5$)	−37.0	−18.87
	FI ($M = 10$)	−1.0	−0.46
	MI ($M = 10$)	−37.1	−18.84

for imputation.

Table 4.2 shows the Monte Carlo relative biases and the Monte Carlo t-statistics of the variance estimators. Note that the imputed values for fractional imputation are selected from the respondent with M closest x-values. Thus, we expect that the fractional imputation can be quite robust against the failure of the imputation model since the conditional expectation of y given x is a smooth function of x. The fractional imputation variance estimator shows reasonably small relative biases for all parameters. Multiple imputation variance estimators shows big biases for all parameters except for θ_1. In particular, the multiple imputation variance estimator underestimates the variance of the imputed estimator of θ_4. A similar phenomenon was also identified by Robins and Wang (2000).

5. CONCLUDING REMARKS

Generally speaking, the imputed value of a missing unit j can be written

$$y_j^* = \widehat{y}_j + \widehat{e}_j^*, \tag{5.1}$$

where $\widehat{y}_j$ is the predicted value of y_j obtained from the imputation model and $\widehat{e}_j^*$ is selected from the set $\{\widehat{e}_i = y_i - \widehat{y}_i; i \in A_R\}$ of the residuals in the respondents.

If the model is such that $E(y_i \mid \mathbf{x}_i) = f(\mathbf{x}_i; \boldsymbol{\beta})$ for some $\boldsymbol{\beta}$ with known $f(\cdot)$, the imputation method using the predictive value $\widehat{y}_i = f(\mathbf{x}_i; \widehat{\boldsymbol{\beta}})$ with consistent $\widehat{\boldsymbol{\beta}}$ provides not only consistent estimate for the marginal mean of y but also consistent estimates for the covariance of y and $\mathbf{x}$.

To make the above imputation method as a hot deck imputation, one can achieve the goal by choosing the donor i for missing j with $\widehat{y}_i = \widehat{y}_j$. However, finding such donors is not always feasible. Fractional hot deck imputation in this case can be naturally implemented as follows:

1. Choose M donors with M closest $\widehat{y}$-values,
2. Compute the fractional weights that satisfy

$$\sum_{i \in A_R} w_{ij}^* (1, \widehat{y}_i) = (1, \widehat{y}_j), \qquad \text{for } j \in A_M.$$

Thus, the constraint (2.3) used in Section 2 is a special case of the above general imputation method where $f(\mathbf{x}_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_i$, which does not require the estimation of $\boldsymbol{\beta}$. If, instead, the logistic regression model $f(\mathbf{x}_i; \boldsymbol{\beta}) = [1 + \exp(-\mathbf{x}_i' \boldsymbol{\beta})]^{-1}$ is used as the imputation model, then $\widehat{\boldsymbol{\beta}}$ is directly used to implement the regression hot deck imputation. In this case, variance estimation would be more complicated. Further extension will not be discussed here and will be a topic of future research.

ACKNOWLEDGEMENTS

I thank two anonymous referees for helpful comments, which greatly improved the presentation of the paper.

REFERENCES

DEVILLE, J.-C. AND SÄRNDAL, C.-E. (1994). "Variance estimation for the regression imputed Horvitz-Thompson estimator", *Journal of Official Statistics*, **10**, 381–394.

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

KIM, J. K. AND FULLER, W. (2004). "Fractional hot deck imputation", *Biometrika*, **91**, 559–578.

ROBINS, J. M. AND WANG, N. (2000). "Inference for imputation estimators", *Biometrika*, **87**, 113–124.

RUBIN, D. B. (1976). "Inference and missing data", *Biometrika*, **63**, 581–592.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

SCHENKER, N. AND WELSH, A. H. (1988). "Asymptotic results for multiple imputation", *The Annals of Statistics*, **16,** 1550–1566.

SHAO, J. AND WANG, H. (2002). "Sample correlation coefficients based on survey data under regression imputation", *Journal of the American Statistical Association*, **97,** 544–552.