J. Statist. Res. Iran **10** (2013): 125–146

مجلهٔ پژوهش‌های آماری ایران
پژوهشکدهٔ آمار
دورهٔ ۱۰، شمارهٔ ۲، پاییز و زمستان ۱۳۹۲، صص ۱۲۵–۱۴۶

An Empirical Comparison of Performance of the Unified Approach to Linearization of Variance Estimation after Imputation with Some Other Methods

Shahrzad Khatibi Nouri* and Hamidreza Navvabpour

Allameh Tabataba'i University

Abstract. Imputation is one of the most common methods to reduce item non-response effects. Imputation results in a complete data set, and then it is possible to use naïve estimators. After using most of common imputation methods, mean and total (imputation estimators) are still unbiased. However their variances (imputation variances) are underestimated by naïve variance estimators. Sampling mechanism and response variable values are variation sources which have been hidden in naïve variance estimators. While missing mechanism and imputation processes are other sources which are created after imputation. The naïve estimator does not account for these new variation sources. In this paper, a recent method of unified approach to linearization imputation variance estimation is explained. In this method, imputation estimator is linearized with respect to nuisance parameters estimators. Then linear estimator is asymptotically equal to imputation estimator. Variance estimators are also asymptotically equal. The unified approach can cover all deterministic and stochastic imputation methods, except nearest neighbors method. By a simulation study, imputation variance estimators of multiple imputation, model-assisted, bootstrap and unified approach are compared when regression imputation has been implemented. Performance of the imputation variance estimators are compared with respect to relative efficiency

* Corresponding author

and coverage probability. Findings of the study show that unified approach and model-assisted are close in values of efficiencies and give more stable results through either increasing sample size or non-response rate.

Keywords. Non-response; multiple imputation; imputation variance; missing mechanism; quasi-likelihood; model-assisted method; bootstrap; linearization; reverse approach.

MSC 2010: 62F40.

1 Introduction

Imputation is one of the most common methods to reduce effects of item non-response. Imputation assigns a value as a substitute for a non-response. Imputed values are like predicted values for missing values. However, imputed values are definitely different form genuine values which could be observed. Statistical inferences validity depends on validity of imputation model assumptions and missing mechanism.

Variance as a measure of dispersion can be used to determine precision of estimators and comparing them. naïve variances estimators tend to underestimate when imputed are considered as observed values in a data set. These estimators are affected by sampling design and variable values. Nevertheless other sources of variation, which have been created after imputation, such as response mechanism (or missing mechanism) and imputation process are involved in the estimation. These new sources of variation are disregarded by naïve variance estimators, so they underestimate variance after imputation. Small variance makes narrow confidence limits and it is more likely to reject the null hypothesis.

Underestimating or overestimating imputation variance can cause difficulty in applications as well. If some imputation methods are compared by using a measure of dispersion (e.g. mean square error or variance of an imputation estimator) then underestimation can lead to incorrect method selection. In this case, MSE of the best method (which has less dispersion) may has been underestimated. Estimators such as mean and total which are used for complete data after imputation are called **imputation estimator** and variances of these estimators are called **imputation variance**.

Having been introduced multiple imputation and its variance by Rubin (1987), statisticians paid more attention to impact of imputation on vari-

ances and inferences. Särndal (1992) worked on variance estimators after hot-deck imputation by considering a model-assisted approach. Deville and Särndal (1994) applied model-assisted approach to estimate variance under regression imputation. Rao and Shao (1992) suggested adjusted jackknife to estimate variance after hot-deck imputation, because it was not possible to implement normal jackknife with missing values. Shao and Sitter (1996) studied estimation of imputation variance via bootstrap. Also, Shao et al. (1998) utilized another replication method, balanced half samples, to estimate imputation variance. Shao and Steel (1999) introduced a linearization procedure to estimate variance under deterministic and stochastic imputations. Brick et al. (2005) compared three methods of imputation variance estimations, model-assisted, adjusted jackknife and multiple imputation, under hot-deck imputation. Kim and Rao (2009) introduced a unified approach to linearization of imputation variance estimation by generalizing the method of Shao and Steel (1999).

In the second section, regression imputation which is used in the simulation study is reviewed. In Sections 3, 4 and 5 multiple imputation, model-assisted approach and bootstrap as methods of imputation variance estimation, are explained. In the sixth section, the unified approach to linearize imputation variance estimation is defined. All imputation variance estimations which have been reviewed through these six sections are compared in a simulation study in Section 7.

2 Missing and Response Mechanisms

There are three missing mechanisms according to reasons that non-responses occur; missing completely at random, missing at random and missing not at random (Lohr 2009).

In this paper, response mechanism is defined by response indicator variable in a selected sample. Let s_m and s_r show non-respondents and respondents sets in the selected sample, s, respectively. Suppose the response indicator is:

$$r_i = \begin{cases} 1 & \text{if} \quad i \in s_r \\ 0 & \text{if} \quad i \in s_m \end{cases} \qquad \text{for} \quad i = 1, ..., n \ .$$

There are two response mechanisms besides missing mechanisms which have been mentioned above. Concepts of missing and response mechanisms are close to each other.

Response mechanism is un-confounded while it depends on the sampling design of s or auxiliary information, $x_s = \{x_k : k \in s\}$ or both but it is independent of response variable, $y_s = \{y_k : k \in s\}$. Therefore un-confounded response mechanism is ignorable. If response probabilities depend on y_s then response mechanism is confounded and is not ignorable (Deville and Särndal 1994).

2.1 Regression Imputation

All imputation methods due to their degree of randomness in their processes can be classified to deterministic and stochastic imputation. In this paper, regression imputation as a deterministic method is used in the simulation. Some variables, which are measured for a unit, are usually correlated in surveys. As a result, missing values of each variable can be predicted by fitting regression models on different groups of variables.

Typically, imputation is applied to partial non-responses (item non-responses). Additionally, these methods are valid while missing mechanism is ignorable or response mechanism is un-confounded. Otherwise, these mechanisms should be modeled (Arnab and Singh, 2002).

Another point is imputation classes that can be constructed by variables which are convenient for categorizing sampling units in the data set. Imputation classes can increase precision of imputation process as long as levels of classification variables are not missing.

We consider $[\boldsymbol{x}_{1l}, \boldsymbol{x}_{2l}, \ldots, \boldsymbol{x}_{kl}]'$ as a vector of auxiliary information in the lth imputation class. The vector of response variable in the lth imputation class, y_l, is observed for the first n_{rl} units and the remaining $n_l - n_{rl}$ units are missing values. Firstly, based on n_{rl} respondents, a regression model for y_l is fitted on $[\boldsymbol{x}_{1l}, \boldsymbol{x}_{2l}, \ldots, \boldsymbol{x}_{kl}]'$. Next, non-responses are imputed by predicted values. Regression model and predicted values in the lth imputation class are as follows:

$$y_{il} = \beta_{0l} + \beta_{1l}x_{1il} + \cdots + \beta_{kl}x_{kil} + \varepsilon_{il}; \qquad i = 1, \ldots, n_{rl}$$

$$\hat{y}_{il} = \hat{\beta}_{0l} + \hat{\beta}_{1l}x_{1il} + \cdots + \hat{\beta}_{kl}x_{kil}; \qquad i = n_{rl} + 1, \ldots, n_l$$

ε_{il} is the error term which satisfies $E(\varepsilon_{il}) = 0$, $V(\varepsilon_{il}) = \sigma_i^2$ and $\text{cov}(\varepsilon_{il}, \varepsilon_{jl}) = 0$ for $i \neq j$, β_{il}'s are regression coefficients and $\hat{\beta}_{il}$'s are their estimations through least squares method in the lth imputation class.

3 Variance Estimation of Multiple Imputation

Multiple imputation is a general method that incorporates the uncertainty into the imputation process. Rubin (1987) recommended implementing imputation at least 5 times to obtain more precision in the estimations. In this view, the overall variance can be decomposed as the variation within imputed data sets and the variation between imputed data sets, i.e., $V_m = V_{\text{between } m \text{ sets}} + V_{\text{within } m \text{ sets}}$.

Multiple imputation is a model-based method and it is sensitive to the model determination. On top of that Bayes theory plays the main role in the imputation step. In this method, it is believed that naïve estimators are the best parameter estimators whereas we can manage imputation process to preserve the validity of estimators. It is assumed that response (non-response) and sampling mechanisms are ignorable while there are more discussions on non-ignorable mechanisms in Rubin (1987).

Multiple imputation is comprised of three stages: firstly, the missing data are imputed m times independently to generate m complete data sets (**imputation stage**). Secondly, standard techniques are applied to each complete data set (**analysis stage**). Thirdly, the results from m complete data sets are combined in order to make a final result that reflects the uncertainty due to imputation (**combined analysis**).

Rubin (1987) discussed regression imputation on the basis of Bayes theory. Multiple regression imputation consists of generating parameters values m times from the posterior distribution by a random process. Consequently missing values are imputed by m different models.

In Section 7, we use multiple regression imputation when the sampling design is stratified. For this design, combined estimators of parameter and its variance are available in Gerami et al. (2005).

Multiple imputation can prevent underestimating in the variance, by using a proper imputation method and a sufficient number of the imputation replication. The imputation method which produces enough variation in the variables values is a proper imputation (Rubin 1987).

4 Model-assisted Imputation Variance Estimation

Särndal (1992) introduced a model-assisted method to estimate the variance of survey estimates when imputation has been implemented. Thus imputation process implies the model and sample selection indicates the random-

ization distribution. He regarded the total error of the survey as the sum of sampling error and imputation error. Consequently, an overall variance is derived as the sum of a sampling variance and imputation variance. Deville and Särndal (1994) emphasize that response mechanism should be un-confounded in this method.

In order to estimate population total of y, $\theta = \sum_{i=1}^{N} y_i$, let $s - s_r$ be the subset of s in which y_i's are imputed. Data after regression imputation are

$$y_{\bullet i} = \begin{cases} y_i, & \text{if} \quad i \in s_r \\ y_{imp,i} = \boldsymbol{x}'_i\boldsymbol{\beta}, & \text{if} \quad i \in s - s_r \end{cases}$$

where the multiple regression model is $y_i = \boldsymbol{x}'_i\boldsymbol{\beta}+\varepsilon_i$ and $V_\xi(\varepsilon_i) = E_\xi(\varepsilon_i^2) = \sigma^2$ (ξ indicates imputation model). The auxiliary variable, $\boldsymbol{x}'_i$, is a p-dimensional vector without missing values and the weighted least squares estimator of the regression coefficient in the self-weighting sampling design ($w_i = w$) is shown as $\hat{\boldsymbol{\beta}} = \left(\sum_{i\in s_r} \boldsymbol{x}_i\boldsymbol{x}'_i\right)^{-1} \sum_{i\in s_r} \boldsymbol{x}_i y_i$.

In the case of complete response, total estimator is $\hat{\theta} = \sum_{i\in s} w_i y_i$. In the presence of non-response and imputation, the imputation estimator of total is $\hat{\theta}_I = \sum_{i\in s} w_i y_{\bullet i} = \sum_{k\in s_r} w_i y_i + \sum_{k\in s-s_r} w_i \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}$. Särndal (1992) assumed that weights w_i are identical to those used when all data are actually observed. It represents an assumption that imputation by the chosen rule causes little or no systematic error in the estimates. The total error of $\hat{\theta}_I$ is decomposed as

$$\hat{\theta}_I - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_I - \hat{\theta}) = \text{sampling error} + \text{imputation error}$$

The imputation error is $\hat{\theta}_I - \hat{\theta} = \sum_{i\in s-s_r} w_i(y_{imp,i} - y_i)$ where $e_i = y_i - y_{imp,i}$ is called imputation residual and cannot be observed for $i \in (s - s_r)$.

Särndal (1992) uses three different probability distributions which their expectations are considered as E_ξ, E_s and E_r. Correspondingly, ξ indicates "with respect to imputation model", s indicates "with respect to the sampling design" and r indicates "with respect to the response mechanism given s" which is usually unknown. The estimator $\hat{\theta}_I$ is overall unbiased in the sense of $E_\xi E_s E_r(\hat{\theta}_I - \theta) = 0$ if the following two conditions hold:

1. The $E_\xi E_s E_r(\cdot)$ can be changed, and

2. Expected value of imputation residuals under imputation model are zero (Särndal 1992).

Finally, overall variance is given by

$$V_{tot} = E_\xi E_s E_r \left\{ (\hat{\theta}_I - \theta)^2 \right\} = E_\xi V_p + E_s E_r V_{\xi c} + 2E_\xi E_s E_r \left\{ (\hat{\theta} - \theta)(\hat{\theta}_I - \hat{\theta}) \right\} \quad (1)$$

where $V_p = E_s \left\{ (\hat{\theta} - \theta)^2 \right\}$ is the design-based variance of $\hat{\theta}$ provided that $\hat{\theta}$ is design unbiased. In addition, $V_{\xi c} = E_\xi \left\{ (\hat{\theta}_I - \hat{\theta})^2 | s, r \right\}$ is the model variance of the imputation error conditioned on the sampling design and response mechanism. Deville and Särndal (1994) stated that the third term in (1) is exactly zero when the sampling design is self-weighting. In brief, $V_{tot} = V_{sam} + V_{imp} + V_{mix}$ where $V_{sam} = E_\xi V_p$ is sampling variance, $V_{imp} = E_s E_r V_{\xi c}$ is imputation variance and $V_{mix} = 2E_\xi E_s E_r \left\{ (\hat{\theta} - \theta)(\hat{\theta}_I - \hat{\theta}) \right\}$ is the mixed term.

If a design unbiased estimator of V_p is $\hat{V}_p$ then the variance estimator after imputation is $\hat{V}_{\bullet p}$ which is subject to underestimate V_{sam}, especially in the presence of high rate of non-response. Therefore, a term must be added to $\hat{V}_{\bullet p}$, so $\hat{V}_{sam} = \hat{V}_{\bullet p} + \hat{V}_{diff}$ where $V_{diff} = E_\xi(\hat{V}_p - \hat{V}_{\bullet p} | s, r)$. Now the objective is variance components estimation, $\hat{V}_{sam}$, $\hat{V}_{imp}$ and $\hat{V}_{mix}$. In this paper, we consider multiple regression imputation and we follow variance components estimation by the method of Deville and Särndal (1994) who revised model-assisted approach in a more general manner. Moreover, their approach in the estimation can be extended to the stratified sampling with a self-weighting design in each stratum.

5 Bootstrap Imputation Variance Estimation

Shao and Sitter (1996) reviewed bootstrap method in the presence of missing data. The naïve bootstrap underestimates variance after imputation because it takes into account the imputed data as they were originally observed. Thus it does not capture the inflation in variance due to non-response. Shao and Sitter (1996) recommended that the bootstrap data set should also be imputed in the same way as the parent data set was imputed.

The exact bootstrap algorithm for multistage sampling is available on Shao and Sitter (1996), which have been used to stratified sampling extent

with regression imputation here. In the case of simple stratified sampling without replacement, if the sampling fraction is ignorable then the mentioned procedure suits the situation. For $b = 1, \ldots, B$ bootstrap samples, bootstrap variance of $\hat{\theta}_I$ is: $v_B(\hat{\theta}_I) \approx \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}_I^{*b} - \bar{\theta}_I^*)^2$ where $\bar{\theta}_I^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_I^{*b}$ and $\hat{\theta}_I^{*b}$ is the imputation estimator of total in the bth bootstrap sample.

In the presence of imputation classes, process of imputing is restricted to each class. Then the imputation estimator is calculated by summation over classes. The bootstrap variance estimator is consistent as long as the number of imputation classes is fixed (Shao and Sitter, 1996).

6 A Unified Approach to Linearization Imputation Variance Estimation

Firstly, Shao and Steel (1999) introduced a method of linearization imputation variance. However, covariances in their method were complicated and could not be calculated straightforwardly. Kim and Rao (2009) presented a unified approach which can be reviewed as an extension of Shao and Steel's method. Kim and Rao (2009) used their sample-response path and reverse approach definitions.

A model-assisted approach (Särndal, 1992) is used in the unified approach to linearize imputation variance estimation. A population model which clarifies imputation model and distribution of y given x is assumed. Let missing mechanism is at random without specifying the distribution of r_i, response indicator. Thus distribution of r_i depends only on $\boldsymbol{x}_i$. The unified approach is a general method which can cover most deterministic and stochastic imputations apart from nearest neighbors.

Under deterministic imputation such as regression imputation, a super population model of y_i given $\boldsymbol{x}_i$ for some p-dimensional vector $\boldsymbol{\beta}_0$ is assumed $E_\xi(y_i|\boldsymbol{x}_i) = m(\boldsymbol{x}_i; \boldsymbol{\beta}_0) = \boldsymbol{x}_i'\boldsymbol{\beta}_0$ where $m(\boldsymbol{x}_i; \boldsymbol{\beta})$ is a known function of $\boldsymbol{x}_i$ for a given $\boldsymbol{\beta}$. We use $\hat{y}_i = m(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ as imputed values for missing y_i where $\hat{\boldsymbol{\beta}}$ is the solution of following estimating equations (quasilikelihood equations)

$$\hat{U}(\boldsymbol{\beta}) \equiv N^{-1} \sum_{i \in s} w_i r_i \{y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})\} h(\boldsymbol{x}_i; \boldsymbol{\beta}) = 0.$$

If variance function is specified as $\mathrm{var}_\xi(y_i|\boldsymbol{x}_i) = \sigma^2 q(\boldsymbol{x}_i, \boldsymbol{\beta}_0)$ for a known function $q(\cdot)$, then $h(\boldsymbol{x}_i; \boldsymbol{\beta}) = \frac{\dot{m}(\boldsymbol{x}_i, \boldsymbol{\beta})}{q(\boldsymbol{x}_i, \boldsymbol{\beta})} \equiv h_i$ where $\dot{m}(\boldsymbol{x}_i, \boldsymbol{\beta}) = \partial m(\boldsymbol{x}_i, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$

(McCullagh and Nelder, 1989). In linear regression model $q(\boldsymbol{x}_i, \boldsymbol{\beta}) = 1$ and $h(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{h}_i = \boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i(1-p)})$.

Imputation estimator of total is $\hat{\theta}_{Id} = \sum_{i \in s} w_i \left\{ r_i y_i + (1 - r_i) m(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) \right\}$. Kim and Rao (2009) proved that $\hat{\theta}_{Id}$ satisfies $n^{1/2} N^{-1} (\hat{\theta}_{Id} - \tilde{\theta}_{Id}) = o_p(1)$ where

$$\tilde{\theta}_{Id} = \sum_{i \in s} w_i \left[m(\boldsymbol{x}_i; \boldsymbol{\beta}_0) + r_i \left\{ 1 + \boldsymbol{c}' h(\boldsymbol{x}_i; \boldsymbol{\beta}_0) \right\} \left\{ y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta}_0) \right\} \right] \equiv \sum_{i \in s} w_i \eta_i$$

and

$$\boldsymbol{c} = \left\{ \sum_{i=1}^{N} r_i \dot{m}(\boldsymbol{x}_i; \boldsymbol{\beta}_0) \boldsymbol{h}'_{io} \right\}^{-1} \sum_{i=1}^{N} (1 - r_i) \dot{m}(\boldsymbol{x}_i; \boldsymbol{\beta}_0)$$

where $\boldsymbol{h}_{io} = h(\boldsymbol{x}_i; \boldsymbol{\beta}_0)$.

They states that $\hat{\theta}_{Id}$ is asymptotically equivalent to $\tilde{\theta}_{Id}$ and consequently $\text{var}(\hat{\theta}_{Id}) \simeq \text{var}(\tilde{\theta}_{Id})$. As a result variance of $\hat{\theta}_{Id}$ can be estimated by its asymptotic equivalence.

Owing to reverse approach, decomposition of $\text{var}(\tilde{\theta}_{Id})$ is as follows:

$$\text{var}(\tilde{\theta}_{Id}) = E_\gamma \left[V_s(\tilde{\theta}_{Id}) \right] + V_\gamma \left[E_s(\tilde{\theta}_{Id}) \right].$$

Due to imputation model and response mechanism which are contained in γ, we can write $V_\gamma \left[E_s(\tilde{\theta}_{Id}) \right] = E_r \left[V_\xi(E_s(\tilde{\theta}_{Id})) \right] + V_r \left[E_\xi(E_s(\tilde{\theta}_{Id})) \right]$. Afterwards $\text{var}(\tilde{\theta}_{Id})$ can be rewritten as $\text{var}(\tilde{\theta}_{Id}) \equiv V_1 + V_2 + V_3$. Reference distribution in the above variances or expectations, which are denoted by s, ξ and r, operate respectively over the sampling design, the imputation model (super population model) and the unknown response mechanism. Kim and Rao (2009) have shown that $V_3 = 0$ then $\text{var}(\tilde{\theta}_{Id}) \equiv V_1 + V_2$.

The first term can be estimated by applying the standard variance estimator to the pseudo- values η_i. In practice $\boldsymbol{\beta}_0$ is replaced by $\hat{\boldsymbol{\beta}}$ then

$$\hat{V}_1 = \sum_{i \in s} \sum_{j \in s} \Omega_{ij} \hat{\eta}_i \hat{\eta}_j \tag{2}$$

where Ω_{ij} is the joint inclusion probability of i and j being in the sample, $\hat{\eta}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + r_i (1 + \hat{\boldsymbol{c}}' \boldsymbol{x}_i)(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}})$ are pseudo-values in regression imputation, and

$$\hat{\boldsymbol{c}} = \left\{ \sum_{i \in s} w_i r_i \boldsymbol{x}_i \boldsymbol{x}_i' \right\}^{-1} \left(\sum_{i \in s} w_i (1 - r_i) \boldsymbol{x}_i \right)$$

In order to estimate the second term V_2, It was assumed that $\text{var}_\xi(y_i|\boldsymbol{x_i}) = \sigma^2 q(\boldsymbol{x_i}, \boldsymbol{\beta}_0)$. So

$$V_2 = E_r \left\{ \text{var}_\xi \left(\sum_{i=1}^{N} \eta_i | R_N \right) \right\} = \sigma^2 E_r \left\{ \sum_{i=1}^{N} r_i (1 + c' \boldsymbol{h}_i)^2 q(\boldsymbol{x}_i, \boldsymbol{\beta}_0) \right\}$$

where $\boldsymbol{h}_i = h(\boldsymbol{x}_i, \boldsymbol{\beta})$ that estimated by $\hat{\boldsymbol{h}}_i = h(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i$. It seems that Kim and Rao (2009) did not consider $\text{var}_\xi \left[m(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}) \right]$ and $\text{cov}_\xi \left[m(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}), y_i \right]$ in the latter statement. However, under regression imputation, by noting that a robust estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i \in s} \{y_i - \hat{y}_i\}^2$ and considering the neglected variance and covariance, a consistent estimator of V_2 is given by

$$\begin{aligned} \hat{V}_2 = \sum_{i \in s} w_i r_i \Bigg[& \left\{ 1 - r_i (1 + \hat{\boldsymbol{c}}' \boldsymbol{x}_i) \right\}^2 \left\{ \boldsymbol{x}'_i \left(\sum_{i \in s} r_i \boldsymbol{x}_i \boldsymbol{x}'_i \right)^{-1} \boldsymbol{x}_i + 1 \right\} \\ & + r_i (1 + \hat{\boldsymbol{c}}' \boldsymbol{x}_i)^2 \Bigg] \frac{(y_i - \hat{y}_i)^2}{(n-p)} \\ & + 2 \sum_{i \in s} w_i r_i \Bigg[\left\{ 1 - r_i (1 + \hat{\boldsymbol{c}}' \boldsymbol{x}_i) \right\} r_i (1 + \hat{\boldsymbol{c}}' \boldsymbol{x}_i) \boldsymbol{x}'_i \\ & \times \left(\sum_{i \in s} r_i \boldsymbol{x}_i \boldsymbol{x}'_i \right)^{-1} \boldsymbol{x}_i \Bigg] \frac{(y_i - \hat{y}_i)^2}{(n-p)} \end{aligned} \tag{3}$$

In fact $\hat{y}_i$ does not exist for the non-respondent unit, as a result r_i comes next to w_i to compensate this matter.

Ultimately, variance of $\hat{\theta}_{Id}$ can be estimated by $\hat{V} = \hat{V}_1 + \hat{V}_2$. Actually, the strong point of applying the reverse approach is entering pseudo-values instead of y_i into $\hat{V}_n$ and making $\hat{V}_1$ accordingly.

7 Simulation Study

In this section, four imputation variances, which have been reviewed through recent sections, are estimated and compared in a simulation study. These estimators are compared by using *relative efficiency* and *coverage probability*. We used 2009 Iran urban household income and expenditure survey data in the simulation study. We used SAS software to conduct the simulation.

7.1 Household Income and Expenditure Survey

Overall objective of the household income and expenditure survey is to estimate the annual average of the income and expenditure of urban and rural households for the whole country as well as the provinces (Statistical Centre of Iran, 1389). We use the data on urban households as the target population in the simulation study.

Generally, the household information consists of five main parts: characteristics of the head of the household, characteristics of the residential location, the household characteristics and the information on the household income and expenditure. In this study, four variables are used: *gross expenditure*, *non-food expenditure*, *income* and *strata* which has three levels. The variables definitions are available in the publication of Statistical Centre of Iran (2010).

The households which were unit non-respondents through the four desired variables have been omitted. Observations with the negative incomes have been removed as well. Consequently, the data set which is used in the study has 18538 urban households with the four desired variables. In order to approximate the sampling distribution of the imputation variance estimators, Monte Carlo method is applied to the mentioned data set which has been considered as the target population in the study.

7.2 Design of the Simulation Study

To study the effect of the sample size on the parameter estimates, three sample sizes 100, 500, and 1000 have been considered. The sampling design is stratified random sampling. The allocation method is proportional allocation. Table 1 shows strata sample sizes for three values of n.

Table 1. Sample sizes in each stratum for $n = 100$, $n = 500$, and $n = 1000$

Stratum	N_h	n_h	n_h	n_h
11	6599	36	178	356
12	6121	33	165	330
21	5818	31	157	314
Total	$N = 18538$	$n = 100$	$n = 500$	$n = 1000$

Where 11, 12 and 21 are strata indicators which were specified through regions. Since sampling fractions are small, we ignore the finite population corrections.

Non-response rate as well as the sample size affects survey estimates. Three values of non-response rate of 10%, 20% and 30% are considered in each sample size. As a result, there are nine comparisons among estimators.

Suppose r_i is response indicator and gets values 1 for the respondent and 0 for the non-respondent. The response indicator is generated from a Bernoulli distribution in each sample size and non-response rate category. In this study, it is assumed that the strata and the imputation classes are the same.

We assumed that the response mechanism is un-confounded and it is constant in every imputation class (stratum). The variables of interest are y_i= gross expenditure of the ith household, x_i= ith household income and z_i= non-food expenditure of the ith household. Item non-response is only generated on the y_i which will be imputed by the regression imputation. The regression imputation model is:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad E(\varepsilon_i, \varepsilon_j) = 0\ (i \neq j), \quad V(\varepsilon_i) = \sigma^2$$

The simulation study consists of four steps. **First;** a stratified sample of size n is selected from $N = 18538$ households. There are three strata which are the imputation classes as well. The sampling method in every stratum is simple random sampling. This sample is called a simulation sample. **Second;** after generating r_i, the non-respondents of y_{hi} are flagged in the simulation sample. Therefore non-responses are imputed by $\hat{y}_{hi} = \hat{\beta}_{0h} + x_{hi}\hat{\beta}_{1h} + z_{hi}\hat{\beta}_{2h}$ in the hth class of the imputation for $h = 1, 2, 3, \ldots$ and for $i = 1, 2, 3, \ldots$. Eventually, the population total can be estimated by the imputation estimators. For instance, the total is estimated in the multiple imputation method with $m = 5$ by $\hat{\theta}_{MI} = \frac{1}{5}\sum_{j=1}^{5}\hat{\theta}_{Istj}$ when $\hat{\theta}_{Istj} = \sum_{h=1}^{3}\sum_{i=1}^{n_h} r_{hi}w_h y_{hi} + (1 - r_{hi})w_h\hat{y}_{hi}$ is the total estimator in the stratified random sampling, where $w_h = \frac{N_h}{n_h}$ is sampling weight. The imputation estimator of the total by 1000 bootstrap samples is:

$$\hat{\theta}_{BOOT} = \frac{1}{1000}\sum_{b=1}^{1000}\hat{\theta}^b_{Ist}$$

when $\hat{\theta}_{Ist}^{b}$ is the estimates of total in the bth bootstrap sample based on the stratified sampling. The imputation estimator of the total for the model-assisted and the unified approach is $\hat{\theta}_{SAR} = \hat{\theta}_{KIM} = \hat{\theta}_{Ist} = \sum_{h=1}^{3}\sum_{i=1}^{n_h} r_{hi} w_h y_{hi} + (1 - r_{hi}) w_h \hat{y}_{hi}$. All mentioned imputation estimators are unbiased.

The imputation variance which can be estimated by the multiple imputation method was reviewed in Section 3. Here, we use symbol MI to indicate this method. The imputation variance estimator with $m = 5$ times imputing process is

$$\hat{V}_{MI}(\hat{\theta}_{Ist}) = \frac{1}{5}\sum_{j=1}^{5} \text{var}(\hat{\theta}_{Istj}) + \left\{ \frac{6}{5} \times \frac{1}{4} \sum_{j=1}^{5} (\hat{\theta}_{Istj} - \hat{\theta}_{MI})^2 \right\}$$

where

$$\text{var}(\hat{\theta}_{Istj}) = \sum_{h=1}^{3} N^2 \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h} = \sum_{h=1}^{3} N_h^2 \frac{s_h^2}{n_h} = \sum_{h=1}^{3} \text{var}(\hat{\theta}_{Ih})$$

The model-assisted method is denoted by SAR and its estimator based on section 4 is:

$$\hat{V}_{SAR}(\hat{\theta}_{Ist}) = \sum_{h=1}^{3} \hat{V}_{\bullet ph} + \hat{V}_{difh} + \hat{V}_{imph} = \sum_{h=1}^{3} \hat{V}_{samh} + \hat{V}_{imph} = \sum_{h=1}^{3} \text{v}_{SAR}(\hat{\theta}_{Ih})$$

The variance components are estimated by the method of Deville and Särndal (1994). The unified approach variance estimator which has been reviewed in section 6 is notated by KIM.

$$\hat{V}_{KIM}(\hat{\theta}_{Ist}) = \sum_{h=1}^{3} \hat{V}_{1h} + \hat{V}_{2h} = \sum_{h=1}^{3} \text{v}_{KIM}(\hat{\theta}_{Ih})$$

In this expression, $\hat{V}_{1h}$ and $\hat{V}_{2h}$ are estimated by (2) and (3) for every stratum. We use the bootstrap algorithm which has been introduced in Section 5 to estimate the imputation variance. naïve estimators depend on the stratified sampling in each bootstrap sample, thus

$$\hat{V}_{BOOT}(\hat{\theta}_{Ist}) = \frac{1}{1000}\sum_{b=1}^{1000} (\hat{\theta}_{Ist}^{b} - \hat{\theta}_{BOOT})^2$$

Third; the first step is repeated 1000 times thus 1000 independent samples are generated from the population. The second step is applied to each simulation sample. For example, in the MI method the estimated parameters in the kth simulated sample are $\hat{\theta}^k_{MI}$ and $\hat{V}^k_{MI}$. Afterwards by using Monte Carlo method, parameters of the sampling distribution of the variance estimator in the MI method are approximated as follows:

$$\hat{\theta}_{MI} = \frac{1}{1000}\sum_{k=1}^{1000}\hat{\theta}^k_{MI},\ \hat{\mathrm{V}}_{MI} = \frac{1}{1000}\sum_{k=1}^{1000}\hat{V}^k_{MI},$$

Monte Carlo approximation was also applied to SAR and KIM methods. Results are presented in the end of this section.

Forth; a 95% confidence interval for the total of gross expenditure of the urban households is calculated in each sample. For instance, 95% confidence interval for the total in the model-assisted method is $\hat{\theta}^k_{SAR} \pm 1.96\sqrt{\hat{V}^k_{SAR}}$ for $k = 1, \ldots, 1000$. The same is conducted for other methods. The pivotal statistic which constructs the confidence limits in the MI method has Student's t distribution with ν degree of freedom. However, in our study, due to the large value of ν, $t_{(\nu,\frac{\alpha}{2})}$ is close to $z_{\frac{\alpha}{2}}$. So, we use 1.96 in the MI's confidence limits as well as the others.

The imputer model can use most of the auxiliary variables and the best model selection would not be done in the imputation stage. We leave it for the analysis stage (analyst model).

7.3 Comparisons Criteria

We know that before generating r_i, the sample data set is complete. Therefore, an unbiased naïve estimator, $\hat{\theta}^k_{st} = \sum_{h=1}^{3}\sum_{i=1}^{n_h} w_h y_{hi}$, can be used for the total gross expenditure of the households in the kth Monte Carlo sample. In the case of completed data set (before the r_i impresses the sample), the naïve estimator is denoted by 100% Response. The variance estimator of $\hat{\theta}^k_{st}$ is estimated by

$$\hat{V}^k_{100\%\ \text{Response}} = N^2\sum_{h=1}^{3}\sum_{i=1}^{n_h}\left(\frac{N_h}{N}\right)^2\frac{s^2_{kh}}{n_h} \tag{4}$$

and $\hat{V}_{100\%\ \text{Response}} = \frac{1}{1000}\sum_{k=1}^{1000}\hat{V}^k_{100\%\ \text{Response}}$.

As we discussed in the beginning, naïve estimators such as (4) tend to underestimate the variance in the presence of imputation. We estimated the variance after the regression imputation by the naïve estimator and denoted by STD in the tables. In other words, the estimator (4) is applied to the imputed data in each replication and $\hat{V}_{STD} = \frac{1}{1000}\sum_{k=1}^{1000} \hat{V}_{STD}^{k}$ is obtained.

The total gross expenditure of urban households in 2010 was $\theta = 16336.491$ (100×10^6 Rials) which is required for computing the biases and coverage probabilities. Consequently, it is possible to report the number of confidence intervals which cover the true parameter. If c denotes the number, then coverage probability is defined as

$$\widehat{\text{coverage}}(\theta) = \frac{c}{1000} \times 100.$$

Here, it is desirable to get the $\widehat{\text{coverage}}(\theta)$ close to 95%, because we considered $\alpha = 0.05$ in each replication.

The other measure which is used in the study is relative efficiency. Whether the imputation estimator is unbiased or biased, $\hat{V}$ or $\widehat{MSE}$ is computed, respectively. For instance, for the unified approach we have $\widehat{MSE}(\hat{\theta}_{Ist}) = \hat{V}_{KIM}$.

The efficiency of $\widehat{MSE}(\hat{\theta}_{st}) = \hat{V}_{100\%\ \text{Response}}$ related to $\widehat{MSE}$ is computed by $\frac{\widehat{MSE}}{\hat{V}_{100\%\ \text{Response}}}$. For instance, Monte Carlo estimation of efficiency of 100% Response relative to a specific estimator ($esp.$) is: $\widehat{reff}(100\%\ \text{Response}, \text{esp.}) = \frac{\widehat{MSE}_{esp.}}{\hat{V}_{100\%\ \text{Response}}}$. We used $\hat{V}_{100\%\ \text{Response}}$ to fix the value of the denominator of $\widehat{reff}$'s. Root mean squared error estimate ($\widehat{rootMSE}$) is provided in tables in order to ease the comparisons and for making a unit less measure of comparisons we also estimated relative MSE denoted by $\widehat{RMSE} = \frac{\widehat{rootMSE}}{\hat{\theta}_{(esp.)}}$.

7.4 Findings

There are nine tables to show the findings of the simulation study as follows (data has been rescaled by 100×10^6 Rials):

Table 2. Sample size of $n = 100$ with 10% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	1660	0.1017	1.03799	93.7%
MI	1670	0.1023	1.05111	94.1%
SAR	1663	0.1019	1.04243	93.8%
BOOT	1615	0.1020	0.98314	88.8%
KIM	1664	0.1019	1.04333	93.8%
100% Response	1629	0.0998	1	—

Table 3. Sample size of $n = 100$ with 20% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	1663	0.1018	1.04212	93.6%
MI	1672	0.1024	1.05367	93.5%
SAR	1671	0.1023	1.05153	93.9%
BOOT	1638	0.1033	1.01051	89.1%
KIM	1671	0.1023	1.05227	93.8%
100% Response	1629	0.0998	1	—

Table 4. Sample size of $n = 100$ with 30% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	1660	0.0321	1.03882	93.5%
MI	1673	0.0324	1.0541	93.9%
SAR	1673	0.0324	1.05432	93.6%
BOOT	1649	0.0329	1.02452	89.2%
KIM	1674	0.0324	1.05533	93.7%
100% Response	1629	0.0316	1	—

Table 5. Sample size of $n = 500$ with 10% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	696	0.0426	0.99924	94.6%
MI	698	0.0427	1.00463	94.8%
SAR	698	0.0427	1.00462	95%
BOOT	692	0.0427	0.98904	93.3%
KIM	698	0.0427	1.00482	95%
100% Response	696	0.0426	1	—

Table 6. Sample size of $n = 500$ with 20% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	696	0.0426	1.00006	94.6%
MI	700	0.0429	1.01293	94.9%
SAR	700	0.0429	1.01127	94.8%
BOOT	695	0.0428	0.99599	92.9%
KIM	700	0.0429	1.01153	94.8%
100% Response	696	0.0426	1	—

Table 7. Sample size of $n = 500$ with 30% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	702	0.0430	1.01743	94.8%
MI	706	0.0433	1.03029	94.5%
SAR	708	0.0434	1.03522	95%
BOOT	703	0.0433	1.0201	92.9%
KIM	708	0.0434	1.03605	95%
100% Response	696	0.0426	1	—

Table 8. Sample size of $n = 1000$ with 10% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	497	0.0304	1.0014	93.7%
MI	498	0.0305	1.00654	94%
SAR	498	0.0305	1.00677	94%
BOOT	497	0.0305	1.00263	92.7%
KIM	498	0.0305	1.03605	94.1%
100% Response	497	0.0304	1	—

Table 9. Sample size of $n = 1000$ with 20% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	498	0.0305	1.00634	94.1%
MI	501	0.0306	1.01632	94.6%
SAR	501	0.0307	1.01778	94.4%
BOOT	500	0.0307	1.0138	93.1%
KIM	501	0.0307	1.01787	94.4%
100% Response	497	0.0304	1	—

Table 10. Sample size of $n = 1000$ with 30% non-response rate

Method	$\widehat{rootMSE}$	$\widehat{RMSE}$	$\widehat{reff}$	$\widehat{\text{coverage}}(\theta)$
STD	500	0.0306	1.01364	94.5%
MI	505	0.0309	1.03189	94.7%
SAR	505	0.0309	1.03214	94.9%
BOOT	504	0.0309	1.02852	93.5%
KIM	505	0.0309	1.03231	94.9%
100% Response	497	0.0304	1	—

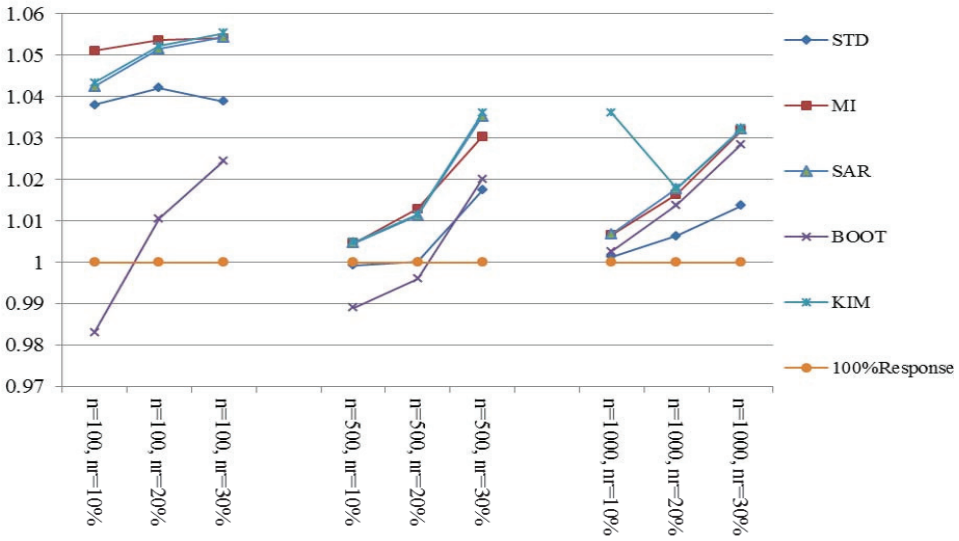

Figure 1. Relative efficiencies of four methods

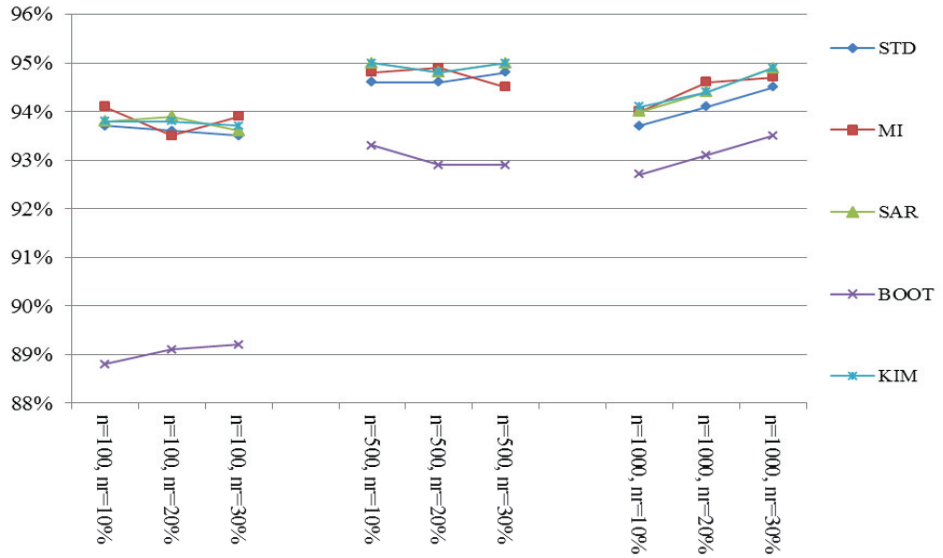

Figure 2. Coverage probabilities of four methods

In the tables, most of the relative efficiencies are greater than 1 which shows that the variance estimation methods produced a larger $\widehat{MSE}$ or $\hat{V}$ than the 100% Response case. In spite of that the differences from 1 are not substantial, the small inflation shows the penalty which is paid in regard to the imputation. There are a few cases that the efficiencies are smaller than 1.

The relative efficiency decreases when the sample size increases. For the sample size of 1000, and 10% non-response rate, values of efficiencies are close to 1. However, by increasing the non-response probability, values of efficiencies become a little bit greater than 1. It means that in the small non-response rates, the imputation variance estimator performs as good as the 100% Response estimator.

For sample size 100 and three non-response rates and for sample size 500 and non-response rates 10% and 20%, the Boot estimator has the least $\widehat{rootMSE}$, so it is more efficient than the other estimators. For sample size 500 and non-response rate 30%, STD has the least $\widehat{rootMSE}$. By increasing sample size from 500 and for all non-response rates, STD estimator has the least $\widehat{rootMSE}$. Generally, the STD has small value of relative efficiency and it is close to 1 which means, it is similar to the 100% Response estimator.

Although efficiency of STD relative to 100% Response is near 1 but it does not guarantee to estimate imputation variance accurately. Particularly, we do know that STD underestimates imputation variance. In addition, we cannot determine r_i in the population to estimate the true value of variance in order to see the differences. Regardless of STD method, the BOOT has the least efficiency comparing to the three remaining methods.

In all sample sizes and non-response rates Boot has the least percentage of covering the parameter θ (total gross urban household expenditures). That happened due to the narrow confidence intervals which were derived from the underestimated variances. The BOOT in some sample sizes works worse than STD.

By noticing the values of coverage probabilities, it can be said that MI, SAR and KIM perform similarly. The coverage probability is quietly the same for SAR and KIM within 9 comparisons. However, the MI is not stable. We would expect more stability in the MI if we used more than 5 imputations. The SAR and KIM have 95% coverage probability for the sample size 500 with respect to 20% and 30% of non-response rates.

8 Conclusion

Regarding values of either relative efficiencies or $\widehat{rootMSE}$'s and the coverage probability, the MI provides acceptable results in the small sample sizes, while all other methods perform well in the larger sample sizes. The KIM and SAR are more stable in the value of the coverage probability whereas the MI is not. The MI method usually needs a large computer memory to perform calculations. Moreover, it will end up with reliable results when imputations are replicated more than 10 (more repetitions tend to gain more precision). On the contrary, KIM and SAR use single imputation and show more stable result through either increasing in sample size or non-response rate.

The simulation study show KIM and SAR methods provide similar results in biases, relative efficiencies, and coverage probabilities, so we recommend them to compensate for variance underestimation after imputation. KIM and SAR methods are straightforward and show stability in their performance. Note that without considering covariance term in (3), KIM estimator could not perform as well as SAR estimator.

References

Arnab, R. and Singh, S. (2002). Estimation of Variance from Missing Data. *SSC Annual meeting*.

Brick, J.M., Jones, M.E., Kalton, G. and Valliant, R. (2005). A Simulation Study of Three Methods of Variance Estimation with Hot Deck Imputation. *Survey methodology*, **31**, 151−159.

Deville, J.C. and Särndal, C.E. (1994). Variance Estimation for Regression Imputed Horvitz-Thompson Estimator. *J. Offic. Statist.*, **10**, 381−394.

Gerami, A., Ganjali, M., Harandi, F., Ranji, H., Hamidi, O. and Tapak, L. (2005). *The Imputation of Missing Data and Its Effects on the Estimates and Analysis*. Statistical research and training center of Iran, Tehran (in Persian).

Kim, J.K. and Rao, J.N.K. (2009). A Unified Approach to Linearization Variance Estimation from Survey Data After Imputation for Item Nonresponse. *Biometrika*, **96**, 917−932.

Lohr, S.L. (2009). *Sampling: Design and Analysis*. Duxbury Press.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London.

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, **79**, 811−822.

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

Särndal, C.E. (1992). Methods for Estimating the Precision When Imputation has been Used. *Survey Methodology*, **18**, 241−252.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced Repeated Replication for Stratified Multistage Survey Data Under Imputation. *JASA*, **93**, 831−819.

Shao, J. and Sitter, R.R. (1996). Bootstrap for Imputed Data. *JASA*, **91**, 1278−1288.

Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions. *JASA*, **94**, 254−265.

Statistical Center of Iran (1389). *The Result of the Urban Household Income and Expenditure Survey in 2009*. Statistical Center of Iran, Tehran, Iran (in Persian).

Shahrzad Khatibi Nouri
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *khtbnr.sh@gmail.com*

Hamidreza Navvabpour
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *h.navvabpour@srtc.ac.ir*