SSC Annual Meeting, June 2011
Proceedings of the Survey Methods Section

VARIANCE ESTIMATION UNDER MULTIPLE CYCLE IMPUTATION

W. Qian[1] and A. Demnati[2]

ABSTRACT

Imputation is a common method for handling item nonresponse in surveys. Given a set of auxiliary variables, we may perform imputation for multiple cycles to impute all the missing values. In each cycle, a different combination of imputation method and auxiliary variables is used, and also the preceding imputed values are treated as observed ones. In this study, we apply a unified linearization approach to obtain the variance estimator for population total estimator under the above set-up. Simulation results are presented to show the performance of the variance estimator in tracking the total variance.

KEY WORDS: Item Non Response, Linearization Method, Variance Estimation

RÉSUMÉ

L'imputation est une méthode commune pour traiter la non-réponse partielle dans les enquêtes. Dans le cas où on a un ensemble de variables auxiliaires donné, nous pouvons effectuer l'imputation pour des cycles multiples afin d'imputer toutes les valeurs manquantes. Pour chaque cycle, différentes combinaisons de méthodes d'imputation et de variables sont utilisées. De plus, les valeurs imputées précédentes sont traitées comme des variables observées. Dans cette étude, nous appliquons une approche linéarisée unifiée afin d'obtenir un estimateur de la variance pour un estimateur d'un total de population selon l'exposé ci-dessus. Les résultats de la simulation sont présentés pour montrer la performance de l'estimateur de la variance à capturer la variance totale.

MOTS-CLÉS : Estimation de la variance; méthode de linéarisation; non-réponse partielle.

1. INTRODUCTION

Imputation is a common method for handling item nonresponse in surveys. Given a set of auxiliary variables, a survey may perform imputation for multiple cycles to impute all the missing values. For example, the Unified Enterprise Survey (UES) conducted by Statistics Canada first substitutes the missing items by values from tax files if they are available, then the UES carries out historical trend imputation, current ratio imputation and other imputation in sequence for the remaining missing values. Since variance estimation under multiple cycle imputation is difficult to derive and compute, surveys usually employ~~s~~ a naïve variance estimator, treating imputed values as observed data. Shao and Steel (1999) show that the naïve variance estimator is not consistent and it can be seriously biased when the sampling fraction is large and the response rate is low.

In this article, we study variance estimation under multiple cycle imputation, using a unified linearization approach by Demnati and Rao (2004). We consider two imputed estimators for the finite population total. We start with an introduction of the multiple cycle imputation in section 2, as well as two imputed estimators. Then, in section 3 we review the Demnati-Rao (D-R) approach for the survey data in the presence of nonresponse. In section 4, associated D-R variance estimators for both total estimators are derived, and a simulation is conducted to illustrate their performance.

2. MULTIPLE CYCLE IMPUTATION

Consider a finite population $U = \{1,2, \dots, N\}$. Let y denote the variable of interest, and let y_k be the value of y for the k^{th} population unit. A probability sample S of size n is selected from the finite population. After data collection, the sample S

[1] W.Qian, Business Survey Methods Division, Statistics Canada, Ottawa, Canada, Wei.Qian@statcan.gc.ca
[2] A.Demnati, Business Survey Methods Division, Statistics Canada, Ottawa, Canada, Abdellatif.Demnati@statcan.gc.ca

consists of two parts: S_o and S_m, where S_o is the subset of all sampled units with observed *y*-values and $S_m = S - S_o$ is the set of all nonrespondents. We impute y_k for $k \in S_m$ and produce a complete sample data set.

When the number of missing values to impute is large, we may not be able to have all the missing values imputed from only one auxiliary variable because it might be not available for some units. Often, surveys have multiple sources of auxiliary information, such as administrative files. Based on them, the imputation can be implemented in many ways to impute all the missing values. In this study, we consider multiple cycle imputation that works as follows.

Given *I* sets of auxiliary variables, say $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_I$, we perform imputation for *I* cycles so that all the missing values are imputed. The order of these variable sets being used depends on the appropriateness of their imputation models. We use one imputation method per cycle that may be different (in our study, we consider substitution and ratio imputation). Also, the precedent imputed values are treated as observed data in future cycles. Therefore, through *I* cycles of imputation, S_m is composed of $S_m = S_{1m} \cup \dots \cup S_{Im}$ where $S_{im},\ i = 1, \dots, I$ is the subset of nonrespondents whose *y*-values are imputed at the i^{th} cycle. Accordingly, we obtain a complete sample *y*-vector $\boldsymbol{y}_n^I = (\boldsymbol{y}_o^T, \widehat{\boldsymbol{y}}_{1m}^T, \dots, \widehat{\boldsymbol{y}}_{Im}^T)^T$, where $\boldsymbol{y}_o = \{y_k; k \in S_o\}$ and $\widehat{\boldsymbol{y}}_{im}, i = 1, \dots, I$ are the set of imputed y-values in S_{im}. The values in $\widehat{\boldsymbol{y}}_{im}$, for $i = 1, \dots, I$, are produced through imputation in cycle *i*, using the i^{th} set of auxiliary variables $\boldsymbol{x}_i$ and $\widehat{\boldsymbol{y}}_{(i)m}$ where $\widehat{\boldsymbol{y}}_{(i)m} = (\boldsymbol{y}_o^T, \widehat{\boldsymbol{y}}_{1m}^T, \dots, \widehat{\boldsymbol{y}}_{i-1,m}^T)^T$.

We are interested in estimating the total in a domain of the finite population: $Y = \sum_{k \in U} \delta_k y_k$, where $\delta_k = 1$ if *k* belongs to the domain and $\delta_k = 0$ otherwise. Based on the complete imputed sample vector, we define an imputed estimator as

$$\hat{Y} = \sum_{k \in U} d_k \delta_k [o_k y_k + (1 - o_k) \hat{y}_k^*], \tag{1}$$

where $d_k = a_k / \pi_k$ with $a_k = 1$ if $k \in S$; $a_k = 0$ otherwise, $\pi_k = \Pr\{k \in S\}$, $o_k = 1$ if y_k is observed; $o_k = 0$ otherwise, and $\hat{y}_k^*$ denotes the imputed value for unobserved y_k.

For some surveys using administrative files in which values are available for most population units, instead of imputing the missing values for only nonrespondents, they also impute values for non-sampled units and then obtain a simulated census. Based on the simulated census, we define another imputed estimator for the population total as

$$\hat{Y}^{(N)} = \sum_{k \in U} \delta_k [a_k o_k y_k + (1 - a_k o_k) \hat{y}_k^*]. \tag{2}$$

The estimator $\hat{Y}^{(N)}$ is biased for the finite population total *Y* (Demnati and Rao 2007). However, if the imputation model holds for the population, the bias is small and $\hat{Y}^{(N)}$ outperforms $\hat{Y}$ in terms of their mean square errors (MSE).

Our objective is to estimate the total variance of both $\hat{Y}$ and $\hat{Y}^{(N)}$ under multiple cycle imputation. Linearization variance estimation method can provide asymptotically unbiased and consistent variance estimators. We apply the unified approach of linearization variance estimation by Demnati and Rao (2004).

3. LINEARIZATION VARIANCE ESTIMATION

Let $\boldsymbol{A}_d$ denote a $2 \times N$ random matrix with the k^{th} column $\boldsymbol{d}_k = (d_k, d_k o_k)^T$. Suppose that $\hat{\theta}$ is an estimator that can be expressed explicitly or implicitly as $f(\boldsymbol{A}_d)$, where *f* is a differentiable function. According to Demnati and Rao (2007), we can approximate the total variance of $\hat{\theta}$, $V_T(\hat{\theta})$, by

$$V_T(\hat{\theta}) \approx V_T(\hat{\Psi}),$$

where $\hat{\Psi} = \sum_{k \in U} d_k \psi_{1k} + \sum_{k \in U} d_k o_k \psi_{2k}$ with the constant vector $\psi_k = (\psi_{1k}, \psi_{2k})^T = \partial f(\boldsymbol{A}_b) / \partial \boldsymbol{b}_k |_{\boldsymbol{A}_b = \boldsymbol{A}_d}$, and $\boldsymbol{A}_b$ is a $2 \times N$ matrix with the *k*th column $\boldsymbol{b}_k = (b_{1k}, b_{2k})^T$, here, ψ_k is the linearized variable associated with the estimator $\hat{\theta}$.

It remains to find an estimator for $V_T(\hat{\Psi})$. We consider the variance estimation of a general linear combination $\hat{U} = \sum_{k \in U} d_k u_{1k} + d_k o_k u_{2k}$ where $u_k = (u_{1k}, u_{2k})^T$ is a constant vector. In the presence of nonresponse, one often decomposes the total variance into sampling variance component and nonresponse variance component, and then estimates each component. In most papers (e,g. Deville and Särndal 1994 and Rao and Sitter 1995), the following sampling-response path is used for variance decomposition, i.e. *population* → *complete sample* → *sample with* *nonrespondents*. Shao and Steel (1999) use another sampling-response path that reverses the order of sampling and response. In this study, we use the reversed sample-response path.

Decomposition: Following the reversed sampling-response approach, i.e. *population* → *census with nonrespondents* → *sample with nonrespondents*, we have:

$$V_T(\hat{U}) = E_r V_p(\hat{U}) + V_r E_p(\hat{U}) \quad (3)$$
$$= V_1(\boldsymbol{u}) + V_2(\boldsymbol{u}),$$

where E_r, V_r denote respectively the expectation and variance with respect to the response mechanism, and E_p, V_p denote respectively the expectation and variance with respect to the sampling design. Here, $V_1(\boldsymbol{u}) = E_r V_p(\hat{U})$, $V_2(\boldsymbol{u}) = V_r E_p(\hat{U})$, and $\boldsymbol{u}$ is a $N \times 2$ constant matrix with the k^{th} row u_k^T.

We obtain the total variance estimator $\hat{V}_T(\hat{U})$ by finding estimators for the sampling variance component, $V_1(\boldsymbol{u})$, and the nonresponse variance component, $V_2(\boldsymbol{u})$.

We assume that the units report their information independently and the sampling design is independent of the response mechanism.

The sampling variance, $V_1(\boldsymbol{u})$, can be estimated by a design-unbiased estimator of $V_p(\hat{U})$ with fixed o_k. Thus, by applying the Horvitz-Thompson variance estimator for any design, we obtain an estimator of $V_1(\boldsymbol{u})$ as

$$\vartheta_1(\boldsymbol{u}) = \vartheta_{HT}(\boldsymbol{u_p}) = \sum_{k\in U}\sum_{t\in U} d_k d_t (1-\omega_{kt}) u_{p;k} u_{p;t}, \quad (5)$$

or applying Sen-Yates-Grundy variance estimator for only fixed size design, we obtain another estimator as

$$\vartheta_1(\boldsymbol{u}) = \vartheta_{SYG}(\boldsymbol{u_p}) = -\frac{1}{2}\sum_{k\in U}\sum_{\substack{t\in U\\ t\neq k}} a_k a_t (1-\omega_{kt}) \left(\frac{u_{p;k}}{\pi_k} - \frac{u_{p;t}}{\pi_t}\right)^2, \quad (6)$$

where $u_{p;k} = u_{1k} + o_k u_{2k}$ for fixed o_k, $d_{kt} = a_k a_t/\pi_{kt}$, $\omega_{kt} = \pi_k \pi_t/\pi_{kt}$ and $\pi_{kt} = E_p[a_k a_t]$ for $k \neq t$.

The nonresponse variance, $V_2(\boldsymbol{u})$, is simplified as $V_2(\boldsymbol{u}) = V_r(\sum_{k\in U} o_k u_{r;k})$ where $u_{r;k} = u_{2k}$. We may use the estimator $\vartheta_2(\boldsymbol{u})$ given by

$$\vartheta_2(\boldsymbol{u}) = \sum d_k o_k (1-\hat{p}_k) {u_{r;k}}^2, \quad (7)$$

where $\hat{p}_k$ is a consistent estimator of $p_k = E_r[o_k]$.

Therefore, we can derive the total variance estimator of $\hat{U}$, $\vartheta_T(\boldsymbol{u})$, as the sum of both terms (5) or (6) and (7), i.e.

$$\vartheta_T(\boldsymbol{u}) = \vartheta_1(\boldsymbol{u}) + \vartheta_2(\boldsymbol{u}). \quad (8)$$

Now we turn to our estimator $\hat{\theta}$. Replacing $\boldsymbol{u}$ in (8) by $\boldsymbol{\psi}$, we obtain the D-R total variance estimator as

$$\vartheta_{DR}(\hat{\theta}) = \vartheta_T(\boldsymbol{\psi}),$$

where $\boldsymbol{\psi}$ be a $N \times 2$ constant matrix with the k^{th} row ψ_k^T.

4. VARIANCE ESTIMATOR FOR THE TWO POPULATION TOTAL ESTIMATORS

In this section, we derive the total variance estimator for the sample based estimator $\hat{Y}$ and the simulated-census based estimator $\hat{Y}^{(N)}$ under multiple cycle imputation. We set *I=3*.

Let $(\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{z})$ denote three auxiliary variables that are partially observed, and let $(\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3)$ denote the variables indicating the availability of $(\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{z})$, for example, if t_k is observed, then $I_{1k} = 1$; otherwise, $I_{1k} = 0$, similar for I_{2k} and I_{3k}. We assume that the vector $(\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3)$ is fix and satisfies $\sum_{i=1}^{3} I_{ik} \geq 1$ for $k = 1, \dots, N$. Using the auxiliary data $(\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{z})$, we carry out 3 cycles of imputation: substitution by $\boldsymbol{t}$, ratio imputation on $\boldsymbol{x}$ and another ratio imputation on $\boldsymbol{z}$ in order.

The imputed value $\hat{y}_k^*$ in (1) and (2) can be written as

$$\hat{y}_k^* = L_{1k}\hat{y}_{1k}^* + L_{2k}\hat{y}_{2k}^* + L_{3k}\hat{y}_{3k}^*, \tag{9}$$

where $(\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3)$ indicate the imputation cycle from which we obtain $\hat{y}_k^*$, i.e. $L_{1k} = I_{1k}$, $L_{2k} = (1 - I_{1k})I_{2k}$, and $L_{3k} = (1 - I_{1k})(1 - I_{2k})I_{3k}$; $\hat{y}_{ik}^*$, $i = 1, \ldots, 3$, are the predicted values from the ith imputation cycle.

For the sample based estimator $\hat{Y}$, $\hat{y}_{1k}^* = t_k$, $\hat{y}_{2k}^* = \hat{\alpha}x_k$ with

$$\hat{\alpha} = \frac{\sum_{k\in U} d_k I_{2k}[o_k y_k + (1 - o_k)L_{1k}\hat{y}_{1k}^*]}{\sum_{k\in U} d_k I_{2k}[o_k + (1 - o_k)L_{1k}]x_k},$$

and $\hat{y}_{3k}^* = \hat{\beta} z_k$ with

$$\hat{\beta} = \frac{\sum_{k\in U} d_k I_{3k}[o_k y_k + (1 - o_k)(L_{1k}\hat{y}_{1k}^* + L_{2k}\hat{y}_{2k}^*)]}{\sum_{k\in U} d_k I_{3k}[o_k + (1 - o_k)(L_{1k} + L_{2k})]z_k}.$$

For the simulated-census based estimator $\hat{Y}^{(N)}$, $\hat{y}_{1k}^* = t_k$, $\hat{y}_{2k}^* = \hat{\alpha}^{(N)}x_k$ with

$$\hat{\alpha}^{(N)} = \frac{\sum_{k\in U} I_{2k}[a_k o_k y_k + (1 - a_k o_k)L_{1k}\hat{y}_{1k}^*]}{\sum_{k\in U} I_{2k}[a_k o_k + (1 - a_k o_k)L_{1k}]x_k},$$

and $\hat{y}_{3k}^* = \hat{\beta}^{(N)} z_k$ with

$$\hat{\beta}^{(N)} = \frac{\sum_{k\in U} I_{3k}[a_k o_k y_k + (1 - a_k o_k)(L_{1k}\hat{y}_{1k}^* + L_{2k}\hat{y}_{2k}^*)]}{\sum_{k\in U} I_{3k}[a_k o_k + (1 - a_k o_k)(L_{1k} + L_{2k})]\, z_k}.$$

Both estimators, $\hat{Y}$ and $\hat{Y}^{(N)}$, can be expressed a differentiable function of $\boldsymbol{A}_d$. Following the preceding section, the D-R approach works and we only need to find their associated linear variables.

Taking the derivatives, we obtain the linear variables ψ_k for the sample based estimator $\hat{Y}$ by

$$\begin{aligned}\psi_{1k} = \delta_k\hat{y}_k^* &+ L_{1k}I_{2k}\big(C_x + \tilde{C}_x D_z\big)(\hat{y}_{1k}^* - \hat{y}_{2k}^*)\\ &+ L_{1k}I_{3k}D_z(\hat{y}_{1k}^* - \hat{y}_{3k}^*) + L_{2k}I_{3k}D_z(\hat{y}_{2k}^* - \hat{y}_{3k}^*),\\ \psi_{2k} &= \big(\delta_k + L_{2k}C_x + L_{3k}D_z + L_{2k}\tilde{C}_x D_z\big)(y_k - \hat{y}_k^*),\end{aligned}$$

where C_x, D_z and $\tilde{C}_x$ are given by

$$C_x = \left\{\sum\nolimits_{k\in U} d_k I_{2k}[o_k + (1 - o_k)L_{1k}]x_k\right\}^{-1}\sum\nolimits_{k\in U}\delta_k d_k(1 - o_k)L_{2k}x_k,$$

$$D_z = \left\{\sum\nolimits_{k\in U} d_k I_{3k}[o_k + (1 - o_k)(L_{1k} + L_{2k})]z_k\right\}^{-1}\sum\nolimits_{k\in U}\delta_k d_k(1 - o_k)L_{3k}z_k, \text{ and}$$

$$\tilde{C}_x = \left\{\sum\nolimits_{k\in U} d_k I_{2k}[o_k + (1 - o_k)L_{1k}]x_k\right\}^{-1}\sum\nolimits_{k\in U}\delta_k d_k I_{3k}(1 - o_k)L_{2k}x_k.$$

Similarly, for the simulated-census based estimator $\hat{Y}^{(N)}$, we apply the D-R approach to obtain the linear data, $\psi_k^{(N)}$, as

$$\psi_{1k}^{(N)} = 0, \qquad \text{and}$$

$$\psi_{2k}^{(N)} = \pi_k\left(\delta_k + L_{2k}C_x^{(N)} + L_{3k}D_z^{(N)} + L_{2k}\tilde{C}_x^{(N)}D_z^{(N)}\right)(y_k - \hat{y}_k^*),$$

where $C_x^{(N)}$, $D_z^{(N)}$, $\tilde{C}_x^{(N)}$ are given by

$$C_x^{(N)} = \left\{\sum\nolimits_{k\in U} I_{2k}[a_k o_k + (1 - a_k o_k)L_{1k}]x_k\right\}^{-1}\sum\nolimits_{k\in U}\delta_k(1 - a_k o_k)L_{2k}x_k,$$

$$D_z^{(N)} = \left\{\sum\nolimits_{k\in U} I_{3k}[a_k o_k + (1 - a_k o_k)(L_{1k} + L_{2k})]z_k\right\}^{-1}\sum\nolimits_{k\in U}\delta_k(1 - a_k o_k)L_{3k}\, z_k, \qquad \text{and}$$

$$\tilde{C}_x^{(N)} = \left\{\sum\nolimits_{k\in U} I_{2k}[a_k o_k + (1 - a_k o_k)L_{1k}]x_k\right\}^{-1}\sum\nolimits_{k\in U}\delta_k I_{3k}(1 - a_k o_k)L_{2k}x_k.$$

Note, for simplicity, that we consider the whole population as one imputation class for ratio imputation. The D-R approach can be also extended to the case with multiple imputation classes.

4.1 Simulation study

Using a synthetic population of size N=9404 with two variables, *total revenue* and *size* whose values are available for all population units, we generate $(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t})$ from the following models: *z=total revenue,* $x_k|z_k = 1.3z_k + 1000z_k^{1/2}\varepsilon_{1k}$, $y_k|x_k = 1.2z_k + 1000x_k^{1/2}\varepsilon_{2k}$, and $t_k|y_k = y_k + 60000\varepsilon_{3k}$ where ε_{1k}, ε_{2k} and ε_{3k} are independent observations generated from *N*(0,1). $\boldsymbol{y}$ is the variable of interest; $\boldsymbol{t}$ is the substitution variable and $\boldsymbol{x}, \boldsymbol{z}$ are auxiliary variables for ratio imputation. Also, we generate $(\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{I}_3)$ from Bernoulli distributions with respectively probability of success (0.2, 0.7, 0.8). The resulting percentages of $(\boldsymbol{L}_1,\ \boldsymbol{L}_2,\ \boldsymbol{L}_3)$ in the population are (20, 56, 24); that is, 20% of the population units are eligible for substitution, 56% are for ratio imputation on $\boldsymbol{x}$, and 24% are for ratio imputation on $\boldsymbol{z}$. The population is stratified by *size* into three strata: TA (*size* $\geq 10^7$), TS1(*size* $\geq 2\times10^4$ and *size* $<10^7$) and TS2(*size* $< 2\times10^4$). The counts of the population units in these strata are (191, 4458, 4755) respectively. Here TA stands for 'take all' and TS stands for 'take some'.

The response indicator o_k is generated from Bernoulli(p_k) independently where p_k is the probability of reporting *y*-value for the k^{th} unit. Two response mechanisms are considered: uniform and logistic model. For the uniform response mechanism, each unit has the same response probability *p*. The values of *p* considered are 0.6, 0.7 and 0.8. For logistic response mechanism, p_k is calculated from a logistic model: $logit(p_k) = 10 \times l_k + 0.5$ where $l_k = (size_k - M)/R$ with *M=median(size)* and *R=range(size),* which gives an average of 63% response rate.

Two sampling schemes are considered: stratified simple random sampling (STSRS) and stratified Poisson sampling (STPOI). For STSRS, the sample sizes for TA, TS1 and TS2 are *n*=(191, 400, 100). For STPOI, all units in TA are selected in the sample. The inclusion probability for the units in TS1 and TS2 is computed by $\pi_k = n_h size_k / \sum_{k\in h} size_k$, here $\Sigma_{k\in h}$ is the sum of all units in the stratum *h*. Also, the sampling is independent of the response mechanism.

The sample based estimator $\hat{Y}$ for the population total is studied. In the simulation, *B*=10000 independent samples are selected from the population. For each sample *b=1,… ,B*, the response probability is estimated for the sample as $\hat{p}_k = \sum_{k\in U} d_k o_k / \sum_{k\in U} d_k$ for uniform mechanism and $logit(\hat{p}_k) \propto l_k$ for the logistic response mechanism. From each sample, we calculate the estimate, $\hat{Y}_b$, for the population total, the associated naïve variance estimate $\vartheta_N(\hat{Y}_b)$ and associated D-R variance estimate $\vartheta_{DR}(\hat{Y}_b)$. We simulated the total variance of $\hat{Y}$ by $V_{MC} = 1/B \sum_{b=1}^{B} \left(\hat{Y}_b - \bar{\hat{Y}}_b\right)^2$ where $\bar{\hat{Y}}_b = 1/B \sum_{b=1}^{B} \hat{Y}_b$. We simulated the relative bias of $\vartheta_N(\hat{Y})$ and $\vartheta_{DR}(\hat{Y})$ as $RB(\vartheta_N) = (\bar{\vartheta}_N - V_{MC})/V_{MC}$ and $RB(\vartheta_{DR}) = (\bar{\vartheta}_{DR} - V_{MC})/V_{MC}$ respectively, where $\bar{\vartheta}_N = 1/B \sum_{b=1}^{B} \vartheta_N(\hat{Y}_b)$ and $\bar{\vartheta}_{DR} = 1/B \sum_{b=1}^{B} \vartheta_{DR}(\hat{Y}_b)$.

The simulation results are presented in Table 1. It is observed that

1. the D-R variance estimator produces approximately unbiased estimates,
2. the naïve variance estimator always underestimates the total variance, and
3. the nonresponse variance component is negligible in comparison to the sampling variance component, because the sampling fraction is small. In this case, one may approximate the total variance by only the sampling variance component and avoid the estimation of the response probabilities.

5. REMARKS AND FUTURE WORK

This study has considered variance estimation under multiple cycle imputation for two estimators of the population total. Using the D-R approach, we obtain consistent and approximately unbiased variance estimators. However, it has been shown that variance estimation under multiple cycle imputation becomes complicated. Thus, we are investigating a single-cycle imputation method so that all the missing values can be imputed in one cycle but the underlying model for multiple cycle imputation is respected. Also, we are investigating variance estimation under both imputation and weight adjustment.

Table 1- Simulation results:
relative bias for variance estimator: Demnati-Rao Vs. Naïve

Sampling Design	*Response Probability*	% *Relative bias* $RB(\vartheta_N)$	$RB(\vartheta_{DR})$	$\bar{\vartheta}_2/\bar{\vartheta}_{DR}$
STSRS	***0.6***	***-15.48%***	***0.08%***	***1.65%***
	0.7	***-11.49%***	***-0.37%***	***1.21%***
	0.8	***-8.67%***	***-1.58%***	***0.79%***
	Logistic	***-7.95%***	***0.92%***	***0.85%***
STPOI	***0.6***	***-22.69%***	***0.38%***	***1.84%***
	0.7	***-16.77%***	***0.25%***	***1.37%***
	0.8	***-11.27%***	***-0.11%***	***0.90%***
	Logistic	***-13.11%***	***0.66%***	***0.99%***

Note: $\bar{\vartheta}_2 = 1/B\sum_{b=1}^{B}\vartheta_{2b}$ where ϑ_2 is defined in (7).

REFERENCE

Demnati, A. and Rao, J. N. K. (2004). Linearization Variance Estimators for Survey Data (with discussion). *Survey Methodology*, **30**, 17-34.

Demnati, A. and Rao, J.N.K. (2007). Linearization Variance Estimators for Survey Data: some recent work (with comments). *Third International Conference on Establishment Surveys*, Montreal, Canada, 916-925.

Deville, J. C. and Särndal, C. E. (1994). Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator, *Journal of Official Statistics*, **10**, 381-394.

Rao, J. N. K. and Sitter, R. R. (1995). Variance Estimation Under Two-Phase Sampling With Application to Imputation for Missing Data. *Biometrika*, **82**, 453-460.

Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, **94**, 254-265.