

# LINEARIZED VARIANCE ESTIMATION FROM SIMULATED CENSUS DATA

A. Demnati<sup>1</sup> and J.N.K. Rao<sup>2</sup>

## ABSTRACT

Simulated census data are often generated from a probability sample by imputing for the non-sampled units and sample non-respondents, using auxiliary variables available for the population units. Composite imputation with two or more different imputation methods is also used; for example, the value from an administrative file when it is available and regression imputation otherwise. A naïve estimator of a total based on the simulated census total and a design-consistent estimator are studied, and associated linearization variance estimators are obtained by using the unified variance estimation approach of Demnati and Rao (*Survey Methodology*, 2004). Simulation results are also presented.

KEY WORDS: Regression imputation; Substitution method; Variance estimation.

## RÉSUMÉ

Il est fréquent de produire des données simulées de recensement à l'aide d'un échantillon probabiliste en faisant une imputation pour les unités non échantillonnées et les non répondants dans l'échantillon, à l'aide de variables auxiliaires disponibles pour les unités de population. L'imputation composite regroupant deux méthodes d'imputation ou plus est également utilisée; par exemple l'utilisation de la valeur provenant d'un fichier administratif quand elle est disponible et l'imputation par la régression autrement. Nous étudions un estimateur naïf d'un total fondé sur le total du recensement simulé et un estimateur convergent sous le plan, et nous obtenons les estimateurs de variance par linéarisation associés en utilisant l'approche unifiée d'estimation de la variance de Demnati et Rao (*Techniques d'enquête*, 2004). Nous présentons aussi les résultats d'une étude par simulation.

MOTS CLÉS : Imputation par régression; méthode de substitution; estimation de la variance.

## 1. INTRODUCTION

For operational convenience, statistical offices often generate simulated census data from a probability sample by imputing the values for the non-sampled units and the values for the missing items in the sample, using auxiliary variables available for the population data such as census and administrative data. Composite imputation involving two or more different methods is also often used; for example, the values from administrative file (e.g., tax file), when it is available and regression imputation otherwise.

In this paper, we study the estimation of a population total or its model expectation (under an assumed superpopulation model) and associated variance estimation. For variance estimation, we have used the unified approach proposed by Demnati and Rao (2004). We have studied naïve estimators based on the simulated census as well as design-consistent estimators. In section 2, we consider the case of complete response and also report the results of a simulation study on estimating totals. The case of missing responses is studied in section 3.

## 2. COMPLETE RESPONSE

In this section, we consider the case where complete responses are obtained from sampled units. Imputation is then performed only on all non-sampled units. The resulting simulated census data is complete. In section 2.1 the naïve estimator of a total based on the simulated census total and a design-based estimator are studied. A limited simulation compares the performance of the two estimators and associated variance estimators. Section 2.2 considers the estimation of model parameters.

---

<sup>1</sup> A. Demnati, Business Survey Methods Division, Statistics Canada, Ottawa, Canada, Abdellatif.Demnati@statcan.gc.ca

<sup>2</sup> J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, JRao@math.carleton.ca

## 2.1. Finite Population Totals

We consider the case where complete responses are obtained from sampled units. Imputation is performed on all non-sampled units. A naïve estimator of the finite population total  $Y = \sum y_k$  is given by

$$\hat{Y}^{(N)} = \sum a_k y_k + \sum (1 - a_k) \hat{y}_k^*, \quad (2.1)$$

where  $a_k = 1$  if unit  $k$  is in the sample  $s$ ,  $a_k = 0$  otherwise, the imputed value  $\hat{y}_k^*$  is given by  $\hat{y}_k^* = \sum_g I_{gk} \hat{y}_{gk}^*$ ,  $I_{gk}$  is the imputation method indicator for unit  $k$  with  $\sum_g I_{gk} = 1$ . We consider the case of two imputation methods:  $\hat{y}_k^* = (1 - I_k) \hat{y}_{1k}^* + I_k \hat{y}_{2k}^*$  with  $\hat{y}_{1k}^* = t_k$  and  $\hat{y}_{2k}^* = \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_a$  where  $\hat{\boldsymbol{\beta}}_a = \hat{\mathbf{Q}}_a^{-1} \sum a_k I_k c_k \mathbf{x}_k y_k$  for some specified  $c_k$ ,  $\hat{\mathbf{Q}}_a = \sum a_k I_k c_k \mathbf{x}_k \mathbf{x}_k^T$ ,  $t_k$  is the value from administrative files,  $\mathbf{x}_k$  is the vector of auxiliary variables, and the constant  $I_k$  is the missing  $t_k$  indicator.

Suppose the parameter of interest is  $\theta_N = E_p(\hat{Y}^{(N)})$ , where  $E_p$  denotes design expectation, then the Demnati–Rao (DR) variance estimator (Demnati and Rao, 2004) is simply given by

$$\mathcal{G}_{DR}(\hat{Y}^{(N)}) = \mathcal{G}(z), \quad (2.2)$$

where  $\mathcal{G}(u)$ , in operation notation, is the variance estimator of the linear estimator  $\hat{U} = \sum d_k u_k$ ,  $z_k = \partial f(\mathbf{b}) / \partial b_k |_{b=d}$ ,  $f(\mathbf{d}) = \hat{Y}^{(N)}$ ,  $\mathbf{d}$  is the  $N \times 1$  vector of sampling weights and  $\mathbf{b}$  is the  $N \times 1$  vector of arbitrary real numbers. The design based variance estimator of the total  $\hat{U}$  is

$$\mathcal{G}(u) = \sum \sum d_k d_l (1 - \omega_{kl}) u_k u_l, \quad (2.3)$$

where  $d_k = a_k / \pi_k$ ,  $\omega_{kl} = \pi_k \pi_l / \pi_{kl}$ ,  $\pi_k$  is the inclusion probability and  $\pi_{kl}$  is the joint inclusion probability. It remains to evaluate  $z_k$ . We have  $f(\mathbf{b}) = \sum b_k \pi_k y_k + \sum (1 - b_k \pi_k) \hat{y}_k^*(\mathbf{b})$  with  $\hat{y}_k^*(\mathbf{b}) = (1 - I_k) t_k + I_k \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_a(\mathbf{b})$ ,  $\hat{\boldsymbol{\beta}}_a(\mathbf{b}) = [\hat{\mathbf{Q}}_a(\mathbf{b})]^{-1} \sum b_k \pi_k I_k c_k \mathbf{x}_k y_k$  and  $\hat{\mathbf{Q}}_a(\mathbf{b}) = \sum b_k \pi_k I_k c_k \mathbf{x}_k \mathbf{x}_k^T$ . Hence,

$$z_k = \partial f(\mathbf{b}) / \partial b_k |_{b=d} = (1 - I_k)(y_k - t_k) g_{1k}^{(N)} + I_k (y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_a) g_{2k}^{(N)}, \quad (2.4)$$

where  $g_{1k}^{(N)} = \pi_k$  and  $g_{2k}^{(N)} = \pi_k [1 + \sum (1 - a_l) I_l \mathbf{x}_l^T \hat{\mathbf{Q}}_a^{-1} \mathbf{x}_l c_l]$ . We may write (2.4) in the form

$$z_k = g_k^{(N)} (y_k - \hat{y}_k^*), \quad (2.5)$$

where  $g_k^{(N)} = (1 - I_k) g_{1k}^{(N)} + I_k g_{2k}^{(N)}$ .

The finite population parameter  $\theta_N$  induced by the estimator  $\hat{Y}^{(N)}$  is given by

$$\theta_N = E_p(\hat{Y}^{(N)}) \approx \sum \pi_k y_k + \sum (1 - \pi_k) y_k^*, \quad (2.6)$$

where  $y_k^* = E_p(\hat{y}_k^*)$ . It is clear from (2.6) that  $\theta_N$  depends on the selection probabilities. For example, the use of  $\hat{Y}^{(N)}$  under simple random sampling induces a mixture of two totals as finite population parameter:  $\theta_N \approx f \sum y_k + (1 - f) \sum y_k^*$  where  $f$  is the sampling fraction. The sampling bias induced by the estimator  $\hat{Y}^{(N)}$  in estimating the finite population total  $Y$  is given by

$$\theta_N - Y \approx - \sum (1 - \pi_k) (y_k - y_k^*) \equiv B. \quad (2.7)$$

In order to remove the conditional sampling bias, one may first estimate the bias by

$$\hat{B} = - \sum d_k (1 - \pi_k) (y_k - \hat{y}_k^*), \quad (2.8)$$

and then adjust  $\hat{Y}^{(N)}$  to get the estimator

$$\hat{Y} = \sum d_k y_k + \sum (1 - d_k) \hat{y}_k^*, \quad (2.9)$$

with  $\hat{y}_{2k}^* = \mathbf{x}_k^T \hat{\boldsymbol{\beta}}$ ,  $\hat{\boldsymbol{\beta}} = \hat{\mathbf{Q}}^{-1} \sum d_k I_k c_k \mathbf{x}_k y_k$  and  $\hat{\mathbf{Q}} = \sum d_k I_k c_k \mathbf{x}_k \mathbf{x}_k^T$ . The design-based estimator  $\hat{Y}$  is approximately unbiased for  $Y$ :  $E_p(\hat{Y}) \approx Y$ .

A variance estimator of  $\hat{Y}$  is given by (2.3) with  $u_k$  replaced by

$$z_k = \partial f(\mathbf{b}) / \partial b_k |_{b=d} = (1 - I_k)(y_k - t_k) g_{1k} + I_k (y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}) g_{2k}, \quad (2.10)$$

where  $f(\mathbf{d}) = \hat{Y}$ ,  $g_{1k} = 1$  and  $g_{2k} = 1 + \sum(1-d_i)I_i x_i^T \hat{\mathbf{Q}}^{-1} \mathbf{x}_k c_k$ . We may write (2.10) as

$$z_k = g_k(y_k - \hat{y}_k^*), \quad (2.11)$$

where  $g_k = (1-I_k)g_{1k} + I_k g_{2k}$ .

### Simulation

We conducted a small simulation study to examine the performances of the estimators  $\hat{Y}^{(N)}$  and  $\hat{Y}$  and associated variance estimators. We first generated a finite population  $\{\mathbf{y}_k\}$ , with  $\mathbf{y}_k = (y_{1k}, y_{2k}, y_{3k})^T$ , of size  $N=393$  from the following models:  $y_{1k} = x_k + x_k^{1/2} \varepsilon_k$ ,  $y_{2k} = 1.2x_k + x_k^{1/2} \varepsilon_k$ , and  $y_{3k} = 5 + 1.2x_k + x_k^{1/2} \varepsilon_k$ , where  $\varepsilon_k$  are independent observations generated from  $N(0,1)$ , and the fixed  $x_k$  are the ‘‘number of beds’’ for the Hospitals population studied in Valliant *et al.* (2000, p.424-427). We stratified the population into two strata with 272 units  $k$  having  $x_k \leq 350$  in stratum 1 and 121 units  $k$  with  $x_k > 350$  in stratum 2. We selected  $R=5,000$  stratified simple random samples of sizes  $n_1 = n_2 = 15$ . Our vector parameter of interest is the finite population total  $\theta_N = (Y_1, Y_2, Y_3)^T$ . Non-sampled units are imputed. For each variable  $l$ ,  $l=1,2,3$ , four estimators are considered. Two estimators used the substitution imputation method: the naïve estimator  $\hat{Y}_{IS}^{(N)} = \sum a_k y_{lk} + \sum(1-a_k)x_k$ , and the corresponding design-based estimator  $\hat{Y}_{IS} = \sum d_k y_{lk} + \sum(1-d_k)x_k$ ,  $l=1,2,3$ . The other two estimators used the ratio imputation method: the naïve estimator  $\hat{Y}_{IR}^{(N)} = \sum a_k y_{lk} + \sum(1-a_k)x_k \hat{\beta}_{la}$ , and the associated design-based estimator  $\hat{Y}_{IR} = \sum d_k y_{lk} + \sum(1-d_k)x_k \hat{\beta}_l$ , where  $\hat{\beta}_{la} = \sum a_k y_{lk} / \sum a_k x_k$  and  $\hat{\beta}_l = \sum d_k y_{lk} / \sum d_k x_k$ . The estimator  $\hat{Y}_{IR}$  reduces to ratio estimator:  $\hat{Y}_{IR} = \hat{Y}(X/\hat{X})$ . Let  $\hat{\theta}$  denotes an estimator of a population total  $\theta_N$  and  $\mathcal{G}(\hat{\theta})$  be the associated variance estimator. We calculated the simulated relative bias of  $\hat{\theta}$  and  $\mathcal{G}(\hat{\theta})$  as  $RB(\hat{\theta}) = (\hat{\theta} - \theta_N) / \theta_N$  and  $RB\{\mathcal{G}(\hat{\theta})\} = \{\mathcal{G}(\hat{\theta}) - MSE(\hat{\theta})\} / MSE(\hat{\theta})$ , where  $\hat{\theta} = R^{-1} \sum_{r=1}^R \hat{\theta}_r$  is the mean of the estimates  $\hat{\theta}_r$  from the simulated samples  $r=1, \dots, R$  and  $MSE(\hat{\theta}) = R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \theta_N)^2$  is the simulated mean squared error (MSE). We calculated  $RB(\hat{\theta})$ ,  $RB\{\mathcal{G}(\hat{\theta})\}$  and mean squared error (MSE) ratios for each component of the vector  $(\hat{Y}_1^T, \hat{Y}_2^T, \hat{Y}_3^T)^T$  with  $\hat{Y}_l = (\hat{Y}_{IS}^{(N)}, \hat{Y}_{IS}, \hat{Y}_{IR}^{(N)}, \hat{Y}_{IR})^T$  ( $l=1,2,3$ ) and those values are reported in Table 1. It is clear from Table 1 that for the variables 2 and 3, which violate the underlying model supporting the substitution method, the naïve substitution estimator  $\hat{Y}_{IS}^{(N)}$  is highly inefficient to  $\hat{Y}_{IS}$  ( $l=2,3$ ) due to large relative bias. But for variable 1, which obeys the underlying model for the substitution method, the relative bias of  $\hat{Y}_{IS}^{(N)}$  is small and  $\hat{Y}_{IS}^{(N)}$  is highly efficient relative to  $\hat{Y}_{IS}$ . Further, the naïve estimator  $\hat{Y}_{IR}^{(N)}$  under ratio imputation is slightly more efficient than the corresponding design-based estimator  $\hat{Y}_{IR}$  for all the variables. Turning to the relative bias of variance estimators, it is clear from Table 1 that the variance estimator for  $\hat{Y}_{IS}^{(N)}$  leads to large underestimation of MSE. On the other hand, the design-based variance estimator for  $\hat{Y}_{IR}^{(N)}$  and  $\hat{Y}_{IR}$  both perform well in terms of relative bias except for variable 3: variance estimator for  $\hat{Y}_{3R}^{(N)}$  leads to 13% underestimation.

### 2.2. Model Parameters

Let  $\hat{\theta}$  denote either the estimator  $\hat{Y}^{(N)}$  or  $\hat{Y}$ , and suppose that the parameter of interest is the model parameter  $\theta = E_m(Y)$  where  $E_m$  denotes model expectation under a model on  $y_k$ . Let  $\mathbf{d}_k = (d_{1k}, d_{2k})^T$  with  $d_{1k} = d_k$ ,  $d_{2k} = d_k y_k$  and let  $\mathcal{G}(\mathbf{u})$  denotes an estimator of total variance of the linear combination  $\hat{U} = \sum \mathbf{u}_k^T \mathbf{d}_k$  where  $\mathbf{u}_k = (u_{1k}, u_{2k})^T$  is a vector of constants. Then a variance estimator of  $\hat{\theta}$  is given by  $\mathcal{G}(z)$  (Demnati and Rao, 2007) with  $\mathbf{z}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d}$  where  $\mathbf{A}_d$  is a  $2 \times N$  matrix with  $k^{th}$  column  $\mathbf{d}_k$ , and  $\mathbf{A}_b$  is a  $2 \times N$  matrix of arbitrary real numbers with  $k^{th}$  column  $\mathbf{b}_k$ . We have

$$\mathcal{G}(\mathbf{u}) = \sum \sum d_k d_i \omega_{ki} \text{cov}_m(y_k, y_i) u_{k:m} u_{i:m} + \sum \sum d_k d_i (1 - \omega_{ki}) u_{k:is} u_{i:is}, \quad (2.12)$$

where  $u_{k:m} = u_{2k}$  and  $u_{k:is} = u_{1k} + u_{2k} y_k$ . In (2.12),  $\text{cov}_m(y_k, y_i)$  is an estimator of the model covariance of  $y_k$  and  $y_i$  under the assumed model. When the model covariance of  $y_k$  and  $y_i$  is zero,  $\text{cov}_m(y_k, y_i)$  is taken as zero. It remains to evaluate

$\mathbf{z}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d}$ . We have  $\mathbf{z}_k = \mathbf{g}_k^* (-\hat{y}_k^*, 1)^T$  where  $\mathbf{g}_k^* = \mathbf{g}_k^{(N)}$  for  $\hat{Y}^{(N)}$ ; and  $\mathbf{g}_k^* = \mathbf{g}_k$  for  $\hat{Y}$ . The DR variance estimator of  $\hat{\theta}$  is given by (2.12) with  $u_{k:m}$  and  $u_{k:s}$  replaced by  $z_{k:m} = \mathbf{g}_k^*$  and  $z_{k:s} = \mathbf{g}_k^* (y_k - \hat{y}_k^*)$ .

If the order of expectation can be interchanged so that  $E_m E_p = E_p E_m$ , then a new estimator of total variance of the linear combination  $\hat{U}$  is given by

$$\mathcal{G}^*(\mathbf{u}) = \sum \sum d_k d_t \text{cov}_m(y_k, y_t) u_{k:m} u_{t:m} + \sum \sum d_k d_t (1 - \omega_{kt}) u_{k:s}^* u_{t:s}^*, \quad (2.13)$$

where  $u_{k:s}^* = u_{1k} + \hat{E}_m(y_k) u_{2k}$  and  $\hat{E}_m$  denotes the estimator of model expectation under a model on  $y_k$ . We get  $z_{k:s}^* = \mathbf{g}_k^* (\hat{E}_m(y_k) - \hat{y}_k^*)$ . If the imputation method provides unbiased estimator of the model mean, then  $z_{k:s}^* = \mathbf{g}_k^* (\hat{y}_k^* - \hat{y}_k^*) = 0$  and  $\text{Cov}_m(y_k, y_t)$  can be estimated by  $(y_k - \hat{y}_k^*)(y_t - \hat{y}_t^*)$ . In this case, the new variance estimator (2.13) for  $\hat{\theta}$  reduces to

$$\mathcal{G}_{DR}^*(\hat{\theta}) = \sum \sum d_k d_t (y_k - \hat{y}_k^*)(y_t - \hat{y}_t^*) z_{k:m} z_{t:m}. \quad (2.14)$$

### 3. MISSING RESPONSES

In this section, we consider the case of missing responses. After imputation for non-respondents as well as non-sampled units, the estimator of the population total  $Y$  is given by

$$\hat{Y}(\tau) = \sum a_k \tau_k o_k y_k + \sum (1 - a_k \tau_k o_k) \hat{y}_k^*, \quad (3.1)$$

where  $o_k = 1$  if  $y_k$  is observed,  $o_k = 0$  otherwise,  $\hat{y}_{2k}^* = \mathbf{x}_k^T \hat{\boldsymbol{\beta}}(\tau)$ ,  $\hat{\boldsymbol{\beta}}(\tau) = [\hat{\mathbf{Q}}(\tau)]^{-1} \sum a_k \tau_k o_k I_k c_k \mathbf{x}_k y_k$  and  $\hat{\mathbf{Q}}(\tau) = \sum a_k \tau_k o_k I_k c_k \mathbf{x}_k \mathbf{x}_k^T$ . We consider two values for the constants  $\tau_k$ :  $\tau_k = 1$  and  $\tau_k = \pi_k^{-1}$ .

Suppose first that the parameter of interest is  $\theta_N = E_p E_r(\hat{Y}(\tau))$  where  $E_r$  denotes expectation under response mechanism. Let  $\mathbf{d}_k = (d_{1k}, d_{2k})^T$  with  $d_{1k} = d_k$ ,  $d_{2k} = d_k o_k$  and let  $\mathcal{G}(\mathbf{u})$  denote an estimator of total variance of the linear combination  $\hat{U} = \sum \mathbf{u}_k^T \mathbf{d}_k$  where  $\mathbf{u}_k = (u_{1k}, u_{2k})^T$  is a vector of constants. Then a variance estimator of  $\hat{Y}(\tau)$  is given by  $\mathcal{G}(\mathbf{z})$  (Demnati and Rao, 2007) with  $\mathbf{z}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d}$  where  $\mathbf{A}_d$  is a  $2 \times N$  matrix with  $k^{\text{th}}$  column  $\mathbf{d}_k$ , and  $\mathbf{A}_b$  is a  $2 \times N$  matrix of arbitrary real numbers with  $k^{\text{th}}$  column  $\mathbf{b}_k$ . We have, assuming that the order of expectation can be interchanged so that  $E_p E_r = E_r E_p$ ,

$$\mathcal{G}(\mathbf{u}) = \sum \sum d_k d_t \omega_{kt} o_k o_t [(\hat{\xi}_{kt} - \hat{\xi}_k \hat{\xi}_t) / \hat{\xi}_{kt}] u_{k:o} u_{t:o} + \sum \sum d_k d_t (1 - \omega_{kt}) u_{k:s}^* u_{t:s}^*, \quad (3.2)$$

where  $\hat{\xi}_k = \hat{E}_r(o_k)$ ,  $\hat{\xi}_{kt} = \hat{E}_r(o_k o_t)$ ,  $u_{k:o} = u_k$ ,  $u_{k:s}^* = o_k u_k$ , and  $\hat{E}_r$  denotes the estimator of model expectation under a model on  $o_k$ . A variance estimator of  $\hat{Y}(\tau)$  is then given by (3.2) with  $u_{1k} = 0$  and  $u_{2k}$  replaced by

$$z_{2k} = (y_k - \hat{y}_k^*) \mathbf{g}_k^*, \quad (3.3)$$

where  $\mathbf{g}_k^* = \mathbf{g}_k^{(N)}$  for  $\tau_k = 1$  and  $\mathbf{g}_k^* = \mathbf{g}_k$  for  $\tau_k = \pi_k^{-1}$ ,  $\mathbf{g}_k^{(N)} = (1 - I_k) \mathbf{g}_{1k}^{(N)} + I_k \mathbf{g}_{2k}^{(N)}$ ,  $\mathbf{g}_{1k}^{(N)} = \pi_k$ ,  $\mathbf{g}_{2k}^{(N)} = \pi_k [1 + \sum (1 - a_t o_t) I_t \mathbf{x}_t^T [\hat{\mathbf{Q}}(1)]^{-1} \mathbf{x}_k c_k]$ , and  $\mathbf{g}_k = (1 - I_k) + I_k [1 + \sum (1 - d_t o_t) I_t \mathbf{x}_t^T [\hat{\mathbf{Q}}(\pi^{-1})]^{-1} c_k \mathbf{x}_k]$ . Note that  $\theta_N = Y$  if  $\hat{Y}(\tau)$  is unbiased for  $Y$  under the assumed response mechanism and the design.

Suppose now that the parameter of interest is  $\theta = E_m(Y)$ . Let  $\mathbf{d}_k = (d_{1k}, d_{2k}, d_{3k})^T$  with  $d_{1k} = d_k$ ,  $d_{2k} = d_k o_k$  and  $d_{3k} = d_k o_k y_k$  then it follows from (3.1) that  $\hat{Y}(\tau)$  is of the form  $f(\mathbf{A}_d)$  where  $\mathbf{A}_d$  is a  $3 \times N$  matrix with  $k^{\text{th}}$  column  $\mathbf{d}_k$ . A variance estimator of  $\hat{Y}(\tau)$  is given by (2.2) where  $\mathcal{G}(\mathbf{u})$  denotes now a variance estimator of the linear combination  $\hat{U} = \sum \mathbf{u}_k^T \mathbf{d}_k$  and  $\mathbf{u}_k = (u_{1k}, u_{2k}, u_{3k})^T$ . We have

$$\begin{aligned} \mathcal{G}(\mathbf{u}) &= \sum \sum d_k d_t \omega_{kt} o_k o_t [(\hat{\xi}_k \hat{\xi}_t / \hat{\xi}_{kt})] \text{cov}_m(y_k, y_t) u_{k:m} u_{t:m} \\ &\quad + \sum \sum d_k d_t \omega_{kt} o_k o_t [(\hat{\xi}_{kt} - \hat{\xi}_k \hat{\xi}_t) / \hat{\xi}_{kt}] u_{k:o} u_{t:o} \\ &\quad + \sum \sum d_k d_t (1 - \omega_{kt}) u_{k:s}^* u_{t:s}^*, \end{aligned} \quad (3.4)$$

where  $u_{k:m} = u_{3k}$ ,  $u_{k:o} = u_{2k} + y_k u_{3k}$ ,  $u_{k:s}^* = u_{1k} + o_k u_{k:o}$ , and  $\text{cov}_m(y_k, y_t)$  is an estimator of the model covariance of  $y_k$  and  $y_t$  under the assumed model on  $y$ . When the model covariance of  $y_k$  and  $y_t$  is zero,  $\text{cov}_m(y_k, y_t)$  is taken as zero. The DR

variance estimator of  $\hat{Y}(\tau)$  is given by (3.4) with  $\mathbf{u}_k$  replaced by  $\mathbf{z}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d} = \mathbf{g}_k^*(0, -\hat{y}_k^*, 1)^T$ , where  $\mathbf{A}_b$  is a  $3 \times N$  matrix of arbitrary real numbers with  $k^{\text{th}}$  column  $\mathbf{b}_k$ . We have  $z_{k,m} = g_k^*$ ,  $z_{k,o} = g_k^*(y_k - \hat{y}_k^*)$ , and  $z_{k,s} = o_k g_k^*(y_k - \hat{y}_k^*)$ . If  $\text{Cov}_m(y_k, y_i)$  is estimated by  $\text{cov}_m(y_k, y_i) = (y_k - \hat{y}_k^*)(y_i - \hat{y}_i^*)$  then  $\mathcal{G}_{DR}(\hat{Y}(\tau))$  reduces to

$$\mathcal{G}_{DR}(\hat{Y}(\tau)) = \sum \sum d_k d_i \omega_{kt} o_k o_i z_{k,o} z_{i,o} + \sum \sum d_k d_i (1 - \omega_{kt}) z_{k,s} z_{i,s}. \quad (3.5)$$

It is interesting to note that the probability of response is absent from (3.5) when the imputation method provides an unbiased estimator of the model mean.

If the order of expectation can be interchanged so that  $E_m E_r E_p = E_p E_r E_m$ , then a variance estimator,  $\mathcal{G}^*(u)$ , of the linear combination  $\sum \mathbf{u}_k^T \mathbf{d}_k$  is given by

$$\begin{aligned} \mathcal{G}^*(u) &= \sum \sum d_k d_i o_k o_i \text{cov}_m(y_k, y_i) u_{k,m} u_{i,m} \\ &\quad + \sum \sum d_k d_i \omega_{kt} o_k o_i [(\hat{\xi}_{kt} - \hat{\xi}_k \hat{\xi}_i) / \hat{\xi}_{kt}] u_{k,o}^* u_{i,o}^* \\ &\quad + \sum \sum d_k d_i (1 - \omega_{kt}) u_{k,s}^* u_{i,s}^*, \end{aligned} \quad (3.6)$$

where  $u_{k,o}^* = u_{2k} + \hat{E}_m(y_k) u_{3k}$ , and  $u_{k,s}^* = u_{1k} + o_k u_{k,o}^*$ . Substituting  $\mathbf{z}_k = \mathbf{g}_k^*(0, -\hat{y}_k^*, 1)^T$  into  $\mathbf{u}_k$  in (3.6), the new variance estimator  $\mathcal{G}_{DR}^*(\hat{Y}(\tau))$  reduces to

$$\mathcal{G}_{DR}^*(\hat{Y}(\tau)) = \sum \sum d_k d_i o_k o_i \text{cov}_y(y_k, y_i) z_{k,m} z_{i,m}, \quad (3.7)$$

since  $z_{k,o}^* = 0$  and  $z_{k,s}^* = 0$  when the imputation method provides an unbiased estimator of the model mean. Note, again, that the probability of response is absent from (3.7).

We now consider the case of  $\tau_k = \pi_k^{-1}$  and assume the order of expectation can be interchanged so that  $E_p E_r = E_r E_p$ . The expectation of  $\hat{Y}(\pi^{-1})$  with respect to the sampling design and response mechanism is approximately given by,

$$E_r E_p \hat{Y}(\pi^{-1}) \approx \sum \xi_k y_k + \sum (1 - \xi_k) y_k^*, \quad (3.8)$$

where  $y_k^* = E_r E_p(\hat{y}_k^*)$ , and  $\xi_k = E_r(o_k)$ . Then, one may adjust the estimator  $\hat{Y}(\pi^{-1})$  to get a design-response-based estimator of the finite population total  $Y$  as

$$\hat{Y}_\xi = \sum d_k (o_k / \hat{\xi}_k) y_k + \sum (1 - d_k (o_k / \hat{\xi}_k)) \hat{y}_k^*, \quad (3.9)$$

with  $\hat{\xi}_k = \xi_k(\hat{\alpha})$ ,  $\hat{y}_{2k}^* = \mathbf{x}_k^T \hat{\beta}_\xi$ ,  $\hat{\beta}_\xi = \hat{\mathbf{Q}}_\xi^{-1} \sum d_k (o_k / \hat{\xi}_k) I_k c_k \mathbf{x}_k y_k$ ,  $\hat{\mathbf{Q}}_\xi = \sum d_k (o_k / \hat{\xi}_k) I_k c_k \mathbf{x}_k \mathbf{x}_k^T$ , and  $\hat{\alpha}$  in  $\xi_k(\hat{\alpha})$  is the solution to the following set of estimating equations  $\hat{\mathbf{I}}(\alpha) = \sum d_k \mathbf{h}_k (o_k - \xi_k) = 0$ , where  $\xi_k = \xi_k(\alpha)$ ,  $\alpha$  denotes the parameter of the model on  $\xi_k$  and  $\mathbf{h}_k$  is the predictor variable. The estimator (3.9) is design-consistent under assumed response mechanism.

Suppose first that the parameter of interest is  $\theta_N = Y$ , then let  $\mathbf{d}_k = (d_{1k}, d_{2k})^T$  with  $d_{1k} = d_k$  and  $d_{2k} = d_k o_k$ . A variance estimator of  $\hat{Y}_\xi$  is given by (3.2) with  $\mathbf{u}_k$  replaced by

$$\mathbf{z}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d} = \begin{cases} \hat{\mathbf{J}}_0(\hat{\alpha}) [\hat{\mathbf{J}}(\hat{\alpha})]^{-1} \mathbf{h}_k \hat{\xi}_k \\ (1 / \hat{\xi}_k) (y_k - \hat{y}_k^*) \mathbf{g}_k - \hat{\mathbf{J}}_0(\hat{\alpha}) [\hat{\mathbf{J}}(\hat{\alpha})]^{-1} \mathbf{h}_k, \end{cases} \quad (3.10)$$

where  $\mathbf{g}_k = (1 - I_k) + I_k [1 + \sum (1 - d_i (o_i / \hat{\xi}_i)) I_i \mathbf{x}_i^T \hat{\mathbf{Q}}_\xi^{-1} c_i \mathbf{x}_i]$ ,  $\hat{\mathbf{J}}_0(\alpha) = -\partial \hat{Y}_\xi(\alpha) / \partial \alpha$  and  $\hat{\mathbf{J}}(\alpha) = -\partial \hat{\mathbf{I}}(\alpha) / \partial \alpha$ .

Suppose the parameter of interest is  $\theta = E_m(Y)$ , then let  $\mathbf{d}_k = (d_{1k}, d_{2k}, d_{3k})^T$  with  $d_{1k} = d_k$ ,  $d_{2k} = d_k o_k$  and  $d_{3k} = d_k o_k y_k$ . A variance estimator of  $\hat{Y}_\xi$  is given by (3.4) with  $\mathbf{u}_k$  replaced by

$$\mathbf{z}_k = \partial f(\mathbf{A}_b) / \partial \mathbf{b}_k |_{\mathbf{A}_b = \mathbf{A}_d} = \begin{cases} \hat{\mathbf{J}}_0(\hat{\alpha}) [\hat{\mathbf{J}}(\hat{\alpha})]^{-1} \mathbf{h}_k \hat{\xi}_k \\ -\hat{y}_k^* \mathbf{g}_k / \hat{\xi}_k - \hat{\mathbf{J}}_0(\hat{\alpha}) [\hat{\mathbf{J}}(\hat{\alpha})]^{-1} \mathbf{h}_k \\ \mathbf{g}_k / \hat{\xi}_k. \end{cases} \quad (3.11)$$

When  $z_k$  is substitute for  $u_k$  in (3.4) or in (3.6), the resulting variance estimator for  $\hat{Y}_\xi$  depends on the probability of response even when the imputation method provides an unbiased estimator of the model mean, whereas (3.5) and (3.7) do not for  $\hat{Y}(\tau)$ .

### CONCLUDING REMARKS

Extension of our results to two-phase sampling estimator of the form  $\hat{Y} = \sum a_k \tau_k y_k + \sum (a_k^{(1)} \tau_k^{(1)} - a_k \tau_k) \hat{y}_k^*$  is not included due to page limit, where  $a_k^{(1)}$  is the first-phase sample membership indicator variable,  $a_k$  is the two-phase sample membership indicator variable,  $(\tau_k^{(1)}, \tau_k) = (1, a_k^{(1)}) / \pi_k^{(1)}$  in case of the naïve estimator,  $(\tau_k^{(1)}, \tau_k) = (1 / \pi_k^{(1)}, 1 / \pi_k)$  for the design-based estimator and  $(\pi_k^{(1)}, \pi_k)$  denote the first and second phase selection probabilities respectively.

### REFERENCES

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*, Montréal, Canada, 916-925.

Valliant R., Dorfman, A.H. and Royall, R.M. (2000), “*Finite Population Sampling and Inference: A Prediction Approach*”, Wiley.

Table 1: Simulation Results

Imputation Method	Estimator	Variable	$RB(\hat{\theta})$	$MSE(\hat{Y}^{(N)}) / MSE(\hat{Y})$	$RB\{g(\hat{\theta})\}$
Substitution	Naïve	1	-.003	.11	-.92
		2	-.14	109.72	-.99
		3	-.16	119.01	-.99
	Design-based	1	.0001	1	.03
		2	-.0000041	1	.02
		3	-.0001	1	.008
Ratio	Naïve	1	-.001	.78	-.001
		2	-.001	.86	-.02
		3	-.001	.83	-.13
	Design-based	1	.0001	1	.02
		2	.0001	1	.01
		3	-.00006	1	-.001