ON ESTIMATION OF POPULATION TOTAL USING GENERALIZED REGRESSION PREDICTOR

Raghunath Arnab[1] and Sarjinder Singh[2]

1. Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban-4000, South Africa **E-mail:** arnab@pixie.udw.ac.za
2. Department of Statistics, St. Cloud State University, 720 Fourth Avenue South, St. Cloud, MN 56301-4498, USA.
E-mail: sarjinder@yahoo.com

SUMMARY

The generalized regression predictor (greg) is used for the estimation of a finite population total when the study variable is well related to the auxiliary variable. Chaudhuri and Roy (1997) provided the lower bound of the mean square error (mse) of variance estimators belonging to a class of non-homogeneous quadratic unbiased estimators. They also found the optimum variance estimator whose mean square error attains the lower bound. In the present paper, we have shown that the derivation of the lower bound in Chaudhuri and Roy (1997)'s paper is incorrect and their proposed optimal estimator does not attain the lower bound as originally claimed. An example is also provided which contradict the result of Chaudhuri and Roy (1997). Model assisted higher order calibration approach has been proposed to investigate the variance of the regression predictor.

Key words: Generalized regression predictor, Auxiliary information, Estimation of total/variance, Optimality

1. INTRODUCTION

The use of auxiliary information in survey sampling plays an eminent role in both the estimation and selection stages. Typical uses are its incorporation at the estimation stage through the use of regression, ratio or product estimators when the study variable y is well related to the auxiliary variable x, which is assumed to be positive. Särndal (1982), Särndal, Swensson and Wretman (1992) recommended the use of the generalized regression predictor (greg) for estimation of the finite population total Y. Almost complete review can be had from Singh (2003). It is well known that the greg predictor is asymptotically design unbiased (ADU) for Y under the Brewer (1979) approach, irrespective of the validity of any model. Several authors including Liu (1974), Särndal (1982), Kott (1990), Särndal (1996) and Zou (1999) proposed variance estimators for the greg to facilitate the estimation of confidence interval for the population total Y. Chaudhuri and Roy (1997) pointed out that although the variance estimators are ADU, under large samples, but little is known about their efficiencies. So, Chaudhuri and Roy (1997) provided the lower bound of the variance estimators belonging to the class of non-homogenous quadratic unbiased estimators for the population total under a certain superpopulation model. They found that the optimal estimator attains the lower bound. The proposed optimum estimator cannot be used in practice since it involves several unknown model parameters. Hence, they modified the optimum estimator by replacing the model parameters by their estimates. In this paper, we show that the derivation of the lower bound of mean square error, presented by Chaudhuri and Roy (1997), is incorrect. Hence, their optimum estimator does not attain the lower bound. So, in our present investigation, we have proposed some alternative estimators by using (i) the calibration approach under a linear superpopulation model passing through the origin and (ii) known population variance of the auxiliary variable x. The efficiencies of the proposed estimators are compared with the existing alternatives by appropriate simulation techniques. Empirical investigations reveal that some of the proposed estimators fare better than the existing alternatives.

1.1 NOTATION AND PRELIMINARIES

Consider a finite population $U=\{1,...,i,...N\}$ of N identifiable units. Let $y_i(x_i)$ be the value of the study (auxiliary) variable of the ith unit of the population. The values of y_i's are unknown before survey but the values of x_i's are assumed to be known and positive. Here we consider the problem of estimation of the finite population total $Y=\sum_{i\in U} y_i$ using a sample s selected by a fixed effective size sampling design p. The inclusion probabilities of units i and the pair of units $i \neq j$ are denoted respectively by π_i and π_{ij}, and assumed to be positive. Chaudhuri and Roy (1997) considered the following superpopulation model

$$\text{Model M: } y_i=\beta \; x_i+\in_i,\; i\in U \tag{1.1}$$

where β is an unknown constant, $\in_i$'s are error component, independently distributed with $E_m(\in_i)=0$ and $V_m(\in_i)=\sigma_i^2, (\sigma_i^2>0, \text{unknown})$. Here E_m, V_m denote respectively expectation and variance with respect to the superpopulation model M. Chaudhuri and Roy (1997) considered the generalized regression predictor (greg) for Y

$$t_g=\sum_{i\in s}\frac{y_i}{\pi_i}+\hat{\beta}_Q(X-\sum_{i\in s}\frac{x_i}{\pi_i}) \tag{1.2}$$

where $I_{si}=1$ if $i\in s$ and 0 if $i\notin s$, $\hat{\beta}_Q=\frac{\sum_{i\in U} Q_i x_i y_i I_{si}}{\sum_{i\in U} Q_i x_i^2 I_{si}}$,

$Q_i(>0)$'s are suitably chosen constants and $X=\sum_{i\in U} x_i$.

The approximate expression for the variance of t_g provided by Sarndal (1982) is given by

$$V_p(t_g)=\sum_{i\neq}\sum_{j\in U}(\frac{E_i}{\pi_i}-\frac{E_j}{\pi_j})^2\Delta_{ij}\pi_{ij} \;=V(y) \text{ (say)}$$

where $E_i = y_i - \beta_Q x_i$, $\beta_Q = \frac{\sum_{i\in U} Q_i x_i y_i \pi_i}{\sum_{i\in U} Q_i x_i^2 \pi_i}$, $\Delta_{ij} = \frac{\pi_i \pi_j - \pi_{ij}}{2\pi_{ij}}$ and V_p denotes variance with respect to the sampling design p.

Chaudhuri and Roy (1997) have given an alternative expression for $V(y)$ as

$$V(y) = \sum_{i\in U} \alpha_i y_i^2 + \sum_{i\neq}\sum_{j\in U} \alpha_{ij} y_i y_j = V \text{ (say)} \qquad (1.3)$$

where

$$\alpha_i = \left(\frac{1}{\pi_i} - 1\right) + \left(Q_i^2 x_i^2 \pi_i^2\right) \frac{V_p\left(\sum_{i\in U} \frac{x_i}{\pi_i} I_{si}\right)}{\left(\sum_{i\in U} Q_i x_i^2 \pi_i\right)^2} - 2Q_i x_i \pi_i \frac{\sum_{k\in U} x_k \left(\frac{\pi_{ik}}{\pi_i \pi_k} - 1\right)}{\left(\sum_{i\in U} Q_i x_i^2 \pi_i\right)}$$

$$\alpha_{ij} = \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) + \left(Q_i Q_j \pi_i \pi_j x_i x_j\right) \frac{V_p\left(\sum_{i\in U} \frac{x_i}{\pi_i} I_{si}\right)}{\left(\sum_{i\in U} Q_i x_i^2 \pi_i\right)^2}$$

$$- Q_j x_j \pi_j \frac{\sum_{k\in U} x_k \left(\frac{\pi_{ik}}{\pi_i \pi_k} - 1\right)}{\left(\sum_{i\in U} Q_i x_i^2 \pi_i\right)} - Q_i x_i \pi_i \frac{\sum_{k\in U} x_k \left(\frac{\pi_{jk}}{\pi_j \pi_k} - 1\right)}{\left(\sum_{i\in U} Q_i x_i^2 \pi_i\right)}$$

and $\pi_{ii} = \pi_i$. Chaudhuri and Roy (1997) considered the class H , of non-homogeneous quadratic unbiased estimator for $V(y)$ of the form

$$v = v(y) = a_s + \sum_{i\in U} b_{si} y_i^2 I_{si} + \sum_{i\neq}\sum_{j\in U} b_{sij} y_i y_j I_{sij} \qquad (1.4)$$

where a_s, b_{si} and b_{sij} are constants free from y_i 's and satisfy the unbiasedness conditions as follows:
$E_p(a_s) = 0$, $E_p(b_{si} I_{si}) = \alpha_i$ and $E_p(b_{sij} I_{sij}) = \alpha_{ij}$ $\forall i \neq j$.
Here E_p denotes expectation with respect to the sampling design p. Chaudhuri and Roy (1997) derived the lower bound of the variance of an estimator belonging to the class H under the super-population model M given in (1.1) and proved that the following optimal estimator

$$v_0 = \sum_{i\in U} \alpha_i \left(y_i^2 - \sigma_i^2 - \mu_i^2\right)\frac{I_{si}}{\pi_i} + \sum_{i\neq j}\sum \alpha_{ij}\left(y_i y_j - \mu_i \mu_j\right)\frac{I_{sij}}{\pi_{ij}} + \sum_{i\in U} \alpha_i \left(\sigma_i^2 + \mu_i^2\right) + \sum_{i\neq}\sum_{j\in U} \alpha_{ij} \mu_i \mu_j \qquad (1.5)$$

(where $\mu_i = E_m(y_i) = \beta x_i$) attains the lower bound.

(In the expression of v_0 of the Theorem 1, page143, of Chaudhuri and Roy (1997)'s paper, " $\sum_{i\in U} \alpha_i \left(\sigma_i^2 + \mu_i^2\right)$" was wrongly written as " $\sum_{i\in U} \alpha_i^2 \left(\sigma_i^2 + \mu_i^2\right)$". We presume it is simply a typographical error). In the present note we have shown that the result concerning the lower bound of the variance estimators provided by Chaudhuri and Roy (1997) is incorrect and also the proposed optimal estimator v_0 does not attain the lower bound as claimed by the authors. The estimator v_0 at (1.5) can not be used in practice since it involves unknown parameters μ_i 's and σ_i^2 's. So, Chaudhuri and Roy proposed the following alternative estimators when $\mu_i = \beta x_i$ and $\sigma_i^2 = \sigma^2 x_i^g$ by replacing β and σ^2 with their suitable estimators as follows:

$$\hat{v}_1^* = \sum_i \alpha_i \left(y_i^2 - \hat{\theta} x_i^2 - \hat{\phi} x_i^g\right)\frac{I_{si}}{\pi_i} + \hat{\phi}\sum_i \alpha_i x_i^g + \hat{\theta}\sum_i \alpha_i x_i^2 + \sum_{i\neq j}\sum \alpha_{ij}\left(y_i y_j - \hat{\theta} x_i x_j\right)\frac{I_{sij}}{\pi_{ij}} + \hat{\theta}\sum_{i\neq j}\sum \alpha_{ij} x_i x_j \qquad (1.6)$$

and

$$\hat{v}_2^* = \left(\sum_i \alpha_i x_i^2\right)\left[\left(\frac{\sum \alpha_i y_i^2 (I_{si}/\pi_i)}{\sum \alpha_i x_i^2 (I_{si}/\pi_i)}\right) - \hat{\phi}\left(\frac{\sum \alpha_i x_i^g (I_{si}/\pi_i)}{\sum \alpha_i x_i^2 (I_{si}/\pi_i)}\right)\right] + \left(\sum_{i\neq j}\sum \alpha_{ij} x_i x_j\right)\left(\frac{\sum_{i\neq j}\sum \alpha_{ij} y_i y_j (I_{sij}/\pi_{ij})}{\sum_{i\neq j}\sum \alpha_{ij} x_i x_j (I_{sij}/\pi_{ij})}\right) + \hat{\phi}\sum_i \alpha_i x_i^g \qquad (1.7)$$

where

$$\hat{\theta} = \left(\frac{\sum_{i\in s} x_i^{1-g} y_i}{\sum_{i\in s} x_i^{2-g}}\right)^2 - \frac{1}{(n-1)}\left[\frac{\sum_{i\in s} \frac{y_i^2}{x_i^g}}{\sum_{i\in s} x_i^{2-g}} - \left(\frac{\sum_{i\in s} x_i^{1-g} y_i}{\sum_{i\in s} x_i^{2-g}}\right)^2\right],$$

$$\hat{\phi} = \frac{1}{(n-1)}\left[\sum_{i\in s} \frac{y_i^2}{x_i^g} - \frac{\left(\sum_{i\in s} x_i^{1-g} y_i\right)^2}{\sum_{i\in s} x_i^{2-g}}\right].$$

2. CHAUDHURI AND ROY'S THEOREMS

The lower bound of the estimator of variance given in Theorem 1 (page 143) by Chaudhuri and Roy (1997) has been restated in the following theorem:

Theorem 1. (Chaudhuri and Roy, 1997) Under model M, and $v \in H$

$$M(v) = E_m E_p (v - V(y))^2 \geq \sum_{i\in U} \alpha_i^2 \left(\frac{1}{\pi_i} - 1\right)\eta_i^2 + \sum_{i\neq}\sum_{j\in U} \alpha_{ij}^2 \left(\frac{1}{\pi_{ij}} - 1\right)\eta_{ij}$$

$$= M_0 \qquad (2.1)$$

where $\eta_i^2 = \delta_i - \left(\sigma_i^2 + \mu_i^2\right)^2$, $\eta_{ij} = \left(\sigma_i^2 + \mu_i^2\right)\left(\sigma_j^2 + \mu_j^2\right) - \mu_i^2 \mu_j^2$ and $\mu_i = E_m(y_i) = \beta x_i$.

The equality is attained in the above if v equals

$$v_0 = \sum_{i\in U} \alpha_i \left(y_i^2 - \sigma_i^2 - \mu_i^2\right)\frac{I_{si}}{\pi_i} + \sum_{i\neq}\sum_{j\in U} \alpha_{ij}\left(y_i y_j - \mu_i \mu_j\right)\frac{I_{sij}}{\pi_{ij}} + \sum_{i\in U} \alpha_i \left(\sigma_i^2 + \mu_i^2\right) + \sum_{i\neq}\sum_{j\in U} \alpha_{ij} \mu_i \mu_j \qquad (2.2)$$

In the following theorem we will show that v_0 does not attain the lower bound M_0, given in (2.1).

Theorem 2.1. The correct expression for the expected variance of v_0 is given by

$$M(v_0) = E_m E_p (v_0 - V(y))^2$$
$$= \sum_{i \in U} \alpha_i^2 \eta_i^2 (d_i - 1) + 2 \sum_{i \neq j \in U} \sum \alpha_{ij}^2 \eta_{ij} (d_{ij} - 1)$$
$$+ 4 \sum\sum\sum_{i \neq j \neq k \in U} \sigma_i^2 \alpha_{ij} \alpha_{ik} \mu_j \mu_k (d_{ij} d_{ik} \pi_{ijk} - 1)$$
$$+ 4 \sum\sum_{i \neq j \in U} \alpha_i \alpha_{ij} \mu_j \left\{ \gamma_i - \mu_i (\sigma_i^2 + \mu_i^2) \right\} (d_i - 1) \quad (2.3)$$

where $\gamma_i = E_m(y_i^3) < \infty$.

Proof. Obvious by following Arnab and Singh (2002).

Theorem 2.2. $M(v)$ can not have lower bound given by Chaudhuri and Roy (1997) as follows:

$$\sum_{i \in U} \alpha_i^2 \left(\frac{1}{\pi_i} - 1 \right) \eta_i^2 + \sum_{i \neq} \sum_{j \in U} \alpha_{ij}^2 \left(\frac{1}{\pi_{ij}} - 1 \right) \eta_{ij} \quad (2.4)$$

Proof. Obvious by following Arnab and Singh (2002).

3. CALIBRATED ESTIMATORS OF VARIANCE

The Horvitz -Thompson type estimator of the variance $V(y)$ of the greg predictor is given by

$$\hat{v}_{ht}(y) = \sum_i \alpha_i d_i y_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} d_{ij} y_i y_j I_{sij} \quad (3.1)$$

where $d_i = 1/\pi_i$ and $d_{ij} = 1/\pi_{ij}$. Now we consider a calibrated estimator of variance of regression predictor as

$$\hat{v}_c(y) = \sum_i \alpha_i w_i y_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} w_{ij} y_i y_j I_{sij} \quad (3.2)$$

where w_i and w_{ij} are calibration weights obtained by minimizing a distance function

$$D_{sp} = \sum_i \frac{(w_i \alpha_i - d_i \alpha_i)^2}{d_i \alpha_i q_i} I_{si} + \sum_{i \neq} \sum_j \frac{(\alpha_{ij} w_{ij} - d_{ij} \alpha_{ij})^2}{d_{ij} \alpha_{ij} q_{ij}} I_{sij} \quad (3.3)$$

subject to various calibration constraints given below; q_i and q_{ij} are suitably chosen weights to form different kinds of variance estimators.

Case 1: Here we choose the calibration constraints as

$$E_m[\hat{v}_c(y)] = E_m[V(y)] \quad (3.4)$$

or equivalently

$$\sum_i \alpha_i w_i E_m(y_i^2) I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} w_{ij} E_m(y_i y_j) I_{sij} = \sum_i \alpha_i E_m(y_i^2) + \sum_{i \neq} \sum_j \alpha_{ij} E_m(y_i y_j) \quad (3.5)$$

For the superpopulation model (1.1) with $\sigma_i^2 = \sigma^2 x_i^g$ ($\sigma^2 > 0$ and $g \geq 0$),

$$V(x) = \sum_i \alpha_i x_i^2 + \sum_{i \neq} \sum_j \alpha_{ij} x_i x_j = 0 \quad (3.6)$$

and hence (3.5) reduces to

$$\sigma^2 \sum_i \alpha_i w_i x_i^g I_{si} + \beta^2 \left\{ \sum_i \alpha_i w_i x_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} w_{ij} x_i x_j I_{sij} \right\} = \sigma^2 \sum_i \alpha_i x_i^g \quad (3.7)$$

On comparing the coefficients of σ^2 and β^2 on both sides of (3.7), the system of calibration equations becomes

$$\sum_i \alpha_i w_i x_i^2 + \sum_{i \neq} \sum_j \alpha_{ij} w_{ij} x_i x_j I_{sij} = 0 \;\&\; \sum_i \alpha_i w_i x_i^g I_{si} = \sum_i \alpha_i x_i^g \quad (3.8)$$

In order to minimize (3.3) subject to (3.8), consider

$$\phi = \sum_i \frac{(w_i \alpha_i - d_i \alpha_i)^2}{d_i \alpha_i q_i} I_{si} + \sum_{i \neq} \sum_j \frac{(\alpha_{ij} w_{ij} - d_{ij} \alpha_{ij})^2}{d_{ij} \alpha_{ij} q_{ij}} I_{sij}$$
$$- 2\lambda \left\{ \sum_i \alpha_i w_i x_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} w_{ij} x_i x_j I_{sij} \right\} - 2\mu \left\{ \sum_i \alpha_i w_i x_i^g I_{si} \right\} \quad (3.9)$$

Now

$$\frac{\partial \phi}{\partial w_i} = 0 \Rightarrow w_i \alpha_i = d_i \alpha_i + d_i \alpha_i q_i (\lambda x_i^2 + \mu x_i^g) \quad (3.10)$$

and

$$\frac{\partial \phi}{\partial w_{ij}} = 0 \Rightarrow w_{ij} \alpha_{ij} = d_{ij} \alpha_{ij} + \lambda d_{ij} \alpha_{ij} q_{ij} x_i x_j. \quad (3.11)$$

On substituting (3.10) and (3.11) in (3.8) we have

$$\lambda = \frac{C\Delta_1 - B\Delta_2}{AC - B^2} \text{ and } \mu = \frac{A\Delta_2 - B\Delta_1}{AC - B^2}. \quad (3.12)$$

were

$A = \sum_i \alpha_i d_i q_i x_i^4 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} d_{ij} q_{ij} x_i^2 x_j^2 I_{sij}$; $B = \sum_i \alpha_i d_i q_i x_i^{g+2} I_{si}$;

$C = \sum_i \alpha_i d_i q_i x_i^{2g} I_{si}$; $\Delta_1 = -\hat{v}_{ht}(x)$; $\Delta_2 = \sum_i \alpha_i x_i^g - \sum_i \alpha_i d_i x_i^g I_{si}$

and $\hat{v}_{ht}(x) = \sum_i \alpha_i d_i x_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} d_{ij} x_i x_j I_{sij}$.

Substituting (3.12) in (3.10) and (3.11) we get

$$w_i = d_i + \frac{C\Delta_1 - B\Delta_2}{AC - B^2} d_i q_i x_i^2 + \frac{A\Delta_2 - B\Delta_1}{AC - B^2} d_i q_i x_i^g = w_i(1) \quad (3.13)$$

and

$$w_{ij} = d_{ij} + \frac{C\Delta_1 - B\Delta_2}{AC - B^2} d_{ij} q_{ij} x_i x_j = w_i(2) \text{ (say)} \quad (3.14)$$

Finally putting (3.13) and (3.14) in (3.2), we get

$$\hat{v}_c(y) = \hat{v}_{ht}(y) - \hat{\beta}^2 \hat{v}_{ht}(x) + \hat{\sigma}^2 \left(\sum_i \alpha_i x_i^g - \sum_i \alpha_i d_i x_i^g I_{si} \right) = \hat{v}_c(1) \quad (3.15)$$

where

$\hat{\beta}^2 = \frac{CP - BQ}{AC - B^2}$ and $\hat{\sigma}^2 = \frac{AQ - BP}{AC - B^2}$ with $Q = \sum_i \alpha_i d_i q_i x_i^g y_i^2 I_{si}$.

and $P = \sum_i \alpha_i d_i q_i x_i^2 y_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} d_{ij} q_{ij} x_i x_j y_i y_j I_{sij}$. Here we note that $\hat{\beta}^2$ and $\hat{\sigma}^2$ are model unbiased estimator for β^2 and σ^2 respectively.

Case II. Using the relation $V(x) = \sum_i \alpha_i x_i^2 + \sum_{i \neq} \sum_j \alpha_{ij} x_i x_j = 0$, we set the calibration constrain assuming $T_1(x)$ is known as:

(i.) $\sum_i w_i \alpha_i x_i^2 I_{si} = \sum_i \alpha_i x_i^2 = T_1(x)$

and (3.16)

(ii) $\sum_{i \neq} \sum_j \alpha_{ij} w_{ij} x_i x_j I_{sij} = \sum_{i \neq} \sum_j \alpha_{ij} x_i x_j = T_2(x) = - T_1(x)$

The equation (3.16) satisfies

$$\hat{v}_c(x) = \sum_i \alpha_i w_i x_i^2 I_{si} + \sum_{i \neq} \sum_j \alpha_{ij} w_{ij} x_i x_j I_{sij} = V(x) = 0$$

Minimization of (3.3) subject to (3.16) yields calibrated weights

$$w_i = d_i + \frac{d_i q_i x_i^2}{\sum_i d_i \alpha_i q_i x_i^4 I_{si}} \left[\sum_i \alpha_i x_i^2 - \sum_i d_i x_i^2 I_{si} \right] = w_i(2) \qquad (3.17)$$

and

$$w_{ij} = d_{ij} + \frac{d_{ij} q_{ij} x_i x_j}{\sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{ij} x_i^2 x_j^2 I_{sij}} \left[\sum\sum_{i \neq j} \alpha_{ij} x_i x_j - \sum\sum_{i \neq j} d_{ij} x_i x_j I_{sij} \right] \qquad (3.18)$$

$$= w_{sij}(2)$$

On putting the values of $w_i(2)$ and $w_{ij}(2)$ in (3.2.), we get an alternative calibrated estimator of the variance of the regression predictor as

$$\hat{v}_c(2) = \hat{v}_{ht}(y) + b_1 \left[T_1(x) - \sum_i d_i \alpha_i x_i^2 I_{si} \right] + b_2 \left[T_2(x) - \sum\sum_{i \neq j} d_{ij} \alpha_{ij} x_i x_j I_{sij} \right] \qquad (3.19)$$

where

$$b_1 = \frac{\sum_i d_i \alpha_i q_i x_i^2 y_i^2 I_{si}}{\sum_i d_i \alpha_i q_i x_i^4 I_{si}} \text{ and } b_2 = \frac{\sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{ij} x_i x_j y_i y_j I_{sij}}{\sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{ij} x_i^2 x_j^2 I_{sij}}.$$

Case III. Here we assume that $T_1(x)$ is unknown and use a constraint of calibration

$$\hat{v}_c(x) = \sum_i \alpha_i w_i x_i^2 I_{si} + \sum\sum_{i \neq j} \alpha_{ij} w_{ij} x_i x_j I_{sij} = V(x) = 0 \qquad (3.20)$$

In this situation calibration weights obtained by minimizing (3.3) subject to (3.15) come out as

$$w_i = d_i + \frac{d_i q_i x_i^2 \{ -\hat{v}_{ht}(x) \}}{\sum_i d_i \alpha_i q_i x_i^4 I_{si} + \sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{ij} x_i^2 x_j^2 I_{sij}} = w_i(3) \qquad (3.21)$$

and

$$w_{ij} = d_{ij} + \frac{d_{ij} q_{ij} x_i x_j \{ -\hat{v}_{ht}(x) \}}{\sum_i d_i \alpha_i q_i x_i^4 I_{si} + \sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{ij} x_i^2 x_j^2 I_{sij}} = w_{ij}(3) \qquad (3.22)$$

(where $\hat{v}_{ht}(x) = \sum_i \alpha_i d_i x_i^2 I_{si} + \sum\sum_{i \neq j} \alpha_{ij} d_{ij} x_i x_j I_{sij}$). The resultant calibrated estimator of variance of the regression predictor is

$$\hat{v}_c(3) = \hat{v}_{ht}(y) + \left[\frac{\sum_i d_i \alpha_i q_i y_i^2 x_i^2 I_{si} + \sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{sij} x_i x_j y_i y_j I_{sij}}{\sum_i d_i \alpha_i q_i x_i^4 I_{si} + \sum\sum_{i \neq j} d_{ij} \alpha_{ij} q_{ij} x_i^2 x_j^2 I_{sij}} \right] \left(-\hat{v}_{ht}(x) \right) \qquad (3.23)$$

4. SIMULATION STUDIES

In this section, we present results of simulation studies to compare performances of the proposed estimators $\hat{v}_c(1), \hat{v}_c(2)$ and $\hat{v}_c(3)$ of the variance of the generalized regression predictor $\hat{Y}_g$ with the conventional estimator $\hat{v}_{ht}(y)$ (given in (3.1)) and $\hat{v}_1^*$ and $\hat{v}_2^*$ proposed by Chaudhuri and Roy (1997). It should be noted that the estimators $\hat{v}_c(1)$, and $\hat{v}_1^*$ are of the similar form where $\hat{\beta}^2$, $\hat{\theta}$ and $\hat{\sigma}^2$, $\hat{\phi}$ are respectively model unbiased estimators for β^2 and σ^2. Both the estimators $\hat{\beta}^2$ and $\hat{\theta}$ involve g which is generally unknown. So, we propose the following alternative variance estimator:

$$\hat{v}_c(4) = \hat{v}_{ht}(y) - b\hat{v}_{ht}(x) + \hat{\phi}\left(\sum_i \alpha_i x_i^g - \sum_i \alpha_i d_i x_i^g I_{si}\right) \qquad (4.1)$$

where $b = \dfrac{\sum\sum_{i \neq j} \alpha_{ij} y_i y_j d_{ij} I_{sij}}{\sum\sum_{i \neq j} \alpha_{ij} x_i x_j d_{ij} I_{sij}}$ is model unbiased for β^2 and free of g, and $\hat{\phi}$ is as given in (1.7). For the present simulation studies, we generate three populations, each of size 200 (= N). First we select x_i's ($i = 1,\ldots, 200$) as a random sample from a gamma population with parameters $\alpha = 15$ and $\beta = 1$ (Mathematika with seed no: 19491000). From each x_i, we generate y_i using the model: $y_i = \beta x_i + \epsilon_i$ for $i = 1,\ldots,200$ where for a given x_i, ϵ_i is a random sample selected independently from a normal population with mean zero and variance $\sigma^2 x_i^g$. Three populations viz. Population 1, Population 2 and Population 3 are generated with $\beta = 8,\ \sigma = 2$ and $g = 1.2$; $\beta = 4,\ \sigma = 1$ and $g = 1.5$ and $\beta = 4,\ \sigma = 1$ and $g = 1.8$, respectively. From each of the three populations, we draw two sets of R (=2000) independent samples, each of sizes 25 and 40 following (i) Simple random sampling without replacement (SRSWOR) where $\pi_i = \dfrac{n}{N}$ and $\pi_{ij} = \dfrac{n(n-1)}{N(N-1)}$ and (ii) Midzuno-Sen (1952-53), (M-S for brief) sampling scheme using x_i's as the measure of size. The first two order inclusion probabilities for M-S sampling schemes are

$$\pi_i = \frac{N-n}{N-1} p_i + \frac{n-1}{N-1} \text{ and}$$

$$\pi_{ij} = \frac{(n-1)(N-n)}{(N-1)(N-2)}(p_i + p_j) + \frac{(n-1)(n-2)}{(N-1)(N-2)}$$

respectively where $p_i = x_i / X$ and $X = \sum_i x_i$. It is already mentioned that for SRSWOR, $\hat{Y}_g$ reduces to a ratio estimator and a regression estimator when $Q_i = 1/x_i$ and $Q_i = 1$, respectively. The estimators for the variance of the ratio and regression estimators are obtained by substituting $q_i = 1/x_i$, $q_{ij} = 1/(x_i x_j)$ and $q_i = 1$, $q_{ij} = 1$, respectively in the proposed variance estimators $\hat{v}_c(j)$'s, $j = 1,2,3,4$. For the M-S sampling scheme, $\hat{Y}_g$ reduces to the ratio estimator and regression estimator when $Q_i = 1/(x_i \pi_i)$ and $Q_i = 1/\pi_i$, respectively. We take $q_i = 1$ and $q_{ij} = 1$ in $\hat{v}_c(j)$'s, $j = 1,2,3,4$ for estimating variance for both the ratio and regression estimators. The relative efficiency of $\hat{v}_c(j)$ compared with the conventional estimator $\hat{v}_{ht}(y)$ is

$$E_c(j) = \frac{\overline{V}}{\overline{V}_c(j)} x100 \quad \text{for } j = 1,2,3,4 \qquad (4.2)$$

where

$\overline{V} = \frac{1}{R}\sum_{k=1}^{R}[\hat{v}_{ht}(y \mid s_k) - V_y]^2$; $\overline{V}_c(j) = \frac{1}{R}\sum_{k=1}^{R}[\hat{v}_c(j \mid s_k) - V_y]^2$

$\hat{v}_{ht}(y \mid s_k)$ = value of $\hat{v}_{ht}(y)$ computed from the sample s_k

$\hat{v}_c(j \mid s_k)$ = value of $\hat{v}_c(j)$ computed from the sample s_k.

Similarly, the efficiency of Chaudhuri and Roy's (1997) estimators $\hat{v}_1^*$ and $\hat{v}_2^*$ are computed as follows:

$$E_j^* = \frac{\overline{V}}{\overline{V}_j^*} x100 \quad \text{for } j = 1,2 \tag{4.3}$$

where $\overline{V}_j^* = \frac{1}{R}\sum_{k=1}^{R}[\hat{v}_j^*(s_k) - V_y]^2$ and $\hat{v}_j^*(s_k)$ = value of $\hat{v}_j^*$ omputed from the sample s_k. Relative efficiencies *$E_c(j)$* and *E_j^*'s* for SRSWOR and M-S sampling schemes for the three populations are presented in Table1, Table-2 and Table-3. From the tables, we note, with the exception of $\hat{v}_c(1)$ and $\hat{v}_c(3)$, that the proposed alternative estimators, including the two by Chaudhuri and Roy (1977), provide remarkable gains in efficiency over the conventional estimator $\hat{v}_{ht}(y)$, irrespective of g values and the sampling design used. The estimator $\hat{v}_c(4)$ performs the best around the true value of g for both the sampling designs (SRSWOR and M-S) except for the Population-1 for estimating the variance of the ratio predictor under SRSWOR sampling with n = 40 and for g ≤ 1.2. Only this situation $\hat{v}_1^*$ performs the best. However, for g > 1.2, $\hat{v}_c(4)$ is the best. The second best is $\hat{v}_c(2)$ except for estimating variance for the ratio predictor under SRSWOR for the Population-1. $\hat{v}_c(2)$ has an additional advantage as it does not require any knowledge of g. There is no definite ordering among $\hat{v}_c(1)$, $\hat{v}_1^*$ and $\hat{v}_2^*$ in general. It appears that $\hat{v}_2^*$ performs better than $\hat{v}_1^*$ for estimating variance of the regression predictor around the true value of g. The estimator $\hat{v}_c(1)$ does not perform well for estimating the variance of the ratio predictor for SRSWOR sampling with smaller value of g. But for M-S sampling, it performs reasonably well. The estimator $\hat{v}_c(3)$ does not perform well for estimating variance of the regression predictor of all populations, however, it works less efficiently for ratio predictors of all populations.

The non-negativity properties of the estimators are also studied (details are not given to save space). It is found that the conventional estimator $\hat{v}_{ht}(y)$ can take very often negative values up to 40%; $\hat{v}_c(1)$ and $\hat{v}_c(3)$ take up to 20%; while $\hat{v}_1^*$ and $\hat{v}_2^*$ and $\hat{v}_c(2)$ take nonnegative values with negligible frequency. The estimator $\hat{v}_c(4)$ does not take any negative values, as determined in this study.

The percentage relative bias ($= \frac{Bias}{\overline{V}} x100$) of the estimators are also investigated (details not presented here). It is found that the conventional estimator $\hat{v}_{ht}(y)$ has up to 40% relative bias. The estimators $\hat{v}_1^*$, $\hat{v}_2^*$, $\hat{v}_c(2)$ and $\hat{v}_c(4)$ have negligible relative bias in all situations whereas $\hat{v}_c(1)$ and $\hat{v}_c(3)$ have very high bias whenever they have low efficiency but in other cases they also have small bias.

5. CONCLUDING REMARKS

The lower bound of the mean-square error of the variance of the regression predictor, provided by Chaudhuri and Roy (1997), is incorrect and consequently the estimator $\hat{v}_0$ given in (1.5) does not attain the lower bound as originally claimed by those workers. The conventional estimator $\hat{v}_{ht}(y)$ should not be used for estimating variance of the greg because (i) it can take very often negative values (ii) it is of low efficiency and (iii) of high bias. The use of $\hat{v}_c(4)$ is recommended when some rough idea about the magnitude of g is available. If nothing is known about g, we can safely use $\hat{v}_c(2)$ for the estimation of variance.

REFERENCES

Arnab, R and Singh, S. (2002). Calibrated estimators of the variance of the regression predictor. Working paper. (complete proofs of the theorems are available on request from the authors)

Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. J.*Amer. Statist. Assoc*., 74, 991-915.

Chaudhuri, A. and Roy, D (1997). Optimal variance estimation for generalized regression predictor. *Jour. Statist. Planning and Inference*, **60**, 139-151.

Horvitz, D.G. and Thompson, D.J. (1952). A generalisation of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc*., **47**, 663-685.

Kott, P.S. (1990). Estimating the conditional variances of a design consistent regression estimator. *J. Statist. Plann. Inference*, **24**, 287-296.

Liu, T.P. (1974). Bayes estimation for variance of a finite population. *Metrika*, 21,127-132

Midzuno, J. (1952). On the estimating system with probabilities proportional to sum of sizes. *Ann. Ins. Statist. Math.,* 3, 99-107

Särndal, C.E. (1982). Implications of survey designs for generalized regression estimators of linear functions. *J. Statist. Plann. Inference*, **7**, 155-170.

Särndal, C.E., Swensson, B.E. and Wretman, J.H. (1992). Model Assisted Survey Sampling. *Springer-Verlag, NY.*

Särndal, C.E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Jour. Amer. Statist. Assoc*., **91**, 1289-1300.

Sen, A.R. (1953). On the estimator of the variance in sampling with varying probabilities. *J..Ind. Soc. Agri. Statist*. 5, 119-127.

Sengupta, S. (1988). Optimality of design unbiased strategy for estimating finite population variance. *Sankhyā*, *Series B, 50, 149-152*.

Singh, S. (2003). *Advanced Sampling theory with applications: How Michael "Selected" Amy.* pp 1-1220 (Vol. 1 and Vol. 2). Kluwer Academic Publishers, The Netherlands. (http://www.wkap.nl/prod/b/1-4020-1689-1)

Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of the regression estimator: Higher level calibration approach. *Survey Methodology, 24*,41-50.

Zou, G. (1999). Variance estimation for unequal probability sampling. *Metrika*, 50,71-82

APPENDIX

Relative Efficiencies of the Variance Estimators for the Population 1 (true value of g = 1.2) for sample size $n = 25$.

(i) SRSWOR Sampling:

g	E_1^*	E_2^*	$E_c(1)$	$E_c(2)$	$E_c(3)$	$E_c(4)$
RATIO PREDICTOR						
0.0	251723	242446	8863	258864	148959	266498
0.5	283719	246970	84	258864	148959	273022
0.8	287309	249365	1	258864	148959	275436
0.9	285275	250187	0	258864	148959	275969
1.0	281605	251015	1	258864	148959	276344
1.1	276354	251835	0	258864	148959	276545
1.2	269619	252623	2	258864	148959	276546
1.3	261538	253346	1	258864	148959	276319
1.4	252275	253961	1	258864	148959	275829
1.5	242016	254411	682	258864	148959	275033
1.6	230961	254624	3277	258864	148959	273883
1.7	219309	254516	8939	258864	148959	272322
1.8	207257	253981	21242	258864	148959	270285
1.9	194986	252900	63820	258864	148959	267700
2.0	182662	251134	258864	258864	148959	264489
2.2	158423	244935	25966	258864	148959	255847
2.5	124630	226653	1038	258864	148959	236100
3.0	78652	168191	0	258864	148959	182611
REGRESSION PREDICTOR						
0.0	258725	195054	10616	221941	7	251598
0.5	241534	195900	97720	221941	7	254184
0.8	220310	196026	128768	221941	7	253984
0.9	212186	196038	137822	221941	7	253599
1.0	203744	196026	146563	221941	7	253037
1.1	195090	195978	155093	221941	7	252280
1.2	186320	195876	163469	221941	7	251309
1.3	177518	195697	171713	221941	7	250100
1.4	168754	195413	179818	221941	7	248627
1.5	160087	194987	187748	221941	7	246857
1.6	151565	194377	195443	221941	7	244755
1.7	143225	193531	202821	221941	7	242281
1.8	135099	192392	209779	221941	7	239390
1.9	127208	190891	216197	221941	7	236036
2.0	119569	188954	221941	221941	7	232168
2.2	105091	183448	230848	221941	7	222686
2.5	85442	169891	235823	221941	7	203413
3.0	58149	130552	218002	221941	7	156989

(ii) Midzuno-Sen Sampling Scheme

g	E_1^*	E_2^*	$E_c(1)$	$E_c(2)$	$E_c(3)$	$E_c(4)$
RATIO PREDICTOR						
0.0	270262	258334	159329	276725	142	282462
0.5	295143	262604	189814	276725	142	288648
0.8	294347	264949	208034	276725	142	290610
0.9	291060	265774	214350	276725	142	290942
1.0	286289	266611	220784	276725	142	291089
1.1	280107	267446	227311	276725	142	291030
1.2	272618	268253	233887	276725	142	290737
1.3	263957	268996	240449	276725	142	290180
1.4	254277	269627	246915	276725	142	289319
1.5	243749	270085	253181	276725	142	288109
1.6	232549	270292	259127	276725	142	286500
1.7	220852	270155	264612	276725	142	284430
1.8	208829	269561	269483	276725	142	281834
1.9	196639	268378	273576	276725	142	278640
2.0	184426	266455	276725	276725	142	274768
2.2	160427	259723	279574	276725	142	264666
2.5	126878	239894	273553	276725	142	242496
3.0	80760	176756	236335	276725	142	185314
REGRESSION PREDICTOR						
0.0	264970	201110	108247	227800	13	256664
0.5	243602	202612	136977	227800	13	259375
0.8	221801	203117	153882	227800	13	259072
0.9	213706	203254	159802	227800	13	258621
1.0	205365	203366	165887	227800	13	257977
1.1	196868	203442	172134	227800	13	257122
1.2	188293	203463	178525	227800	13	256036
1.3	179706	203406	185029	227800	13	254695
1.4	171166	203241	191602	227800	13	253073
1.5	162720	202931	198182	227800	13	251138
1.6	154409	202431	204689	227800	13	248854
1.7	146263	201687	211026	227800	13	246181
1.8	138309	200637	217080	227800	13	243078
1.9	130565	199211	222719	227800	13	239498
2.0	123045	197327	227800	227800	13	235394
2.2	108717	191840	235680	227800	13	225419
2.5	89074	177976	239644	227800	13	205418
3.0	61292	136849	220507	227800	13	158126

(iii) Remark: Note that similar results for different true values of model parameter g =1.5, 1.8 and different sample sizes $n = 25, 40$ are available from the authors, but are not cited due to space limit of six pages.