ON ESTIMATION OF POPULATION TOTAL USING GENERALIZED REGRESSION PREDICTOR

Raghunath Arnab¹ and Sarjinder Singh²

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 Yᵢ = (Xᵢ, eᵢ), i ∈ U be the value of the study (auxiliary) variable of the ith unit of the population. The values of Yᵢ’s are unknown before survey but the values of Xᵢ’s are assumed to be known and positive. Here we consider the problem of estimation of the finite population total Y = ∑ᵢ Yᵢ 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 ≠ j are denoted respectively by πᵢ and πᵢj, and assumed to be positive. Chaudhuri and Roy (1997) considered the following superpopulation model

Model M: Yᵢ = β Xᵢ + eᵢ, i ∈ U  (1.1)

where β is an unknown constant, eᵢ’s are error component, independently distributed with Eₑᵢ = 0 and Varₑᵢ = σₑᵢ², (σₑᵢ² > 0, unknown). Here Eₑᵢ, Varₑᵢ denote respective expectation and variance with respect to the superpopulation model M. Chaudhuri and Roy (1997) considered the generalized regression predictor (greg) for Y

\[ t₉ = \sum_{i\in S} \frac{X_i\hat{\beta}_Q}{\pi_i} \]  (1.2)

where Iᵢ = 1 if i ∈ S and 0 if i ∉ S , \[ \hat{\beta}_Q = \sum_{i\in U} \frac{Q_i X_i Y_i I_{si}}{\sum_{i\in U} Q_i X_i^2 I_{si}}, \]

Qᵢ(>0)’s are suitably chosen constants and X = ∑ᵢ Xᵢ.

The approximate expression for the variance of t₉ provided by Särndal (1982) is given by

\[ V_p(t₉) = \sum_{i\in S} \sum_{j\in S} \frac{E_p}{\pi_i} \frac{E_p}{\pi_j} \Delta_{ij} \Delta_{ij} = V(y) \text{ (say)} \]
where \( E_i = y_i - \beta Q x_i \), \( \beta = \frac{\sum \alpha_i x_i y_i \pi_i}{\sum \alpha_i x_i^2 \pi_i} \), \( \Delta ij = \frac{\pi_i x_i - \pi_j x_j}{2\pi_j} \) 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

\[
V(Y) = \sum_{i \in U} \alpha_i y_i^2 + \frac{\sum \sum_{i \neq j} \alpha_i y_i y_j}{\pi_i} = V \text{ (say)}
\]

where

\[
\alpha_i = \left( \frac{1}{\pi_i} - 1 \right) + \left( \frac{Q x_i^2 \pi_i}{\sum_{i \in U} Q x_i^2 \pi_i} - 2Q x_i \pi_i \right) \left( \frac{\sum_{i \in U} Q x_i^2 \pi_i}{\sum_{i \in U} Q x_i \pi_i} \right)^2
\]

\[
\alpha_{ij} = \left( \frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) + \left( \frac{Q \pi_i \pi_j x_i x_j}{\sum_{i \in U} Q x_i^2 \pi_i} - 2Q x_i \pi_i \right) \left( \frac{\sum_{i \in U} Q x_i^2 \pi_i}{\sum_{i \in U} Q x_i \pi_i} \right)^2
\]

\[
- Q x_i \pi_i \left( \frac{\pi_{ik}}{\pi_i \pi_k} - 1 \right) - Q x_i \pi_i \left( \frac{\sum_{k \in U} x_k (\pi_{ik} \pi_i \pi_k)}{\sum_{i \in U} x_k \pi_i \pi_k} - 1 \right)
\]

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

\[
v = v(y) = a_s + \sum_{i \in U} b_i y_i^2 I_{si} + \sum_{i \neq j} b_{ij} y_i y_j I_{ij}
\]

where \( a_s, b_{ij} \) and \( b_{ij} \) are constants free from \( y_i \)'s and satisfy the unbiasedness conditions as follows:

\[
E_p(a_s) = 0, E_p(b_{ij}) = a_i \quad \text{and} \quad E_p(b_{ij} I_{ij}) = a_{ij} \quad \forall i \neq j.
\]

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

\[
v_0 = \sum_{i \in U} \alpha_i (y_i^2 - \sigma_i^2 - \mu_i^2) I_{si} + \sum_{i \neq j} \alpha_{ij} (y_i y_j - \mu_i \mu_j) I_{ij} \tag{1.5}
\]

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

(In the expression of \( v_0 \) of the Theorem 1, page 143, of Chaudhuri and Roy (1997)'s paper, “ \( \sum \alpha_i (\sigma_i^2 + \mu_i^2) \)” was wrongly written as “ \( \sum \alpha_i^2 (\sigma_i^2 + \mu_i^2) \)” . 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 \( \mu_i \)'s and \( \sigma_i^2 \)'s. So, Chaudhuri and Roy proposed the following alternative estimators when \( \mu_i = \beta x_i \) and \( \sigma_i^2 = \sigma^2 x_i^p \) by replacing \( \beta \) and \( \sigma^2 \) with their suitable estimators as follows:

\[
v_1^* = \sum_{i} \alpha_i y_i^2 \left( y_i^2 - \hat{\theta} x_i^2 - \phi \sigma_i^2 \frac{I_{si}}{\pi_i} + \phi \sum \alpha_i x_i \right) + \hat{\theta} \sum \alpha_i x_i^2 + \hat{\phi} \sum \alpha_i y_i y_j \frac{I_{ij}}{\pi_j} \tag{1.6}
\]

and

\[
v_2^* = \left( \sum_{i} \alpha_i x_i^2 \right) \left( \sum \frac{\sum a_{ij} y_i y_j (I_{si}/\pi_i)}{\pi_j} - \frac{\phi \sum \alpha_i x_i^2 (I_{si}/\pi_i)}{\pi_j} \right) + \hat{\phi} \sum \alpha_i x_i^2 \tag{1.7}
\]

where

\[
\theta = \left. \left( \frac{\sum \alpha_i x_i^2 - g y_i}{\sum \alpha_i x_i^2} \right)^2 \right|_{i \in x} - \frac{1}{(n-1)} \left( \frac{\sum \alpha_i x_i^2}{\sum \alpha_i x_i^2} \right)^2 \left( \frac{\sum \alpha_i x_i^2 - y_i}{\sum \alpha_i x_i^2} \right)^2 \right|_{i \in x},
\]

\[
\phi = \frac{1}{(n-1)} \left( \frac{\sum \alpha_i x_i^2}{\sum \alpha_i x_i^2} \right)^2
\]

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 \( v \in H \)

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

\[
= M_0 \tag{2.1}
\]

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

The equality is attained in the above if \( v \) equals

\[
v_0 = \sum_{i \in U} \alpha_i (y_i^2 - \sigma_i^2 - \mu_i^2) \frac{I_{si}}{\pi_i} + \sum_{i \neq j \in U} \alpha_{ij} (y_i y_j - \mu_i \mu_j) \frac{I_{ij}}{\pi_j} \tag{2.2}
\]

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
On comparing the coefficients of \( \sigma^2 \) and \( \beta^2 \) on both sides of (3.7), the system of calibration equations becomes

\[
\Sigma_i \alpha_i w_i x_i^2 + \Sigma \Sigma_i \alpha_i g_i y_i x_j I_{ij} = 0 & \Sigma_i \alpha_i w_i x_i^2 I_{si} = \Sigma \alpha_i x_i^2 \quad (3.8)
\]

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

\[
\phi = \sum_i \frac{(w_i \alpha_i - d_{ij} \alpha_i)^2}{d_i \alpha_i q_i} I_{si} + \sum_i \frac{(\alpha_i g_i - d_{ij} \alpha_i)^2}{d_i \alpha_i q_i} I_{ij}
\]

\[
= 2.2 \sum_i \alpha_i w_i x_i^2 I_{si} + \sum_i \alpha_i g_i y_i x_j I_{ij} \beta = \Sigma \Sigma_i \alpha_i w_i x_i^2 I_{si}
\]

\[
\frac{\partial \phi}{\partial w_i} = 0 \Rightarrow w_i \alpha_i = d_{ij} \alpha_i + \lambda d_{ij} \alpha_i g_i y_i x_j \quad .
\]

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

\[
\lambda = \frac{\Sigma A_{i1} - B A_{22}}{A C - B^2} \quad \quad \text{and} \quad \mu = \frac{A A_{21} - B A_{12}}{A C - B^2} \quad .
\]

were

\[
A_1 = \Sigma \alpha_i d_{i2} q_i x_i^2 I_{si} + \Sigma \Sigma_i \alpha_i d_{ij} q_i y_i x_j I_{ij} \quad ; \quad B = \Sigma \alpha_i d_{i1} q_i x_i^2 I_{si} \quad ;
\]

\[
C = \Sigma \alpha_i d_{i1} q_i x_i^2 I_{si} \quad ; \quad \Delta_1 = - \Sigma \alpha_i x_i^2 \quad ; \quad \Delta_2 = \Sigma \alpha_i x_i^2 - \Sigma \alpha_i d_{ij} x_j
\]

\[
\text{Substituting (3.12) in (3.10) and (3.11) we get}
\]

\[
\frac{\partial \phi}{\partial w_i} = 0 \Rightarrow w_i \alpha_i = d_{ij} \alpha_i + \lambda d_{ij} \alpha_i g_i y_i x_j \quad (3.13)
\]

and

\[
\frac{\partial \phi}{\partial w_j} = 0 \Rightarrow \frac{\Sigma A_{i2} B_{i2}}{A C - B^2} = d_{ij} q_i y_i x_j = w_j \quad (3.14)
\]

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

\[
\hat{\sigma}^2 = \frac{CP - BQ}{A C - B^2} \quad \text{and} \quad \hat{\beta}^2 = \frac{AQ - BP}{A C - B^2}
\]

\[Q = \Sigma \alpha_i d_{i3} q_i x_i^2 I_{si} + \Sigma \Sigma_i \alpha_i d_{ij} q_i y_i x_j I_{ij} \quad .
\]

Here we note that \( \hat{\beta}^2 \) and \( \hat{\sigma}^2 \) are model unbiased estimator for \( \beta^2 \) and \( \sigma^2 \) respectively.

**Case II.** Using the relation \( V(x) = \Sigma \alpha_i x_i^2 + \Sigma \Sigma_i \alpha_i g_i x_j = 0 \), we set the calibration constrain assuming \( T_i(x) \) is known as:

\[
(i) \quad \Sigma \alpha_i w_i x_i^2 I_{si} = \Sigma \alpha_i x_i^2 = T_i(x)
\]

and

\[
(ii) \quad \Sigma \Sigma_i \alpha_i g_i y_i x_j I_{ij} = \Sigma \Sigma_i \alpha_i g_i x_j = T_2(x) = T_1(x)
\]

The equation (3.16) satisfies

\[
\hat{V}_i(x) = \Sigma \alpha_i w_i x_i^2 I_{si} + \sum \Sigma_i \alpha_i g_i y_i x_j I_{ij} = V(x) = 0
\]
Minimization of (3.3) subject to (3.16) yields calibrated weights
\[ w_i = d_i + \frac{d_i q_i x_i^2}{\sum_i d_i q_i x_i^4 I_{si}} \left[ \sum_i \alpha_i x_i^2 - d_i x_i^4 I_{si} \right] = w_i(2) \] (3.17)
and
\[ w_{ij} = d_{ij} + \frac{d_{ij} q_{ij} x_{ij}}{\sum_{i,j} d_{ij} q_{ij} x_{ij}^2 I_{sij}} \left[ \sum_{i,j} \alpha_i x_{ij} - \sum_{i,j} d_{ij} x_{ij} I_{sij} \right] = w_{ij}(2) \] (3.18)
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
\[ \hat{\sigma}_c(2) = \hat{\sigma}_h(y) + b_2 \left[ \tilde{T}_2(x) - \sum_{i,j} d_{ij} \alpha_i x_{ij} I_{sij} \right] \] (3.19)
where
\[ b_1 = \sum_{i} d_i \alpha_i x_i^2 I_{si} \quad \text{and} \quad b_2 = \sum_{i,j} d_{ij} \alpha_i x_{ij}^2 I_{sij} \]

Case III. Here we assume that \( T_1(x) \) is unknown and use a constraint of calibration
\[ \hat{\sigma}_c(x) = \sum_i \alpha_i w_i x_i^2 I_{si} + \sum_{i,j} \alpha_{ij} w_i x_{ij} I_{sij} = V(x) = 0 \] (3.20)
In this situation calibration weights obtained by minimizing (3.3) subject to (3.15) come out as
\[ w_i = d_i + \frac{d_i q_i x_i^2 \left[ \hat{v}_{hi}(x) \right]}{\sum_i d_i q_i x_i^4 I_{si}} \] (3.21)
and
\[ w_{ij} = d_{ij} + \frac{d_{ij} q_{ij} x_{ij} \left[ \hat{v}_{hij}(x) \right]}{\sum_{i,j} d_{ij} q_{ij} x_{ij}^2 I_{sij}} \] (3.22)
where \( \hat{v}_{hi}(x) = \sum_i \alpha_i d_i x_i^2 I_{si} + \sum_{i,j} \alpha_{ij} d_{ij} x_{ij} I_{sij} \). The resultant calibrated estimator of variance of the regression predictor is
\[ \hat{\sigma}_c(3) = \hat{\sigma}_h(y) + \left[ \sum_{i} d_i \alpha_i q_i x_i^2 I_{si} + \sum_{i,j} d_{ij} \alpha_{ij} q_{ij} x_{ij} x_{ij} I_{sij} \right] \left[ \hat{v}_{hi}(x) \right] \] (3.23)

4. SIMULATION STUDIES

In this section, we present results of simulation studies to compare performances of the proposed estimators \( \hat{\sigma}_c(1), \hat{\sigma}_c(2) \) and \( \hat{\sigma}_c(3) \) of the variance of the generalized regression predictor \( \hat{Y}_g \) with the conventional estimator \( \hat{\sigma}_h(y) \) (given in (3.1)) and \( \hat{\sigma}_1^* \) and \( \hat{\sigma}_2^* \) proposed by Chaudhuri and Roy (1997). It should be noted that the estimators \( \hat{\sigma}_c(1) \), and \( \hat{\sigma}_c(3) \) are of the similar form where \( \hat{\beta}^2 \), \( \hat{\phi} \) and \( \hat{\sigma}^2 \) are respectively model unbiased estimators for \( \beta^2 \) and \( \sigma^2 \).
Both the estimators \( \hat{\beta}^2 \) and \( \hat{\phi} \) involve \( g \) which is generally unknown. So, we propose the following alternative variance estimator:
\[ \hat{\sigma}_c(4) = \hat{\sigma}_h(y) - b_4 \hat{v}_{hi}(x) + \hat{v}_2(3) \] (4.1)
where \( b = \sum_{i} \alpha_i q_i y_i d_{ij} I_{sij} \) is model unbiased for \( \beta^2 \) and \( \sigma^2 \) 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 \( i = 1, \ldots, 200 \) and for a given \( x_i \) an is a random sample selected independently from a normal population with mean zero and variance \( \sigma^2 x_i^2 \). 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 \( \pi_i = \frac{n}{N} \) and \( \pi_{ij} = \frac{n(n-1)}{N(N-1)} \) and (ii) Midzuno-Sen (1952-53), (M-S for brief) sampling scheme using \( x_i \) as the measure of size. The first two order inclusion probabilities for M-S sampling schemes are
\[ \pi_i = \frac{N-n}{N-1} p_i + \frac{n-1}{N-1} \]
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 \( X = \sum x_i \). It is already mentioned that for SRSWOR, \( \hat{\sigma}_g \) 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 \( q_i = 1 / x_i \) and \( q_i = 1 \). For the proposed variance estimators \( \hat{\sigma}_c(j) \)’s, \( j = 1,2,3,4 \). For the M-S sampling scheme, \( \hat{\sigma}_g \) 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 \( q_i = 1 \) in \( \hat{\sigma}_c(j) \)’s, \( j = 1,2,3,4 \) for estimating variance for both the ratio and regression estimators. The relative efficiency of \( \hat{\sigma}_c(j) \) compared with the conventional estimator \( \hat{\sigma}_h(y) \) is
\[ E_c(j) = \frac{\hat{\sigma}_c(j)}{\hat{\sigma}_c(j)} 100 \quad \text{for} \quad j = 1,2,3,4 \] (4.2)
where

\[ \bar{V} = \frac{1}{R} \sum_{k=1}^{R} [\hat{v}_{ht}(y | s_k) - V_y]^2 ; \]

\[ \bar{V}_c(j) = \frac{1}{R} \sum_{k=1}^{R} [\hat{v}_c(j | s_k) - V_y]^2 . \]

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

\( \hat{v}_c(j | s_k) \) = 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:

\[ E_j^* = \frac{\bar{V}}{\bar{V}_j} \times 100 \quad \text{for } j = 1, 2 \]  \hspace{1cm} (4.3)

where

\[ \bar{V}_j^* = \frac{1}{R} \sum_{k=1}^{R} [\hat{v}_j^*(s_k) - V_y]^2 \]

and \( \hat{v}_j^*(s_k) = \) value of \( \hat{v}_j^* \) computed 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 \( \hat{v}_1(1) \) 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 \( g \leq 1.2 \). 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 \( \hat{v}_c(1) \) and \( \hat{v}_c(3) \) 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 variance of the regression predictor for SRSWOR sampling with smaller values 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 (\( = \frac{\text{Bias}}{V} \times 100 \)) 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 (1977), 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 regression predictor around the true value of \( g \). 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)


**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:**

<table>
<thead>
<tr>
<th>$g$</th>
<th>$E_1^2$</th>
<th>$E_2^2$</th>
<th>$E_r(1)$</th>
<th>$E_r(2)$</th>
<th>$E_r(3)$</th>
<th>$E_r(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>251723</td>
<td>242446</td>
<td>8863</td>
<td>258864</td>
<td>148959</td>
<td>266498</td>
</tr>
<tr>
<td>0.5</td>
<td>283791</td>
<td>246970</td>
<td>84</td>
<td>258864</td>
<td>148959</td>
<td>273022</td>
</tr>
<tr>
<td>0.8</td>
<td>287309</td>
<td>249365</td>
<td>1</td>
<td>258864</td>
<td>148959</td>
<td>275436</td>
</tr>
<tr>
<td>0.9</td>
<td>285275</td>
<td>250187</td>
<td>0</td>
<td>258864</td>
<td>148959</td>
<td>275969</td>
</tr>
<tr>
<td>1.0</td>
<td>281605</td>
<td>251015</td>
<td>1</td>
<td>258864</td>
<td>148959</td>
<td>276344</td>
</tr>
<tr>
<td>1.1</td>
<td>276354</td>
<td>251835</td>
<td>0</td>
<td>258864</td>
<td>148959</td>
<td>276545</td>
</tr>
<tr>
<td>1.2</td>
<td>269619</td>
<td>252623</td>
<td>2</td>
<td>258864</td>
<td>148959</td>
<td>276546</td>
</tr>
<tr>
<td>1.3</td>
<td>261538</td>
<td>253346</td>
<td>1</td>
<td>258864</td>
<td>148959</td>
<td>276319</td>
</tr>
<tr>
<td>1.4</td>
<td>252275</td>
<td>253961</td>
<td>1</td>
<td>258864</td>
<td>148959</td>
<td>275829</td>
</tr>
<tr>
<td>1.5</td>
<td>242016</td>
<td>254411</td>
<td>682</td>
<td>258864</td>
<td>148959</td>
<td>275033</td>
</tr>
<tr>
<td>1.6</td>
<td>230961</td>
<td>254624</td>
<td>3277</td>
<td>258864</td>
<td>148959</td>
<td>273883</td>
</tr>
<tr>
<td>1.7</td>
<td>219309</td>
<td>254516</td>
<td>8939</td>
<td>258864</td>
<td>148959</td>
<td>272322</td>
</tr>
<tr>
<td>1.8</td>
<td>207257</td>
<td>253981</td>
<td>21242</td>
<td>258864</td>
<td>148959</td>
<td>270285</td>
</tr>
<tr>
<td>1.9</td>
<td>194986</td>
<td>252900</td>
<td>63820</td>
<td>258864</td>
<td>148959</td>
<td>267700</td>
</tr>
<tr>
<td>2.0</td>
<td>182662</td>
<td>251134</td>
<td>258864</td>
<td>148959</td>
<td>264489</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>158423</td>
<td>249935</td>
<td>25966</td>
<td>258864</td>
<td>148959</td>
<td>258474</td>
</tr>
<tr>
<td>2.5</td>
<td>124630</td>
<td>226653</td>
<td>1038</td>
<td>258864</td>
<td>148959</td>
<td>236100</td>
</tr>
<tr>
<td>3.0</td>
<td>78652</td>
<td>168191</td>
<td>0</td>
<td>258864</td>
<td>148959</td>
<td>182611</td>
</tr>
</tbody>
</table>

(ii) **Midzuno-Sen Sampling Scheme**

<table>
<thead>
<tr>
<th>$g$</th>
<th>$E_1^2$</th>
<th>$E_2^2$</th>
<th>$E_r(1)$</th>
<th>$E_r(2)$</th>
<th>$E_r(3)$</th>
<th>$E_r(4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>270262</td>
<td>258334</td>
<td>159329</td>
<td>276725</td>
<td>142</td>
<td>282462</td>
</tr>
<tr>
<td>0.5</td>
<td>295143</td>
<td>262604</td>
<td>189814</td>
<td>276725</td>
<td>142</td>
<td>288648</td>
</tr>
<tr>
<td>0.8</td>
<td>294347</td>
<td>264949</td>
<td>208034</td>
<td>276725</td>
<td>142</td>
<td>290610</td>
</tr>
<tr>
<td>0.9</td>
<td>291064</td>
<td>265774</td>
<td>214350</td>
<td>276725</td>
<td>142</td>
<td>290942</td>
</tr>
<tr>
<td>1.0</td>
<td>286289</td>
<td>266611</td>
<td>220784</td>
<td>276725</td>
<td>142</td>
<td>291089</td>
</tr>
<tr>
<td>1.1</td>
<td>280107</td>
<td>267446</td>
<td>227311</td>
<td>276725</td>
<td>142</td>
<td>291030</td>
</tr>
<tr>
<td>1.2</td>
<td>272618</td>
<td>268253</td>
<td>233887</td>
<td>276725</td>
<td>142</td>
<td>290737</td>
</tr>
<tr>
<td>1.3</td>
<td>263957</td>
<td>268996</td>
<td>240449</td>
<td>276725</td>
<td>142</td>
<td>290180</td>
</tr>
<tr>
<td>1.4</td>
<td>252477</td>
<td>269627</td>
<td>246915</td>
<td>276725</td>
<td>142</td>
<td>289319</td>
</tr>
<tr>
<td>1.5</td>
<td>243749</td>
<td>270085</td>
<td>253181</td>
<td>276725</td>
<td>142</td>
<td>285109</td>
</tr>
<tr>
<td>1.6</td>
<td>232549</td>
<td>270292</td>
<td>250127</td>
<td>276725</td>
<td>142</td>
<td>285600</td>
</tr>
<tr>
<td>1.8</td>
<td>208829</td>
<td>269561</td>
<td>264983</td>
<td>276725</td>
<td>142</td>
<td>281834</td>
</tr>
<tr>
<td>1.9</td>
<td>196639</td>
<td>268378</td>
<td>273576</td>
<td>276725</td>
<td>142</td>
<td>278640</td>
</tr>
<tr>
<td>2.0</td>
<td>184426</td>
<td>266455</td>
<td>276725</td>
<td>276725</td>
<td>142</td>
<td>273746</td>
</tr>
<tr>
<td>2.2</td>
<td>160427</td>
<td>259723</td>
<td>279574</td>
<td>276725</td>
<td>142</td>
<td>264666</td>
</tr>
<tr>
<td>2.5</td>
<td>126878</td>
<td>239984</td>
<td>273553</td>
<td>276725</td>
<td>142</td>
<td>242496</td>
</tr>
<tr>
<td>3.0</td>
<td>80760</td>
<td>176756</td>
<td>236335</td>
<td>276725</td>
<td>142</td>
<td>185314</td>
</tr>
</tbody>
</table>

(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.