www.sciencevision.info

Original Research

Sci Vis 11 (1), 2-6
January-March, 2011
ISSN (print) 0975-6175
ISSN (online) 2229-6026

Some generalized classes of double sampling regression type estimators using auxiliary information

Shashi Bhushan[1]*, R. Karan Singh[2] and Arvind Pandey[1]

[1] *Department of Statistics, Pachhunga University College, Mizoram University, Aizawl 796 001, India*
[2] *Department of Statistics, University of Lucknow, Lucknow 226 007, India*

Received 17 December 2010 | Revised 17 January 2011 | Accepted 27 January 2011

Abstract

In the present study, a double sampling regression type estimator representing a class of estimators is proposed. The bias and mean square error (MSE) of the proposed estimator is obtained. A more generalized class of double sampling regression type estimator utilizing the auxiliary information available at first phase in the form of mean and variance of the auxiliary variable is also proposed. The bias and MSE of the proposed class were obtained for this class. The concluding remarks show that the proposed classes of estimators are better than that of double sampling estimator based on earlier proposition. An empirical study is included for illustration.

Key words: Auxiliary information; bias; double sampling regression type estimator; mean square error.

Mathematics Subject Classification 2000: Primary 62D05 Secondary 62F10

Introduction

If the supplementary information on the auxiliary variable is not known then the double sampling ratio and regression strategies are very well known. Many biased double sampling ratio type, double sampling regression type transformed estimators and the biased double sampling estimators obtained through parametric linear combination of ratio or regression estimator and the usual unbiased estimators are available for estimating the population mean.[1-7] The use of population mean and variance of auxiliary variable for increasing the efficiency of the sampling strategy has been discussed recently by Singh,[8] Bhushan[1] and Bhushan *et al.*[9] among others. Let y be the characteristic under study and x be the auxiliary variable. Thus for a finite population of size N, we denote by

Y_i: the observation on the i th unit of the population for the characteristic y under study ($i = 1,2,\ldots,N$),

X_i: the observation on the i th unit of the population for the auxiliary characteristic x under study ($i = 1,2,\ldots,N$),

Corresponding author: S. Bhushan
Phone.
E-mail: shashi.py@gmail.com

$$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i \qquad \bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$$

$$S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2,$$

$$S_x^2=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2$$

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2$$

$$S_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})=\rho S_x S_y$$

$$\beta=\frac{S_{xy}}{S_x^{\ 2}}=\rho\frac{S_y}{S_x}$$

(where ρ is the population correlation coefficient between x and y) and

$$\mu_{pq}=\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^p(Y_i-\bar{Y})^q$$: the (p,q) th product moment about mean between x and y.

If the information about the mean and variance of the auxiliary variable is not known then we resort to double sampling. Let the auxiliary characteristic x be observed on a large preliminary simple random sample of size n' drawn without replacement from a population of size N in the first phase. Also the characteristic of interest y and the auxiliary characteristic x be observed on the second phase sample of size n drawn from first phase sample by simple random sampling without replacement.

$$\bar{x}'=\frac{1}{n'}\sum_{i=1}^{n'}x_i \text{ and } \hat{\sigma}_x^{'2}=\frac{1}{n'}\sum_{i=1}^{n'}(x_i-\bar{x}')^2$$

be the sample mean and sample variance of auxiliary characteristic x based on first phase sample of size n'. Also based on second phase sub-sample of size n, let

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad \text{and} \quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

be the sample mean of auxiliary characteristic x and characteristic y under study, respectively,

$$s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \text{ and } s_y^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

be the sample variance of characteristic x and characteristic y, respectively,

$$s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

be the sample covariance between x and y, and

$$b=\frac{s_{xy}}{s_x^2}$$

be the sample regression coefficient of y on x.

A Proposed Double Sampling Estimator

A proposed double sampling estimator of population mean $\bar{Y}$ is $\hat{\bar{Y}}$

$$\hat{\bar{Y}}_{\theta D}=\bar{y}\left\{1+\frac{\theta(\hat{\sigma}_x^2-\hat{\sigma}_x^{'2})}{\hat{\sigma}_x^{'2}}\right\}+b(\bar{x}'-\bar{x})$$

$$=\bar{y}+\theta\bar{y}\left(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_x^{'2}}-1\right)+b(\bar{x}'-\bar{x}) \tag{1}$$

where θ is the characterizing scalar to be determined suitably. Note that for θ=0 the proposed estimator reduces to the usual double sampling linear regression estimator

$$\bar{y}_{lrD}=\bar{y}+b(\bar{x}'-\bar{x}) \tag{2}$$

Let $\bar{y}=\bar{Y}+e_0$ $\quad \bar{x}=\bar{X}+e_1$ $\quad \bar{x}'=\bar{X}+e_1'$

$s_{xy}=S_{xy}+e_2$ $\quad s_x^2=S_x^2+e_3$

$\hat{\sigma}_x^2=\sigma_x^2+e_4$ $\quad \hat{\sigma}_x^{'2}=\sigma_x^2+e_4'$

with

$$E(e_0)=E(e_1)=E(e_1^{'})=E(e_2)=E(e_3)=E(e_4)=E(e_4^{'})=0 \quad (3)$$

Then by using (1), we have

$$\hat{\hat{Y}}_{\theta D}=\bar{Y}+e_0+\frac{\theta\bar{Y}}{\sigma_x^2}\left(e_4-e_4^{'}+\frac{e_0e_4}{\bar{Y}}-\frac{e_0e_4^{'}}{\bar{Y}}-\frac{e_4e_4^{'}}{\sigma_x^2}+\frac{e_4^{'2}}{\sigma_x^2}\right)+\beta\left(e_1^{'}-e_1+\frac{e_1^{'}e_2}{S_{xy}}-\frac{e_1e_2}{S_{xy}}-\frac{e_1^{'}e_3}{S_x^2}+\frac{e_1e_3}{S_x^2}\right) \quad (4)$$

Using the results given in Sukhatme and Sukhatme,[4] Bhushan[1] and Bhushan *et al.*[9]

$$E(e_0e_4)=\gamma_n\mu_{21} \qquad E(e_0e_4^{'})=\gamma_{n'}\mu_{21}$$
$$E(e_1e_2)=\gamma_n\mu_{21} \qquad E(e_1^{'}e_2)=\gamma_{n'}\mu_{21}$$
$$E(e_1e_3)=\gamma_n\mu_{30} \qquad E(e_1^{'}e_3)=\gamma_{n'}\mu_{30}$$
$$E(e_4^{'2})=E(e_4e_4^{'})=\gamma_{n'}(\mu_{40}-\mu_{20}^{2})$$
$$E(e_0^{2})=\gamma_nS_y^{2} \qquad E(e_1^{2})=\gamma_nS_x^{2}$$
$$E(e_0e_1)=\gamma_nS_{xy} \qquad E(e_4^{2})=\gamma_n(\mu_{40}-\mu_{20}^{2})$$
$$E(e_1e_4)=\gamma_n\mu_{30} \qquad E(e_0e_4)=\gamma_n\mu_{21}$$
$$E(e_0e_1^{'})=\gamma_{n'}S_{xy} \qquad E(e_1^{'2})=E(e_1e_1^{'})=\gamma_{n'}S_x^{2}$$
$$E(e_4^{'2})=E(e_4e_4^{'})=\gamma_{n'}(\mu_{40}-\mu_{20}^{2})$$
$$E(e_0e_4^{'})=\gamma_{n'}\mu_{21}$$
$$E(e_1e_4^{'})=E(e_1^{'}e_4)=E(e_1^{'}e_4^{'})=\gamma_{n'}\mu_{30} \quad (5)$$

where $\gamma_n=(N-n)/Nn$ and $\gamma_{n'}=(N-n')/Nn'$.

Using (4) and (5) it can be seen that

$$Bias(\hat{\hat{Y}}_{\theta D})=E(\hat{\hat{Y}}_{\theta D})-\bar{Y}=(\gamma_n-\gamma_{n'})\left\{\frac{\theta\mu_{21}}{\sigma_x^2}-\beta\left(\frac{\mu_{21}}{S_{xy}}-\frac{\mu_{30}}{S_x^2}\right)\right\} \quad (6)$$

showing that $\hat{\hat{Y}}_{\theta D}$ is a biased estimator of population mean. Using (4) and neglecting terms of e_i's having powers greater than two, we get the MSE given by

$$MSE\left(\hat{\hat{Y}}_{\theta D}\right)=(\gamma_n-\gamma_{n'})[(1-\rho^2)S_y^{2}+\frac{\theta^2\bar{Y}^2}{\sigma_x^4}(\mu_{40}-\mu_{20}^{2})+\frac{2\theta\bar{Y}}{\sigma_x^2}\mu_{21}-\frac{2\beta\theta\bar{Y}}{\sigma_x^2}\mu_{30}]+\gamma_{n'}S_y^{2} \quad (7)$$

which is minimum for the optimum value of θ given by

$$\theta_{opt}=\frac{(\beta\mu_{30}-\mu_{21})\mu_{20}}{\bar{Y}(\mu_{40}-\mu_{20}^{2})} \quad (8)$$

and the minimum mean square error of $\hat{\hat{Y}}_{\theta D}$ is given by

$$MSE\left(\hat{\hat{Y}}_{\theta D}\right)_{\min}=(\gamma_n-\gamma_{n'})[(1-\rho^2)S_y^{2}-\frac{(\beta\mu_{30}-\mu_{21})^2}{\mu_{20}^{2}(\beta_2-1)}]+\gamma_{n'}S_y^{2} \quad (9)$$

A More Generalized Class of Double Sampling Estimators

A more generalized estimator of population mean $\bar{Y}$ is proposed to be

$$\hat{\hat{Y}}_{gd}=\bar{y}g(w)+b(\bar{x}^{'}-\bar{x}) \quad (10)$$

where $w=\frac{\hat{\sigma}_x^{2}}{\hat{\sigma}_x^{'2}}$; $g(w)$ is a function of w such that $g(w)$=1 at w=1 satisfying the following conditions:

1. Whatever be the sample chosen, w assumes values in the bounded closed interval I of the real line containing the point unity.
2. Within the interval I the function $g(w)$ is continuous and bounded.
3. The first, second and third partial derivatives of $g(w)$ exist and are continuous and bounded in I.

By expanding $g(w)$ about the point w=1 in the third order Taylor's series, we have

$$\hat{\hat{Y}}_{gd} = \bar{y}\left\{g(1) + (w-1)g'(1) + \frac{(w-1)^2}{2!}g''(1) + \frac{(w-1)^3}{3!}g'''(w^*)\right\} + b(\bar{x}' - \bar{x}) \quad (11)$$

where $w^* = 1+\psi(w-1), 0<\psi<1$ and ψ may depend on w; $g'(1)$, $g''(1)$ and $g'''(w^*)$ denote the first, second and third order derivatives of $g(w)$ at the point w=1,1 and w^*, respectively.

Using the notations given in (3), we have

$$w = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_x'^2} = \left(1 + \frac{e_4}{\sigma_x^2}\right)\left(1 + \frac{e_4'}{\sigma_x^2}\right)^{-1} \quad \text{and}$$

$$w - 1 = \frac{\hat{\sigma}_x^2 - \hat{\sigma}_x'^2}{\hat{\sigma}_x'^2} = \left(\frac{e_4 - e_4'}{\sigma_x^2}\right)\left(1 + \frac{e_4'}{\sigma_x^2}\right)^{-1}$$

Putting these values in (11) and neglecting the terms of e_i's (i=0,1,2,3,4) and e_i''s (i=2,4) having powers greater than two, we get

$$\hat{\hat{Y}}_{gd} = \bar{Y} + e_0 + \left(e_4 - e_4' - \frac{e_4 e_4'}{\sigma_x^2} + \frac{e_4'^2}{\sigma_x^2}\right)\frac{\bar{Y}g'(1)}{\sigma_x^2} + (e_0e_4 - e_0e_4')\frac{g'(1)}{\sigma_x^2} + (e_4^2 + e_4'^2 - 2e_4e_4')\frac{\bar{Y}g''(1)}{2\sigma_x^4} + \beta\left(e_1' - e_1 + \frac{e_1'e_2}{S_{xy}} - \frac{e_1e_2}{S_{xy}} - \frac{e_1'e_3}{S_x^2} + \frac{e_1e_3}{S_x^2}\right) \quad (12)$$

Taking expectation, we get

$$Bias(\hat{\hat{Y}}_{gd}) = (\gamma_n - \gamma_{n'})\left[\frac{g'(1)}{\sigma_x^2}\mu_{21} + \frac{\bar{Y}g''(1)}{2\sigma_x^4}(\mu_{40} - \mu_{20}{}^2) + \beta\left(\frac{\mu_{30}}{S_x^2} - \frac{\mu_{21}}{S_{xy}}\right)\right] \quad (13)$$

showing that $\hat{\hat{Y}}_{gd}$ is a biased estimator of population mean. Also, the mean square error of the estimator is given by

$$MSE(\hat{\hat{Y}}_{gd}) = \gamma_{n'}S_y^2 + (\gamma_n - \gamma_{n'})[(1-\rho^2)S_y^2 + \frac{\bar{Y}^2\{g'(1)\}^2}{\sigma_x^4}(\mu_{40} - \mu_{20}{}^2) + \frac{2\bar{Y}g'(1)}{\sigma_x^2}\mu_{21} - \frac{2\beta\bar{Y}g'(1)}{\sigma_x^2}\mu_{30}] \quad (14)$$

(14) is minimized when

$$g'(1) = \frac{(\beta\mu_{30} - \mu_{21})\mu_{20}}{\bar{Y}(\mu_{40} - \mu_{20}{}^2)} \quad (15)$$

and the minimum mean square error of $\hat{\hat{Y}}_{gd}$ is same as that of expression given by (9).

Concluding Remarks

The proposed generalized classes of estimators $\hat{\hat{Y}}_{\theta D}$ and $\hat{\hat{Y}}_{gd}$ are biased and their biases are given by (6) and (13) respectively. The minimum mean square errors of the proposed generalized classes are equal and are given by (9). Therefore, the proposed generalized classes of estimators $\hat{\hat{Y}}_{\theta D}$ and $\hat{\hat{Y}}_{gd}$ are preferred to usual linear regression estimator, ratio estimator, mean per unit estimator and product estimator in the sense of lesser mean square error. In the class $\hat{\hat{Y}}_{gd}$ of estimators, there exists a subclass of optimum estimators satisfying (15) such that every member of the subclass attains the same minimum mean square error given by (9). For example, for

$$\theta = g'(1) = \frac{(\beta\mu_{30} - \mu_{21})\mu_{20}}{\bar{Y}(\mu_{40} - \mu_{20}{}^2)} = g'(1)_{opt}$$

the estimator $\hat{\hat{Y}}_{\theta D}$ belongs to the sub-class of estimators attaining the minimum mean square error given by (9). Further, a double sampling ratio estimator based on Singh[8] is

$$\bar{y}_{sd} = \bar{y}\frac{(\bar{x}' + \sigma_x)}{(\bar{x} + \sigma_x)}$$ having MSE given by

$$MSE(\bar{y}_{sd}) = (\gamma_n - \gamma_{n'})[S_y{}^2 + R^2\delta^2 S_x{}^2 - 2R\delta S_{xy}] + \gamma_{n'}S_y{}^2$$

where $\delta = \dfrac{\bar{X}}{\bar{X} + \sigma_x}$ such that

$$MSE(\bar{y}_{sd}) - MSE(\hat{\hat{Y}}_{\theta D})_{\min} = (\gamma_n - \gamma_{n'})[(\rho S_y - R\delta S_x)^2 + \frac{(\beta\mu_{30} - \mu_{21})^2}{\mu_{20}^{\ 2}(\beta_2 - 1)}] \geq 0$$

showing that the proposed class of estimators are better than the double sampling Singh[8] estimator. Also, the parameters involved in the $g'(1)_{opt}$ may be estimated by the corresponding sample values in order to get a class of estimators depending upon estimated optimum value.

Empirical Study

The gain in precision of the proposed estimator(s) versus the usual double sampling linear regression estimator is studied for thirty two populations as provided in Bhushan[1] and Bhushan *et al*.[9] The table given below provides the gain in precision when the proposed estimator(s) are used over the double sampling linear regression estimator.

Table 1. Gain in precision of the proposed estimators over linear regression estimator.

Population Number	**Gain** $(\hat{\hat{Y}}_{\theta D} / \hat{\hat{Y}}_{gd})$	**Population Number**	**Gain** $(\hat{\hat{Y}}_{\theta D} / \hat{\hat{Y}}_{gd})$
1	2.58583	17	300.001
2	9.08774	18	232.916
3	0.224016	19	10.8319
4	1.36995	20	11.1082
5	6.46651	21	0.0658345
6	25.9388	22	5.58773
7	1.04419	23	13.1132
8	0.035816	24	4.36055
9	6.54509	25	5.7562
10	65.492	26	2.78404
11	1.09776	27	0.0496923
12	2.58583	28	0.323048
13	0.0787519	29	0.586345
14	4.43361	30	11.9976
15	68.5157	31	5.47894
16	10.3939	32	2.16561

References

1. Bhushan S (2007). *Some Improved Sampling Strategies in Finite Population*. Ph.D. thesis, University of Lucknow, Lucknow, U.P., India.
2. Singh D & Chaudhary FS (2002). *Theory and Analysis of Sampling Survey Design*. New Age Pvt. Ltd., New Delhi.
3. Mukhopadhyaya P (1998). *Theory and Methods of Survey Sampling*. Prentice-Hall India, New Delhi.
4. Sukhatme PV & Sukhatme BV (1997). *Sampling Theory of Surveys with Applications*. Piyush Publications, Delhi.
5. Kumar R (1995). *Some Contributions to Survey Sampling*. Unpublished thesis submitted and accepted by Lucknow University, U.P., India.
6. Murthy MN (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.
7. Cochran WG (1977). *Sampling Techniques* 3rd edn. John Wiley and Sons, New York.
8. Singh GN (2003). On the improvement of product method of estimation in sample surveys, *J Ind Soc Agr Stat*, **56**, 267-275.
9. Bhushan S, Mishra P & Singh RK (2010). A class of regression type estimators using mean and variance of auxiliary variable (under communication).

