Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 255-260

Variance Estimation for the Regression Estimator of the Mean in Stratified Sampling

Sat Gupta[1*] and Javid Shabbir[2]

[1]*Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC 27402, USA*

[2]*Department of Statistics, Quaid-i-Azam University, Islamabad 45320, Pakistan*

Received 18 January 2010; Revised 29 April 2010; Accepted 30 April 2010

SUMMARY

In this paper, we propose a class of estimators for variance of separate regression estimator of mean in stratified sampling and derive its properties under large sample approximation. The proposed class of estimators performs better than the traditional regression estimator and the Wu (1985) estimator. Mean square errors of different estimators are compared numerically also using three different data sets from the literature.

Keywords: Separate regression estimator, Stratification, Bias, Mean square error, Efficiency.

1. INTRODUCTION

Let U be a finite population of size N. The study variable and the auxiliary variable are denoted by y and x respectively and the population is partitioned into L non-overlapping strata according to some characteristics. The size of the h^{th} stratum is $N_h (h = 1, 2, ..., L)$ such that $\sum_{h=1}^{L} N_h = N$. A stratified sample of size n is drawn from this population and let n_h be sample size from h^{th} stratum such that $\sum_{h=1}^{L} n_h = n$. The observations on y and x corresponding to i^{th} unit of h^{th} stratum ($h = 1, 2, ..., L$) are y_{hi} and x_{hi} respectively. Let $\bar{y}_h$ and $\bar{x}_h$ be sample means and $\bar{Y}_h$ and $\bar{X}_h$ be population means of y and x respectively in h^{th} stratum. Suppose $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ are stratified sample means and $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$ and $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$ are population means of y and x respectively, where $W_h = N_h/N$ is known stratum weight. Let $s_{yh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ and $s_{xh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ be sample variances and $S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$ and $S_{xh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_h)^2$ be population variances of y and x respectively in h^{th} stratum. Finally, let $s_{yxh} =$

*Corresponding author : Sat Gupta
E-mail address : sngupta@uncg.edu

$\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)(x_{hi}-\bar{x}_h)$ and $S_{yxh} = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)(x_{hi}-\bar{X}_h)$ be sample and population covariances respectively in h^{th} stratum. We assume that all parameters corresponding to auxiliary variable x are known. In subsequent presentation, we ignore finite population correction term $(1 - n_h/N_h)$ for computational ease.

A well-known estimator of $\bar{Y}$ is separate regression estimator $\bar{y}_S = \sum_{h=1}^{L} W_h\left\{\bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)\right\}$, where b_h is sample regression coefficient. Variance of $\bar{y}_S$ is given by

$$V(\bar{y}_s) = \sum_{h=1}^{L} W_h^2 S_{yh}^2(1-\rho_h^2)/n_h \tag{1}$$

where $\rho_h = S_{yxh}\big/\left(S_{yh}S_{xh}\right)$ is population correlation coefficient between y and x in h^{th} stratum.

The primary objective of this paper is to present an estimator of $V(\bar{y}_S)$ and compare it with some of the known estimators. An obvious estimator of $V(\bar{y}_S)$ is given by

$$v_s = \sum_{h=1}^{L} W_h^2 s_{yh}^2(1-r_h^2)/n_h \tag{2}$$

where $r_h = s_{yxh}\big/\left(s_{yh}s_{xh}\right)$ is sample correlation coefficient between y and x in h^{th} stratum. Properties of v_s can be derived easily once we define the following error terms:

$$\varepsilon_{0h} = \frac{s_{yh}^2 - S_{yh}^2}{S_{yh}^2},\ \varepsilon_{1h} = \frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h}$$

$$\varepsilon_{2h} = \frac{s_{xh}^2 - S_{xh}^2}{S_{xh}^2} \text{ and } \varepsilon_{3h} = \frac{s_{yxh} - S_{yxh}}{S_{yxh}}$$

It can be verified that $E(\varepsilon_{ih}) = 0$ $(i = 0, 1, 2, 3)$. Also up to first order of approximation, we have the following expectations that can be derived easily on the lines of Sukhatme *et al.* (1997):

$$E(\varepsilon_{0h}^2) = \frac{1}{n_h}(\lambda_{40h}-1),\ E(\varepsilon_{1h}^2) = \frac{1}{n_h}C_{xh}^2$$

$$E(\varepsilon_{2h}^2) = \frac{1}{n_h}(\lambda_{04h}-1),\ E(\varepsilon_{3h}^2) = \frac{1}{n_h}\left(\frac{\lambda_{22h}}{\rho_h^2}-1\right)$$

$$E(\varepsilon_{0h}\varepsilon_{ih}) = \frac{1}{n_h}\lambda_{21h}C_{xh},\ E(\varepsilon_{0h}\varepsilon_{2h}) = \frac{1}{n_h}(\lambda_{22h}-1)$$

$$E(\varepsilon_{0h}\varepsilon_{3h}) = \frac{1}{n_h}\left(\frac{\lambda_{31h}}{\rho_h}-1\right),\ E(\varepsilon_{1h}\varepsilon_{2h}) = \frac{1}{n_h}\lambda_{03h}C_{xh}$$

$$E(\varepsilon_{1h}\varepsilon_{3h}) = \frac{1}{n_h}\left(\frac{\lambda_{12h}C_{xh}}{\rho_h}\right)$$

$$E(\varepsilon_{2h}\varepsilon_{3h}) = \frac{1}{n_h}\left(\frac{\lambda_{13h}}{\rho_h}-1\right)$$

where

$$\lambda_{pqh} = \frac{\mu_{pqh}}{\mu_{20h}^{p/2}\mu_{02h}^{q/2}}$$

and $\mu_{pqh} = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y}_h)^p(x_{hi}-\bar{X}_h)^q$

Now writing v_s in term of ε's, we have

$$v_s = \sum_{h=1}^{L}\frac{1}{n_h}W_h^2\left[S_{yh}^2(1+\varepsilon_{0h}) - \frac{S_{yxh}^2(1+\varepsilon_{3h})^2}{S_{xh}^2(1+\varepsilon_{2h})}\right] \tag{3}$$

From (3), the bias and *MSE* of v_s are given by

$$B(V_h) \cong -\sum_{h=1}^{L} W_h^2 S_{yh}^2\rho_h^2 A_h/n_h^2 \tag{4}$$

where

$$A_h = (\lambda_{04h}-1) + \left(\frac{\lambda_{22h}}{\rho_h^2}-1\right) - 2\left(\frac{\lambda_{13h}}{\rho_h}-1\right)$$

and

$$MSE(v_s) \cong \sum_{h=1}^{L} W_h^4 S_{yh}^4$$

$$\left[(\lambda_{40h}-1)+\rho_h^4 B_h + 2\rho_h^2 C_h\right]\Big/ n_h^3 \qquad (5)$$

where

$$B_h = (\lambda_{04h}-1)+4\left(\lambda_{22h}/\rho_h^2 - 1\right)$$

$$-4\left(\lambda_{13h}/\rho_h - 1\right)$$

and $\quad C_h = (\lambda_{22h}-1)-2\left(\lambda_{31h}/\rho_h - 1\right)$

Many authors have presented estimators that exploit a variety of information available from the auxiliary variable. These include Das and Tripathi (1981), Srivastava and Jhajj (1980, 1983), Wu (1985), Prasad and Singh (1990, 1992). In particular, Wu (1985) has presented an estimator of $V(\bar{y}_S)$ which is given by

$$v_W = \sum_{h=1}^{L} W_h^2 s_{yh}^2 (1-r_h^2)\left(\bar{X}_h/\bar{x}_h\right)^{g_h} \Big/ n_h \qquad (6)$$

where g_h is an appropriately chosen constant. Properties of this estimator will be discussed in the next section.

2. PROPOSED ESTIMATOR

Our proposed estimator of $V(\bar{y}_S)$ is an extension of the Wu (1985) estimator where we exploit information on both $\bar{X}_h$ and S_{xh}^2. The new estimator is given by

$$v_p = \sum_{h=1}^{L} \frac{1}{n_h} W_h^2 s_{yh}^2 (1-r_h^2)\left(\frac{\bar{X}_h + a_k}{\bar{x}_h + a_k}\right)^{k_{1h}} \left(\frac{S_{xh}^2 + b_k}{s_{xh}^2 + b_k}\right)^{k_{2h}} \qquad (7)$$

where k_{ih} (i = 1, 2) are constants whose values are to be determined and a_k and b_k are some known unit free constants, such as the coefficient of variation (C_{xh}), coefficient of kurtosis (λ_{04h}) or coefficient of correlation (ρ_h). Note that

(i) For $k_{1h} = 0$ and $k_{2h} = 0$, we have $v_P = v_S$.

(ii) For $k_{1h} = g_h$, $a_k = 0$ and $k_{2h} = 0$, we have $v_P = v_W$.

Now we derive the properties of v_P under large sample approximation. Writing v_P in term of ε's, we have

$$v_P = \sum_{h=1}^{L} \frac{1}{n_h} W_h^2 \left[S_{yh}^2 (1+\varepsilon_{0h}) - \frac{S_{yxh}^2 (1+\varepsilon_{3h})^2}{S_{xh}^2 (1+\varepsilon_{2h})} \right]$$

$$\left[(1+\phi_{1h}\varepsilon_{1h})^{-k_{1h}} (1+\phi_{2h}\varepsilon_{2h})^{-k_{2h}}\right] \qquad (8)$$

where

$$\phi_{1h} = \bar{X}_h \Big/ \left(\bar{X}_h + a_k\right)$$

and $\quad \phi_{2h} = S_{xh}^2 \Big/ \left(S_{xh}^2 + b_k\right)$

From (8), the bias of v_P is given by

$$B(v_p) \cong \sum_{h=1}^{L} \frac{1}{n_h^2} W_h^2 S_{yh}^2 \Bigg[(1-\rho_h^2) \Big\{ k_{1h} k_{2h} \phi_{1h} \phi_{2h}$$

$$(\lambda_{03h} C_{xh}) + \frac{1}{2} k_{1h}(k_{1h}+1)\phi_{1h}^2 C_{xh}^2$$

$$+\frac{1}{2} k_{2h}(k_{2h}+1)(\lambda_{04h}-1)\Big\}$$

$$+\rho_h^2 \Big\{ 2k_{1h}\phi_{1h}\left(\frac{\lambda_{12h} C_{xh}}{\rho_h}\right) + 2k_{2h}\phi_{2h}\left(\frac{\lambda_{13h}}{\rho_h} - 1\right)$$

$$-k_{1h}\phi_{1h}\lambda_{03h} C_{xh} - k_{2h}(\lambda_{04h}-1)$$

$$+2\left(\frac{\lambda_{13h}}{\rho_h} - 1\right) - (\lambda_{04h}-1) - \left(\frac{\lambda_{22h}}{\rho_h^2} - 1\right)\Big\}$$

$$-\left\{k_{1h}\phi_{1h}\lambda_{21h} C_{xh} + k_{2h}\phi_{2h}(\lambda_{22h}-1)\right\}\Bigg] \qquad (9)$$

Substituting $k_{1h} = g_h$, $k_{2h} = 0$ and $a_k = 0 = b_k$ i.e. $(\phi_{1h} = \phi_{2h} = 1)$ in (9), we can get the bias expression for the Wu (1985) estimator as

$$B(v_w) \cong \sum_{h=1}^{L} \frac{1}{n_h^2} W_h^2 S_{yh}^2 \left[(1-\rho_h^2)\frac{1}{2} g_h(g_h+1)\, C_{xh}^2 \right.$$

$$+\rho_h^2 \left\{ 2g_h \frac{\lambda_{12h} C_{xh}}{\rho_h} - g_h \lambda_{03h} C_{xh} + 2\left(\frac{\lambda_{13h}}{\rho_h} - 1 \right) \right.$$

$$\left. \left. -(\lambda_{04h}-1) - \left(\frac{\lambda_{22h}}{\rho_h^2} - 1 \right) \right\} - g_h \lambda_{21h} C_{xh} \right] \quad (10)$$

Also from (8), the MSE of v_p is given by

$$MSE(v_P) \cong MSE(v_s) + \sum_{h=1}^{L} \frac{1}{n_h^3} W_h^4 S_{yh}^4 (1-\rho_h^2)^2$$

$$\left[\left\{ k_{1h}^2 \phi_{1h}^2 C_{xh}^2 + k_{2h}^2 \phi_{2h}^2 (\lambda_{04h}-1) \right. \right.$$

$$\left. + 2k_{1h} k_{2h} \phi_{1h} \phi_{2h} \lambda_{03h} C_{xh} \right\}$$

$$\left. - \frac{2}{(1-\rho_h^2)} \left\{ k_{1h}\phi_{1h} D_h + k_{2h} \phi_{2h} E_h \right\} \right] \quad (11)$$

where

$$D_h = \lambda_{21h} C_{xh} - \rho_h^2 C_{xh} \left\{ 2\frac{\lambda_{12h}}{\rho_h} - \lambda_{03h} \right\}$$

and $$E_h = (\lambda_{22h}-1) - \rho_h^2 \left\{ 2\left(\frac{\lambda_{13h}}{\rho_h} - 1 \right) - (\lambda_{04h}-1) \right\}$$

From (11), optimum values of k_{ih} (i = 1, 2) are given by

$$k_{1h} = \frac{(\lambda_{04h}-1) D_h - \lambda_{03h} C_{xh} E_h}{\phi_{1h} C_{xh}^2 (1-\rho_h^2)(\lambda_{04h} - \lambda_{03h}^2 - 1)}$$

and $$k_{2h} = \frac{C_{xh} E_h - \lambda_{03h} C_{xh} D_h}{\phi_{2h} C_{xh}^2 (1-\rho_h^2)(\lambda_{04h} - \lambda_{03h}^2 - 1)}$$

Substituting optimum values of k_{ih} (i = 1, 2) in (11), we get minimum MSE of v_P as

$$MSE(v_P)_{\min} \cong MSE(v_s) - \sum_{h=1}^{L} \frac{1}{n_h^3} W_h^4 S_{yh}^4$$

$$\left[\frac{D_h^2}{C_{xh}^2} + \frac{(C_{xh}E_h - \lambda_{03h} D_h)^2}{C_{xh}^2 \left\{ \lambda_{04h} - \lambda_{03h}^2 - 1 \right\}} \right] \quad (12)$$

Note that substituting $k_{1h} = g_h$, $k_{2h} = 0$ and $a_k = 0 = b_k$, i.e. $(\phi_{1h} = \phi_{2h} = 1)$ in (11), we can get the MSE of v_W as

$$MSE(v_W) \cong MSE(v_s) + \sum_{h=1}^{L} \frac{1}{n_h^3} W_h^4 S_{yh}^4 (1-\rho_h^2)^2$$

$$\left\{ g_h^2 C_{xh}^2 - 2g_h \frac{D_h}{(1-\rho_h^2)} \right\} \quad (13)$$

From (13), the optimum value of g_h is $g_h^* = D_h \Big/ \left[C_{xh}^2 (1-\rho_h^2) \right]$. Putting optimum value of g_h in (13), we can get the minimum MSE of v_W as

$$MSE(v_W)_{\min} \cong MSE(v_S) - \sum_{h=1}^{L} W_h^4 S_{yh}^4 D_h^2 \Big/ \left(n_h^3 C_{xh}^2 \right) \quad (14)$$

Note that the optimum values of k_{ih} (i = 1, 2) involve unknown parameters. Wu (1985) pointed out that, if computational ease is not an issue and sample size is such that the asymptotic results become effective, then estimators based on estimated optimal values of k_{ih} (i = 1, 2) can be used.

3. COMPARISON OF ESTIMATORS

We now compare the proposed estimator with two other estimators discussed above. It is easy to verify that

(i) $MSE(v_P)$min < $MSE(v_S)$ if

$$\sum_{h=1}^{L} \frac{1}{n_h^3} W_h^2 S_{yh}^4 \left[\frac{D_h^2}{C_{xh}^2} + \frac{(C_{xh}E_h - \lambda_{03h} D_h)^2}{C_{xh}^2 \left\{ \lambda_{04h} - \lambda_{03h}^2 - 1 \right\}} \right] > 0$$

(ii) $MSE(v_P)_{\min} < MSE(v_W)_{\min}$ if

$$\sum_{h=1}^{L} \frac{1}{n_h^3} W_h^2 S_{yh}^4 \left[\frac{(C_{xh}E_h - \lambda_{03h}D_h)^2}{C_{xh}^2 \left\{ \lambda_{04h} - 1 - \lambda_{03h}^2 \right\}} \right] > 0$$

Conditions (i) and (ii) always hold true because $(\lambda_{04h} - \lambda_{03h}^2 - 1) \geq 0$ (see Jhajj *et al.* 2005). We use the following expression for obtaining the percent relative efficiencies with respect to v_S:

$$\text{PRE} = \frac{MSE(v_S)}{MSE(v_i)} \times 100, \; i = S, W, P$$

The results reported in Table 1 are based on three data sets given in the Appendix.

Table 1. Percent relative efficiency of different estimators with respect to v_S

Estimator	Data 1	Data 2	Data 3
v_S	100.000	100.000	100.000
v_W	128.460	142.713	111.970
v_P	164.431	156.052	136.557

Results in Table 1 show that the performance of proposed estimator is better than the other competing estimators. This was clearly expected based on Conditions (i) and (ii) above which always hold true. Thus utilizing the information on population variance in addition to the population mean of x can further improve efficiency of the Wu (1985) estimator.

ACKNOWLEDGEMENT

The authors are thankful to the referees for their valuable suggestions that resulted in significant improvement in the presentation of this paper.

REFERENCES

Das, A.K. and Tripathi, T.P. (1981). A class of sampling strategies for population mean using information on mean and variance of an auxiliary character. Technical Report No. 23/81, Stat. and Math. Division , ISI, Calcutta.

Jhajj, H.S., Sharma, M.K. and Grover, L.K. (2005). An efficient class of chain estimators of population variance under sub-sampling scheme. *J. Jap. Stat. Soc.*, **35**, 273-286.

Prasad, B. and Singh, H.P. (1990). Some improved ratio-type estimators of finite population variance in sample survey. *Comm. Statist.- Theory Methods*, **19**, 1127-1139.

Prasad, B. and Singh, H.P. (1992). Unbiased estimators of finite population variance using auxiliary information in sample survey. *Comm. Statist.- Theory Methods*, **21**, 1367-1376.

Singh, R. and Mangat, N.S. (1996). *Elements of Survey Sampling*. Kluwer Academic Publisher.

Srivastava, S.K. and Jhajj, H.S. (1980). A class of estimators using auxiliary information for estimating the finite population variance. *Sankhya*, **C42**, 87-96.

Srivastava, S.K. and Jhajj, H.S. (1983). A class of estimators of mean and variance using auxiliary information when correlation coefficient is known. *Biom. J.*, **25**, 401-407.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1997). *Sampling Theory of Surveys with Applications*. 3rd ed. Iowa State University Press, Ames, Iowa, U.S.A. and Indian Society of Agricultural Statistics, New Delhi, India.

Wu, C.F.J. (1985). Variance estimation for the combined ratio and combined regression estimator. *J. Roy. Statist. Soc.*, **47**, 147-154.

APPENDIX

Data 1 : [Source: Singh and Mangat (1996, p. 212)]

y : leaf area for the newly developed strain of wheat and x: weight of leaves.

$N = 39$, $N_1 = 12$, $N_2 = 13$, $N_3 = 14$, $L = 3$, $n = 14$, $n_1 = 4$, $n_2 = 5$, $n_3 = 3$

Stratum		Values of Parameters											
No.	$\bar{Y}_h$	$\bar{X}_h$	S_{yh}	C_{xh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	25.75	103.40	6.066460	0.111893	0.920237	0.429305	0.509785	0.235990	1.912346	2.274823	1.939455	2.025797	1.879711
2	28.94	110.90	5.291523	0.073375	0.915402	0.996098	0.081555	1.034165	2.970998	3.436904	2.981927	3.096674	2.930390
3	25.84	104.30	6.496130	0.119378	0.966819	0.205762	0.297117	0.083846	2.513437	2.895550	2.344898	2.675952	2.398860

Data 2 : [Source: Singh and Mangat (1996, p. 219)]

y : juice quantity and x: weight of cane (Kg).

$N = 25$, $N_1 = 6$, $N_2 = 12$, $N_3 = 7$, $L = 3$, $n = 10$, $n_1 = 3$, $n_2 = 4$, $n_3 = 3$

Stratum		Values of Parameters											
No.	$\bar{Y}_h$	$\bar{X}_h$	S_{yh}	C_{xh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	135.00	366.67	8.16496	0.141888	0.945563	0.576173	0.649223	0.459841	2.261624	2.286563	2.343750	2.258679	2.288691
2	99.17	310.83	14.40968	0.139532	0.948196	0.985721	0.973885	0.946518	3.379509	3.268973	3.792407	3.277747	3.548659
3	80.71	317.14	10.15191	0.169523	0.753223	1.035401	0.891565	0.858180	2.311771	3.130635	2.329428	2.487514	2.217034

Data 3 : [Source: Singh and Mangat (1996, p. 220)]

y : total number of milch cows in 1993 and x: total number of milch cows in 1990.

$N = 24$, $N_1 = 7$, $N_2 = 12$, $N_3 = 5$, $L = 3$, $n = 10$, $n_1 = 3$, $n_2 = 5$, $n_3 = 2$

Stratum		Values of Parameters											
No.	$\bar{Y}_h$	$\bar{X}_h$	S_{yh}	C_{xh}	ρ_h	λ_{12h}	λ_{21h}	λ_{03h}	λ_{22h}	λ_{04h}	λ_{40h}	λ_{13h}	λ_{31h}
1	17.43	15.29	3.88613	0.299114	0.765459	– 0.441840	– 0.449445	0.038284	1.134807	1.849759	1.655367	1.092983	1.316937
2	19.58	17.25	3.90424	0.318601	0.406654	– 0.276272	– 0.244895	0.150792	0.569598	2.312027	2.750973	0.834902	0.674840
3	20.69	17.80	3.26190	0.183769	0.494577	– 0.811980	– 0.284742	– 0.569229	1.346146	1.833392	1.592543	1.112349	1.070460