Jour. Ind. Soc. Ag. Statistics
48(1), 1996 : 98–104

A Class of Almost Unbiased Regression-type Estimators in Two Phase Sampling Applying Quenouille's Method

Housila P. Singh, N.P. Katyar* and D.K. Gangwar*
School of Studies in Statistics, Vikram University, Ujjain - 456010
(Received : July, 1988)

SUMMARY

A class of almost unbiased regression-type estimators is proposed with the help of the Jack-knife technique developed by Quenouille [1] for simple random sampling in two phases. The mean square error/variance expressions of the resulting estimator is derived. Optimum estimator in the proposed class of estimators is also investigated and its mean square error/variance is compared with the usual biased linear regression estimator and it is found that they are approximately the same..

Keywords : Almost unbiased regression-type estimators, Simple random sampling, Optimum estimator, Variance/mean square error.

Introduction

If the necessary auxiliary information is not readily available for the population before sampling, such information is collected for a larger preliminary sample and then more precise information is collected for the variable under study on a final or second phase sample. The technique known as two phase sampling, consists of taking a larger sample of size n′ by simple random sampling without replacement (SRSWOR) to estimate $\overline{X} = N^{-1} \sum_{i=i}^{N} x_i$ the population mean of auxiliary character x while a subsample (or a second phase sample) of size n out of n′ units is drawn by SRSWOR to abserve the characteristic y under study. Denote $\overline{x}' = \sum_{i=1}^{n'} x_i / n'$ the sample mean of x based on n′ and $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ and $\overline{y} = n^{-1} \sum_{i=1}^{n} y_i$ the sample means of x and y, respectively. For estimating the population mean $\overline{Y} = N^{-1} \sum_{i=1}^{N} y_i$ of the study character y, the double (or two phase) sampling regression estimator is defined as

$$\overline{y}_{ld} = \overline{y} + \hat{\beta} (\overline{x}' - \overline{x}) \tag{1.1}$$

* J.N. Agriculture University, Jabalpur, M.P.

Income & Expenditure Account for the year ending 31st March 1995

Previous Year	Expenditure	Current Year
44,755.00	To Establishment Expenses	46,755.00
5,651.22	To Printing & Stationery	3,468.63
700.00	To Audit Fees	700.00
1,779.00	To Travelling & Conveyance	2,657.00
182.75	To Cycle Repairs	103.00
—	To Legal Exps.	3,300.00
1,362.00	To Depreciation on Assets	921.00
54,429.97		57,904.63

Previous Year	Income	Current Year
40,000.00	By Grant in Aid from Govt. of India, Ministry of Agriculture, Dept. of Agricultural Co-operation	40,000.00
150.00	By Sale of Cycle	—
14,279.97	By Deficit for the year trfd. to Balance Sheet	17,904.63
54,429.97		57,904.63

where $\hat{\beta} = S_{yx} / S_x^2$ is the sample regression coefficient of y on x,

$$S_{yx} = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \text{ and } S_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The bias and mean square error (MSE) of $\bar{y}_{ld}$, to the first degree of approximation, are respectively given by

$$B(\bar{y}_{ld}) = -\frac{N}{(N-2)} \cdot \frac{(n'-n)}{n'} \cdot \frac{\beta}{n} \left(\frac{\mu_{21}}{\mu_{11}} - \frac{\mu_{30}}{\mu_{20}} \right) \tag{1.2}$$

and

$$\text{MSE}(\bar{y}_{ld}) = \left[\left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 (1 - \rho^2) + \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 \right] \tag{1.3}$$

[See Sukhatme et.al. [2], page 245]

where $\mu_{rs} = \sum_{i=1}^{N} (x_i - \bar{X})^r (y_i - \bar{Y})^s / N$, $r \geq 0, s \geq 0$; (r, s) being non-negative integers, $S_y^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / (N-1)$ and ρ is the correlation coefficient between x and y.

It follows from (1·2) that the estimator $\bar{y}_{ld}$ is biased. To obtain almost unbiased estimator, divide the second phase sample at random, into k (≥ 2) sub-samples, each of size n/k (assumed to be an integer m) and consider the estimator of the form

$$\hat{Y}'_{Ld} = \frac{1}{k} \sum_{j=1}^{k} \bar{y}^*_{(ld)j} \tag{1·4}$$

where $\bar{y}^*_{(ld)j} = [\bar{y}^*_j + \hat{\beta}^* (\bar{x}' - \bar{x}^*_j)]$ is the standard regression estimates computed from the with group j omitted; j = 1, 2, ... k ; $\bar{y}^*_j = \dfrac{(n\bar{y} - m\bar{y}_j)}{(n-m)}$,

$\bar{x}^*_j = \dfrac{(n\bar{x} - m\bar{x}_j)}{(n-m)}$; $(\bar{y}_j, \bar{x}_j)$ being the j^{th} subsample means of y and x respectively;

$$\hat{\beta}^*_j = [(n-1) S_{yx} - (m-1) S_{yxj}] / [(n-1) S_x^2 - (m-1) S_{xj}^2],$$

$$S_{yxj} = \frac{1}{(m-1)} \sum_{i=1}^{m} (y_{ji} - \bar{y}_j)(x_{ji} - \bar{x}_j), \quad S_{xj}^2 = \frac{1}{(m-1)} \sum_{i=1}^{m} (x_{ji} - \bar{x}_j)^2$$

and m = n/k

The bias of $\hat{Y}'_{Ld}$ to the first degree of approximation is obtained as follows :

$$B(\hat{Y}'_{Ld}) = -\frac{N}{(N-2)} \cdot \frac{(n'-n+m)}{n'(n-m)} \beta \left(\frac{\mu_{21}}{\mu_{11}} - \frac{\mu_{30}}{\mu_{20}} \right) \tag{1.5}$$

The mean square error of $\hat{Y}'_{Ld}$ to the first degree of approximation is obtained as follows :

Define

$$\bar{y}_j^* = \bar{Y}(1+e_{0j}^*), \quad \bar{x}_j^* = \bar{X}(1+e_{1j}^*), \quad \bar{x}' = \bar{x}(1+e'_1) \text{ and } \hat{\beta}_j^* = \beta(1+e_{2j}^*)$$

such that $E(e_{0j}^*) = E(e_{1j}^*) = E(e'_1) = 0$

and

$$\left.\begin{aligned} E(\hat{\beta}_j^*) &= \beta + 0(n^{-1}) \\ \Rightarrow \quad E(e_{2j}^*) &= 0(n^{-1}) \end{aligned}\right\}$$

Expressing (1.4) in terms of e^*'s and e_1', we have

$$\hat{Y}'_{Ld} = \frac{1}{k}\sum_{j=1}^{k}[\bar{Y}(1+e_{0j}^*) + \beta\bar{X}(1+e_{2j}^*)(e_1' - e_{1j}^*)]$$

$$= \bar{Y} + \frac{\bar{Y}}{k}\sum_{j=1}^{k}\left[e_{0j}^* - \left(\frac{\beta\bar{X}}{\bar{Y}}\right)(e_{1j}^* - e_1' + e_{2j}^* e_{1j}^* - e_{2j}^* e_1')\right]$$

or

$$(\hat{Y}'_{Ld} - \bar{Y}) = \frac{\bar{Y}}{k}\sum_{j=1}^{k}\left[e_{0j}^* - \left(\rho\frac{C_y}{C_x}\right)(e_{1j}^* - e_1' + e_{2j}^* e_{1j}^* - e_1' e_{2j}^*)\right] \tag{1.6}$$

where $C_y = Sy/\bar{Y}$ and $C_x = S_x/\bar{X}$

Squaring both sides of (1.6) and ignoring terms involving e's having power greater than two, we have

$$(\bar{Y}'_{Ld} - \bar{Y})^2 = \frac{\bar{Y}^2}{k^2}\left[\sum_{j=1}^{k}\left\{e_{0j}^* - \left(\rho\frac{C_y}{C_x}\right)(e_{1j}^* - e'_1)\right\}\right]^2$$

$$= \frac{\bar{Y}^2}{k^2}\left[\sum_{j=1}^{k}\left\{e_{0j}^{*2} + \left(\rho\frac{C_y}{C_x}\right)^2(e_{1j}^{*2} - 2e_{1j}^* e'_1 + e_1'^2) - 2\left(\rho\frac{C_y}{C_x}\right)(e_{0j}^* e_{1j}^* - e_{0j}^* e'_1)\right\}\right.$$

$$+ \sum_{j \neq l = 1}^{k}\left\{e_{0j}^* e_{0l}^* - \left(\rho\frac{C_y}{C_x}\right)(e_{0j}^* e_{1l}^* - e_{0j}^* e'_1) - \left(\rho\frac{C_y}{C_x}\right)(e_{1j}^* e_{0l} - e'_1 e_{0l}^*)\right.$$

$$\left.\left. + \left(\rho\frac{C_y}{C_x}\right)^2(e_{1j}^* e_{1l}^* - e_{1j}^* e'_1 - e_{1l}^* e'_1 + e_1'^2)\right\}\right] \tag{1.7}$$

The following results can easily be established :

$$E(e_{0j}^{*2}) = \lambda C_y^2, \quad E(e_{1j}^{*2}) = \lambda C_x^2, \quad E(e_{0j}^{*} e_{1j}^{*}) = \lambda \rho C_y C_x,$$

$$E(e_1'^2) = E(e_{1j}^{*} e_1') = E(e_{11}^{*} e_1') = \lambda^* C_x^2,$$

$$E(e_{0j}^{*} e_1') = E(e_{01}^{*} e_1') = \lambda^* \rho C_y C_x,$$

$$E(e_{0j}^{*} e_{01}^{*}) = A C_y^2, E(e_{0j}^{*} e_{11}^{*}) = E(e_{01}^{*} e_{1j}^{*}) = A \rho C_y C_x,$$

$$E(e_{1j}^{*} e_{11}^{*}) = A C_x^2,$$

where

$$\lambda = \frac{(N-n+m)}{N(n-m)}, \lambda^* = \frac{(N-n')}{n'N} \text{ and } A = \frac{1}{(k-1)^2}\left[(k^2-2k)\left(\frac{1}{n}-\frac{1}{N}\right)-\frac{1}{N}\right]$$

Taking expectation of both sides of (1.7), and using the above results and after simplification, we get the MSE of Y'_{Ld} to the first degree of approximation as

$$\text{MSE}(\hat{Y}'_{Ld}) = \left[\left(\frac{1}{n}-\frac{1}{n'}\right)S_y^2(1-\rho^2)+\left(\frac{1}{n'}-\frac{L}{N}\right)S_y^2\right] = \text{MSE}(\bar{y}_{ld}) \quad (1.8)$$

2. *The Class of Estimators*

Taking the linear combination of $\bar{y}, \bar{y}_{ld}$ and $\hat{Y}'_{Ld}$, define the following class of estimators of $\bar{Y}$ as

$$\hat{Y} = w_1 \bar{y} + w_2 \bar{y}_{ld} + w_3 \hat{Y}'_{Ld} \quad (2.1)$$

where w_1, w_2 and w_3 are constants such that

$$w_1 + w_2 + w_3 = 1 \quad (2.2)$$

It follows from, (1.2) and (1.5) that an estimator in the class (2.1) is unbiased if and only if

$$\delta w_2 + w_3 = 0 \quad (2.3)$$

where $$\delta = \frac{(n'-n)(n-m)}{(n'-n+m)n} \quad (2.4)$$

If we set $w_2 = \alpha, w_3 = \theta$ and $w_1 = (1-\alpha-\theta)$ then unbiasedness condition in (2.3) reduces to

$$\theta = -\delta\alpha \quad (2.5)$$

(α, θ) being the chosen constants.

Thus, obtain a general class of almost unbiased regression-type estimators

$$\hat{\bar{Y}}_\alpha = \left[\{1 - \alpha(1-\delta)\}\bar{y} + \alpha\,\bar{y}_{ld} - \alpha\,\delta\,\frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j} \right] \quad (2.6)$$

for population mean $\bar{Y}$.

Remark 1.1. (i) For $\alpha = 0$, $\hat{\bar{Y}}_\alpha$ yields the usual unbiased estimator $\bar{y}$ while for $\alpha = (1-\delta)^{-1}$ it reduces to

$$\hat{\bar{Y}}_0 = \left[\frac{(n'-n+m)}{n'} \cdot k \cdot \bar{y}_{ld} - \frac{(n'-n)}{n'} \cdot \frac{(k-1)}{k} \cdot \sum_{j=1}^{k} \bar{y}^*_{(ld)j} \right], \quad (2.7)$$

(ii) For $\alpha = \delta^{-1}$, $\hat{\bar{Y}}_\alpha$ reduces to another almost unbiased estimator

$$\hat{\bar{Y}}_1 = \left[\frac{(2\delta-1)}{\delta}\bar{y} + \delta^{-1}\bar{y}_{ld} - \frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j} \right], \quad (2.8)$$

(iii) For $\alpha = -\delta^{-1}$, $\hat{\bar{Y}}_\alpha$ boils down to

$$\hat{\bar{Y}}_2 = \left[\delta^{-1}(\bar{y} - \bar{y}_{ld}) + \frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j} \right] \quad (2.9)$$

which is almost unbiased regression-type estimator of $\bar{Y}$.

Remark 2.2. For $\left(\hat{\beta} = \frac{\bar{y}}{\bar{x}},\ \hat{\beta}^*_j = \frac{\bar{y}^*_j}{\bar{x}^*_j} \right)$, $\hat{\bar{Y}}_\alpha$ reduces to a class of almost unbiased ratio-type estimators of $\bar{Y}$ designated by

$$\hat{\bar{Y}}_{\alpha r} = \left[\{1 - \alpha(1-\delta)\}\bar{y} + \alpha\,\bar{y}\,(\bar{x}'/\bar{x}) - \alpha\,\delta\,\frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_j\,(\bar{x}'/\bar{x}^*_j) \right] \quad (2.10)$$

while for $\left(\hat{\beta} = -\frac{\bar{y}}{\bar{x}'},\ \hat{\beta}^*_j = -\frac{\bar{y}^*_j}{\bar{x}'} \right)$, $\hat{\bar{Y}}_\alpha$ gives

$$\hat{\bar{Y}}_{\alpha p} = \left[\{1 - \alpha(1-\delta)\}\bar{y} + \alpha\,\bar{y}\,(\bar{x}/\bar{x}') - \alpha\,\delta\,\frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_j\,(\bar{x}^*_j/\bar{x}') \right] \quad (2.11)$$

which is a class of almost unbiased product-type estimator of $\bar{Y}$.

It is to be noted that for different choices of α many almost unbiased ratio and product-type estimators from $\hat{\bar{Y}}_{\alpha r}$ and $\hat{\bar{Y}}_{\alpha p}$ defined in (2.10) and (2.11) respectively, can be generated. Further, it is to be remarked that many regression-type (almost unbiased) estimators can be had from $\hat{Y}_\alpha$ in (2.6) just by substituting different values of α.

3. *Optimum Estimator in the Class $\hat{Y}_\alpha$ in (2.6)*

To obtain the variance of $\hat{Y}_\alpha$, use the following results :

$$\left.\begin{aligned} V(\bar{y}) &= \frac{(N-n)}{Nn} S_y^2 \\ V(\bar{y}_{ld}) &= V\left(\frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right) = \left[\left(\frac{1}{n}-\frac{1}{n'}\right) S_y^2 (1-\rho^2) + \left(\frac{1}{n'}-\frac{1}{N}\right) S_y^2\right] \\ Cov(\bar{y}, \bar{y}_{ld}) &= Cov\left(\bar{y}, \frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right) = Cov\left(\bar{y}_{ld}, \frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right) = V(\bar{y}_{ld}) \end{aligned}\right\} \tag{3.1}$$

From (2.6), we have

$$\begin{aligned} V(\hat{Y}_\alpha) = \Bigg[& \{1-\alpha(1-\delta)\}^2 V(\bar{y}) + \alpha^2 V(\bar{y}_{ld}) + \alpha^2\delta^2 V\left(\frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right) \\ & + 2\alpha\{1-\alpha(1-\delta)\} Cov(\bar{y}, \bar{y}_{ld}) \\ & - 2\alpha\delta\{1-\alpha(1-\delta)\} Cov\left(\bar{y}, \frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right) \\ & - 2\alpha^2\delta\, Cov\left(\bar{y}_{ld}, \frac{1}{k}\sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right)\Bigg] \end{aligned} \tag{3.2}$$

Using (3.1) in (3.2), we get

$$V(\hat{Y}_\alpha) = \left[\left(\frac{1}{n}-\frac{1}{N}\right) S_y^2 + \alpha(1-\delta)\{\alpha(1-\delta)-2\}\left(\frac{1}{n}-\frac{1}{n'}\right) S_y^2 \rho^2\right] \tag{3.3}$$

which is minimized for

$$\alpha = (1-\delta)^{-1} \tag{3.4}$$

Substituting $\alpha = (1-\delta)^{-1}$ in (2.6), we get the optimum estimator in the class (2.6) as

$$\hat{Y}_0 = \left[\frac{(n'-n+m)}{n'} \cdot k \cdot \bar{y}_{ld} - \frac{(n'-n)}{n'}\frac{(k-1)}{k} \sum_{j=1}^{k} \bar{y}^*_{(ld)j}\right] \tag{3.5}$$

Substituting $\alpha = (1-\delta)^{-1}$ in (3.3), we get the variance of $\hat{Y}_0$ as

$$V(\hat{Y}_0) = \left[\left(\frac{1}{n}-\frac{1}{n'}\right)S_y^2(1-\rho^2)+\left(\frac{1}{n}-\frac{1}{N}\right)S_y^2\right] \tag{3.6}$$

ACKNOWLEDGEMENT

Authors are thankful to the referee for his valuable suggestions regarding the improvement of the paper.

REFERENCES

[1] Quenouille, M.H., 1956. Notes on bias in estimation. *Biometrika*, 43, 353-360.

[2] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C., 1984. Sampling theory of surveys with applications. Third Revised edition, Iowa State University Press, Ames, Iowa (U.S.A.) and Indian Society of Agricultural Statistics, New Delhi, India.