ON SOME QUADRATIC AND HIGHER DEGREE RATIO AND PRODUCT ESTIMATORS

By
P. C. Gupta*
University of Rajasthan, Jaipur
(Received : September, 1975)

1. Introduction

It is a well known result that the ratio estimates are the best among a wide class of estimates when the relation between y_i and x_i, the character under study and the auxiliary character respectively, is a straight line through origin and the variance of y_i about this line is proportional to x_i (Cochran [1]). In many practical situations the regression line does not pass through origin. Considering this fact Srivastava [2] suggested a ratio type estimator which will be more efficient than the conventional ratio estimate in some situations. His estimate, however, requires a prior knowledge of $\rho c_y/c_x$ and it is equivalent to linear regression estimate if $\rho c_y/c_x$ is known in advance.

In this paper an attempt has been made to study the nature of ratio and product estimates if they are represented by a polynomial in $(\bar{X}_N/\bar{x}_n)$ and $(\bar{x}_n/\bar{X}_N)$ respectively. It has been studied that the proposed estimates though require a prior knowledge of $\rho c_y/c_x$ are superior to conventional ratio and product estimator respectively, as long as the difference between the optimum weight and the chosen weight do not exceed $|1-\rho c_y/c_x|$. In optimum case they are equivalent to linear regression estimate.

2. Sampling Procedure and the Estimate

We select a sample of size 'n' out of N by simple random sampling without replacement and study variates (y_i, x_i), $i=1, 2,..., N$; y_i and x_i are the values of the character under study. Y and the auxiliary variables X, which is highly correlated to Y, for ith unit, respectively. Further the total information on X is also available.

Let

$$\bar{y}_n=\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{x}_n=\frac{1}{n}\sum_{i=1}^{n} x_i$$

*Now at South Gujarat University, Surat.

be the sample means for Y and X respectively. We propose the following estimate for the population mean of Y

$$\hat{\bar{Y}}_R^* = a\bar{y}_n(\bar{X}_N/\bar{x}_n) + (1-a)\bar{y}_n\,(\bar{X}_N/\bar{x}_n)^2 \quad ...(2)$$

where 'a' is chosen so that the variance of $\hat{\bar{Y}}_R^*$ is minimum.

It is evident that the conventional ratio estimate

$$\hat{\bar{Y}}_R = (\bar{y}_n/\bar{x}_n)\,\bar{X}_N$$

is a particular case of $\hat{\bar{Y}}_R^*$ if $a=1$. ...(3)

3. The Bias and the Variance of the $\hat{\bar{Y}}_R^*$

The bias and the variance of the proposed estimate $\hat{\bar{Y}}_R^*$ will be obtained as discussed by Sukhatme [3]. The expression for the bias and the variance are approximately given as follows :

$$\text{Bias}\left(\hat{\bar{Y}}_R^*\right) = \bar{Y}_N\left(\frac{1}{n}-\frac{1}{N}\right)\left[\left(c_x^2 - c_{xy}\right)+(1-a)\left(2c_x^2 - 2c_{xy}\right)\right] \quad ...(4)$$

$$\text{Var}\left(\hat{\bar{Y}}_R^*\right) = \bar{Y}_n^2\left(\frac{1}{n}-\frac{1}{N}\right)\left[a^2\left(c_y^2 - 2c_{xy}+c_x^2\right)+(1-a)^2\right.$$
$$\left.\left(c_y^2 - 4c_{xy}+4c_x^2\right)+2a(1-a)\left(c_y^2 - 3c_{xy}+2c_x^2\right)\right] \quad ...(5)$$

where c_x, c_y, c_{xy} represent the coefficients of variations. The optimum value of a, which is obtained by minimising the $V\left(\hat{y}_R^*\right)$, is given by

$$a^* = 2 - \rho\,\frac{c_y}{c_x} \quad ...(6)$$

The bias and the variance of the proposed estimate under (6) are given by

$$\text{Bias} = -\bar{Y}_N\left(\frac{1}{n}-\frac{1}{N}\right)(\rho c_y - c_x)^2 \quad ...(7)$$

$$\text{opt. } V\left(\hat{y}_R^*\right) = \bar{Y}_N^2\left(\frac{1}{n}-\frac{1}{N}\right)c_y^2\,(1-\rho^2) \quad ...(8)$$

where ρ is the population correlation coefficient between character X and Y.

4. Comparison of $\hat{\bar{Y}}^*_R$ with Conventional Ratio and Linear Regression Estimate

The variance of the usual ratio estimate can be obtained by putting $a=1$ in (5). Thus

$$V\left(\hat{\bar{Y}}_R\right)=\bar{Y}^2_N\left(\frac{1}{n}-\frac{1}{N}\right)\left(c^2-2c_{xy}+c^2_x\right) \quad \text{... (9)}$$

It is evident that $\hat{\bar{Y}}^*_R$ with optimum weight a^* given in (6) is always superior to conventional ratio estimate as the variance of the former is always less than the latter and is equal to the variance of the linear regression estimator. We shall, however, examine the case when we do not use optimum weight a^* as it requires a prior knowledge of $\rho c_y/c_x$ which is usually not known in advance. If a good guess of this quantity can be obtained from past experience, the proposed estimate can be used more efficiently. Consider

$$V\left(\hat{\bar{Y}}^*_R\right)-V\left(\hat{\bar{Y}}_R\right)=(1-a)\ \bar{Y}^2_N\left(\frac{1}{n}-\frac{1}{N}\right)\left[\left(3c^2_x-2c_{xy}\right)-ac^2_x\right] \quad \text{...(10)}$$

and it will be always negative if 'a' lies between 1 and $\dfrac{3c^2_x-2c_{xy}}{c^2_x}$,

which suggests that the proposed estimate will always have a smaller variance than conventional ratio estimate as long as

$$|\,a-a^*\,| < \left|\,1-\rho\frac{c_y}{c_x}\right| \quad \text{...(11)}$$

This difference between 'a' and its optimum value provide quite a good range for its choice as $0 < a \leqslant 1$.

5. Higher Degree Polynomials in $(\bar{X}_N/\bar{x}_n)$

In this section we shall study the efficiency of higher degree polynomial in $(\bar{X}_N/\bar{x}_n)$. Let the estimate for the population mean be given by

$$\hat{\bar{Y}}^{**}_R=\bar{y}_n\ [a_1\ (\bar{X}_N/\bar{x}_n)+a_2\ (\bar{X}_N/\bar{x}_n)^2+a_3\ (\bar{X}_N/\bar{x}_n)^3] \quad \text{...(12)}$$

where $a_1+a_2+a_3=1$ and the constant a's are so chosen that $V\left(\hat{\bar{Y}}^{**}_R\right)$ is minimum. ...(13)

The bias and the variance of $\hat{\bar{Y}}^{**}_R$ to first degree approximation

$$\text{Bias}\left(\hat{\bar{Y}}^{**}_R\right)=\bar{Y}_N\left(\frac{1}{n}-\frac{1}{N}\right)\left[\left(c_x^2-c_{xy}\right)-(a_2+a_3)\,c_{xy}+c_x^2\,(2a_2+5a_3)\right] \quad \ldots(14)$$

$$V\left(\hat{\bar{Y}}^{**}_R\right)=\bar{Y}^2\left(\frac{1}{n}-\frac{1}{N}\right)\left[\left(c_y^2-2c_{xy}+c_x^2\right)+a_2^2\left(3c_x^2-2c_{xy}\right)+a_3^2\left(8c_x^2-4c_{xy}\right)+2a_1a_2\left(c_x^2-c_{xy}\right)+2a_1a_3\left(2c_x^2-2c_{xy}\right)+2a_2a_3\left(5c_x^2-3c_{xy}\right)\right] \quad \ldots(15)$$

The optimum values of a_i $(i=1, 2, 3)$ are obtained by minimising (15) subjected to (13) as a solution of the following system of linear equations :

$$\left.\begin{array}{l} a_1+a_2+a_3=1 \\ \left(c_x^2-c_{xy}\right)a_2+2\left(c_x^2-c_{xy}\right)a_3=\lambda/2 \\ \left(c_x^2-c_{xy}\right)a_1+\left(3c_x^2-2c_{xy}\right)a_2+\left(5c_x^2-3c_{xy}\right)a_3=\lambda/2 \\ 2\left(c_x^2-c_{xy}\right)a_1+\left(5c_x^2-3c_{xy}\right)a_2+4\left(2c_x^2-c_{xy}\right)a_3=\lambda/2 \end{array}\right\} \ldots(16)$$

where λ is the Langrange's multipler.

It can be seen that no unique solution exist as these equations do not have full rank and as such infinite number of optimum set of solutions for $\{a_1\}$ exist. Let us consider one of them by putting $a_2=a_3$, we have

$$a_1^*=\left(5c_x^2-2c_{xy}\right)\Big/3c_x^2$$

$$a_2^*=a_3^*=\left(c_{xy}-c_x^2\right)\Big/3c_x^2 \quad \ldots(17)$$

Substituting these opt $\{a_i\}$ in (14) and (15) we get,

$$\text{Bias}\left(\hat{\bar{Y}}^{**}_R\right)=\frac{7\left(c_{xy}-c_x^2\right)c_x^2}{c_{xy}}\,\bar{Y}_N\left(\frac{1}{n}-\frac{1}{N}\right) \quad \ldots(18)$$

and

$$\text{opt}\left(\hat{\bar{Y}}^{**}_R\right)=\bar{Y}_N\left(\frac{1}{n}-\frac{1}{N}\right)c_y\,(1-\rho^2) \quad \ldots(19)$$

which is same as the optimum variance of $\hat{\bar{Y}}^*_R$ or that of the linear regression estimator. This result suggests that to first degree approximation we cannot improve the ratio estimates beyond the linear regression estimate even if we take third degree polynomial in $(\bar{X}_N/\bar{x}_n)$. Further it may be stated that for the known value of β_{yx}, *viz.*, population regression coefficient of Y and X, the linear regression estimate is unbiased.

6. Double Sampling Estimates

When the detailed information on the auxiliary character is not known, it is often obtained on a large but relatively cheaper sample and this procedure is the usual double sampling. We shall consider both the cases (*i*) when the second sample is a sub-sample from the preliminary large sample, (*ii*) when the second sample is not the sub-sample from the first.

Case (i) When Second Phase Sample is a Sub-Sample From First Phase Sample

We propose the following double sampling estimate

$$\hat{\bar{Y}}^*_{RD} = a \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_{n'} + (1-a)\, \bar{y}_n \left(\frac{\bar{x}_{n'}}{\bar{x}_n}\right)^2 \qquad \ldots(20)$$

where a is chosen so that variance of $\hat{\bar{Y}}^*_{RD}$ is minimum. Neglecting the terms of order $1/N$, the bias and the variance of (20) to first degree approximation is given by

$$\text{Bias}\left(\hat{\bar{Y}}^*_{RD}\right) = \bar{Y}_N \left[(1-a)\left(\frac{1}{n} - \frac{1}{n'}\right)\left(3c_x^2 - 2c_{xy}\right) + a\left(\frac{1}{n} - \frac{1}{n'}\right)\left(c_x^2 - c_{xy}\right)\right] \qquad \ldots(21)$$

$$V\left(\hat{\bar{Y}}^*_{RD}\right) = \bar{Y}_N^2\left(\frac{1}{n} - \frac{1}{n'}\right)\left(c_y^2 - 2c_{xy} + c_x^2\right) + \frac{\bar{Y}_N^2\, c_y^2}{n'} + \bar{Y}_N^2\left(\frac{1}{n} - \frac{1}{n'}\right)\left[(1-a)^2\left(3c_x^2 - 2c_{xy}\right) + 2a(1-a)\left(c_y^2 - c_{xy}\right)\right] \qquad \ldots(22)$$

The optimum value of 'a' which is obtained by minimising (22) comes out to be same as given in (6). The optimum variance of (20) can be obtained as

$$\text{opt } V\left(\hat{\bar{Y}}^*_{RD}\right)=\bar{Y}^2_N\left(\frac{1}{n}-\frac{1}{n'}\right)\left(c^2_y-2c_{xy}+c^2_x\right)$$

$$+\frac{\bar{Y}^2_N \cdot c^2_y}{n'}+\bar{Y}^2_N\left(\frac{1}{n}-\frac{1}{n'}\right)(\rho c_y-c_x)^2 \qquad \ldots(23)$$

which is same as the variance of conventional double sampling linear regression estimate. Hence with the prior knowledge of $\rho c_y/c_x$ the proposed estimate is always superior to double sampling ratio estimate.

(ii) When Second Phase Sample is Not a Sub-Sample From the First Phase Sample

When n is drawn independently of n' the bias and the variance of the proposed estimate in (20) is given as follows :

$$\text{Bias}\left(\hat{\bar{Y}}^*_{RD}\right)_I=\bar{y}_N\, a\left(\frac{c^2_x-c_{xy}}{n}\right)+(1-a)\,\bar{Y}_N\left(\frac{c^2_x}{n'}+\frac{3c^2_x-2c_{xy}}{n}\right) \qquad \ldots(24)$$

$$V\left(\hat{\bar{Y}}^*_{RD}\right)_I=\bar{Y}^2_N\left(\frac{c^2_y-2c_{yx}+c^2_x}{n}+\frac{c^2_x}{n'}\right)+(1-a)^2\,\bar{Y}^2_N$$

$$\left(\frac{3c^2_x-2c_{xy}}{n}+\frac{3c^2_x}{n'}\right)+2a\,(1-a)\bar{Y}^2_N$$

$$\left(\frac{c^2_x-c_{xy}}{n}+\frac{c^2_x}{n'}\right) \qquad \ldots(25)$$

The optimum 'a' in this case is obtained by minimising (25). Thus

$$a^*=\left(2-\frac{n'}{n+n'}\,\rho\,\frac{c_y}{c_x}\right) \qquad \ldots(26)$$

and the optimum variance of (20) is obtained by substituting value of opt. a from (26) in (25)

$$\text{opt. } V\left(\hat{\bar{Y}}^*_{RD}\right)=\frac{\bar{y}^2_N\, c^2_y}{n}\left(1-\frac{n'}{n+n'}\,\rho^2\right) \qquad \ldots(27)$$

7. Comparison of $\hat{\bar{Y}}^*_{RD}$ with Various Estimates

We know that variance of double sampling linear regression estimate with independent samples.

$$V\left(\hat{\bar{Y}}_{d/r}\right)_{\mathrm{I}}=\frac{\bar{Y}_N^2\, C_y^2}{n}\left(1-\frac{n'-n}{n'}\,\rho^2\right) \qquad \ldots(28)$$

$$\text{Consider } V\left(\hat{\bar{Y}}_{d/r}\right)_{\mathrm{I}}-V\left(\hat{\bar{Y}}^*_{RD}\right)_{\mathrm{I}}=\frac{S_y^2\,\rho^2}{n}\left(\frac{n'}{n+n'}-\frac{n'-n}{n'}\right)$$

$$=\frac{n}{n+n'}\,\frac{\bar{Y}_N^2\, c_y^2}{n'}\,\rho^2 \qquad \ldots(29)$$

which is always positive and hence the proposed estimate is superior even to double sampling linear regression estimate which is always superior to double sampling ratio estimate.

Comparison of $\bar{Y}^*_{RD}$ with Double Sampling Ratio Estimate

The variance of the double sampling ratio estimate when the second sample is not a sub-sample from the first can be obtained by putting $a=1$ in (25)

$$V\left(\hat{\bar{Y}}_{RD}\right)_{\mathrm{I}}=\bar{Y}_N^2\left(\frac{c_y^2-2c_{xy}+c_x^2}{n}+\frac{c_x^2}{n'}\right) \qquad \ldots(30)$$

Consider

$$V\left(\hat{\bar{Y}}^*_{RD}\right)_{\mathrm{I}}-V\left(\hat{\bar{Y}}_{RD}\right)_{\mathrm{I}}=(1-a)\left\{(1-a)\left(\frac{3c_x^2-2c_{xy}}{n}+\frac{3c_x^2}{n'}\right)+2a\left(\frac{c_x^2-c_{xy}}{n}+\frac{c_x^2}{n'}\right)\right\} \qquad \ldots(31)$$

which will be always negative if

$$a<\frac{2n'}{(n+n')}\,\rho\,\frac{c_y}{c_x}-3 \qquad \ldots(32)$$

Thus the proposed estimate will be superior to conventional double sampling estimate if

$$\left|a-a^*\right|<\left|\left(1-\frac{n'}{n+n'}\,\frac{c_y}{c_x}\right)\right| \qquad \ldots(33)$$

Comparison with Mean Per Unit for a Given Cost

We shall consider the following cost function

$$c = nc_1 + (n+n')\, c_2 \qquad \ldots(34)$$

where c is the total cost, c_1 and c_2 are cost of evaluating a single unit for character Y and X respectively. We have

$$\text{opt. var}\left(\hat{\bar{Y}}^*_{RD}\right) = \frac{v_1}{n} + \frac{v_2}{n+n'} \qquad \ldots(35)$$

where

$$v_1 = \bar{Y}_N^2 \; c_y^2 \; (1-\rho^2) \text{ and } v_2 = \bar{Y}_N^2 \; c_y^2 \; \rho^2.$$

The minimum of (35) subjected to (34) is given by

$$\bar{Y}_N^2 \; c_y^2 \left(\frac{\sqrt{(1-\rho^2)\, c_1} + \sqrt{c_2\, \rho^2}}{c} \right)^2 \qquad \ldots(36)$$

If all the resources are diverted towards study of character Y only we have

$c = n^* . c_1$, where n^* is opt. sample size and

$$\text{opt. } V\left(\bar{Y}_n\right)_{single} = \frac{c_1 \; \bar{Y}_N^2 \; c_y^2}{c} \qquad \ldots(37)$$

Hence the proposed double sampling estimate is more efficient than mean per unit if (37)−(36) is positive *i.e.*,

$$\rho^2 > \frac{4c_1\, c_2}{(c_1+c_2)^2} \qquad \ldots(38)$$

8. Product Estimates

The theory developed in all the previous sections can also be developed for corresponding quadratic and higher degree product estimates. Here we shall consider the following estimate, its application in double sampling will exactly be on similar lines as discussed in sections 6 and 7.

$$\bar{Y}_P^* = a\bar{y}_n \left(\frac{\bar{x}_n}{\bar{X}_N} \right) + (1-a)\, \bar{y}_n \left(\frac{\bar{x}_n}{\bar{X}_N} \right)^2 \qquad \ldots(39)$$

where the weight 'a' is chosen so that $V\left(\bar{Y}_P^* \right)$ is minimum.

The variance and the Bias of the estimate to first degree approximation are given as follows;

$$\text{Bias}\left(\bar{Y}_P^* \right) = \bar{Y}_N \left(\frac{1}{n} - \frac{1}{N} \right) c_{xy} + (1-a) \left(c_x^2 + c_{xy} \right) \ldots(40)$$

$$\text{Var}\left(\bar{Y}^*_P\right)=\bar{Y}^2_N\left(\frac{1}{n}-\frac{1}{N}\right)a^2\left(c_y^2+2c_{yx}+c_x^2\right)$$

$$+(1-a)^2.\left(c_y^2+4c_x^2+4c_{xy}\right)+2a(1-a)\left(c_y^2+3c_{xy}+2c_x^2\right) \quad \ldots(41)$$

The opt. value of 'a' is obtained by minimising (41) and is given by

$$a^*=2+\rho\frac{c_y}{c_x} \quad \ldots(42)$$

The optimum variance of the estimate is obtained by substituting the opt. value of 'a' from (42) in (41) and it comes out to be same as (8) which is same as the variance of linear regression estimator.

It is obvious that in the optimum case the proposed estimate is superior to conventional product estimate and it will continue to be superior if

$$V\left(\hat{\bar{Y}}_P\right)-V\left(\hat{\bar{Y}}^*_P\right)>0$$

or

$$\bar{Y}^2_N\left(\frac{1}{n}-\frac{1}{N}\right)(1-a)\left[\left(3c_x^2+2c_{xy}\right)(1-a)+2a\left(c_x^2+c_{xy}\right)\right]<0$$

i.e.

$$a<\frac{3c_x^2+2c_{xy}}{c_x^2}$$

Hence the suggested estimate will be superior to conventional product estimate as long as

$$\mid a-a^*\mid<\mid 1-\rho\frac{c_y}{c_x}\mid$$

which is same as obtained in section 4.

It can now easily be seen that we shall not be able to improve, to first degree approximation even if we consider higher degree product estimates with the same reasoning as discussed in section 5.

SUMMARY

Some ratio and product type estimators involving higher powers of $(\bar{X}_N/\bar{x}_n)$ and $(\bar{x}_n/\bar{X}_N)$ respectively are suggested. In the optimum case they are as efficient as the linear regression estimator. Further they will be superior to corresponding ratio and product estimates if the difference between weights taken and optimum weights

does not exceed $\left| 1-\rho\frac{C_y}{C_x} \right|$. It has been further studied that in double sampling case the proposed estimates are even superior to double sampling linear regression estimate in their optimum cases provided second sample is drawn independently of the first. They have been also compared with mean per unit taking suitable linear cost function.

References

[1] Cochran W.G. (1963) : 'Sampling techniques' 2nd Ed., John Wiley and Sons, Inc., New York.

[2] Srivastava S.K. (1970) : 'Two-phase sampling estimators in sample surveys'. *Aust. J. Statist.* **12**, 23-27.

[3] Sukhatme, P.V. and Sukhatme B.V. (1970) : Sampling theory of surveys with application. Asia Publishing House, New Delhi.