Jour. Ind. Soc. Ag. Statistics
Vol. XXXVII, No. 2 (1985), pp. 158-162

ON TWO-PHASE MULTIVARIATE SAMPLING ESTIMATOR

PRADEEP KUMAR BEDI
Department of Statistics, University of Rajasthan, Jaipur-302004

(Received: October, 1982)

SUMMARY

In two-phase sampling, when the two samples are drawn independently, the suggested multivariate regression estimator and generalised two-phase estimator have been shown to have smaller mean square error than the corresponding usual multivariate regression estimator and Srivastava's [5] estimator. When the coefficients of the proposed estimators, are estimated, Expected mean square error under a suitable model are also derived.

Keywords: Multivariate Regression estimator; Generalised two-phase estimator; Minimum variance; Minimum asymptotic mean square error.

1. Introduction

Sometimes, information on multi-auxiliary variables $x_1, \ldots, x_p$ each of size N are available though their population mean vector $\overline{X}$ is unknown. To utilise this information, the application of two-phase sampling is well-known in the literature. Srivastava [5] assumed that all the p-auxiliary variables are measured on each individual in the first-phase sample of size n_1 ($\mathbf{x}'$ denotes mean vector) and then a second-phase sample of size n is drawn independently of the first-phase sample on each member of which the character under study y and the auxiliary variables are measured ($\bar{y}$ and $\bar{\mathbf{x}}$ denotes respective mean and mean vector). His proposed generalised two-phase estimator for $\overline{Y}(\overline{Y_N})$ is superior than the corresponding usual regression estimator. Rao [3] dealing with one auxiliary variable has suggested two estimates i.e. $\bar{x}_w$ (the best linear combination of the two independent samples) and $\bar{x}_v$ (mean based on v distinct units in two independent samples) for $\overline{X}$. Srivastava's [5] estimator with one auxiliary variable is as precise as the regression estimator when $\overline{X}$ is estimated by $\bar{x}_w$. Further, Rao [3] has shown that the efficiency of the regression estimator of $\bar{Y}$ will increase when $\bar{X}$ is estimated by $\bar{x}_v$ instead of $\bar{x}_w$ or $\bar{x}'$.

In this paper, an attempt has been made to extend the two-phase regression estimator for $\bar{Y}$ due to Rao [5], when the information on more than one-auxiliary variable is available. In Section 2, a more generalised two-phase estimator than that of Srivastava's [5] has also been considered, when two-samples are drawn independently. When the coefficients in the proposed multivariate regression estimators are estimated, expected mean square error of the estimator under a suitable model is given in Section 3. In this study, samples have been drawn according to simple random sampling without replacement. Henceforth x will denotes the vector of auxiliary variables.

2. Multivariate Regression Estimator and Generalised Estimator

It can easily be seen that the multivariate regression estimator for $\bar{Y}$ when the $\bar{X}$ is estimated by $\bar{x}_w$ is as precise as that of Srivastava's [1981] estimator. So, here we have considered the multivariate regression estimator for $\bar{Y}$ when $\bar{X}$ is estimated by $\bar{x}_v$ i.e.

$$\hat{\bar{y}}_v = \bar{y} + B' (\bar{x}_v - \bar{x}) \tag{2.1}$$

where B' is a column vector of p constants to be determined so that the variance of the estimator $\hat{\bar{y}}_v$ is minimal. Clearly, for fixed B, $\hat{\bar{y}}_v$ provides an unbiased estimator of $\bar{Y}$. Under the usual notations

$$\text{Cov}\,(\bar{x}_{vi}, \bar{x}_i) = V(\bar{x}_{vi}) = \left[E\left(\frac{1}{v}\right) - \frac{1}{N} \right] S^2_{xi},$$

$$\text{Cov}\,(\bar{y}, \bar{x}_{vi}) = \left[E\left(\frac{1}{v}\right) - \frac{1}{N} \right] S_{yxi}$$

$$\text{Cov}\,(\bar{x}_i\, \bar{x}_{vj}) = \text{Cov}\,(\bar{x}_{vi}, \bar{x}_j) = \left[E\left(\frac{1}{v}\right) - \frac{1}{N} \right] S_{x_i x_j}\quad i \neq j = 1, 2, \ldots, p$$

and

$$E\left(\frac{1}{v}\right) = \sum_{k=0}^{n} D_k \qquad \text{where } D_0 = 1/(n_1 + n)$$

and

$$\frac{D_{k+1}}{D_k} = \frac{(n-k)(n_1-k)}{(n+n_1-k-1)(N-k)}.$$

Though, in a particular sample, the number of distinct units vary with auxiliary variable, the expectation of the reciprocals of these will be the same because the sample sizes at first-phase and second-phase are same for all auxiliary variables. Therefore, the variance of $\hat{\bar{y}}_v$ is

$$V(\hat{\bar{y}}_v) = \left[\frac{1}{n} - \frac{1}{N}\right] S_y^2 + \left[\frac{1}{n} - E\left(\frac{1}{\nu}\right)\right] S_y^2 \,[B'AB - B'd] \qquad (2.2)$$

where $A = [a_{ij}]$ be $p \times p$ positive definite matrix with $a_{ij} = S_{xi}\, S_{xj}/S_y^2$ and $d' = (d_1, \ldots, d_p)$ with $d_1 = S_{yx_i}/S_y^2$. The variance in (2.2) is minimized for

$$B = A^{-1}\, d \qquad (2.3)$$

and the minimum variance is given by

$$V_0\,(\hat{\bar{y}}_v) = \left[\frac{1}{n} - \frac{1}{N}\right] S_y - \left[\frac{1}{n} - E\left(\frac{1}{\nu}\right)\right] R^2\, S_y^2 \qquad (2.4)$$

where R^2 is the multiple correlation coefficient between y and $x_1, \ldots, x_p$. Clearly, $V_0\,(\hat{\bar{y}}_v)$ is smaller than the variance of the Srivastava [5] estimator.

If $\bar{X}$ is estimated by $\bar{x}_v$, than Srivastava's [5] type estimator for $\bar{Y}$ is

$$t_v = \bar{y}\; h\,(\boldsymbol{u}) \qquad (2.5)$$

where $h\,(\boldsymbol{u})$ is any function of $\boldsymbol{u}$ which is a column vector with elements $u_i = \bar{x}_{vi} \sqrt{\bar{x}_i}$ $(i = 1, \ldots, p)$. Clearly, Rao's [4] ratio estimator is a particular case of (2.5). Following the approach similar to that adopted by Srivastava [5] it can be easily seen that minimum asymptotic mean square error of the estimator t_v is equal to (2.4).

The class of estimators (2.5) does not include the regression type estimator such as (2.1). However, even if we consider a wider class of estimators, i.e.

$$t_g = g\,(\bar{y}, \boldsymbol{u}) \qquad (2.6)$$

of $\bar{Y}$, which includes the estimator (2.1) and where g is a function of $\bar{y}$ and $\boldsymbol{u}$, such that

$$g\,(\bar{Y}, 1) = \bar{Y} \qquad (2.7)$$

The minimum asymptotic mean square error of the t_g is equal to (2.4) and is not reduced.

3. Multivariate Regression Estimator when Coefficients are Estimated

Khan and Tripathi [2] have given expected mean square error of the multivariate regression estimator when B is estimated by b, the least square estimates obtained from the second-phase sample and $\overline{X}$ is estimated by $\bar{x}'$. When B is estimated by b, the estimator $\hat{\bar{y}}_v$ becomes

$$\hat{\bar{y}}_v = \bar{y} + b' (\bar{x}_v - \bar{x}) \tag{3.1}$$

The mean square error of $\hat{\bar{y}}_v$ for the finite population is

$$M (\hat{\bar{y}}_v) = \frac{1}{u'c} \Sigma \Sigma (\hat{\bar{y}}_v - \overline{Y})^2 \tag{3.2}$$

where

$$u' = \binom{N}{n_1} \text{ and } c = \binom{N}{n}$$

The model that has been considered is

$$y_j = \alpha + B x_j + e_j \quad (j = 1, \ldots, N) \tag{3.3}$$

with $E (e_j \mid x_j) = 0$, $E (e_j e_j \mid x_j x_j) = 0$ and $V (e_j \mid x_j) = S_y^2 (1 - R^2)$. Further, it is assumed that x_j are drawn from a multivariate population with mean vector μ and variance covariance matrix Σ. From (3.3)

$$\hat{\bar{y}}_v - \overline{Y} = (b - B)' (\bar{x}_v - \bar{x}) + B' (\bar{x}_v - \overline{X}) + \bar{e}_n - \bar{e}_N \tag{3.4}$$

By averaging over the distribution of e's, it follows from (3.4) that under (3.3) $\hat{\bar{y}}_v$ is unbiased for fixed $\bar{x}$'s. Further, from (3.4)

$$(\hat{\bar{y}}_v - \overline{Y})^2 = (\bar{e}_n - \bar{e}_N)^2 + B' (\bar{x}_v - \overline{X}) (\bar{x}_v - \overline{X})' B + (\bar{x}_v - \bar{x})' [a^{-1}. \{\Sigma e_j^2 (x_j - \bar{x}) (x_j - \bar{x})'\} a^{-1}] (\bar{x}_v - \bar{x}) + \text{terms whose expectation is zero} \tag{3.5}$$

where a is the sample variance covariance matrix of auxiliary variables. Now

$$EE (\bar{e}_n - \bar{e}_N)^2 = \left[\frac{1}{n} - \frac{1}{N}\right] S_y^2 (1 - R^2) \tag{3.6}$$

$$\text{and } EE [B' (\bar{x}_v - \overline{X}) (\bar{x}_v - \overline{X})' B] = \left[E\left(\frac{1}{v}\right) - \frac{1}{N}\right] R^2 S_y^2 \tag{3.7}$$

The expectation over the finite population and then over the superpopulation of the last terms in (3.5) has been

$$\left[\frac{1}{n} - E\left(\frac{1}{v}\right)\right] \frac{S_y^2 (1 - R) \cdot p}{n - p - 2} \tag{3.8}$$

following the approach similar to that adopted by Rao [3]. We get

$$EM (\hat{\bar{y}}_v) = \left[E\left(\frac{1}{v}\right) - \frac{1}{N}\right] S_y^2 + \left[\frac{1}{n} - E\left(\frac{1}{v}\right)\right] \frac{S_y^2 (1 - R^2) (n - 2)}{n - p - 2} \tag{3.9}$$

Similarly, the expected mean square error of the multivariate regression estimator when $\overline{X}$ is estimated by $\bar{x}_w$ can be obtained by replacing $E(1/\nu)$ in (3.9) by $1/n_1$. Clearly, $EM(\hat{\bar{y}}_v)$ is smaller than that of expected mean square error of the corresponding regression estimator when $\overline{X}$ is estimated by $\bar{x}_w$ or $\bar{x}'$.

The unbiased estimate of the expression in (3.9) is

$$m(\hat{\bar{y}}_v) = \left[\frac{1}{\nu} - \frac{1}{N}\right] s_y^2 + \left(\frac{1}{n} - \frac{1}{\nu}\right) \frac{n-2}{n-p-2} s_e^2$$

where

$$s_y^2 = \Sigma (y_j - \bar{y})^2/(n-1),$$

and

$$s_e^2 = \Sigma [(y_j - \bar{y}) - b'(x_j - \bar{x})]^2/(n-p-1)$$

REFERENCES

[1] Cochran, W. G. (1977): *Sampling Techniques* (3rd ed). John Wiley and Sons, New York.

[2] Khan, S. and Tripathi, T. P. (1967): The use of Multivariate auxiliary information in double sampling, *JISA* **5**; 42-48.

[3] Rao, P. S. R. S. (1972): On two-phase regression estimator, *Sankhya A*, **34**: 473-76.

[4] Rao, P. S. R. S. (1975): On two-phase ratio estimator in finite population, *JASA* **70**; 839-45.

[5] Srivastava, S. K. (1981): A generalized two-phase sampling estimator, *JISAS* **33**: 37-46.

[6] Srivastava, S. K. and Jhajj, H. S. (1980): A class of estimators using auxiliary information for estimating finite population variance, *Sankhya C*, **42**: 87-96.