

A Regression Type Estimator with Two Auxiliary Variables for Two-Phase Sampling

Naqvi Hamad¹, Muhammad Hanif², Najeeb Haider¹

¹Department of Statistics, Government Postgraduate College, Dera Ghazi Khan, Pakistan ²National College of Business Administration and Economics, Lahore, Pakistan Email: naqvihamad@hotmail.com, drmianhanif@gmail.com, haiderdr@yahoo.co.uk

Received September 9, 2012; revised October 12, 2012; accepted October 25, 2012

Copyright © 2013 Naqvi Hamad *et al.* This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

This paper is an extension of Hanif, Hamad and Shahbaz estimator [1] for two-phase sampling. The aim of this paper is to develop a regression type estimator with two auxiliary variables for two-phase sampling when we don't have any type of information about auxiliary variables at population level. To avoid multi-collinearity, it is assumed that both auxiliary variables have minimum correlation. Mean square error and bias of proposed estimator in two-phase sampling is derived. Mean square error of proposed estimator shows an improvement over other well known estimators under the same case.

Keywords: Mean Square Error; Precision; Two-Phase Sampling; Auxiliary Variable; Regression Type Estimator; Simple Random Sampling without Replacement

1. Introduction

It is fact that precision of estimators of the mean of study variable "y" is increased by proper attachment of highly correlated auxiliary variables. In some situations where auxiliary information is available at population level and cost per unit of collecting study variable "y" is affordable then single-phase sampling is more appropriate. But in a situation where prior information of auxiliary variable is lacking then it is neither practical nor economical to conduct a census for this purpose. The appropriate technique used to get estimates of those auxiliary variables on the basis of samples is two-phase sampling. In such cases we take large preliminary sample and from that auxiliary variables are computed. The main sample is independently sub-sampled from that large sample.

Two-phase sampling is a powerful technique which was firstly introduced by Neyman [2] for the stratification purpose. Two-phase sampling is based on the idea of a sampling design in which nature (specifically the size) of sampling units does not differ at any phase of sampling. "Two-phase sampling is generally employed when number of units, required to give the desired precision on different items, is widely different. This technique is employed to utilize the information collected at the first phase in order to improve the precision of the information to be collected at the second phase" [3]. In two-phase sampling, regression and ratio estimation techniques are used to estimate the finite population mean. Ratio estimator incorporates the prior information closely related to study variable and regression technique is used when relation between study variable and auxiliary variable(s) is linear. Regression estimator is considered to be more useful than ratio estimator except when regression line does not pass through origin otherwise these two estimators have almost same significance and analyst has to decide intuitively.

Let the population consist of N units, y_i, x_i and z_i (i = 1, 2, ..., n) denote the values of the *i*-th unit of the character Y, X and Z respectively. Here y_i is our variable of interest, x_i is main auxiliary variable and z_i is second auxiliary variable. The two auxiliary variables are highly correlated with variable of interest. Let S_1 be first phase sample of size n_1 from the population of size N according to a simple random sampling without replacement and $\overline{x_1}$, $\overline{z_1}$ the sample means of two auxiliary variables are observed. Let S_2 be second phase sample of size n_2 from first phase sample and $\overline{y_2}, \overline{x_2}, \overline{z_2}$ are observed. The notations used in this paper are:

$$\overline{y}_{2} = \overline{Y} + \overline{\varepsilon} y_{2}, \theta_{1} = (1/n_{1} - 1/N),$$

$$E\left(\overline{\varepsilon}_{x_{1}}\overline{\varepsilon}_{x_{2}}\right) = \theta_{1}S_{x}^{2} = \theta_{1}\overline{X}^{2}C_{x}^{2},$$

$$E\left(\overline{\varepsilon}_{x_{1}}\overline{\varepsilon}_{y_{2}}\right) = \theta_{1}\overline{X}\overline{Y}C_{x}\rho_{xy}.$$
(1.1)

Cochran [4] appears to be the first to use auxiliary information in Ratio estimator when there is highly positive correlation between study variable and auxiliary variables. Hansen and Hurwitz [5] were first to suggest the use of auxiliary information in selecting the population with varying probabilities. Robson [6] gave the idea of product estimator when there is highly negative correlation. Two-phase sampling version of [6] is:

$$T_{l(2)} = \overline{y}_2 \, \frac{\overline{x}_2}{\overline{x}_1} \,, \tag{1.2}$$

$$MSE\left(T_{1(2)}\right)$$

= $\overline{Y}^{2}\left[\theta_{2}C_{y}^{2} + \left(\theta_{2} - \theta_{1}\right)\left(C_{x}^{2} + 2C_{x}C_{y}\rho_{xy}\right)\right].$ (1.3)

Sukhatme [3] used auxiliary variable in his ratio type estimators for two-phase sampling. One of his estimators was:

$$T_{2(2)} = \frac{\overline{y}_2}{\overline{x}_2} \,\overline{x}_1, \qquad (1.4)$$

$$MSE(T_{2(2)})$$

= $\overline{Y}^{2} \Big[\theta_{2}C_{y}^{2} + (\theta_{2} - \theta_{1})(C_{x}^{2} - 2C_{x}C_{y}\rho_{xy}) \Big].$ (1.5)

Raj [7] proposed a method of using information on several variates to achieve higher precision in two-phase sampling. The two-phase sampling version of [7] is:

$$T_{3(2)} = wU + (1 - w)V \tag{1.6}$$

where $U = \overline{y} + b_{yx} (\overline{x_1} - \overline{x_2}), V = \overline{y} + b_{yz} (\overline{z_1} - \overline{z_2})$ and "w" is a suitably chosen constant.

$$MSE(T_{3(2)}) = \theta_{2}\overline{Y}^{2}C_{y}^{2} + (\theta_{2} - \theta_{1})\overline{Y}^{2}C_{y}^{2}$$

$$\times \left[1 - \rho_{yx}^{2} - \frac{\rho_{yz}^{2}(\rho_{yz} - \rho_{yx}\rho_{xz})^{2}}{(\rho_{yx}^{2} + \rho_{yz}^{2} - 2\rho_{yx}\rho_{yz}\rho_{xz})}\right].$$
(1.7)

Mohanty [8] demonstrated that precision of study variable in two-phase sampling can be increased by combining the regression and ratio estimators using two auxiliary variables.

$$T_{4(2)} = \left[\overline{y}_2 + b_{yx} \left(\overline{x}_1 - \overline{x}_2\right)\right] \frac{\overline{z}_1}{\overline{z}_2}, \qquad (1.8)$$

$$MSE(T_{4(2)}) = \overline{Y}^{2} \bigg[\theta_{2} C_{y}^{2} + (\theta_{2} - \theta_{1}) \Big\{ \rho_{xz}^{2} C_{z}^{2} - (\rho_{xy} C_{y} - \rho_{xz} C_{z})^{2} + (C_{z} - C_{y} \rho_{yz})^{2} - C_{y}^{2} \rho_{yz}^{2} \Big\} \bigg].$$
(1.9)

Srivastava [9] developed a following class of ratio

type estimators:

$$T_{5(2)} = \overline{y}_2 \left(\frac{\overline{x}_1}{\overline{x}_2}\right)^{\alpha}, \qquad (1.10)$$

$$MSE(T_{5(2)}) = \overline{Y}^2 C_y^2 \Big[\theta_2 - (\theta_2 - \theta_1) \rho_{xy}^2 \Big]. \quad (1.11)$$

Mukerjee *et al.* [10] developed three regression type estimators. One was for the situation when no auxiliary information was available.

$$T_{6(2)} = \overline{y}_2 + b_{yx} \left(\overline{x}_1 - \overline{x}_2 \right) + b_{yz} \left(\overline{z}_1 - \overline{z}_2 \right), (1.12)$$

$$MSE(T_{6(2)}) = \overline{Y}^{2}C_{y}^{2} \Big[\theta_{2} - (\theta_{2} - \theta_{1})(\rho_{xy}^{2} + \rho_{yz}^{2} - 2\rho_{xy}\rho_{yz}\rho_{xz})\Big].$$
(1.13)

Samiuddin and Hanif [11] developed two-phase sampling version of Sukhatme *et al.* [12] regression estimator when population means \overline{X} and \overline{Z} are not known.

$$T_{7(2)} = \overline{y} + \alpha_1 \left(\overline{x}_1 - \overline{x}_2 \right) + \alpha_2 \left(\overline{z}_1 - \overline{z}_2 \right)$$
(1.14)

$$MSE(T_{7(2)}) = \overline{Y}^{2}C_{y}^{2} \left[\theta_{2}\left(1 - \rho_{y \cdot xz}^{2}\right) + \theta_{1}\rho_{y \cdot xz}^{2}\right] \quad (1.15)$$

where $\alpha_1 = \frac{C_y \left(\rho_{xy} - \rho_{xz} \rho_{yz} \right)}{C_x \left(1 - \rho_{xz}^2 \right)}, \ \alpha_2 = \frac{C_y \left(\rho_{yz} - \rho_{xy} \rho_{xz} \right)}{C_z \left(1 - \rho_{xz}^2 \right)}$

and $\rho_{y,xz}^2$ is the partial correlation coefficient of y given x and z.

Singh and Espejo [13] extended their own work of single phase sampling ratio-product estimator suggested in (2003) to two-phase sampling.

$$T_{8(2)} = \overline{y}_2 \left\{ k \frac{\overline{x}_1}{\overline{x}_2} + \left(1 - k\right) \frac{\overline{x}_2}{\overline{x}_1} \right\}, \qquad (1.16)$$

$$MSE(T_{8(2)}) = \bar{Y}^{2}C_{y}^{2}\left[\theta_{2}\left(1-\rho_{xy}^{2}\right)+\theta_{1}\rho_{xy}^{2}\right] \quad (1.17)$$

where $k = \frac{1}{2} \left(1 + \frac{C_y}{C_x} \rho_{xy} \right)$ and $0 \le k \le 1$.

Hanif *et al.* [14] developed regression type estimators of population mean in two-phase sampling. One of those estimators was:

$$T_{9(2)} = \left(\overline{y}_2 + b_{yx}\left(\overline{x}_1 - \overline{x}_2\right)\right) \left\{ K \frac{\overline{z}_1}{\overline{z}_2} + \left(1 - K\right) \frac{\overline{z}_2}{\overline{z}_1} \right\}, \quad (1.18)$$

$$MSE(T_{9(2)}) = \overline{Y}^{2}C_{y}^{2}\left[\theta_{2} - (\theta_{2} - \theta_{1})\left\{\rho_{xy}^{2} + (\rho_{yz} - \rho_{xy}\rho_{xz})^{2}\right\}\right].$$
(1.19)

2. Proposed Regression Type Estimator for Two-Phase Sampling

We propose following estimator using two auxiliary

OJS

variables for two-phase sampling when we don't have any information of auxiliary variables *i.e.* both \overline{x}_1 and \overline{z}_1 are unknown.

$$T_{H(2)} = \left\{ \overline{y}_2 + K_1 \left(\overline{x}_1 - \overline{x}_2 \right) \right\} \left\{ K_2 \frac{\overline{z}_1}{\overline{z}_2} + \left(1 - K_2 \right) \frac{\overline{z}_2}{\overline{z}_1} \right\}$$
(2.1)
$$-\infty \le K_1 \le \infty, 0 \le K_2 \le 1.$$

Putting the notations of (1.1) in (2.1), squaring and taking expectation, we can obtain mean square as:

$$MSE(T_{H(2)}) = \left[\theta_{2}\bar{Y}^{2}C_{y}^{2} + K_{1}^{2}\bar{X}^{2}C_{x}^{2}(\theta_{2} - \theta_{1}) + \bar{Y}^{2}C_{z}^{2}(\theta_{2} - \theta_{1}) + 4K_{2}^{2}\bar{Y}^{2}C_{z}^{2}(\theta_{2} - \theta_{1}) - 2K_{1}\bar{X}\bar{Y}C_{x}C_{y}\rho_{xy}(\theta_{2} - \theta_{1}) + 2\bar{Y}^{2}C_{y}C_{z}\rho_{yz}(\theta_{2} - \theta_{1}) - 4K_{2}\bar{Y}^{2}C_{y}C_{z}\rho_{yz}(\theta_{2} - \theta_{1}) - 2K_{1}\bar{X}\bar{Y}C_{x}C_{z}\rho_{xz}(\theta_{2} - \theta_{1}) + 4K_{1}K_{2}\bar{X}\bar{Y}C_{x}C_{z}\rho_{xz}(\theta_{2} - \theta_{1}) + 4K_{1}K_{2}\bar{X}\bar{Y}C_{x}C_{z}\rho_{xz}(\theta_{2} - \theta_{1}) - 4K_{2}\bar{Y}^{2}C_{z}^{2}(\theta_{2} - \theta_{1})\right].$$

$$(2.2)$$

In order to get optimum value of K_1 and K_2 we differentiate (2.2) with respect to K_1 and equating to zero we get:

$$K_{2} = \frac{1}{2} + \frac{1}{2} \frac{C_{y}}{C_{z}} \rho_{yz} - \frac{1}{2} K_{1} \frac{\overline{X}C_{x}}{\overline{Y}C_{z}} \rho_{xz}.$$
 (2.3)

Putting the value of (2.3) in (2.2) and differentiating with respect to K_1 , we get:

$$K_{1} = \frac{\overline{Y}C_{y}\left(\rho_{xy} - \rho_{xz}\rho_{yz}\right)}{\overline{X}C_{x}\left(1 - \rho_{xz}^{2}\right)} = \beta_{yx \cdot z}$$
(2.4)

where β_{yxz} is the partial regression coefficient of y on x keeping z constant.

Putting the value of (2.4) in (2.3) we get:

$$K_{2} = \frac{1}{2} \left(1 + \frac{C_{y}}{C_{z}} \left(\frac{\rho_{yz} - \rho_{xy} \rho_{xz}}{\left(1 - \rho_{xz}^{2}\right)} \right) \right)$$

$$= \frac{1}{2} \left[1 + \frac{\overline{Z}}{\overline{Y}} \beta_{yz} - \frac{\overline{Z}}{\overline{Y}} \beta_{xz} \beta_{yx \cdot z} \right].$$
 (2.5)

Putting the values of (2.4) and (2.5) in (2.2) and on simplification we have:

$$MSE(T_{H(2)}) = \overline{Y}^{2}C_{y}^{2} \Big[\theta_{2} - (\theta_{2} - \theta_{1}) \Big\{ 1 - (1 - \rho_{yz}^{2}) (1 - \rho_{xyz}^{2}) \Big\} \Big].$$
(2.6)

Expressing the proposed estimator in terms of (1.1) and taking the assumption that $\overline{\varepsilon}$ is very small and expanding $\left(1 + \frac{\overline{\varepsilon}_{z_1}}{\overline{Z}}\right)^{-1}$ and $\left(1 + \frac{\overline{\varepsilon}_{z_2}}{\overline{Z}}\right)^{-1}$ up to second de-

gree, we obtain bias of above estimator as follows

Copyright © 2013 SciRes.

Bias

.

$$= \left(\theta_2 - \theta_1\right) C_z \left\{ K_2 \overline{Y} C_z + \left(\overline{Y} C_y \rho_{yz} - K_1 \overline{X} C_x \rho_{xz}\right) \left(1 - 2K_2\right) \right\}.$$

$$(2.7)$$

Putting (2.4) and (2.5) in (2.7) and after simplification, the optimized bias is

Optimum Bias

$$= (\theta_2 - \theta_1) \overline{Y} C_z^2$$

$$\times \left\{ \frac{1}{2} + \frac{C_y \left(\rho_{yz} - \rho_{xz} \rho_{xy} \right)}{C_z \left(1 - \rho_{xz}^2 \right)} \left(\frac{1}{2} - \frac{C_y \left(\rho_{yz} - \rho_{xz} \rho_{xy} \right)}{C_z \left(1 - \rho_{xz}^2 \right)} \right) \right\}.$$
(2.8)

3. Mathematical Comparison of Proposed Estimator over Other Estimators

In this section, an improvement of our proposed estimator is shown over well-known estimators of two-phase sampling. In each case no information about population characteristics of auxiliary variables is available. It is proved through mathematical comparison that our proposed estimator outperforms the other estimators. We have compared our estimator with [3,6-11,13,14] estimators. The mathematical efficiency of our proposed estimator is given as:

a) Comparison with Robson [6] Estimator

$$MSE(T_{1(2)}) - MSE(T_{H(2)})$$

= $(\theta_2 - \theta_1) \left((C_x + C_y \rho_{xy})^2 + C_y^2 \frac{(\rho_{yz} - \rho_{xy} \rho_{xz})^2}{(1 - \rho_{xz}^2)} \right)$ (3.1)
 $\geq 0.$

b) Comparison with Sukhatme [3] Estimator

$$MSE(T_{2(2)}) - MSE(T_{H(2)})$$

= $(\theta_2 - \theta_1) \left((C_x - C_y \rho_{xy})^2 + C_y^2 \frac{(\rho_{yz} - \rho_{xy} \rho_{xz})^2}{(1 - \rho_{xz}^2)} \right)$ (3.2)
 $\geq 0.$

c) Comparison with Raj [7] Estimator

$$MSE(T_{3(2)}) - MSE(T_{H(2)})$$

$$= \frac{(\theta_2 - \theta_1)\overline{Y}^2 C_y^2 (\rho_{yz} - \rho_{yx}\rho_{xz})^2 (\rho_{yx} - \rho_{yz}\rho_{xz})^2}{(1 - \rho_{xz}^2)(\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz})} (3.3)$$

 $\geq 0.$

d) Comparison with Mohanty [8] Estimator

$$MSE\left(T_{4(2)}\right) - MSE\left(T_{H(2)}\right)$$

= $\left(\theta_{2} - \theta_{1}\right)$
× $\left(\left(C_{z} - C_{y}\left(\rho_{yz} - \rho_{xy}\rho_{xz}\right)\right)^{2} + C_{y}^{2}\left(\rho_{yz} - \rho_{xy}\rho_{xz}\right)^{2}\frac{\rho_{xz}^{2}}{\left(1 - \rho_{xz}^{2}\right)}\right)$
≥ 0. (3.4)

e) Comparison with Srivastava [9] Estimator

$$MSE(T_{5(2)}) - MSE(T_{H(2)})$$

= $(\theta_2 - \theta_1)C_y^2 \frac{(\rho_{yz} - \rho_{xy}\rho_{xz})^2}{(1 - \rho_{xz}^2)} \ge 0.$ (3.5)

f) Comparison with Mukerjee et al. [10] Estimator

$$MSE(T_{6(2)}) - MSE(T_{H(2)})$$

= $(\theta_2 - \theta_1) \left((\rho_{xy} - \rho_{xz} \rho_{yz})^2 \frac{\rho_{xz}^2}{(1 - \rho_{xz}^2)} + \rho_{xz}^2 \rho_{yz}^2 \right)$ (3.6)
 $\geq 0.$

g) Comparison with Sammiuddin and Hanif [11] Estimator

Our proposed estimator gives identical result to [11] because

$$MSE(T_{7(2)}) - MSE(T_{H(2)}) = 0.$$
 (3.7)

But our estimator is more preferable than [11] if we have the estimate of $K_1 = \beta_{yx\cdot z}$, in this way we have to find only one unknown value whereas in [11] estimator we have to find two unknown values. Following special cases give another reason for the suitability of our estimator. Our estimator:

1) becomes classical ratio estimator for $k_1 = 0$ and $k_2 = 0$;

2) converts into Robson [6] estimator for $k_1 = 0$ and $k_2 = 0$;

3) emerges into Mohanty [8] estimator for $k_1 = b_{YX}$ and $k_2 = 0$;

4) reduces to estimator given by Singh and Espejo [13] for $k_1 = 0$;

5) turns into Hanif *et al.* [14] estimator for $k_1 = b_{YX}$.

h) Comparison with Singh and Espejo [13] Estimator

$$MSE\left(T_{8(2)}\right) - MSE\left(T_{H(2)}\right)$$
$$= \left(\theta_2 - \theta_1\right) \frac{\left(\rho_{yz} - \rho_{xy}\rho_{xz}\right)^2}{\left(1 - \rho_{xz}^2\right)} \ge 0.$$
(3.8)

i) Comparison with Hanif et al. [4] Estimator

$$MSE(T_{9(2)}) - MSE(T_{H(2)})$$

= $(\theta_2 - \theta_1) \frac{(\rho_{xz} \rho_{yz} - \rho_{xy} \rho_{xz}^2)^2}{(1 - \rho_{xz}^2)} \ge 0.$ (3.9)

4. Conclusion

In this paper we have proposed a regression type estimator for two-phase sampling when we don't have any advance knowledge of auxiliary variables. [6,8,13,14] are the special cases of our estimator. From Equations (3.1) to (3.9) one can readily see that our proposed estimator is more precise than all other competing estimators discussed in Section 1, so we can say that our estimator provides more accurate estimate about the population parameters.

REFERENCES

- M. Hanif, N. Hamad and M. Q. Shahbaz, "A Modified Regression Type Estimator in Survey Sampling," *World Applied Sciences Journal*, Vol. 7, No. 12, 2009, pp. 1559-1561.
- [2] J. Neyman, "Contribution to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, Vol. 33, No. 201, 1938, pp. 101-116. doi:10.1080/01621459.1938.10503378
- [3] B. V. Sukhatme, "Some Ratio Type Estimators in Two-Phase Sampling," *Journal of the American Statistical Association*, Vol. 57, No. 299, 1962, pp. 628-632. doi:10.1080/01621459.1962.10500551
- [4] W. G. Cochran Cochran, "The Estimation of the Yields of the Cereal Experiments by Sampling for the Ratio of Grain to Total Produce," *Journal of Agricultural Science*, Vol. 30, No. 2, 1940, pp. 262-275. doi:10.1017/S0021859600048012
- [5] M. H. Hansen and W. N. Hurwitz, "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, Vol. 14, No. 4, 1943, pp. 333-362. doi:10.1214/aoms/1177731356
- [6] D. S. Robson, "Application of Multivariate Polykays to the Theory of Unbiased Ratio Type Estimators," *Journal* of the American Statistical Association, Vol. 52, No. 280, 1957, pp. 511-522. doi:10.1080/01621459.1957.10501407
- [7] D. Raj, "On a Method of Using Multi-Auxiliary Information in Sample Surveys," *Journal of the American Statistical Association*, Vol. 60, No. 309, 1965, pp. 270-277. doi:10.1080/01621459.1965.10480789
- [8] S. Mohanty, "Combination of Regression and Ratio Estimate," *Journal of Indian Statistical Association*, Vol. 5, 1967, pp. 16-19.
- [9] S. K. Srivastava, "A Generalized Estimator for the Mean of a Finite Population Using Multi Auxiliary Information," *Journal of the American Statistical Association*, Vol. 66, No. 334, 1971, pp. 404-407.

doi:10.1080/01621459.1971.10482277

- [10] R. Mukerjee, T. J. Rao and K. Vijayan, "Regression Type Estimators Using Multiple Auxiliary Information," *Australian Journal of Statistics*, Vol. 29, No. 3, 1987, pp. 244-254. doi:10.1111/j.1467-842X.1987.tb00742.x
- [11] M. Samiuddin and M. Hanif, "Estimation of Population Mean in Single Phase and Two-Phase Sampling with or without Additional Information," *Pakistan Journal of Statistics*, Vol. 23, No. 2, 2007, pp. 99-118.
- [12] P. V. Sukhatme, B. V. Sukhatme, S. Sukhatme and C. Asok, "Sampling Theory of Surveys with Applications,"

Iowa State University Press, Ames, 1984.

- [13] H. P. Singh and M. R. Espejo, "Double Sampling Ratio-Product Estimator of a Finite Population Mean in Sample Surveys," *Journal of Applied Statistics*, Vol. 34, No. 1, 2007, pp. 71-85. <u>doi:10.1080/02664760600994562</u>
- [14] M. Hanif, N. Hamad and M. Q. Shahbaz, "Some New Regression Type Estimators in Two-Phase Sampling," *World Applied Sciences Journal*, Vol. 8, No. 7, 2010, pp. 799-803.