BULLETIN *of the*
MALAYSIAN MATHEMATICAL
SCIENCES SOCIETY
http://math.usm.my/bulletin

Bull. Malays. Math. Sci. Soc. (2) **28***(1) (2005), 81–86*

A Chain Regression Estimator in Two Phase Sampling Using Multi-auxiliary Information

B.K. PRADHAN

Department of Statistics, Utkal University, Bhubaneswar-751004, India
bkpradhanuu@rediffmail.com

Abstract. In this paper the population mean of the study variable y is estimated in a two phase sampling setup using three auxiliary variables with chain regression concept when the population mean of one of the auxiliary variables is unknown and other auxiliary population mean are known.

2000 Mathematics Subject Classification: 62D05

Key words and phrases: Chain regression, multi-auxiliary information, two phase sampling.

1. Introduction

Information on variables correlated with the main variable under study is popularly known as auxiliary information which may be fruitfully utilised either at planning stage or at design stage or at the information stage to arrive at improved estimator compared to those, not utilising auxiliary information. Use of auxiliary information for forming ratio and regression method of estimation were introduced during the 1930's with a comprehensive theory provided by Cochran [1]. Assuming knowledge of multi-auxiliary variables, multivariate ratio estimator was suggested by Olkin [5], multivariate difference estimator by Raj [7], multiple regression estimator by Shukla [11], weighted regression estimator by Srivastava [13, 14, 15] and Ratio-cum-product estimator by Singh [12]. Extension of these estimators to different sampling designs were taken up by Tripathy [17, 18]. Further contribution are due to Rao and Mudholkar [8], Wright [19] and many others.

When information on any auxiliary variable x highly correlated with y is readily available on all units of the population, it is well known that ratio and regression estimators provide more efficient estimates of population mean of y, envisaging advance information on population mean $\overline{X}$ of x. However, in certain practical situation when $\overline{X}$ is unknown, information on auxiliary variables Z and W are readily available on all the units of the population, which may also be incorporated in the method of estimation.

Received: January 3, 2004; *Revised:* June 24, 2004.

2. Two phase sampling set up

Consider a finite population U of size N indexed by quadruplet characters (y, x, z, w). Our purpose is to estimate the population mean $\overline{Y}$ of a study variable y in the presence of three auxiliary variables x, z and w, when the population mean $\overline{X}$ of x is unknown but information on z and w are available on all the units of the population.

Let us now consider a two phase sampling where in the first phase a large sample $S'(S' \subset U)$ of fixed size n' is drawn following SRSWOR and observe three auxiliary variables x, z and w to estimate $\overline{X}$, while in the second phase a sub-sample $S \subset S'$ of fixed size n is drawn by SRSWOR to observe the characteristic y under study.

3. Use of one auxiliary variable

The two phase regression estimators in this case will be

$$\bar{t}_{1(\mathrm{Re\,g})} = \overline{y}_n + b_{yx}(\overline{x}_{n'} - \overline{x}_n) \tag{3.1}$$

where b_{yx} is the sample regression coefficient of y on x calculated from data based on S and

$$\overline{x}_{n'} = \frac{1}{n'}\sum_{i\in S'} x_i \quad \overline{x}_n = \frac{1}{n}\sum_{i\in S} x_i \quad \text{and} \quad \overline{y}_n = \frac{1}{n}\sum_{i\in S} y_i.$$

The mean square error (MSE) of $\bar{t}_{1(\mathrm{Re\,g})}$ by first order approximation is

$$MSE(\bar{t}_{1(\mathrm{Re\,g})}) = \left(\frac{1}{n} - \frac{1}{N}\right)(1 - \rho_{yx}^2)S_y^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{yx}^2 S_y^2 \tag{3.2}$$

where

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$$

and ρ_{yx} is the correlation coefficient between y and x.

4. Use of second auxiliary variable

Swain [16], Kiregyera [2], Mukherjee *et al.* [4], Sahoo *et al.* [9] and Mishra *et al.* [3] used a second auxiliary variable z closely related to x to suggest different improved estimators assuming that the information on z is available on all the units of the population.

Kiregyera [2] has suggested a regression type estimator

$$\bar{t}_{2(\mathrm{Re\,g})}) = \overline{y}_n + b_{yx}[\overline{x}_{n'} + b_{xz}(\overline{Z} - \overline{z}_{n'}) - \overline{x}_n] \tag{4.1}$$

a technique earlier suggested by Swain [16] which yields

$$MSE(\bar{t}_{2(\mathrm{Re\,g})}) = \left(\frac{1}{n} - \frac{1}{N}\right)(1 - \rho_{yx}^2)S_y^2 + (\rho_{yx}^2 + \rho_{yx}^2\rho_{xz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz})S_y^2. \tag{4.2}$$

Sahoo *et al.* [9] considered a chain regression type estimator

$$\bar{t}_{3(\mathrm{Re\,g})} = \overline{y}_n + b_{yx}(\overline{x}_{n'} - \overline{x}_n) + b_{yz}(\overline{Z} - \overline{z}_{n'}). \tag{4.3}$$

The MSE of $\bar{t}_{3(\mathrm{Reg})}$ to the first order of approximation is

$$MSE(\bar{t}_{3(\mathrm{Reg})}) = \left(\frac{1}{n} - \frac{1}{N}\right)(1-\rho_{yx}^2)S_y^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)(\rho_{yx}^2 - \rho_{yz}^2). \tag{4.4}$$

Another regression type estimator suggested by Mukherjee *et al.* [4] is

$$\bar{t}_{4(\mathrm{Reg})} = \overline{y}_n + b_{yx.z}\left[\overline{x}_{n'} + b_{xz}(\overline{Z} - \overline{z}_{n'}) - \overline{x}_n\right] + b_{yz.x}(\overline{Z} - \overline{z}_n). \tag{4.5}$$

Using a generalized method, a regression type estimator suggested by Mishra and Rout [3] is

$$\bar{t}_{5(\mathrm{Reg})} = \overline{y}_n + b_{yx.z}(\overline{x}_{n'} - \overline{x}_n) + b_{yz.x}(\overline{Z} - \overline{z}_n) + (b_{yz} - b_{yz.x})(\overline{Z} - \overline{z}_{n'}). \tag{4.6}$$

In fact, on simplification $\bar{t}_{4(\mathrm{Reg})} = \bar{t}_{5(\mathrm{Reg})}$. It may be seen that

$$MSE(\bar{t}_{4(\mathrm{Reg})}) = MSE(\bar{t}_{5(\mathrm{Reg})}) = \left(\frac{1}{n} - \frac{1}{N}\right)(1-\rho_{y.xz}^2)S_y^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)(1-\rho_{yz}^2)S_y^2. \tag{4.7}$$

5. Suggested estimators

Since $\widehat{\overline{X}}$ based on n' unit is an unbiased estimator of $\overline{X}$, a regression type estimator

$$\widehat{\overline{X}} = \overline{x}_{n'} + \beta_{xz.w}(\overline{Z} - \overline{z}_{n'}) + \beta_{xw.z}(\overline{W} - \overline{w}_{n'}) \tag{5.1}$$

is found by considering the estimator

$$\widehat{\overline{X}} = \lambda_1\overline{x}_{n'} + \lambda_2\overline{z}_1 + \lambda_3\overline{Z} + \lambda_4\overline{w}_{n'} + \lambda_5\overline{W} \tag{5.2}$$

which is unbiased for $\overline{X}$ by considering $\lambda_1 = 1$, $\lambda_2 = -\lambda_3$ and $\lambda_4 = -\lambda_5$ and then by minimizing the variance $\widehat{\overline{X}}$ given by

$$\widehat{\overline{X}} = \overline{x}_{n'} + \lambda_2(\overline{z}_{n'} - \overline{Z}) + \lambda_4(\overline{w}_{n'} - \overline{W}). \tag{5.3}$$

Here

$$\overline{x}_{n'} = \frac{1}{n'}\sum_{i\in S'} x_i, \quad \overline{z}_{n'} = \frac{1}{n'}\sum_{i\in S'} z_i, \quad \overline{w}_{n'} = \frac{1}{n'}\sum_{i\in S'} w_i$$

and $\beta_{xz.w}$ and $\beta_{xw.z}$ are usual partial regression coefficients.

Let us consider a chain regression type estimator of $\overline{Y}$ given by

$$\bar{t}^* = \overline{y}_n + \lambda_1^*(\widehat{\overline{X}} - \overline{x}_n) + \lambda_2^*(\overline{Z} - \overline{z}_n) + \lambda_3^*(\overline{W} - \overline{w}_n). \tag{5.4}$$

where $\overline{y}_n$, $\overline{x}_n$, $\overline{z}_n$ and $\overline{w}_n$ are the sample mean based on n observations of the second phase and λ_1^*, λ_2^* and λ_3^* are suitable constants.

The optimum values of λ_1^*, λ_2^* and λ_3^* are obtained by minimising $V(\bar{t}^*)$ and we find

$$\lambda_1^* = \beta_{yx.zw}, \quad \lambda_2^* = \beta_{yz.xw} \text{ and } \lambda_3^* = \beta_{yw.xz} \tag{5.5}$$

where $\beta_{yx.zw}$, $\beta_{yz.xw}$ and $\beta_{yw.xz}$ are usual partial regression coefficients. When the partial regression coefficients are known, $\bar{t}^*$ is an unbiased estimator of $\overline{Y}$ with

$$V(\bar{t}^*) = \left(\frac{1}{n} - \frac{1}{N}\right)(1-\rho_{y.xzw}^2)S_y^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)(1-\rho_{y.zw}^2)\rho_{yx.zw}^2 S_y^2 \tag{5.6}$$

where $\rho_{y.xzw}$ and $\rho_{y.zw}$ are usual multiple correlation coefficients and $\rho_{yx.zw}$ is the usual partial correlation coefficient.

The estimators under consideration require advance knowledge of the population regression coefficients and partial regression coefficients, which are usually unknown. However, in practice the consistent estimators $b_{yx.zw}$, $b_{yz.xw}$ and $b_{yw.xz}$ of the population parameters $\beta_{yx.zw}$, $\beta_{yz.xw}$ and $\beta_{yw.xz}$ may be substituted for the purpose. Although the estimators will turn out to be biased, this bias would be negligible in large samples and the approximate mean square errors to $O(1/n)$ will be equivalent to those derived and for large sample, the difference would be minimal.

6. Comparison of efficiency

Sahoo *et al.* [9] has established that $\bar{t}_{1(\mathrm{Re\,g})}$ and $\bar{t}_{2(\mathrm{Re\,g})}$ are less efficient than $\bar{t}_{3(\mathrm{Re\,g})}$. Mishra and Rout [3] has proved that

$$MSE(\bar{t}_{5(\mathrm{Re\,g})}) < MSE(\bar{t}_{3(\mathrm{Re\,g})}). \tag{6.1}$$

Now, from (4.7) and (5.6), we find

$$MSE(\bar{t}_{5(\mathrm{Re\,g})}) - V(\bar{t}^*) = \left[\left(\frac{1}{n} - \frac{1}{N}\right) A + \left(\frac{1}{n'} - \frac{1}{N}\right) B\right] S_y^2 \tag{6.2}$$

where $A = \rho^2_{y.xzw} - \rho^2_{y.xz}$ and $B = (1-\rho^2_{yz})\rho^2_{yx.z} - (1-\rho^2_{y.zw})\rho^2_{yx.zw}$. On simplification, we find

$$A + B = (1 - \rho^2_{yz})\rho^2_{yw.z} \geq 0. \tag{6.3}$$

Since

$$\left(\frac{1}{n} - \frac{1}{N}\right) > \left(\frac{1}{n'} - \frac{1}{N}\right),$$

we have from (6.3)

$$\left(\frac{1}{n} - \frac{1}{N}\right) A + \left(\frac{1}{n'} - \frac{1}{N}\right) B \geq \left(\frac{1}{n'} - \frac{1}{N}\right)(A + B) \geq 0. \tag{6.4}$$

Hence

$$MSE(\bar{t}_{5(\mathrm{Re\,g})}) \geq V(\bar{t}^*). \tag{6.5}$$

The inequality (6.5) shows that $\bar{t}^*$ is an improved regression estimator compared to $\bar{t}_{1(\mathrm{Re\,g})}$, $\bar{t}_{2(\mathrm{Re\,g})}$, $\bar{t}_{3(\mathrm{Re\,g})}$ and $\bar{t}_{5(\mathrm{Re\,g})}$.

7. Numerical illustration

Percent relative efficiency of different estimators compared to mean per unit estimator are presented in Table 2.

Table 1. Description of Population

	Population I	Population II
Source	"Spray congealing: Particle size relationships using a centrifugal wheel automizer" by Scott, Robinson, Pauls and Lantz (1964)	"Measurement of four characters of: Flucus Religiousament" by Pradhan, B.K. (2000)
y	Mean surface-volume particle size of product	Length of petiole
x	Feed rate per unit whetted wheel periphery (gm/sec/cm)	Length of pamina(blade) of the leaf
z	Peripheral wheel velocity(cm/sec)	Width of the leaf at its widest paint
w	Feed Viscosity (poise)	Width of leaf half way along the blade
size	N=35	N=160
ρ_{yx}	0.712296	0.5423
ρ_{yz}	-0.8070192	0.6166
ρ_{yw}	-0.1623959	0.2704
ρ_{xz}	-0.2633457	0.8568
ρ_{xw}	-0.0781118	0.7424
ρ_{zw}	0.1335984	0.8027

Table 2. Relative efficiency of different estimators of population variance with respect to S_y^2 under comparison

Estimator	Auxiliary variables used	Percent Relative Efficiency of Population I: $N = 35$, $n' = 12$, $n = 8$	Percent Relative Efficiency of Population II: $N = 160$, $n' = 50$, $n = 20$
$\overline{y}_n$	None	100	100
$\overline{t}_{1(\mathrm{Re\,g})}$	X	128.08	125.26
$\overline{t}_{2(\mathrm{Re\,g})}$	X, Z	159.03	145.75
$\overline{t}_{3(\mathrm{Re\,g})}$	X, Z	243.36	147.31
$\overline{t}_{4(\mathrm{Re\,g})} = \overline{t}_{5(\mathrm{Re\,g})}$	X, Z	378.98	161.47
$\overline{t}^*_{(\mathrm{Re\,g})}$	X, Z, W	436.27	204.28

Remark 7.1. $\overline{t}^*_{(\mathrm{Re\,g})}$ has substantial gain in efficiency compared to $\overline{t}_{5(\mathrm{Re\,g})}$, $\overline{t}_{4(\mathrm{Re\,g})}$, $\overline{t}_{3(\mathrm{Re\,g})}$, $\overline{t}_{2(\mathrm{Re\,g})}$, $\overline{t}_{1(\mathrm{Re\,g})}$ and $\overline{y}_n$ for the population under consideration.

Acknowledgment. The author wishes to express his sincere gratitude to the referee for his valuable suggestions in improving the manuscript.

References

[1] W. G. Cochran, Sampling theory when the sampling-units are of unequal sizes, *J. Amer. Statist. Assoc.* **37** (1942), 199–212.

[2] B. Kiregyera, Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations, *Metrika* **31**(3–4) (1984), 215–226.

[3] G. Mishra and K. Rout, A regression estimator in two-phase sampling in presence of two auxiliary variables, *Metron* **55**(1–2) (1997), 177–186.

[4] R. Mukerjee, T. J. Rao and K. Vijayan, Regression type estimators using multiple auxiliary information, *Austral. J. Statist.* **29**(3) (1987), 244–254.

[5] I. Olkin, Multivariate ratio estimation for finite populations, *Biometrika* **45** (1958), 154–165.

[6] B. K. Pradhan, Some problems of estimation in multi-phase sampling, Ph.D. Dissertation (unpublished), Utkal University, Orissa, India (2000).

[7] Des Raj, On a method of using multi-auxiliary information in sample surveys, *J. Amer. Statist. Assoc.* **60** (1965), 270–277.

[8] P. S. R. S. Rao and G. S. Mudholkar, Generalized multivariate estimator for the mean of finite populations, *J. Amer. Statist. Assoc.* **62** (1967), 1009–1012.

[9] J. Sahoo,L. N. Sahoo, and S. Mohanty, A regression approach to estimation in two phase sampling using two auxiliary variables, *Current Science* **65**(1) (1993), 73–75.

[10] M. W. Scott, M. J. Rabinson, J. F. Pauls,and R. J. Lantz, Spray congealing: Particle size relationships using a centrifugal wheel atomizer, *J. Pharmaceutical Sci.*, **53**(6) (1964), 670-675.

[11] G. K. Shukla, Multivariate regression estimate, *J. Indian Statist. Assoc.* **3** (1965), 202–211.

[12] M. P. Singh, Ratio cum product method of estimation, *Metrika* **12** (1967), 34–42.

[13] S. K. Srivastava, An estimate of the mean of a finite population using several auxiliary variables, *J. Indian Statist. Assoc.* **3** (1965), 189–194.

[14] S. K. Srivastava, An estimator using auxiliary information in sample surveys, *Calcutta Statist. Assoc. Bull.* **16** (1967), 121–132.

[15] S. K. Srivastava, A generalized estimator for the mean of a finite population using multi-auxiliary information, *J. Amer. Statist. Assoc.* **66** (1971), 404–407.

[16] A. K. P. C. Swain, A note on the use of multiple auxiliary variables in sample surveys, *Trabajos de Estedistica*, **21** (1970), 135–141.

[17] T. P. Tripathy, Contribution to sampling theory using multivariate information, Ph.D. Thesis (unpublished), Punjab University, Patiala, India, (1970).

[18] T. P. Tripathy, On double sampling for multivariate ratio and difference methods of estimation, *J. Indian Soc. Ag. Stat.* **28** (1976), 35–54.

[19] R. L. Wright, Finite population sampling with multivariate auxiliary information, *J. Amer. Statist. Assoc.* **78**(384) (1983), 879–884.