International Journal of Mathematics Trends and Technology Volume 67 Issue 3, 51-56, March, 2021
ISSN: 2231 – 5373 /doi:10.14445/22315373/IJMTT-V67I3P507

A Class of Regression Estimators for Finite Population Mean under Two-Phase Sampling

Priyaranjan Dash[#1], Bishnupriya Behera[*2]

[#]*Department of Statistics, Utkal University, Vani Vihar, Bhubaneswar – 751004, India*

Abstract - *In this paper, we have suggested two different classes of regression-type estimators in two phase sampling using SRSWOR scheme at all the phases. We have seen that one of the suggested class of estimator is more efficient than some existing estimators as it has a minimum mean square error in three phase sampling..*

Keywords — *Multi-phase sampling, regression estimator, bias, mean square error (MSE).*

I. INTRODUCTION

The survey samplers explore the use of auxiliary information at several stages in order to design better estimators for estimating the population parameters. More specifically, the auxiliary information is used at estimation stage for developing more efficient estimators like ratio estimator in case of high degree of positive correlation between study variable (y) and auxiliary variable (x) and product estimator in case of high degree of negative correlation between study variable and auxiliary variable. But in both the cases, the estimators are optimum when the regression line of y on x is linear passing through the origin. But, when the linear regression line of y on x is does not pass-through origin, difference estimator is appropriate to use for estimating the finite population mean. For a detail review on these estimators, one can go through Cochran (1977), Tamhane (1978), Kiregyera (1980, 1984), Rao (1987), Bisht and Sisodia (1990), Naik and Gupta (1991), Updhyaya and Singh (1999) and Singh and Tailor (2005a,b), Singh et. al. (2006), Swain (2012), Khare et. al. (2013), Singh and Majhi (2014), Khan (2016) and many more.

When the population mean of the auxiliary variable X¯is not known to us, the use of double sampling procedure was studied by Bose (1943) and Cochran (1963). In a double sampling procedure, the second phase sample is selected from the first phase sample was initially discussed but Cochran (1963) suggested the independent selection of second phase sample directly from the population.

II. REGRESSION ESTIMATOR IN TWO PHASE SAMPLING

Consider a finite population with N distinct and identifiable units with y and x be the study variable and auxiliary variable taking value y_i and x_i for the i^{th} unit of the population. The classical regression estimator assumes the knowledge of population mean $\bar{X}$ of the auxiliary variable which is not sometimes available to us in advance. In such cases, in order to take the advantage of auxiliary information x, we use double sampling or two-phase sampling method. Here, we select a large preliminary sample $S'$of size n' from N units of the population by SRSWOR and study only x variable which require a very little cost. This sample is known as the first phase sample. From this selected first phase sample, we select a second phase sample S of size $n, (n < n')$, using SRSWOR scheme and study both y and x. The classical regression estimator for estimating population mean $\bar{Y}$ of y is given by

$$t_1 = \bar{y}_n + b_{yx}(\bar{x}_{n'} - \bar{x}_n) \tag{2.1}$$

where $\bar{y}_n$, $\bar{x}_n$ are the sample means of y and x respectively and $b_{yx} = \frac{s_{yx}}{s_x{}^2}$ is the sample regression coefficient of y on x calculated on the basis of second phase sample S and $\bar{x}_{n'}$is the sample mean of x basing on the units of first phase sample S'. The mean square error (MSE) of this estimator upto first order of approximation is given by

$$MSE_1 = MSE(t_1) = \theta\left(1 - \rho^2{}_{yx}\right)S_y{}^2 + \theta_1\rho^2{}_{yx}S_y{}^2 \tag{2.2}$$

Where $\theta = \frac{1}{n} - \frac{1}{N}$, $\theta_1 = \frac{1}{n'} - \frac{1}{N}$, $S_y{}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ is the population mean square of y and ρ_{yx} is the population correlation coefficient between y and x. The notation used for the variables x and z are similar to those of y variable.
Kiregyera (1984) suggested an estimator of $\bar{Y}$ as

$$t_2 = \bar{y}_n + b_{yx}[\bar{x}_{n'} + b_{xz}(\bar{Z} - \bar{z}_{n'}) - \bar{x}_n] \tag{2.3}$$

The MSE of this estimator up to first order of approximation is given by

$$\begin{aligned} MSE_2 = MSE(t_2) &= \theta\left(1-\rho^2{}_{yx}\right){S_y}^2 + \theta_1\left(\rho^2{}_{yx} + \rho^2{}_{yx}\rho^2{}_{xz} - 2\rho_{yx}\rho_{xz}\rho_{yz}\right){S_y}^2 \\ &= \theta\left(1-\rho^2{}_{yx}\right){S_y}^2 + \theta_1[\rho^2{}_{yx} + \rho_{yx}\rho_{xz}\{\rho_{yx}\rho_{xz} - 2\rho_{yz}\}]{S_y}^2 \end{aligned} \tag{2.4}$$

Mukerjee et. al. (1987) suggested another regression estimator of $\bar{Y}$ as

$$t_3 = \bar{y}_n + b_{yx}(\bar{x}_{n'} - \bar{x}_n) + b_{xz}(\bar{z}_{n'} - \bar{z}_n) \tag{2.5}$$

Its MSE up to first order of approximation is given by

$$MSE_3 = MSEW(t_3) = \theta\left(1-\rho^2{}_{y.xz}\right){S_y}^2 + \theta_1\rho^2{}_{y.xz}{S_y}^2 \tag{2.6}$$

Sahoo et. al. (1993) suggested an estimator for estimating $\bar{Y}$ as

$$t_4 = \bar{y}_n + b_{yx}(\bar{x}_{n'} - \bar{x}_n) + b_{xz}(\bar{Z} - \bar{z}_{n'}) \tag{2.7}$$

The MSE of this estimator up to first order of approximation is given by

$$MSE_4 = MSE(t_4) = \theta\left(1-\rho^2{}_{yx}\right){S_y}^2 + \theta_1(\rho^2{}_{yx} - \rho^2{}_{yz}){S_y}^2 \tag{2.8}$$

From equations (2.2), (2.4) and equation (2.8), the performance of estimators t_{1d}, $t_{2d}$and t_{3d} depends upon the correlation coefficients between y, x and z.

III. MODIFIED REGRESSION-IN-REGRESSION ESTIMATOR

We consider a class of estimator of $\bar{Y}$ as

$$t_{c1} = \bar{y}_n + \lambda[\{\bar{x}_{n'} + b_{xz}(\bar{z}_n - \bar{z}_{n'})\} - \bar{x}_n] \tag{3.1}$$

where λ is a constant whose value is to be determined so that the MSE $M(t_{c1})$ will be minimized. Hence,

$$\begin{aligned} MSE_{c1} &= MSE(t_{c1}) \\ &= \theta{S_y}^2 + \lambda^2{\theta_1}'{\rho_{xz}}^2{S_x}^2 + \lambda^2{\theta_1}'{S_x}^2 + 2\lambda{\theta_1}'\rho_{xz}\rho_{yz}S_yS_x - 2\lambda{\theta_1}'\rho_{yx}S_yS_x \\ &\quad - 2\lambda^2{\theta_1}'{\rho_{xz}}^2{S_x}^2 \end{aligned} \tag{3.2}$$

Where ${\theta_1}' = \frac{1}{n} - \frac{1}{n'} = \theta - \theta_1$. Now differentiating MSE_{c1} in equation (3.2) with respect to λ and equating to 0, we get

$$\begin{aligned} \frac{\partial MSE_{c1}}{\partial\lambda} &= 2\lambda{\theta_1}'{\rho_{xz}}^2{S_x}^2 + 2\lambda{\theta_1}'{S_x}^2 + 2{\theta_1}'\rho_{xz}\rho_{yz}S_yS_x - 2{\theta_1}'\rho_{yx}S_yS_x - 4\lambda{\theta_1}'{\rho_{xz}}^2{S_x}^2 = 0 \\ &\Rightarrow \lambda(1-{\rho_{xz}}^2) = (\rho_{yx} - \rho_{yz}\rho_{xz})S_yS_x \Rightarrow \lambda = \frac{\rho_{yx} - \rho_{yz}\rho_{xz}}{1-{\rho_{xz}}^2}\ \frac{S_y}{S_x} = \beta_{yx.z} \end{aligned} \tag{3.3}$$

Hence the modified regression-in-regression estimator in equation (3.1) takes the form

$$t_{c1(opt)} = \bar{y}_n + \beta_{yx.z}\left[\hat{\bar{X}} - \bar{x}_n\right], where\ \hat{\bar{X}} = \bar{x}_{n'} + b_{xz}(\bar{z}_n - \bar{z}_{n'}) \tag{3.4}$$

It can be verified that the estimator $t_{c1(opt)}$is unbiased for estimating $\bar{Y}$ and the minimum MSE is given by

$$MSE_{c1(opt)} = MSE\left(t_{c1(opt)}\right) = \theta\left[1 - (1-{\rho_{Yz}}^2)\rho^2{}_{yx.z}\right]{S_y}^2 + \theta_1(1-{\rho_{Yz}}^2)\rho^2{}_{yx.z}{S_y}^2 \tag{3.5}$$

If the population size N is very large, then $\theta \approx \frac{1}{n}$ and $\theta_1 \approx \frac{1}{n'}$, so we can write

$$MSE_{c1(opt)} = MSE\left(t_{c1(opt)}\right) \approx \left[1-\left(1-\rho_{yz}{}^2\right)\rho_{yx.z}{}^2\right]\frac{S_y{}^2}{n} + \left(1-\rho_{yz}{}^2\right)\rho_{yx.z}{}^2.\ \frac{S_y{}^2}{n'} \qquad (3.6)$$

IV. MODIFIED CHAINED REGRESSION ESTIMATOR

Consider another regression-in-regression estimator of the population mean $\bar{Y}$ as

$$t_{c2} = \bar{y}_n + \lambda_1[\{\bar{x}_{n'} + \beta_{xz}(\bar{z}_n - \bar{z}_{n'})\} - \bar{x}_n] + \lambda_2(\bar{z}_{n'} - \bar{z}_n) \qquad (4.1)$$

Where λ_1 and λ_2 are two constants which are to be determined in order to minimize the MSE of the estimator t_{c2}. Now, the MSE of t_{c2} is given by

$$\begin{aligned} MSE(t_{c2}) = \theta S_y{}^2 - \lambda_1{}^2\,\theta_1{}'\rho_{xz}{}^2 S_x{}^2 + \lambda_1{}^2\theta_1{}' S_x{}^2 + \lambda_2{}^2\,\theta_1{}' S_z{}^2 \\ + 2\lambda_1\theta_1{}'\rho_{xz}\rho_{yz}S_yS_x - 2\lambda_1\theta_1{}'\rho_{yx}S_yS_x - 2\lambda_2\theta_1{}'\rho_{yz}S_yS_z \end{aligned} \qquad (4.2)$$

Now, in order to find the optimum values of λ_1 and λ_2 that minimizes $MSE(t_{c2})$ can be obtained by partially differentiating equation (4.2) with respect to $\lambda_1$and λ_2 and equation to zero, which gives

$$\lambda_1(opt.) = \frac{\rho_{yx} - \rho_{yz}\rho_{xz}}{1-\rho_{xz}{}^2}\,\frac{S_y}{S_x} = \beta_{yx.z}\ and\ \lambda_2(opt.) = \rho_{yz}\frac{S_y}{S_z} = \beta_{yz} \qquad (4.3)$$

Using these optimum values in equation (4.1), the estimator t_{c2} reduces to

$$t_{c2(opt)} = \bar{y}_n + \beta_{yx.z}\left[\hat{\hat{X}} - \bar{x}_n\right] + \beta_{yz}(\bar{z}_{n'} - \bar{z}_n) \qquad (4.4)$$

and the minimum MSE is given by

$$MSE_{c2(opt)} = MSE\left(t_{c2(opt)}\right) = \theta\left(1-\rho_{y.xz}{}^2\right)S_y{}^2 + \theta_1\rho_{y.xz}{}^2 S_y{}^2 \qquad (4.5)$$

where $\rho_{y.xz}$ is the multiple correlation coefficient of y on x and z. If the population size N is very large, then we can write

$$MSE_{c2(opt)} = MSE\left(t_{c2(opt)}\right) = \left(1-\rho_{y.xz}{}^2\right)\frac{S_y{}^2}{n} + \rho_{y.xz}{}^2\frac{S_y{}^2}{n'} \qquad (4.6)$$

V. SIMULATION STUDY

The performance of proposed classes of estimators t_{c1}, t_{c2} are studied along with some competitive estimators like the classical regression estimator t_1, t_2 proposed by Kiregyera (1984), t_3 proposed by Mukerjee et. al. (1987), t_4 proposed by Sahoo et. al. (1993), t_5 proposed by Mukerjee et. al. (1987).

We have considered 10 different natural populations available from different text books for the comparison between these estimators. Table 1 gives the description of these populations and Table 2 gives the values of the population size (N), population mean square of y values $S_y{}^2$ and the simple correlation coefficients between $y, x\ and\ z$ i.e.; $\rho_{yx}, \rho_{xz}, \rho_{yz}$, partial correlation coefficient between $y\ and\ x, i.e.;$ $\rho_{yx.z}$ and the multiple correlation coefficient of y on x and z, i.e.; $\rho_{y.xz}$. Table 3 gives the first phase and second phase sample sizes $(n'\ and\ n)$,mean square error of these estimators. The observation are noted as follows:

1. The class of estimators t_{c1} has greater mean square error than the other classes t_{c2} and also unbiased in its optimum case.
2. The class of estimator t_{c2} in its optimum case reduces to the estimator proposed by Mukerjee et. al. (1987), which is minimum for all these populations under two phase sampling scheme.

VI. CONCLUSION

In the present paper, we have reviewed some regression type estimators using two auxiliary variables under two phase sampling schemes using SRSWOR at all the phases. Also, we have proposed two new classes of estimators for estimating the finite population mean. We have found that the class of estimator t_{c2} in its optimum case reduces to the estimator suggested by Mukerjee et. al. (1987).

Table -1: Description of the Population and their Sources

Pop No.	Source	y	x	z
1	Gujarati (2004), p.238-239.	Per capita consumption of chickens, lb	Real disposable income per capita, $	Real retail price of chicken per lb, ¢.
2	Gujarati (2004), p.238-239.	Per capita consumption of chickens, lb	Real retail price of pork per lb, ¢	Real retail price of chicken per lb, ¢
3	Gujarati (2004), p.238-239.	Billions of Drachmas at constant 1970 prices	Thousands of workers per year.	Amount Invested
4	Gujarati (2004), p.238-239.	Billions of Drachmas at constant 1970 prices	Thousands of workers per year	Capital to Labor Ratio
5	Chaterjee and Hadi (2006), p.55-56.	Overall rating of job being done by supervisor	Handles employee complaints	Raises based on performance
6	Chaterjee and Hadi (2006), p.55-56.	Overall rating of job being done by supervisor	Opportunity to learn new things	Raises based on performance
7	Chaterjee and Hadi (2006), p.55-56.	Scores in the Final	Scores in Second Preliminary	Scores in First Preliminary
8	Chaterjee and Hadi (1988), p.128-129.	Time (in seconds) in a one-mile run	Time (in seconds) in a I/4- mile ma1 run	Resting pulse rate per minute
9	Chaterjee and Hadi (1988), p.128-129.	Time (in seconds) in a one-mile run	Time (in seconds) in a I/4- mile ma1 run	Arm and leg strength
10	Chaterjee and Hadi (1988), p.207-208.	infant deaths per 1,000 live births	Number of inhabitants per physician	Gross national product per capita, 1957 U.S. dollars

Table-2: Value of Different Population Parameters

P. No.	N	${S_y}^2$	ρ_{yx}	ρ_{yz}	ρ_{xz}	${\rho_{yx.z}}^2$	${\rho_{y.xz}}^2$
1	23	54.360	0.947	0.932	0.840	0.835	0.911
2	23	54.360	0.912	0.970	0.840	0.741	0.867
3	27	1382.729	0.947	0.955	0.989	0.062	0.979
4	27	1382.729	0.947	0.997	0.943	0.295	0.898
5	30	148.171	0.825	0.669	0.590	0.718	0.684
6	30	148.171	0.624	0.640	0.590	0.396	0.451
7	22	124.338	0.927	0.884	0.896	0.652	0.886
8	30	4824.547	0.848	0.539	0.501	0.793	0.722
9	30	4542.547	0.848	0.400	0.445	0.816	0.732
10	49	1263.437	0.568	−0.484	−0.53	0.42	0.408

Table 3: Sample Sizes and MSE of Different Estimators under Two Phase Sampling

P. No.	n'	n	MSE_1	MSE_2	MSE_3	MSE_4	MSE_5	MSE_{C1}	MSE_{C2}
1	16	4	2.083	1.528	2.097	1.353	1.213	9.134	1.943
2	16	4	2.742	2.136	2.778	2.012	1.659	9.579	2.389
3	18	5	46.073	19.149	48.164	21.002	4.693	225.318	29.764
4	18	5	46.073	23.552	46.090	23.322	23.193	223.396	45.943
5	20	6	7.979	7.581	7.985	7.118	7.075	13.948	7.934
6	20	6	13.032	12.325	13.184	12.172	11.105	17.989	11.966
7	12	3	9.06	5.546	9.184	5.282	4.465	33.181	8.243
8	20	5	283.597	278.628	283.817	263.404	261.424	463.287	281.61
9	20	5	283.597	280.131	284.659	267.696	258.092	417.512	274.034
10	30	8	94.746	100.424	96.132	90.159	80.333	117.454	84.92

REFERENCES

[1] Bisht, K. K. S., Sisodia, B. V. S., Efficient estimators of mean of finite populations with known coefficient of variation, Journal of Indian Society of Agricultural Statistics, 42(1) (1990) 131-139.

[2] Bose, C., Note on the sampling error in the method of double sampling. Sankhya, The Indian Journal of Statistics, 6 (1943) 329-330.

[3] Chaiterjee, S. and Hadi, A.S., Sensitivity Analysis in Linear Regression, John Wiley & Sons, (1988).

[4] Chaterjee, S. and Hadi, A.S., Regression Analysis by Example, Fourth Edition, John Wiley & Sons, (2006).

[5] Cochban, W. G., Sampling Techniques, John Wiley and Sons, New York, (1963).

[6] Cochban, W. G., Sampling Techniques, John Wiley and Sons, New York, (1977).

[7] Gujarati, D.N., Basic Econometrics, Fourth Edition, McGraw-Hill Companies, (2004).

[8] Khan, M., A ratio chain-type exponential estimator for finite population mean using double sampling. Springer Plus, 5 (2016) 1–9.

[9] Khare, B.B., Srivastava, U., Kumar, K., A generalized chain ratio in regression estimator for population mean using two auxiliary characters in sample survey, Jour Sci Res Banaras Hindu Univ Varanasi, 57 (2013) 147–153.

[10] Kiregyera, B., A chain ratio-type estimator in finite population mean in double sampling using two auxiliary variables, Metrika, 27 (1980) 217–223.

[11] Kiregyera, B., Regression-type estimator using two auxiliary variables and model of double sampling from finite populations, Metrika, 31(1984) 215–223.

[12] Mukerjee, R., Rao, T. J. and Vijayan, K., Regression type estimators using multiple auxiliary information, Australian Journal of Statistics, 29(3) (1987) 244–254.

[13] Naik, V. D., Gupta, P. C., A general class of estimators for estimating population mean using auxiliary information. Metrika, 38 (1991) 11-17.

[14] Rao, P.S.R.S., Ratio and regression estimates with sub sampling the non-respondents, Paper presented at a special contributed session of the International Statistical Association Meeting, Sept., Tokyo, Japan, (1987) 2-16.

[15] Rawlings, J.O., Pantula, S.G. , Dickey, D. A., Applied Regression Analysis: A Research Tool, Second Edition, Springer-Verlag, (1998).

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

[17] Singh, G.N., Majhi, D., Some chain-type exponential estimators of population mean in two-phase sampling. Statistical Transitions, 15(2) (2014) 221–230.

[18] Singh, H. P., Tailor, R., Estimation of finite population mean with known coefficient of variation of an auxiliary character. Statistica, LXV(3) (2005) 407-418.

[19] Singh, H. P., Tailor, R., Estimation of finite population mean using known correlation coefficient between auxiliary characters. Statistica, LXV(3) (2005) 301- 313.

[20] Singh, H.P., Singh, S., Kim, J.M., General families of chain ratio type estimators of the population mean with known coefficient of variation of the second auxiliary variable in two phase sampling, Jour Korean Stat Soc, 35(4) (2006) 377–395.

[21] Swain, A.K.P.C., On Classes of Modified Ratio type and Regresion-cum-Ratio type estimators in Sample Surveys using Two Auxiliary Variables, Statistics in Transition-new series, 13(3) (2012) 473-494.

[22] Tamhane, A. C., Inference Based on Regression Estimator in Double Sampling, Biometrika, 65(2) (1978) 419-427.

[23] Upadhyaya, L. N., Singh, H. P., An estimator for population variance that utilizes the kurtosis of an auxiliary variable in sample surveys, Vikram Mathematical Journal, 19 (1999) 14-17.