A GENERALIZED CLASS OF REGRESSION TYPE ESTIMATORS IN TWO PHASE SAMPLING

Subhash Kumar Yadav[1], Sant Sharan Mishra[2] and Alok Kumar Shukla[3]

[1,2] Department of Mathematics & Statistics, R M L Avadh University, Faizabad, India, [3] Department of Statistics,
D. A. V. College, Kanpur, India
[1]e-mail: drskystats@gmail.com, [2] e-mail: sant_x2003@yahoo.co.in, [3] e-mail: alokshukladav@gmail.com

Abstract. *In the present investigation, a generalized class of linear regression model of estimators has been proposed for estimating the population mean and population total when auxiliary information is not available in survey sampling. The proposed model has lesser mean squared error as compared to ordinary regression method of estimation under two phase sampling scheme. The improvement of the previous estimator has been validated with the help of an empirical data under the aforesaid sampling scheme.*

Keywords: *Auxiliary variable, Mean Squared Error, Regression type estimators, two phase sampling.*

1. INTRODUCTION

The auxiliary information has been in use in sampling theory since the development of the theory and application of modern sample surveys. It is well established that the intelligent use of auxiliary information improves the efficiency of the sampling design by increasing the precision of the estimates. The auxiliary information has been used for the purposes of stratification in stratified sampling, measures of sizes in PPS (Probability Proportional to Size) sampling. In regression method of estimation, one uses auxiliary information so as to improve precision of the estimate of population parameters like population mean and population total etc.

Let y and x be the study and the auxiliary variables respectively. When the variable y under study and the auxiliary variable x is highly correlated and the line of regression of y on x passes through origin, the ratio and product type estimators are used to estimate the parameter under study. When y and x are highly positively correlated and the line of regression of y on x passes through origin, ratio estimator is find better to estimate the parameter under study while the product estimator is used when y and x are negatively correlated. When the regression line does not passes through origin or its neighborhoods, regression estimator is better one to use. In practice it has been seen that the regression line hardly passes through the neighborhoods of the origin, in this situation ratio estimator is not as good as regression estimator. So it is better to seek for some other improved regression type estimator having lesser mean squared error.

Let $U=(U_1,U_2,............,U_N)$ be the finite population of size N out of which a sample of size n is drawn with simple random sampling without replacement technique. Let y and x be the variable under study and the auxiliary variables respectively.Let $\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$ & $\bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$ be the population means of study and the auxiliary variables and $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ & $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ be the respective sample means. When $\bar{X}$ is not known, double sampling or two phase sampling is used to estimate the population mean of the study variable y. Under This sampling technique a large sample of size n' in first phase is taken with simple random sampling without replacement (srswor), without increasing the cost of the survey to estimate $\bar{X}$ and to estimate $\bar{Y}$, a sample of required size n in second phase is drawn.

The linear regression estimate of population mean of variable Y is defined as:

$$\bar{y}_{lr}=\bar{y}+b(\bar{X}-\bar{x}) \tag{1.1}$$

Where $\bar{y}$ *and* $\bar{x}$ are sample mean of variable Y and X respectively. $\bar{X}$ is population mean of auxiliary variable X and is supposed to be known. $\bar{y}_{lr}$ is linear regression estimate of population mean and b is a constant. Sukhatme *et.al.* (1984) have described in detail, the procedures for deriving estimates of population parameters along with their biases, mean square error etc.

But when $\bar{X}$ is unknown, the double sampling version of estimator (1.1) is defined as

$$\bar{y}_{lrd}=\bar{y}-b(\bar{x}-\bar{x}') \tag{1.2}$$

Where $\bar{x}'=\frac{1}{n'}\sum_{i=1}^{n'}x_i$ is the sample mean of auxiliary variable x based on sample of size n'.

The mean square error (MSE) of the estimator (1.2) is

$$MSE(\bar{y}_{lrd})=\lambda S_y^2+\hat{\beta}_1^2\lambda^* S_x^2-2\hat{\beta}_1\lambda^* S_{xy} \tag{1.3}$$

2. A PROPOSED MODEL

Following Misra et.al (2010), we are proposing a generalized class of linear regression type estimator under two phase sampling as

$$\bar{y}_{gld}=\bar{y}-\hat{\beta}_1(\bar{x}-\bar{x}')-\hat{\beta}_2(\bar{z}-\bar{z}') \tag{2.1}$$

Where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates of β_1 and β_2 respectively with $\bar{z}'$ as mean of Z based on sample of size n'. $Z = f(X)$ is a function of auxiliary variable X. When $Z = X^2$, it assumes the double sampling version of relationship considered by Ekpenyong et.al (2008), If $Z = \frac{1}{X}$, it takes the form of Misra et.al. (2009) and many more functions may be considered. It has been shown by Misra et.al (2009) that their estimators of population mean and total are more efficient as compared to estimators of Ekpenyong et.al (2008) and ordinary linear regression estimator. U is independently and identically distributed random variable with mean zero and fixed variance σ^2.

The estimator of population mean based on (2.1) is given by:

$$\bar{y}_{gld} = \bar{y} - \hat{\beta}_1(\bar{x} - \bar{x}') - \hat{\beta}_2(\bar{z} - \bar{z}') + u \qquad (2.2)$$

Where $\bar{y}_{gld}$ is a general regression type estimator of population mean based on proposed model defined in relation (2.1), $\bar{z}$ and $\bar{z}'$ are sample means of the variable Z based on samples of size n and n' respectively.

3. BIAS AND MEAN SQUARED ERROR OF $\bar{y}_{gld}$

To calculate the bias and MSE of $\bar{y}_{gld}$ estimator of population mean, we define.

Let $\bar{y} = \bar{Y}(1+e_0)$,

$\bar{x} = \bar{X}(1+e_1)$,

$\bar{x}' = \bar{X}(1+e_1^{'})$

$\hat{\beta}_1 = \beta_1(1+e_2)$,

$\hat{\beta}_2 = \beta_2(1+e_3)$,

$\bar{z} = \bar{Z}(1+e_4)$ and $\bar{z}' = \bar{Z}(1+e_4^{'})$

Such that $E(e_i) = E(e_i^{'}) = 0 \quad \forall\ i = 0,1,2,3,4$ and

$E(e_0^2) = \lambda C_y^2$, $E(e_1^2) = \lambda C_x^2$, $E(e_1^{'2}) = \lambda^{'} C_x^2$

$E(e_4^2) = \lambda C_z^2$, $E(e_4^{'2}) = \lambda^{'} C_z^2$ $E(e_0 e_1) = \lambda C_{xy}$,

$E(e_0 e_1^{'}) = \lambda^{'} C_{xy}$, $E(e_1 e_1^{'}) = \lambda^{'} C_x^2$, $E(e_0 e_4) = \lambda C_{yz}$,

$E(e_0 e_4^{'}) = \lambda^{'} C_{yz}$, $E(e_4 e_4^{'}) = \lambda^{'} C_z^2$, $E(e_1 e_4) = \lambda C_{zx}$,

$E(e_1 e_4^{'}) = \lambda^{'} C_{zx}$, $E(e_1^{'} e_4) = \lambda^{'} C_{zx}$, $E(e_1^{'} e_4^{'}) = \lambda^{'} C_{zx}$

Where $\lambda = \frac{1}{n} - \frac{1}{N}$ and $\lambda^{'} = \frac{1}{n'} - \frac{1}{N}$

Putting these values in equation (2.1), we get

$$\bar{y}_{gld} = \bar{Y}(1+e_0) - \beta_1(1+e_2)[\bar{X}(1+e_1) - \bar{X}(1+e_1^{'})] -$$
$$\beta_2(1+e_3)[\bar{Z}(1+e_4) - \bar{Z}(1+e_4^{'})]$$

$$\bar{y}_{gld} = \bar{Y}(1+e_0) - \beta_1(1+e_2)[\bar{X}(e_1 - e_1^{'})] -$$
$$\beta_2(1+e_3)[\bar{Z}(e_4 - e_4^{'})]$$

$$\bar{y}_{gld} = \bar{Y} + \bar{Y}e_0 - \beta_1\bar{X}(e_1 - e_1^{'} + e_1e_2 - e_1^{'}e_2) -$$
$$-\beta_2\bar{Z}(e_4 - e_4^{'} + e_3e_4 - e_3e_4^{'})$$

Taking expectation both sides, we get

$$E(\bar{y}_{gld} - \bar{Y}) = \bar{Y}(e_0) - \beta_1\bar{X}[E(e_1e_2) - E(e_1e_2^{'})] -$$
$$\beta_2\bar{Z}[E(e_3e_4) - E(e_3e_4^{'})]$$
$$= -\beta_1\bar{X}[E(e_1e_2) - E(e_1^{'}e_2)] - \beta_2\bar{Z}[E(e_3e_4) - E(e_3e_4^{'})]$$

$$Bias(\bar{y}_{gld}) = -[Cov(\bar{x}, \hat{\beta}_1) - Cov(\bar{x}', \hat{\beta}_1)] -$$
$$[Cov(\bar{z}, \hat{\beta}_2) - Cov(\bar{z}', \hat{\beta}_2)] \qquad (3.1)$$

which is negligible for large sample size. For large samples usually $Cov(\bar{x}, \hat{\beta}_1)$ and $Cov(\bar{x}', \hat{\beta}_1)$ decreases and it becomes zero if the joint distribution of y and x is bivariate normal.

Similarly $Cov(\bar{z}, \hat{\beta}_2)$ and $Cov(\bar{z}', \hat{\beta}_2)$ vanishes if the joint distribution of y and z follows bivariate normal distribution. In this case the proposed regression estimator is exactly unbiased.

To the first order of approximation, we have

$$\bar{y}_{gld} - \bar{Y} = e_0\bar{Y} - (e_1 - e_1^{'})\beta_1\bar{X} - (e_4 - e_4^{'})\beta_2\bar{Z}$$

Therefore

$$MSE(\bar{y}_{gld}) = E[e_0\bar{Y} - (e_1 - e_1^{'})\beta_1\bar{X} - (e_4 - e_4^{'})\beta_2\bar{Z}]^2$$
$$= E[e_0^2\bar{Y}^2 + \beta_1^2\bar{X}^2(e_1 - e_1^{'})^2 + \beta_2^2\bar{Z}^2(e_4 - e_4^{'})^2 -$$
$$2\beta_1\bar{X}\bar{Y}(e_0e_1 - e_0e_1^{'}) + 2\beta_2\overline{YZ}(e_0e_4 - e_0e_4^{'}) -$$
$$2\beta_1\beta_2\bar{Z}\bar{X}(e_1e_4 - e_1^{'}e_4 - e_1e_4^{'} + e_1^{'}e_4^{'})]$$

Applying expectation, putting different values of expectations and simplifying, we get

$$MSE(\bar{y}_{gld}) = [\lambda s_y^2 - 2\beta_1\lambda^* s_{xy} + \beta_1^2\lambda^* s_x^2 -$$
$$2\beta_2\lambda^* s_{yz} + \beta_2^2\lambda^* s_z^2 + 2\beta_1\beta_2\lambda^* s_{xz}] \qquad (3.2)$$

And we know that

$$MSE(\bar{y}_{lrd}) = \lambda S_y^2 + \hat{\beta}_1^2\lambda^* S_x^2 - 2\hat{\beta}_1\lambda^* S_{xy}$$

Where $\lambda^* = \frac{1}{n} - \frac{1}{n'} = \lambda - \lambda^{'}$ and s_{xy}, s_{yz} & s_{zx} are estimators of the population covariances, S_{XY}, S_{YZ} and S_{ZX} respectively, while the variances s_x^2, s_y^2 and s_z^2 are unbiased estimators of population variances S_X^2, S_Y^2 and S_Z^2 respectively.

Now it is required to estimate β_1 and β_2 in such a way that $MSE(\bar{y}_{gld})$ is a minimum. Using the method of ordinary least square, we differentiate partially (3.2) with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$ and obtain following normal equations.

$$\hat{\beta}_2 s_{zx} + \hat{\beta}_1 s_x^2 = s_{xy} \qquad (3.3)$$

$$\hat{\beta}_2 s_x^2 + \hat{\beta}_1 s_{zx} = s_{yz} \qquad (3.4)$$

Solving (3.3) and (3.4) simultaneously, we get

$$\hat{\beta}_1 = \frac{(s_{yz}s_{zx} - s_{yx}s_z^2)}{(s_{xz}^2 - s_x^2 s_z^2)} \qquad (3.5)$$

$$\hat{\beta}_2 = \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)} \qquad (3.6)$$

Thus for these values of $\hat{\beta}_1$ and $\hat{\beta}_2$, the $MSE(\bar{y}_{gld})$ will be minimum.

The estimate of population total (y_{gld}) and its variance using proposed estimator $\bar{y}_{gld}$, are as follows:

$$y_{gld} = N\,\bar{y}_{gld}$$

$$MSE(y_{gld}) = N^2 MSE(\bar{y}_{gld})$$

4. EFFICIENCY COMPARISON:

In view of (1.3) and (3.2), we have

$$MSE(\bar{y}_{lrd}) - MSE(\bar{y}_{gld}) = \lambda^* \left[2\hat{\beta}_2 s_{yz} - \hat{\beta}_2^2 s_z^2 - 2\hat{\beta}_1\hat{\beta}_2 s_{zx} \right]$$

Now

$$2\hat{\beta}_2 s_{yz} - \hat{\beta}_2^2 s_z^2 - 2\hat{\beta}_1\hat{\beta}_2 s_{zx} = \hat{\beta}_2(2s_{yz} - \hat{\beta}_2 s_z^2 - 2\hat{\beta}_1 s_{zx})$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[2s_{yz} - \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)} s_z^2 - 2\frac{(s_{yz}s_{zx} - s_{yx}s_z^2)}{(s_{xz}^2 - s_x^2 s_z^2)} s_{zx} \right]$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[2s_{yz}(s_{xz}^2 - s_x^2 s_z^2) - (s_{yx}s_{zx} - s_{yz}s_x^2)s_z^2 - 2(s_{yz}s_{zx} - s_{yx}s_z^2)s_{zx} \right]$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[2s_{yz}s_{xz}^2 - 2s_{yz}s_x^2 s_z^2 - s_{yx}s_{zx}s_z^2 + s_{yz}s_x^2 s_z^2 - 2s_{yz}s_{zx}s_{zx} + 2s_{yx}s_z^2 s_{zx} \right]$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[-s_{yz}s_x^2 s_z^2 + s_{yx}s_{zx}s_z^2 \right]$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)}{(s_{xz}^2 - s_x^2 s_z^2)^2}\left[s_{yx}s_{zx} - s_{yz}s_x^2 \right] s_z^2$$

$$= \frac{(s_{yx}s_{zx} - s_{yz}s_x^2)^2 s_z^2}{(s_{xz}^2 - s_x^2 s_z^2)^2} > 0$$

$$\Rightarrow \quad Var(\bar{y}_{lrd}) > Var(\bar{y}_{gld})$$

which shows that the estimator $\bar{y}_{gld}$ is more efficient than the estimator $\bar{y}_{lrd}$ as it has lesser mean squared error. This result has also been verified with the help of an empirical example.

5. NUMERICAL VALIDATION

For numerical validation, we have considered the data given in Des Raj (1972). The size of the population has been considered as 100 and a random sample without replacement of size 30 on first phase has been drawn from it and on second phase a sample of size 8 has been drawn from the sample of first phase.

Table no. 1

Estimates / Methods	Mean ($\bar{y}$)	$V(\bar{y})$	Total (y)	$V(y)$
Model ($z = 1/x^2$)	4.0353	0.1312	403.53	1312.00
Model ($z = 1/x$)	4.3075	0.1396	430.75	1396.00
Model ($z = \sqrt{x}$)	4.0367	0.1571	403.67	1571.00
Model ($z = x^2$)	4.3242	0.1682	432.42	1682.00
Linear Regression	4.2158	0.1858	421.58	1858.00

It is observed that estimates of population mean and population total obtained from $\bar{y}_{gld}$ are more efficient as compared to estimates of population mean and population total obtained from $\bar{y}_{lrd}$.

6. CONCLUSION

The precision of the estimates of population parameters such as population mean and population total etc are improved by including a linear term in ordinary linear regression estimator even in two phase sampling. The proposed model is in general form in which it includes the double sampling versions of models of Ekpenyong et al (2008) and also Misra et al (2009) as particular cases. It has been shown with the help of an example that the proposed model provides précised estimates of population mean and population total as compared to ordinary linear regression estimator of population mean and population total under double sampling.

The additional linear term i.e., z , which is a function of auxiliary variable x has been considered as $\frac{1}{x^2}$, $\frac{1}{x}$, $\sqrt{x}$,

x^2 in the present study and many more functions may be formed and theoretically we have shown that it is improved over ordinary linear regression estimator.

7. REFERENCES

[1] Cochran, W.G. (1999), *Sampling Technques*, John Wiley & Sons.

[2] Des Raj (1972), *The design of sampling surveys,* New York*:* McGraw-Hill.

[3] Ekpenyong, E. J., Okonnah, M.I. and John, E.D. (2008), *Polynomial (Nonlinear) Regression Method for Improved Estimation Based on Sampling,* Journal of Applied Sciences, vol. 8(8), pp.1597-1599.

[4] Misra, G.C., Shukla, A.K. and Yadav, S.K. (2009), *A Comparison of Regression Methods for Improved Estimation in Sampling*, Journal of Reliability and Statistical Studies, Vol. 2, Issue 2, pp. 85-90.

[5] Misra, G.C., Shukla, A.K. and Yadav, S.K. (2010), A class of regression type Estimators in survey sampling. Communicated to "*Statistics in transition-new Series*" Journal. BaNoCoSS conference papers.

[6] Sampath, S. (2005), *Sampling Theory and Methods*, Narosa Publishing House, India.

[7] Sukhatme,P.V., Sukhatme, B. V., Sukhatme S. and Asok, C. (1984), *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.