An Improved Regression Type Estimator of Finite Population Mean Using Information on Auxiliary Variable

Peeyush Misra
Department of Statistics, D.A.V.(P.G.) College, Dehradun- 248001, Uttarakhand (India)

N. K. Kamboj
Department of Mathematics, D I T University, Dehradun- 248009, Uttarakhand (India)

Abstract: The present paper considers the problem of estimating the finite population mean using auxiliary information. A regression type improved estimator using auxiliary information is proposed for the purpose. The bias and mean squared error expressions to the first degree of approximation are derived for the proposed estimator. Its comparative study with some of the well known estimators available in the literature is also carried out. An empirical study is also carried out to judge the efficiency of the proposed estimator over others.

Keywords: Auxiliary Variable, Bias, Mean Squared Error and Efficiency.

1. Introduction:

It is well known fact that the proper use of auxiliary information in sample surveys results in substantial improvement in the precision of the estimators of the population parameters. Using auxiliary information, it is possible to increase the efficiency of the usual estimators of population parameters of the study variable which are available in the literature. For details one may see Cochran (1977), Des Raj (1968), Mukhopadhyaya (2012), Murthy (1967), Singh & Chaudhary (1997) and Shukhatme, Sukhatme, Sukhatme & Asok (1984). The auxiliary information may be known in advance or it may be collected while the survey is going on without increasing the cost or less increased. The information so collected on the auxiliary character x may be used at the time of estimation or selection. Here we deal with using the auxiliary information in the form of known parameters of auxiliary character or in case of unknown parameters of auxiliary variable. Estimation procedures for the estimation of parameter mean are considered by so many authors and are available in the literature.

Let $(y_i, x_i), i = 1, 2..., n$ be the n pair of sample observations for the study variable and auxiliary variable respectively drawn from the population of size N using simple random sampling without replacement.

Let us denote by $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ be the population mean of study variable y, $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the population mean of auxiliary variable x.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the sample mean of y and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ be the sample mean of auxiliary variable X .

$S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$ be the population variance of study variable y and $S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$ be the population variance of auxiliary variable x .

$\rho = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{S_Y S_X}$ be the population correlation coefficient between y and x .

Also let $\mu_{rs} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^r\left(X_i - \bar{X}\right)^s$, $C_Y^2 = \frac{S_Y^2}{\bar{Y}^2}$, $C_X^2 = \frac{S_X^2}{\bar{X}^2} = \frac{\mu_{02}}{\bar{X}^2}$, $\rho = \frac{\mu_{11}}{S_Y S_X}$

$\beta_2 = \frac{\mu_{04}}{\mu_{02}^2}$, $\beta_1 = \frac{\mu_{03}^2}{\mu_{02}^3}$, $\gamma_1 = \sqrt{\beta_1}$ and $\beta = \frac{S_{YX}}{S_X^2}$.

For simplicity, it is assumed that N is large enough as compared to n so that finite population correction terms may be ignored. A new regression type improved estimator represented by $\hat{\bar{y}}$ for estimating the population mean is proposed as

$$\hat{\bar{y}} = \bar{y} + b\left(\bar{X} - \bar{x}\right) + k_1\left(\frac{s_x^2}{C_X^2} - \bar{x}^2\right) + k_2\left(\frac{s_y^2}{C_Y^2} - \bar{y}^2\right) \qquad (1.1)$$

where b is an estimate of the change in y when x is increased by unity.

2. Bias and Mean Square Error of the Proposed Estimator:

In order to obtain bias and mean square error of the proposed estimator, let us denote by

$$\bar{y} = \bar{Y}\left(1 + e_0\right)$$

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

$$s_{yx} = e_2 + S_{YX}$$

$$s_x^2 = e_3 + S_X^2$$

$$s_y^2 = e_4 + S_Y^2 \qquad (2.1)$$

so that ignoring finite population correction, for simplicity we have

$$E\left(e_0\right) = E\left(e_1\right) = E\left(e_2\right) = E\left(e_3\right) = E\left(e_4\right) = 0 \qquad (2.2)$$

$$E\left(e_0^2\right)=\frac{\mu_{20}}{n\bar{Y}^2}=\frac{1}{n}C_Y^2$$

$$E\left(e_1^2\right)=\frac{\mu_{02}}{n\bar{X}^2}=\frac{1}{n}C_X^2$$

$$E\left(e_3^2\right)=\left(\frac{\beta_2(x)-1}{n}\right)S_X^4=\frac{\mu_{02}^2}{n}\left(\frac{\mu_{04}}{\mu_{02}^2}-1\right)$$

$$E\left(e_4^2\right)=\left(\frac{\beta_2(y)-1}{n}\right)S_Y^4=\frac{\mu_{20}^2}{n}\left(\frac{\mu_{40}}{\mu_{20}^2}-1\right)$$

$$E\left(e_0e_1\right)=\frac{\mu_{11}}{n\bar{Y}\bar{X}}=\frac{1}{n}\rho C_Y C_X$$

$$E\left(e_0e_3\right)=\frac{\mu_{12}}{n\bar{Y}}$$

$$E\left(e_1e_2\right)=\frac{\mu_{12}}{n\bar{X}}$$

$$E\left(e_1e_3\right)=\frac{\mu_{03}}{n\bar{X}}$$

$$E\left(e_0e_4\right)=\frac{\mu_{30}}{n\bar{Y}}$$

$$E\left(e_1e_4\right)=\frac{\mu_{21}}{n\bar{X}}$$

$$E\left(e_3e_4\right)=\frac{1}{n}\left(\mu_{22}-\mu_{20}\mu_{02}\right) \qquad (2.3)$$

The proposed regression type estimator represented by $\hat{\bar{y}}$ for estimating the population mean given in (1.1) is

$$\hat{\bar{y}}=\bar{y}+b\left(\bar{X}-\bar{x}\right)+k_1\left(\frac{s_x^2}{C_X^2}-\bar{x}^2\right)+k_2\left(\frac{s_y^2}{C_Y^2}-\bar{y}^2\right) \qquad (2.4)$$

In terms of e_i's, $i=0,1,2,3,4$; the above proposed estimator up to terms of order $O(1/n)$ reduces to

$$\hat{\bar{y}}-\bar{Y}=\bar{Y}e_0-\beta\bar{X}e_1-2k_1\bar{X}^2e_1-2k_2\bar{Y}^2e_0+\frac{k_1\bar{X}^2e_3}{S_X^2}+\frac{k_2\bar{Y}^2e_4}{S_Y^2}-k_1\bar{X}^2e_1^2-k_2\bar{Y}^2e_0^2$$

$$+\frac{\beta\bar{X}e_1e_3}{S_X^2}-\frac{\bar{X}e_1e_2}{S_X^2} \qquad (2.5)$$

Taking expectation on both the sides of (2.5), we have bias in $\hat{\bar{y}}$ given by $E\left(\hat{\bar{y}}\right)-\bar{Y}$ up to terms of order $O(1/n)$ to be

$$\text{Bias}\left(\hat{\bar{y}}\right)=\left\{E\left(\hat{\bar{y}}\right)-\bar{Y}\right\}=-\frac{1}{n}\left\{\left(\mu_{02}k_1+\mu_{20}k_2\right)+\frac{1}{S_X^2}\left(\beta\mu_{03}+\mu_{12}\right)\right\} \qquad (2.6)$$

Now squaring both sides of (2.5) and taking expectation, we have mean square error of $\hat{\bar{y}}$ given by $\left\{E\left(\hat{\bar{y}}\right)-\bar{Y}\right\}^2$ up to terms of order $O(1/n)$ to be

$$\text{MSE}\left(\hat{\bar{y}}\right) = \left\{E\left(\hat{\bar{y}}\right)-\bar{Y}\right\}^2$$

$$=\bar{Y}^2E\left(e_0^2\right)+\beta^2\bar{X}^2E\left(e_1^2\right)+4k_1^2\bar{X}^4E\left(e_1^2\right)+4k_2^2\bar{Y}^4E\left(e_0^2\right)+\frac{k_1^2\bar{X}^4}{S_X^4}E\left(e_3^2\right)+\frac{k_2^2\bar{Y}^4}{S_Y^4}E\left(e_4^2\right)$$

$$-2\beta\bar{Y}\bar{X}E\left(e_0e_1\right)-4\bar{Y}\bar{X}^2k_1E\left(e_0e_1\right)-4\bar{Y}^3k_2E\left(e_0^2\right)+\frac{2k_1\bar{Y}\bar{X}^2}{S_X^2}E\left(e_0e_3\right)+\frac{2k_2\bar{Y}^3}{S_Y^2}E\left(e_0e_4\right)$$

$$+4\beta\bar{X}^3k_1E\left(e_1^2\right)+4\beta k_2\bar{Y}^2\bar{X}E\left(e_0e_1\right)-2\beta\frac{\bar{X}^3}{S_X^2}k_1E\left(e_1e_3\right)-2\beta\frac{\bar{Y}^2\bar{X}}{S_Y^2}k_2E\left(e_1e_4\right)$$

$$+8\bar{Y}^2\bar{X}^2k_1k_2E\left(e_0e_1\right)-4\frac{\bar{X}^4k_1^2}{S_X^2}E\left(e_1e_3\right)-4\frac{\bar{Y}^2\bar{X}^2}{S_Y^2}k_1k_2E\left(e_1e_4\right)-4\frac{\bar{Y}^2\bar{X}^2}{S_X^2}k_1k_2E\left(e_0e_3\right)$$

$$-4\frac{\bar{Y}^4}{S_Y^2}k_2^2E\left(e_0e_4\right)+2\frac{\bar{Y}^2\bar{X}^2}{S_Y^2S_X^2}k_1k_2E\left(e_3e_4\right)$$

using values of the expectation given in (2.2) and (2.3), we have

$$\text{MSE}\left(\hat{\bar{y}}\right)=\frac{1}{n}\left(\mu_{20}+\beta^2\mu_{02}-2\beta\mu_{11}\right)+\frac{1}{n}\left[\bar{X}^2\left\{\bar{X}^2\left\{\beta_2(x)-1\right\}-\frac{4\bar{X}}{S_X^2}\mu_{03}+4\mu_{02}\right\}k_1^2\right.$$

$$+\bar{Y}^2\left\{\bar{Y}^2\left\{\beta_2(y)-1\right\}-\frac{4\bar{Y}}{S_Y^2}\mu_{30}+4\mu_{20}\right\}k_2^2-2\bar{X}\left\{\frac{\beta\bar{X}\mu_{03}}{S_X^2}-\frac{\bar{X}\mu_{12}}{S_X^2}-2\beta\mu_{02}+2\mu_{11}\right\}k_1$$

$$\left.-2\bar{Y}\left\{\frac{\beta\bar{Y}\mu_{21}}{S_Y^2}-\frac{\bar{Y}}{S_Y^2}\mu_{30}-2\beta\mu_{11}+2\mu_{20}\right\}k_2+2\bar{Y}\bar{X}\left\{\frac{\bar{Y}\bar{X}}{S_Y^2S_X^2}\left(\mu_{22}-\mu_{20}\mu_{02}\right)-2\frac{\bar{X}}{S_X^2}\mu_{12}-2\frac{\bar{Y}}{S_Y^2}\mu_{21}+4\mu_{11}\right\}k_1k_2\right]$$

$$\text{MSE}\left(\hat{\bar{y}}\right)=\frac{1}{n}\left(\mu_{20}+\beta^2\mu_{02}-2\beta\mu_{11}\right)+\frac{1}{n}\left[\bar{X}^2\delta_1k_1^2+\bar{Y}^2\delta_2k_2^2-2\bar{X}\delta_3k_1+\right.$$

$$\left.-2\bar{Y}\delta_4k_2+2\bar{Y}\bar{X}\delta_5k_1k_2\right]. \qquad (2.7)$$

where $\delta_1=\left\{\bar{X}^2\left\{\beta_2(x)-1\right\}-\frac{4\bar{X}}{S_X^2}\mu_{03}+4\mu_{02}\right\}$

$$\delta_2=\left\{\bar{Y}^2\left\{\beta_2(y)-1\right\}-\frac{4\bar{Y}}{S_Y^2}\mu_{30}+4\mu_{20}\right\}$$

$$\delta_3 = \left\{ \frac{\beta \overline{X} \mu_{03}}{S_X^2} - \frac{\overline{X} \mu_{12}}{S_X^2} - 2\beta\mu_{02} + 2\mu_{11} \right\}$$

$$\delta_4 = \left\{ \frac{\beta \overline{Y} \mu_{21}}{S_Y^2} - \frac{\overline{Y}}{S_Y^2} \mu_{30} - 2\beta\mu_{11} + 2\mu_{20} \right\}$$

$$\delta_5 = \left\{ \frac{\overline{YX}}{S_Y^2 S_X^2} (\mu_{22} - \mu_{20}\mu_{02}) - 2\frac{\overline{X}}{S_X^2} \mu_{12} - 2\frac{\overline{Y}}{S_Y^2} \mu_{21} + 4\mu_{11} \right\}.$$

which attains the minimum for the optimum values

$$k_1 = \frac{(\delta_4\delta_5 - \delta_2\delta_3)}{\overline{X}(\delta_5^2 - \delta_1\delta_2)} \qquad (2.8)$$

$$k_2 = \frac{(\delta_3\delta_5 - \delta_1\delta_4)}{\overline{Y}(\delta_5^2 - \delta_1\delta_2)} \qquad (2.9)$$

Substituting the values of k_1 and k_2 given by (2.8) and (2.9) in (2.7), we get the minimum mean square error of $\hat{\bar{y}}$ to be

$$\mathrm{MSE}(\hat{\bar{y}}) = \frac{1}{n}(\mu_{20} + \beta^2\mu_{02} - 2\beta\mu_{11}) - \frac{1}{n}\left(\frac{\delta_2\delta_3^2 + \delta_1\delta_4^2 - 2\delta_3\delta_4\delta_5}{\delta_1\delta_2 - \delta_5^2} \right) \qquad (2.10)$$

3. Efficiency Comparison:

(i) General estimator of mean in case of SRSWOR:

The general estimator of Mean in case of SRSWOR is $\hat{\bar{y}}_{wor} = \bar{y}$ with $MSE(\hat{\bar{y}}_{wor}) = \frac{\mu_{20}}{n}$

(3.1)

It is clear that the proposed estimator is more efficient than the estimator $\hat{\bar{y}}_{wor}$ based on simple random sampling when no auxiliary information is used.

(ii) Usual regression estimator:

The usual regression estimator is $\bar{y}_{lr} = \bar{y} + b(\overline{X} - \bar{x})$ with

$$\mathrm{MSE}(\bar{y}_{lr})_{\min} = \frac{1}{n}(\mu_{20} + \beta^2\mu_{02} - 2\beta\mu_{11}) \qquad (3.2)$$

It is clear that the proposed estimator is more efficient than the usual regression estimator of mean where the auxiliary information already is in use.

4. Empirical Study:

To illustrate the performance of the proposed estimator, let us consider the following data **Population I:** Cochran (1977, Page Number- 181)

y : Paralytic Polio Cases ‘placebo’ group

x : Paralytic Polio Cases in not inoculated group

μ_{02} = 71.8650173, μ_{20} = 9.889273356, μ_{11} = 19.4349481, μ_{12} = 346.3174191,

μ_{03} = 1453.077703, μ_{40} = 424.1846721, μ_{21} = 94.21286383, μ_{22} = 3029.312542,

μ_{30} = 47.34479951, μ_{04} = 46132.5679, $\bar{y}$ = 2.588235294, $\bar{x}$ = 8.370588235,

S_x = 8.477323711, S_y = 3.144721507, ρ =0.729025009, $\beta_2(y)$ = 4.337367369,

$\beta_2(x)$ = 8.932490454, C_X = 1.012751251, C_Y = 1.215006037, β = 0.270436839,

n = 34.

$MSE(\hat{\bar{y}}_{wor})$ = 0.290860981, $MSE(\bar{y}_{lr})$ = 0.136274924 and $MSE(\hat{\bar{y}})\min$ = 0.106359844

PRE of the proposed estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 273.4687916.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\bar{y}_{lr}$ = 128.1262918.

Population II: Mukhopadhyay (2012, Page Number - 104)

y : Quality of raw materials (in lakhs of bales)

x : Number of labourers (in thousands)

μ_{02} = 9704.4475, μ_{20} = 90.95, μ_{11} = 612.725, μ_{12} = 93756.3475, μ_{03} = 988621.5173,

μ_{40} = 35456.4125, μ_{21} = 11087.635, μ_{22} = 2893630.349, μ_{30} = 1058.55, μ_{04} = 341222548.2,

$\bar{y}$ = 41.5, $\bar{x}$ = 441.95, S_x = 98.51115419, S_y = 9.536770942, ρ =0.652197067,

$\beta_2(y)$ = 4.286367314, $\beta_2(x)$ = 3.623231573, C_X = 0.22290113, C_Y =0.229801709,

β =0.063138576, n = 20.

$MSE(\hat{\bar{y}}_{wor})$ = 4.5475, $MSE(\bar{y}_{lr})$ = 2.613170788 and $MSE(\hat{\bar{y}})\min$ = 2.305558928.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 197.2406753.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\bar{y}_{lr}$ = 113.3421816.

Population III: Murthy (1967, Page Number - 398)

y : Number of absentees

x : Number of workers

μ_{02} = 1299.318551, μ_{20} = 42.13412655, μ_{11} = 154.6041103, μ_{12} = 5086.694392,

μ_{03} = 32025.12931, μ_{40} = 11608.18508, μ_{21} = 1328.325745, μ_{22} = 148328.4069,

μ_{30} = 425.9735118, μ_{04} = 4409987.245, $\bar{y}$ = 9.651162791, $\bar{x}$ = 79.46511628,

S_x = 36.04606151, S_y = 6.491080538, ρ = 0.660763765, $\beta_2(y)$ = 6.53877409,

$\beta_2(x)$= 2.612197776, C_X = 0.453608617, C_X = 0.672569791, β = 0.118988612 and n = 43.

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 0.979863408, $MSE\left(\bar{y}_{lr}\right)$= 0.552046468 and $MSE\left(\hat{\bar{y}}\right)$min = 0.528767534.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 185.3108115.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\bar{y}_{lr}$ = 104.4024893.

Population IV: Singh and Chaudhary (1997, Page Number - 176)

y : Total number of guava trees

x : Area under guava orchard (in acres)

μ_{02} = 12.50056686, μ_{20} = 187123.9172, μ_{11} = 1377.39858, μ_{12} = 4835.465464,

μ_{03} = 37.09863123, μ_{40} = 1.48935E+11, μ_{21} = 712662.4414, μ_{22} = 8747904.451,

μ_{30} = 100476814.5, μ_{04} = 540.1635491, $\bar{y}$ = 746.9230769, $\bar{x}$ = 5.661538462,

S_x = 3.535614072, S_y = 432.5782209, ρ = 0.900596235, $\beta_2(y)$= 4.253426603,

$\beta_2(x)$= 3.456733187, C_X = 0.624497051, C_Y = 0.579146949, β = 110.1868895, n = 13.

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 14394.14747, $MSE\left(\bar{y}_{lr}\right)$= 2719.434771 and $MSE\left(\hat{\bar{y}}\right)$min = 2344.570461.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 613.9353758.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\bar{y}_{lr}$ = 115.9886135.

Population V: Singh and Chaudhary (1997, Page Number: 154-155)

y : Number of milch animals in survey

x : Number of milch animals in census

μ_{02} = 431.5847751, μ_{20} = 270.9134948, μ_{11} = 247.3944637, μ_{12} = 3119.839406,

μ_{03} = 5789.778954, μ_{40} = 154027.4827, μ_{21} = 2422.297374, μ_{22} = 210594.3138,

μ_{30} = 2273.46265, μ_{04} = 508642.4447, $\bar{y}$ = 1133.294118, $\bar{x}$ = 1140.058824,

S_x = 20.77461853, S_y = 16.45945002, ρ = 0.723505104, $\beta_2(y)$= 2.098635139,

$\beta_2(x)$= 2.730740091, C_X = 0.018222409, C_Y = 0.014523547, β = 0.573223334, n = 17.

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 15.93609, $MSE\left(\bar{y}_{lr}\right)$= 7.594189 and $MSE\left(\hat{\bar{y}}\right)$min = 6.717366955.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\hat{\bar{y}}_{wor}$ = 237.2371204.

PRE of the proposed estimator $\hat{\bar{y}}$ over $\bar{y}_{lr}$ = 113.0530638.

5. Conclusions:

(i) From (2.10) it is clear that the proposed new regression type sampling estimator is more efficient than the estimator $\hat{\bar{y}}_{wor}$ based on simple random sampling when no auxiliary information is used and also more efficient than the usual regression estimator $\bar{y}_{lr}$ of mean where the auxiliary information already is in use.

(ii) From (2.8) and (2.9), the mean square error MSE $\left(\hat{\bar{y}}\right)$ of the estimator $\hat{\bar{y}}$ is minimized for the optimum values

$$k_1 = \frac{\left(\delta_4\delta_5 - \delta_2\delta_3\right)}{\bar{X}\left(\delta_5^2 - \delta_1\delta_2\right)} \tag{5.1}$$

$$k_2 = \frac{\left(\delta_3\delta_5 - \delta_1\delta_4\right)}{\bar{Y}\left(\delta_5^2 - \delta_1\delta_2\right)} \tag{5.2}$$

The optimum values involving some unknown parameters may not be known in advance for practical purposes; hence the alternative is to replace the unknown parameters of the optimum values by their unbiased estimators giving estimators depending upon estimated optimum values.

Acknowledgement: The authors are thankful to the referees and the editor in chief for their valuable suggestions regarding improvement of the paper.

References:

1. Cochran, W.G. (1977): Sampling Techniques, 3rd edition, John Wiley and Sons, New York.
2. Des Raj (1968): Sampling Theory, McGraw- Hill, New York.
3. Murthy, M (1967): Sampling Theory and Methods, 1st edition, Calcutta Statistical Publishing Society, Kolkata, India.
4. Mukhopadhyay, Parimal (2012): Theory and Methods of Survey and Sampling, 2nd edition, PHI Learning Private Limited, New Delhi, India.
5. Singh, Daroga and Chaudhary, F. S. (1997): Theory and Analysis of Sampling Survey Designs, New Age International Publishers, New Delhi, India.
6. Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. And Asok, C. (1984): Sampling Theory of Surveys with Applications, 3rd Edition, Ames, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi, India.