Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2582-2845.8487

ISSN: 2582 – 2845

Ind. J. Pure App. Biosci. (2020) *8*(6), 515-522

Indian Journal of Pure & Applied Biosciences

Peer-Reviewed, Refereed, Open Access Journal

Research Article

Improved Linear Regression Using Auxillary Informations

Shabnum Gul*, P.P. Singh and Nisar Ahmad Khan

J.S. University Shikohabad Firozabad (283135) U.P India

*Corresponding Author E-mail: khalidzara2018@gmail.com

Received: 15.11.2020 | Revised: 17.12.2020 | Accepted: 26.12.2020

ABSTRACT

The present study was taken under consideration in order to propose improved linear regression using auxiliary information's of coefficient of regression, coefficient of skewness, coefficient of variation in order to achieve more precision in estimates than the already existing estimators. The properties associated with the proposed estimators are assessed by mean square error and bias and compared with the existing estimators. In the support of the theoretical proposed work we have given numerical illustration.

Keywords: *Auxiliary Information, Linear regression, Mean square error, Bias, Efficiency.*

INTRODUCTION

A Regression type estimator using known coefficient of variation is considered and its properties are studied. There are several instances in physical, biological and agricultural sciences where the mean is Proportional to standard deviation and consequently the coefficient of variation is known although the mean and standard deviation may not be known. Some such situations may be seen in Snedecor (1946), Hald (1952), Davies and Goldsmith (1976) and Gleser and Healy (1976). The well-known Weber's law of psychophysics (see Guilford (1975), Chapter II provides instances where coefficient of variation is known and one such example is given in Singh (1998) also.

Sometimes simple a priori information in the form of coefficient of variation is available to the experimenters in the fields of biology, agriculture, psychophysics etc. Long association of the experimenters with the experimental information concerning the coefficient of variation. This information concerning coefficient of variation. This information concerning coefficient of variation is frequently used to plan experiments, estimate sample size, average, total, etc. (See Searles (1964) also. Further coefficient of variation may be seen in Cochran (1977, 3rd edition) on page 77 and Page 79 of Chapter 4. A good description about knowledge of coefficient of variation is given in Sukhatme et al. (1984) also on page 42.

The objective of the paper is to propose modified estimators for estimating the population mean by using the improved linear regression using auxiliary imformation with the coefficient of regression and coefficient of skewness of the auxiliary variables.

Cite this article: Gul, S., Singh, P. P., & Khan, N. A. (2020). Improved Linear Regression Using Auxillary Informations, *Ind. J. Pure App. Biosci. 8*(6), 515-522. doi: http://dx.doi.org/10.18782/2582-2845.8487

Notations Used

The following are the notations used in the paoer

N Population size

n Sample size

$f = n/N$ Sampling fraction

Y Study variable

X Auxiliary variable

$\overline{X}, \overline{Y}$ Population means

$\bar{x}, \bar{y}$ Sample means

x, y Sample totals

S_x, S_y Population standard deviations

S_{xy} Population covariance between variables

C_x, C_y Population coefficient of variation

ρ Population correlation coefficient

$B(.)$ Bias of the estimator

$MSE(.)$ Mean square error of the estimator

$\hat{\overline{Y}}_i$ Existing modified ratio estimator of $\overline{Y}$

$\hat{\overline{Y}}_{pj}$ Proposed modified ratio estimator of $\overline{Y}$

M_d Population median of X

β_2 Population kurtosis

β_1 Population skewness

$TM = \frac{Q_1 + 2Q_2 + Q_3}{4}$ Tri-Mean

$HL = median((X_j + X_k)/2, 1 \le j \le k \le N)$ Hodges-Lehmann estimator

$MR = \frac{X_{(1)} + X_{(N)}}{2}$ Population mid-range

Subscript

i For existing estimators j For proposed estimators

Procedure and Definitions

Let the Variable of interest be y and the auxiliary variable be x taking the values Y_1, and X_1 respectively for the ith (i = 1,2, ……. N) unit of the population of size N.

Further, Let

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad , \quad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \, , \quad \mu_{rs} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \bar{X})^{\mathrm{r}} \, (Y_i - \bar{Y})^{\mathrm{s}}$$

$$C_{\mathrm{y}} = \frac{\sigma_{\mathrm{y}}}{\overline{Y}} \, , \qquad C_{\mathrm{x}} = \frac{\sigma_{\mathrm{x}}}{\overline{X}}, \beta_2 = \frac{\mu_{04}}{\mu_{02}^2}, \qquad \gamma_1 = \frac{\mu_3}{\mu_{02}^{3/2}}$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 , \qquad \sigma_y^2 \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

$$\sigma_{\text{xy}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) \quad , \qquad \rho = \frac{\sigma_{\text{xy}}}{\sigma_{\text{x}}\sigma_{\text{y}}} , \qquad B = \frac{\sigma_{\text{xy}}}{\sigma_x^2}$$

Also, let

Where $\boldsymbol{y_1, y_2}, \ldots\ldots\ldots, \boldsymbol{y_n}$ are the observation on y and $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots\ldots\ldots, \boldsymbol{x_n}$ are the observations on auxiliary variable x for a simple random sample of size n.

For estimating the population mean using regression method of estimation, the proposed estimator is

$$\bar{y}_{lrh} = [\bar{y} + b(\bar{X} - \bar{x})] + \omega\left(\frac{S_y^2}{C_y^2\frac{G}{QD}} - \bar{y}^2\right)$$

Where ω is the characterizing scalar chosen suitably.

Comparison of Bias and Mean Square Error of $\bar{y}_{lrh}$ of the proposed estimator via-A-vis the competing estimator

For simplicity, it is assumed that the population size N is large enough as compared to the sample size n so that finite population correction term may be ignored.

Let

$$\bar{y} = \bar{Y}(1 + e_0) , \ \bar{x} = \bar{X}(1 + e_1), S_y^2 = \sigma_y^2(1 + e_2)$$

$$S_y^2 = \sigma_y^2(1 + e_3), \ S_{\text{xy}} = \sigma_{\text{xy}}(1 + e_4) \text{ so that}$$

$$E(e_0) = E(e_1) = E(e_2) = E(e_3) = E(e_4) = 0 \text{ and}$$

$$E(e_0^2) = \frac{\sigma_y^2}{n\bar{Y}^2} = \frac{C_y^2}{n} , \qquad E(e_1^2) = \frac{\sigma_x^2}{n\bar{X}^2} = \frac{C_x^2}{n} \quad , \qquad E(e_2^2) = \frac{\beta_2 - 1}{\text{n}}$$

$$E(e_0e_1) = \frac{\sigma_{\text{xy}}}{n\bar{X}\bar{Y}^2} = \frac{\rho C_{\text{x}}C_{\text{y}}}{n} , E(e_0e_2) = \frac{\mu_{03}}{n\sigma_y^2\bar{Y}}, E(e_1e_2) = \frac{\mu_{12}}{n\sigma_y^2\bar{X}}$$

$$E(e_1e_3) = \frac{\mu_{30}}{n\sigma_y^2\bar{X}}, E(e_1e_2) = \frac{\mu_{21}}{n\sigma_{\text{xy}}\bar{X}}$$

Also, we have

$$b = \frac{S_{\text{xy}}}{S_x^2} = \frac{\sigma_{\text{xy}}(1+e_4)}{\sigma_x^2(1+e_3)} = B(1 + e_4)(1 + e_3)^{-1}$$

$$= B(1 + e_4)(1 - e_3 + e_3^2 - \cdots)$$

$$= B(1 - e_3 + e_4 + e_3^2 - e_3e_4 + \cdots)$$

$$\bar{\boldsymbol{y}}_{\boldsymbol{lrh}} = \left[\bar{\boldsymbol{Y}}(\boldsymbol{1} + \boldsymbol{e_0}) + \boldsymbol{B}\left(\boldsymbol{1} - \boldsymbol{e_3} + \boldsymbol{e_4} + \boldsymbol{e_3^2} - \boldsymbol{e_3e_4} + \cdots\right)\{\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}(\boldsymbol{1} + \boldsymbol{e_0})\}\right] +$$

$$\boldsymbol{\omega}\left\{\frac{\overline{\boldsymbol{Y}}^{\boldsymbol{2}}\boldsymbol{\sigma}_{\boldsymbol{x}}^{\boldsymbol{2}}(\mathbf{1}+\boldsymbol{e}_{\mathbf{2}})}{\boldsymbol{\sigma}_{\boldsymbol{x}}^{\boldsymbol{2}}}-\overline{\boldsymbol{Y}}^{\boldsymbol{2}}(\mathbf{1}+\boldsymbol{e}_{\mathbf{0}})^{\mathbf{2}}\right\}$$

$$=\overline{Y}+\bar{Y}_{e_0}+B(1-e_3\ +e_4+e_3^2-e_3e_4+\cdots)\left(-\bar{X}_{e_2}\right)+\omega\overline{Y}^2\{(1+e_2)-(1+e_0)^2\}$$

Or

$$\bar{y}_{lrh}-\overline{Y}=\left[\bar{Y}_{e_0}-B\bar{X}_{e_1}-B\bar{X}_{e_1e_4}+B\bar{X}_{e_1e_3}+\cdots\right]+\omega\overline{Y}^2\{e_2-e_0^2-2e_0\}$$

Taking expectations on both sides, we have bias up to term of order O (1/n) as follows,

$$\text{Bias }(\bar{y}_{lre})=E(\bar{y}_{lr\omega}-\overline{Y})$$

$$\overline{Y}E(e_0)-B\bar{X}E(e_1)-B\bar{X}E(e_1e_4)+B\bar{X}E(e_1e_3)=\omega\bar{Y}\{E(e_2^2)-E(e_0^2)-3E(e_0e_2)-2E(e_0)\}$$

$$\rho=\frac{\sigma_{\mathrm{y}}\mu_{21}}{n\sigma_{\mathrm{x}}\mu_{11}}+\rho\gamma_1\frac{\sigma_{\mathrm{y}}}{n}-\omega\frac{\sigma_{\mathrm{y}}}{n}$$

MSE $(\bar{y}_{lrh})=\left[\left\{\bar{Y}_{e_0}-B\bar{X}_{e_1}\right\}+\omega\overline{Y}^2(e_2-2e_0)\right]^2$ value of ω minimizing the men square error of $\bar{y}_{lrh}$

And the minimum mean square error is given by:

$$\text{MSE}(\bar{y}_{lr\omega})=(1-\rho^2)\frac{\sigma_y^2}{\mathrm{n}}-\frac{\left\{\gamma_1C_y\times\frac{G}{QD}-2C_y^2\times\frac{G}{QD}-\rho\frac{\mu_{12}}{\bar{Y}\sigma_x\sigma_y}+2\rho^2C_y^2\times\frac{G}{QD}\right\}}{\bar{Y}(\beta_2-1)+4C_y^2\times\frac{G}{QD}-4\gamma_1C_y\times\frac{G}{QD}}$$

Estimator based on estimated optimum $\hat{C}$.

If the exact or good guess of β_2,γ_2,ρ and μ_{12} are not available, we can replace these quantities by their consistent sample estimates $\widehat{\beta}_2,\widehat{\gamma}_2,\hat{\rho},\widehat{\mu}_{12}$ respectively and $\widehat{\overline{Y}}=\bar{y}$ in (4.2.4) and get the estimated optimum value of ω denoted by $\hat{\mathrm{c}}$ as.

$$\hat{\mathrm{c}}=-\frac{\left\{\widehat{\gamma}_1C_y\times\frac{G}{QD}-2C_y^2\times\frac{G}{QD}-\hat{\rho}\frac{\widehat{\mu}_{12}}{\bar{Y}s_xs_y}+2\hat{\rho}^2C_y^2\times\frac{G}{QD}\right\}}{\bar{y}\left[(\hat{\beta}_2-1)+4C_y^2\times\frac{G}{QD}-4\widehat{\gamma}_1C_y\times\frac{G}{QD}\right]}$$

$$=-\frac{\left\{\frac{\widehat{\mu}_{03}}{s_y^2}C_y\times\frac{G}{QD}-2C_y^2\times\frac{G}{QD}-\hat{\rho}\frac{\widehat{\mu}_{12}}{\bar{Y}s_xs_y}+2\hat{\rho}^2C_y^2\times\frac{G}{QD}\right\}}{\bar{y}\left[\left(\frac{\widehat{\mu}_{04}}{s_y^2}-1\right)+4C_y^2\times\frac{G}{QD}-4\frac{\widehat{\mu}_{03}}{s_y^2}C_y\times\frac{G}{QD}\right]}$$

Where

$$\beta_2=\frac{\widehat{\mu}_{04}}{s_y^4}\ with\ \widehat{\mu}_{04}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^4,$$

$$\widehat{\gamma}_1=\frac{\widehat{\mu}_{03}}{s_y^3}\ with\ \widehat{\mu}_{03}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3,$$

$$\hat{\rho} = \frac{s_{xy}}{s_x s_y}, \qquad \widehat{\mu}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})^2,$$

we get the estimator based on the estimated optimum $\hat{c}$ as.

$$\bar{y}_{lre} = [\bar{y} + b(\bar{X} - \bar{x})] + \hat{c}\left(\bar{y} - \frac{C_y^2 \times \frac{G}{QD}}{S_y^2}\right)$$

Let,

$$\hat{\mu}_{03} = \mu_{03}(1 + e_5), \quad \hat{\mu}_{04} = \mu_{04}(1 + e_6), \; \hat{\mu}_{12} = \mu_{12}(1 + e_7)$$

Also, we have

$$\hat{c} = \frac{\left\{\begin{array}{c}\frac{\widehat{\mu}_{03}((1+e_5)}{\sigma_y^2(1+e_2)^{\frac{3}{2}}} C_y \times \frac{G}{QD} - 2C_y^2 \times \frac{G}{QD} - \frac{\sigma_{xy}^2(1+e_4)\mu_{12}(1+e_7)}{\bar{Y}\sigma_x^2\sigma_x^2(1+e_3)(1+e_2)(1+e_0)} \\ +2C_y^2 \times \frac{G}{QD} - \frac{\sigma_{xy}^2(1+e_4)^2}{\sigma_x^2\sigma_x^2(1+e_3)(1+e_2)} + \end{array}\right\}}{\bar{Y}(1+e_0)\left\{\frac{\mu_{04}((1+e_6)}{\sigma_y^4(1+e_2)^2} - 1 + 4C_y^2 \times \frac{G}{QD} - 4\frac{\mu_{03}(1+e_5)}{\sigma_y^3(1+e_2)^{3/2}} C_y \times \frac{G}{QD}\right\}}$$

$$\bar{y}_{lre} - \overline{Y} = \left(\bar{Y}_{e_0} - B\bar{X}_{e_2}\right) - \frac{\left\{\gamma_1 C_y \times \frac{G}{QD} - 2C_y^2 \times \frac{G}{QD} - \rho\frac{\mu_{12}}{\sigma_x\sigma_y\overline{Y}} + 2\rho^2 C_y^2 \times \frac{G}{QD}\right\}}{\left\{(\beta_2 - 1) - 4C_y^2 \times \frac{G}{QD} - 4\gamma_1 C_y \times \frac{G}{QD}\right\}}(e_2 - 2e_0 + \cdots)$$

An Illustration

We observe that the conditions discussed in the introduction for known coefficient of variation are satisfied for the data given in Walpoole, R. E., Myres, R. H., Myres, S. L., and Ye, K. (2005, page 473) dealing with measure of aerobic fitness is the oxygen consumption in volume per unit body weight per unit time. Thirty-one individuals were used in an experiment in order to be able to model oxygen consumption (y) against time to run one and half miles (x). consumption of required values has been done and we have the following.

For the population I and II we use the Data Sets data of Singh and Chaudhary (1986) page 177 and for the population III we use the data of Murthy (1967) page 228 in which fixed capital is denoted by X (auxiliary variable) and output of 80 factories are denoted by Y (study variable). For the population IV, the data is of cultivation and production of apple in district Baramulla of Kashmir (Jammu and Kashmir) in which the apple production (in tons) is denoted by *Y* (study variable) and number of apple trees are denoted by *X* (auxiliary variable, 1unit = 100 trees) in 117 villages of the Baramulla region of Jammu and Kashmir in 2010-2011 (Source: RC Mproject, pilot survey for estimation of cultivation andproduction of apple in district Baramulla, RC Mapproved project).

Table 1: Characteristics of these populations

Parameters	Population 3	Population 4
N	80	117
n	20	15
$\overline{Y}$	5182.637	2179
$\overline{X}$	1126.463	560.0
ρ	0.941	0.991703

S_y	1835.659	862
C_y	0.354193	0.9728
S_x	845.610	235.5
C_x	0.7506772	0.7395
β_2	-0.063386	1.10
β_1	1.050002	0.20
MR	1795.5	550.5
HL	1040.5	500.6
QD	588.125	200.45
G	901.081	205.142
D	801.381	150.600
S_{pw}	791.364	98.67
DM	1150.7	594.465

$$G = 155.446,\ QD = 80,\ \text{n} = 20, \beta_2 = 5.31, \sigma_y^2 = 18.1206, \text{N}=80$$

$$\sigma_y^2 = 18.1206, \quad \sigma_x = 2.759, \quad \bar{y} = 2.8513, \bar{x} = 51.8264,$$

$$\bar{x} = 51.8264, \quad C_y = 0.8561, \quad \rho = 0.4491, \mu_{12} = 2.35772, \quad \gamma = 21$$

For the purpose of empirical investigation, we have considered the following three sets of data which are taken from various sources followed by singh and chaudary (1986).

DATA 1:[$\boldsymbol{Source}$: $\boldsymbol{Kadilar\ and\ Cingi}$]

The data is consisted of 80 villages in the Murthy (1967).The variables of interest are as:

Y: The level of apple production (in 100 tones)

and

X: The number of apple trees.

For this data, we have

$N = 104, n = 20, \theta = 0.04038, \bar{Y} = 6.254, \bar{X}$ =13931.683,

$S_y = 11.67, S_x = 23029.072, \beta_2(y) = 16.523, \beta_2(x) = 17.516, \delta_{22} = 14.398, \lambda^k = 0.8112.$

DATA 2: [$\boldsymbol{Source}$: $\boldsymbol{Das}$]

The data is consisted of 117 villages/towns/wards village of Baramulla in (2010-2011) The variables of interest are as:

Y: The number of agricultural labourers

and

X: The number of agricultural labourers

For this data, we have

$N = 278, n = 30, \theta = 0.02974, \bar{Y}$ =39.068, $\bar{X} = 25.111,$

$S_y = 56.457167, S_x = 40.674797, \beta_2(y) = 25.8969, \beta_2(x) = 38.8898, \delta_{22} = 26.8142, \lambda^k =$ 0.6812.

DATA 3: [$\boldsymbol{Source}$: $\boldsymbol{Cochran}$ [$\boldsymbol{1, p. 325}$]]

The data is consisted of 100 blocks in a large city. The variables of interest are as:

Y: The number of persons per block

and

X: The number of rooms per block.

For this data, we have

$N = 100, n = 10, \theta = 0.09, \bar{Y} = 101.1,\ \bar{X} = 58.8, S_y = 14.6595, S_x = 7.53228, \beta_2(y) = 2.3523, \beta_2(x) = 2.2387, \delta_{22} = 1.5432, \lambda^k = 0.4385.$

We provide the Percentage Gain in Efficiency of the proposed estimator with respect to its compititors in the following Table 1.

Table 1: Percentage gain in efficiency of the proposed estimator $s_R^{2(k)}$ with respect to s_R^2 and s_y^2.

Data set	Percentage gain in efficiency of $s_R^{2(k)}$ with respect to s_y^2	Percentage gain in efficiency of $s_R^{2(k)}$ with respect to s_R^2
Data 1	260.8	19.6
Data 2	286.78	79.8
Data 3	34.67	67,8

The above table well indicates the supremacy of the proposed estimators $s_R^{2(k)}$ over s_R^2 and s_y^2.

CONCLUSION

Concluding Remarks

a) for the optimum value of ω, the estimator $\bar{y}_{lr\omega}$ attains the minimum mean square error given by

$$\textbf{MSE}\ (\bar{y}_{lre}) = \frac{\sigma_y^2}{\mathbf{n}}(\mathbf{1}-\boldsymbol{\rho}^{\mathbf{2}})\frac{\bar{Y}^2\left\{\gamma_1 C_y \times \frac{G}{QD} - 2C_y^2 \times \frac{G}{QD} - \rho\frac{\mu_{12}}{\sigma_x \sigma_{\bar{y}}} + 2\rho^2 C_y^2 \times \frac{G}{QD}\right\}^2}{n\left\{(\beta_2 - 1) - 4C_y^2 \times \frac{G}{QD} - 4\gamma_1 C_y \times \frac{G}{QD}\right\}} \quad (4.4.2)$$

b) for the optimum value of ωdepending upon estimated optimumĉ has the mean square error.

$$\text{MSE}\ (\bar{y}_{lre}) = \frac{\sigma_y^2}{\mathbf{n}}(1-\rho^2)\frac{\bar{Y}^2\left\{\gamma_1 C_y \times \frac{G}{QD} - 2C_y^2 \times \frac{G}{QD} - \rho\frac{\mu_{12}}{\sigma_x \sigma_{\bar{y}}} + 2\rho^2 C_y^2 \times \frac{G}{QD}\right\}^2}{n\left\{(\beta_2 - 1) - 4C_y^2 \times \frac{G}{QD} - 4\gamma_1 C_y \times \frac{G}{QD}\right\}}$$

From (4.4.3), we see that the estimator $\bar{y}_{lre}$ depending on estimated optimum value is always more efficient than the usual linear regression estimator $\bar{y}_{lr} = \bar{y} - b(\bar{X} - \bar{x})$ in the sense having lesser mean square error.

The use of proposed estimator is limited for the situations when coefficient of variation is known. However, in case of unknown coefficient of variations its estimated value may be used after studying the performance of the estimator (robustness) against different values of CV, if the guess is in error say 5%, 10%, 15%. 20%, 25%, 50%. Further work is being done in this direction.

REFERENCES

Agrawal, M. C., & Panda, K. B. (1999). A predictive justification for variance estimation using auxiliary information. *Jour. Ind. Ag. Statistics*, *52*(2), 192-200.

Basu, D. (1971). An essay on the logical foundations of statistical inference, Part I, Foundations of Statistical Inference, Ed. By Godambe, V. P. & Sportt, D. A. New York, 203-233.

Cochran, W. G. (1977). Sampling Techniques, 3rdedn. (Wiley & Sons).

Das, A. K. (1988). Contribution to the theory of sampling strategies based on auxiliary information (Ph.D. thesis

submitted to Bidhan Chandra Krishi Vishwa vidyalaya, Mohanpur, Nadia, West Bengal, India).

Kadilar, C., & Cingi, H. (2006). Improvement in variance estimation using auxiliary information, *Hacettepe Journal of Mathematics and Statistics*, *35*(1), 111-115.

Panda, K. B., & Sahoo, N. (2015). Systems of exponential ratio-based and exponential product-based estimators with their efficiency. *ISOR Journal of Mathematics*, *11*(3), Ver. I (May-June), PP 73-77.

Panda, K. B., & Das, P. (2018). Efficient hierarchic multivariate product-based estimator. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, *5*(2), pp. 65-69.

Panda, K. B., & Das, P. (2018). Efficient hierarchic predictive multivariate product estimator based on harmonic mean. *International Journal of Mathematics Trends and Technoloy (IJMTT) – 56*(6), pp. 14-18.

Gupta, Sat & Shabbir, Javid (2008). Variance estimation in simple random sampling using auxiliary information. *Hacettepe Journal of Mathematics and Statistics*, *37*(1), 57-67.