Asian Journal of Statistical Sciences
Vol. 2, No-2, (2022), pp. 141-149
© ARF India, Gurgaon, India

ARF INDIA
Academic Open Access Publishing
Journal homepage link:
http://arfjournals.com/ajss

Efficient Hierarchic Predictive Weighted Linear Regression Estimator with Efficiency

K.B. Panda [a] and P.P. Mohanty [b]

[a] *Dept. of Statistics, Central University of Jharkhand, Ranchi, Jharkhand* [b]*Dept. of Statistics, Utkal University, Bhubaneswar, Odisha*

ARTICLE HISTORY
Compiled March 25, 2023

Received 03 December 2022; Accepted 06 February 2023

ABSTRACT
We have, in this paper, developed a sequence of weighted linear regression estimators. The proposed weighted linear regression estimator of order k, besides being endowed with the predictive character, is found to be more efficient than the simple mean estimator in one hand and the weighted linear regression estimator on the other under optimality of k. Based on the theoretical developments, empirical illustrations involving real-population data have been considered.

KEYWORDS
Hierarchic estimation; predictive estimation; weighted linear regression estimator; bias; mean square error

1. Introduction

When the study variable y is positively correlated with the auxiliary variable x and the regression line of y on x passes through the origin, then ratio estimator is used to estimate the population mean $\overline{Y}$ or the population total Y provided complete information is available for the auxiliary variable. But, if the study variable y is negatively correlated with the auxiliary variable x, then product estimator is used to estimate the population mean $\overline{Y}$ or the population total Y. When the regression line of y on x is linear but the regression line does not pass through the origin, then linear regression estimator is more appropriate than either the ratio or the product estimator from the standpoint of efficiency. The regression estimator is originated from the difference estimator given by

$$\overline{y}_d = \overline{y} + \beta(\overline{X} - \overline{x}), \tag{1}$$

where $\overline{y}$ is the sample mean of y - variable and $\overline{X}$ and $\overline{x}$ are, respectively, the population mean and the sample mean of x - variable and β is a preassigned constant. The estimator $\overline{y}_d$ is unbiased for the population mean $\overline{Y}$ and it will attain its minimum

CONTACT K.B. Panda[a]. Email:kbpanda62@yahoo.co.in variance given by

$$V(\overline{y}_d) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \left(1 - \rho_{yx}^2\right) \tag{2}$$

when β coincides with the population regression coefficient of y on x.

Usually, β, the population quantity is unknown and is estimated by its corresponding sample quantity b_{yx}, the sample regression coefficient. In the circumstances, the usual linear regression estimator given by

$$\overline{y}_{lr} = \overline{y} + b_{yx}(\overline{X} - \overline{x}) \tag{3}$$

is attained.

Its bias and mean square error, to the first degree of approximation, i.e., to o(n^{-1}) have been expressed, respectively, as

$$B(\overline{y}_{lr}) = \frac{N(N-n)}{(N-1)(N-2)} \frac{\beta_{yx}}{n} \left[\frac{\mu_{300}}{S_{yx}} - \frac{\mu_{210}}{S_x^2}\right] \tag{4}$$

and

$$M(\overline{y}_d) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \left(1 - \rho_{yx}^2\right), \tag{5}$$

where $\mu_{pqr} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^p(y_i - \overline{Y})^q(z_i - \overline{Z})^r$, S_y^2 is the population mean square of y and ρ_{yx} is the population correlation coefficient between y and x.

Following Agrawal and Jain (1989), Sahoo et. al. (2007) have proposed a new linear regression estimator by defining $z = x^{-1}(x > 0)$ as a transformed auxiliary variable, which is given by

$$\overline{y}_{lr}^{*} = \overline{y} + b_{yz}(\overline{Z} - \overline{z}), \tag{6}$$

its bias and mean square error, to o$(n^{-1}$), being

$$B(\overline{y}_{lr}^{*}) = \frac{N(N-n)}{(N-1)(N-2)} \frac{\beta_{yz}}{n} \left[\frac{\mu_{003}}{S_{yz}} - \frac{\mu_{012}}{S_z^2}\right] \tag{7}$$

and

$$M(\overline{y}_{lr}^{*}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \left(1 - \rho_{yz}^2\right), \tag{8}$$

where ρ_{yz} is the population correlation coefficient between y and $z = x^{-1}(x > 0)$.

Again, combining these two regression estimators Panda and Chattapadhyay (2022) have considered a new weighted linear regression estimator, given by

$$\overline{y}_{wlr} = w_1 \overline{y}_{lr} + w_2 \overline{y}^*_{lr}$$

$$\Rightarrow \overline{y}_{wlr} = \overline{y} + w_1 b_{yx} \left(\overline{X} - \overline{x} \right) + w_2 b_{yz} \left(\overline{Z} - \overline{z} \right), \tag{9}$$

where w_1 and w_1 are weights such that $w_1 + w_2 = 1$ and $x > 0$.

The estimator is biased and its bias and mean square error, to o(n^{-1}), are given, respectively, by

$$B(\overline{y}_{wlr}) = \frac{N(N-n)}{n(N-1)(N-2} \left[w_1 \beta_{yx} \left(\frac{\mu_{300}}{S_{yx}} - \frac{\mu_{210}}{S_x^2} \right) + w_2 \beta_{yz} \left(\frac{\mu_{003}}{S_{yz}} - \frac{\mu_{012}}{S_z^2} \right) \right] \tag{10}$$

and

$$M(\overline{y}_{wlr}) = \left(\frac{1}{n} - \frac{1}{N} \right) \left[1 + w_1^2 \rho_{yx}^2 + w_2^2 \rho_{yz}^2 - 2w_1 \rho_{yx}^2 - 2w_2 \rho_{yz}^2 + 2w_1 w_2 \rho_{yx} \rho_{yz} \rho_{xz} \right], \tag{11}$$

where ρ_{xz} is the population correlation coefficient between x and z.

Minimization of (11) subject to variations in w's yields

$$w_{1\ opt} = \frac{\rho_{yx}^2 - \rho_{yx}\rho_{yz}\rho_{xz}}{\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}} = \frac{A}{A+B} = 1 - w_{1\ opt}, \tag{12}$$

where $A = \rho_{yx}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}$ and $B = \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}$.

This proposed estimator performs better than the usual linear regression estimator and the estimator due to Sahoo et. al. under certain conditions.

In this paper, invoking the predictive approach due to Basu(1971) followed by Smith(1976) for a fixed population set-up and then, with recursive use of this intuitive predictive format coupled with the technique due to Agarwal and Sthapit (1997), we develop a hierarchic predictive weighted linear regression estimator which under certain condition performs better than the customary weighted linear regression estimator. Numerical investigations have been carried out to illustrate the application of the work proposed here.

2. Predictive character of the proposed estimator

Under the predictive set-up, the population total Y is expressed as

$$Y = \sum_{i \in s} y_i + \sum_{i \in \overline{s}} y_i, \tag{13}$$

where s denotes the sample and $\overline{s}$ is its complement. To estimate the population total Y, we have to predict the second component of the right-hand side of equation (2.1), which is unknown.

The usual predictive format for estimating Y, the population total, is

$$\widehat{Y} = \sum_{i \in s} y_i + \sum_{i \in \overline{s}} \widehat{y_i}, \tag{14}$$

where $\widehat{y_i}$ is the implied predictor of y_i,$(i \in \overline{s})$.

Thus,

$$\begin{aligned}
\widehat{\overline{Y}} &= \sum_{i \in s} y_i + \sum_{i \in \overline{s}} \widehat{\overline{y}}_i \\
&= \frac{n}{N}\overline{y} + \frac{1}{N} \sum_{i \in \overline{s}} \left[\overline{y} + w_1 b_{yx}\left(x_i - \overline{x}\right) + w_2 b_{yz}\left(z_i - \overline{z}\right)\right] \\
&= \overline{y} + \frac{1}{N} + \left[w_1 b_{yx}\left\{\left(N\overline{X} - n\overline{x}\right) - \left(N - n\right)\overline{x}\right\} + w_2 b_{yz}\left\{\left(N\overline{Z} - n\overline{z}\right) - \left(N - n\right)\overline{z}\right\}\right] \\
&= \overline{y} + w_1 b_{yx}\left(\overline{X} - \overline{x}\right) + w_2 b_{yz}\left(\overline{Z} - \overline{z}\right) \\
&\Rightarrow \widehat{\overline{Y}} = \overline{y}_{wlr}
\end{aligned}$$

So, $\overline{y}_{wlr}$ is predictive in form.

3. A sequence of predictive weighted linear regression estimators and their performance

Using weighted linear regression estimator $\overline{y}_{wlr}$ as an intuitive predictor of y_i,$(i \in \overline{s})$, we reach

$$\widehat{Y} = \sum_{i \in s} y_i + (N - n)\overline{y}_{wlr}$$

or,

$$\widehat{\overline{Y}} = \frac{1}{N} \sum_{i \in s} y_i + \frac{1}{N}(N - n)\overline{y}_{wlr} = \overline{y}_{wlr}{}^{(1)},\ \textit{say},$$

where

$$\overline{y}_{wlr}{}^{(1)} = \emptyset_1 \overline{z}_{wlr} + \overline{y}_{wlr}$$

with

$$\emptyset_1 = 1 + \lambda\emptyset_0,\ \emptyset_0 = 0,\ \lambda = 1 - \frac{n}{N} \tag{15}$$

and

$$\overline{z}_{wlr} = -\frac{n}{N}\left\{\overline{y} + w_1 b_{yx}\left(\overline{X} - \overline{x}\right) + w_2 b_{yz}\left(\overline{Z} - \overline{z}\right)\right\}.$$

A second iteration with $\overline{y}_{wlr}{}^{(1)}$ as an intuitive predictor of y_i,($i \in \overline{s}$), in (14) would culminate in $\overline{y}_{wlr}{}^{(2)}$ given by

$$\overline{y}_{wlr}{}^{(2)} = \emptyset_2 \overline{z}_{wlr} + \overline{y}_{wlr},$$

where $\emptyset_2 = 1 + \lambda\emptyset_1$.
Continuing in this way, we would, at the k^{th} iteration, obtain as

$$\overline{y}_{wlr}{}^{(k)} = \emptyset_k \overline{z}_{wlr} + \overline{y}_{wlr},$$

where $\emptyset_k = 1 + \lambda\emptyset_{k-1} = \frac{1-\lambda^k}{1-\lambda}$.

Thus, $\overline{y}_{wlr}{}^{(k)}$ can also be expressed as

$$\overline{y}_{wlr}{}^{(k)} = \left(1 - \lambda^k\right)\overline{y} + \lambda^k \overline{y}_{wlr}, \tag{16}$$

where $\overline{y}_{wlr}{}^{(k)}$ is called as the weighted linear regression estimator of order k. For k=0, $\overline{y}_{wlr}{}^{(k)}$=$\overline{y}_{wlr}$ i.e., $\overline{y}_{wlr}{}^{(k)}$is the weighted linear regression estimator and for $k \to \infty$, we have $\lambda^k \to 0$ and $\overline{y}_{wlr}{}^{(k)}$=$\overline{y}$. Again, if we draw samples of fixed sizes from an infinite population, then $\frac{n}{N} \to 0$. Hence $\overline{y}_{wlr}{}^{(k)}$ becomes $\overline{y}_{wlr}$.

The bias of $\overline{y}_{wlr}{}^{(k)}$ to $o\left(\frac{1}{n}\right)$ can be written as

$$B(\overline{y}_{wlr}{}^{(k)}) = \lambda^k \frac{N(N-n)}{n(N-1)(N-2}\left[w_1\beta_{yx}\left(\frac{\mu_{300}}{S_x^2} - \frac{\mu_{210}}{S_{yx}}\right) + w_2\beta_{yz}\left(\frac{\mu_{003}}{S_z^2} - \frac{\mu_{012}}{S_{yz}}\right)\right]. \tag{17}$$

If $k \geq 1$, then this hierarchic weighted linear regression estimator possesses smaller bias than that of the customary weighted linear regression estimator. The mean square error of $\overline{y}_{wlr}{}^{(k)}$ to $o\left(\frac{1}{n}\right)$ can be written as

$$\begin{aligned} MSE(\overline{y}_{wlr}{}^{(k)}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2[1 + \lambda^{2k}\left(w_1^2\rho_{yx}^2 + w_2^2\rho_{yz}^2 + 2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}\right) \\ - 2\lambda^k\left(w_1\rho_{yx}^2 + w_2\rho_{yz}^2\right)]. \end{aligned} \tag{18}$$

Again, by obtaining the optimum value of k, we can minimize $V(\overline{y}_{wlr}{}^{(k)})$. So

$$\lambda^k = \frac{w_1\rho_{yx}^2 + w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2 + w_2^2\rho_{yz}^2 + 2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}}. \tag{19}$$

By putting the optimum value of k, in the mean squared error of hierarchic predictive weighted linear regression estimator, we have

$$M(\overline{y}_{wlr}{}^{(k)}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \left[1 - \frac{w_1\rho_{yx}^2 + w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2 + w_2^2\rho_{yz}^2 + 2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}}\right]. \qquad (20)$$

4. Efficiency comparison of the proposed estimator vis-à-vis the competing estimators

$\overline{y}_{wlr}{}^{(k)}$ will be more efficient than $\overline{y}_{wlr}$, if

$$\frac{w_1\rho_{yx}^2 + w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2 + w_2^2\rho_{yz}^2 + 2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}} < \frac{1+\lambda^k}{2} \qquad (21)$$

and $\overline{y}_{wlr}{}^{(k)}$ will be more efficient than $\overline{y}$, if

$$\frac{w_1\rho_{yx}^2 + w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2 + w_2^2\rho_{yz}^2 + 2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}} < \frac{\lambda^k}{2}. \qquad (22)$$

Thus, combining equation (21) and (22), we find that the estimator $\overline{y}_{wlr}{}^{(k)}$ will be more efficient than $\overline{y}_{wlr}$ and $\overline{y}$ if

$$\frac{\lambda^k}{2} < \frac{w_1\rho_{yx}^2 + w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2 + w_2^2\rho_{yz}^2 + 2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}} < \frac{1+\lambda^k}{2}. \qquad (23)$$

The bounds on $\frac{w_1\rho_{yx}^2+w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2+w_2^2\rho_{yz}^2+2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}}$ given in equation(23) are called the efficiency bounds. By choosing values of the sampling fraction $f(=\frac{n}{N})$ and hence $\lambda(=1-f)$, we have prepared and presented in the Appendix a Table which gives the bound on $\frac{w_1\rho_{yx}^2+w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2+w_2^2\rho_{yz}^2+2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}}$ for which equation (23) will be satisfied, i.e., $\overline{y}_{wlr}{}^{(k)}$ will be more efficient than $\overline{y}_{wlr}$ and $\overline{y}$.

Furthermore, with a view to finding the percentage gain in efficiency of $\overline{y}_{wlr}$ and $\overline{y}_{wlr}{}^{(k)}$ with respect to $\overline{y}$ and $\overline{y}_{wlr}$ when k is optimally determined, the following formulae are considered:

$$G_1 = \left[\frac{V(\overline{y})}{M(\overline{y}_{wlr})} - 1\right]\times 100,\ G_2 = \left[\frac{V(\overline{y})}{M(\overline{y}_{wlr}{}^{(k)})} - 1\right]\times 100\ and\ G_3 = \left[\frac{V(\overline{y}_{wlr}))}{M(\overline{y}_{wlr}{}^{(k)})} - 1\right]\times 100.$$

5. Numerical Illustration

For the purpose of numerical illustrations, we have considered 8 natural populations from various sources as detailed in the following Table:

Table 1. *Population data sets*

Population	N	n	ρ_{yx}	ρ_{yz}	$w_1 = 1 - w_2$
I: Gujarati (1978) Y: Telephone Ownership in Singapore. X: Per capita GDP in Singapore.	22	6	0.972	-0.850	2.271
II: Gujarati (1978) Y: GDP deflator for domestic goods, X: GDP deflator for imports	15	5	0.989	-0.976	0.708
III: Gujarati (1978) Y: Real gross product X: Real capital income	15	5	0.886	-0.952	-0.756
IV: Gujarati (1978) Y: Plant expenditures X: Sales	22	6	0.990	-0.891	1.182
V: Cochran (1977) Y: Sizes of 49 large U.S. Cities in 1930 X: Sizes of 49 large U.S. Cities in 1930	49	8	0.981	-0.182	1.011
VI: Maddala and Lahiri(2012) Y: Imports X: Consumption	18	3	0.984	-0.926	1.620
VII: DNase (R Dataset) Y: concentration of the protein, X: optical density	176	23	0.931	-0.319	1.047
VIII: Swiss (R Dataset) Y: Education X: Examination.	47	9	0.698	-0.428	1.221

For assessing the performance of the proposed estimator over the competing estimators, we have prepared the following Table:

Table 2. *MSE of the competing estimators*

Estimators	$\overline{y}$	$\overline{y}_{wlr}$	$\overline{y}_{wlr}{}^{(k)}$
Population I	949.00	22.95	16.73
Population II	15944.39	245.52	242.70
Population III	3167676	207861.70	195576.90
Population IV	338.32	4.59	4.31
Population V	1585.49	58.78	58.61
Population VI	43.27	0.48	0.38
Population VII	0.62	0.0820	0.08
Population VIII	8.31	4.17	4.12

Table 3. *Gain in efficiency of different estimators*

Estimators	G1	G2	G3
Population I	4034.39	5569.51	37.13
Population II	6393.88	6469.44	1.16
Population III	1423.93	1519.66	6.28
Population IV	7263.38	7748.52	6.58
Population V	2597.02	2604.77	0.28
Population VI	8863.93	11169.69	25.72
Population VII	658.97	666.31	0.97
Population VIII	99.23	101.35	1.06

The above Table gives the percentage gain in efficiency of the proposed estimator with respect to its competing estimators, implying thereby that, there is substantial gain in efficiency of the proposed estimator over its competing estimators.

6. Conclusion:

The proposed weighted linear regression estimator of order k introduced in this paper is not only endowed with predictive character but also found to be more efficient than the weighted linear regression estimator and the simple unweighted estimator under conditions that hold good in practice quite often. Empirical study based on several natural population datasets provides sufficient ground in support of the estimator from the standpoint of its practical use in a suitable survey sampling situation.

Acknowledgement

The authors are highly grateful to the reviewers for their valuable suggestions leading to improvement in the original manuscript.

References

[1] Agrawal, M.C. and Jain, N (1989). A new predictive product estimator, Biometrika 76, 822-823.

[2] Agrawal, M.C. and Sthapit, A.B. (1997). Hierarchic Predictive Ratio-based and Product-based Estimators and their Efficiencies. Journal of Applied Statistics 24(1), 97-104.

[3] Basu, D. (1971). An essay on the logical foundations of statistical inference, Part I, Foundations of statistical inference, Ed. By V.P. Godambe and D.A. Sportt, New York.

[4] Cochran, W.G. (1977). Sampling Techniques, Third Edition, A Wiley Publication in Applied Statistics.

[5] Gujarati, D. (1995). Basic Econometrics, Mc-Graw Hill, India, International Editions.

[6] Maddala, G.S. and Lahiri, K. (2012). Introduction to Econometrics, Wiley India(P) Ltd, Fourth Edition.

[7] Panda, K.B. and Chhatopadhyay, G. (2022). On Efficient Regression Method of Estimation. Accepted for publication in the International Journal of Mathematics and Statistics.

[8] Sahoo, L. N., Dalabehra, M., Mangaraj A. K. (2007). A regression estimator using harmonic mean of the auxiliary variable, The Philippine Statistician 56(3-4), 31-36.

[9] Smith, T.M.F. (1976). The foundations in survey sampling, a review, Jour. R. Statist. Soc., Series A 139, 183-204.

[10] R dataset

Appendix

Table 4. *Efficiency bounds of* $\frac{w_1\rho_{yx}^2+w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2+w_2^2\rho_{yz}^2+2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}}$ *for various values of f and k.*

	k					
f	1	2	5	10	15	20
0.05	(0.475, 0.975)	(0.451, 0.951)	(0.387, 0.887)	(0.299, 0.799)	(0.232, 0.732)	(0.179. 0.679)
0.10	(0.450, 0.950)	(0.405, 0.905)	(0.295, 0.795)	(0.174, 0.674)	(0.103, 0.603)	(0.061, 0.561)
0.15	(0.425, 0.925)	(0.361, 0.861)	(0.221, 0.721)	(0.098, 0.598)	(0.044, 0.544)	(0.019, 0.519)
0.20	(0.400, 0.900)	(0.320, 0.820)	(0.164, 0.664)	(0.054,0.554)	(0.017, 0.517)	(0.005, 0.505)
0.25	(0.375. 0.875)	(0.281, 0.781)	(0.118, 0.618)	(0.028. 0.528)	(0.006. 0.506)	(0.001, 0.501)
0.30	(0.350, 0.850)	(0.245, 0.745)	(0.084, 0.584)	(0.014, 0.514)	(0.002, 0.502)	(0.000, 0.500)
0.40	(0.300, 0.800)	(0.180, 0.680)	(0.039, 0.539)	(0.003, 0.503)	(0.000, 0.500)	(0.000, 0.500)
0.50	(0.250, 0.750)	(0.125, 0.625)	(0.015, 0.515)	(0.000, 0.500)	(0.000, 0.500)	(0.000, 0.500)
0.60	(0.200, 0.700)	(0.080, 0.580)	(0.005, 0.505)	(0.000, 0.500)	(0.000, 0.500)	(0.000, 0.500)
0.75	(0.125, 0.625)	(0.031, 0.531)	(0.000, 0.500)	(0.000, 0.500)	(0.000, 0.500)	(0.000, 0.500)

Table 4 is relevant in view of locating a suitable value of k for given values of $\frac{w_1\rho_{yx}^2+w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2+w_2^2\rho_{yz}^2+2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}}$ and f. As regards knowledge of the pivotal quantity, it can be said that the knowledge of $\rho_{yx},\ \ \rho_{yz}\ \ and\ \ \rho_{xz}$ are known in advance from a pilot survey or from the past experience, if any, which will remain stable over a period of time. The above table gives more than one values of k which results in better performance of the proposed estimator over its competing estimators. The optimal value of k which is given in equation (19), provided that $\frac{w_1\rho_{yx}^2+w_2\rho_{yz}^2}{w_1^2\rho_{yx}^2+w_2^2\rho_{yz}^2+2w_1w_2\rho_{yx}\rho_{yz}\rho_{xz}} < 1$. Even if the exact optimal value of k is not available, a satisfactory value of k that offers the superiority of our proposed estimator might still be found as exhibited by Table 4.