The Philippine Statistician, 2007
Vol. 56, Nos. 3-4, pp. 31-36

A Regression Estimator Using Harmonic Mean of The Auxiliary Variable

Lokanath N. Sahoo, Alok Kumar Mangaraj and Manmohan Dalabehera[1]

Received: November, 2007; Revised: March, 2008

ABSTRACT

This paper considers a regression estimator of the finite population mean by making use of the simple harmonic mean of the auxiliary variable.

Key Words: Auxiliary variable, Difference estimator, Harmonic mean, Mean square error, Precision, Regression estimator.

I. INTRODUCTION

Let y_k and $x_k : x_k > 0$ respectively be the values of the study variable y and an auxiliary variable x on the kth unit of a finite population defined by $U = \{1,2,...,N\}$. Suppose that x_k's are at our disposal and an estimate is needed for the population mean $\bar{Y}$ of y. As is well known, regression method of estimation is a type of estimation that attempts to make an efficient use of auxiliary information about U. One of the main advantages associated with a regression estimator is that it can be used for both the situations of positive and negative correlations between y and x. Basing on a sample s of fixed size n drawn from U by simple random sampling without replacement, the classical regression estimator of $\bar{Y}$ is defined by

$$\bar{y}_{RG} = \bar{y} - b_{yx}(\bar{x} - \bar{X}),$$

where $\bar{y} = n^{-1}\sum_{k\in s} y_k$, $\bar{x} = n^{-1}\sum_{k\in s} x_k$, $b_{yx} = \dfrac{\sum_{k\in s}(y_k - \bar{y})(x_k - \bar{x})}{\sum_{k\in s}(x_k - \bar{x})^2}$ and $\bar{X}$ is the known population mean of $x-$values. This estimator is useful if a strong linear relationship exists between y and x, and the regression line of y on x intercepts $y-$axis at some distance from the origin. To terms of order n^{-1}, the mean square error of $\bar{y}_{RG}$ is given by

$$M(\bar{y}_{RG}) = \frac{N-n}{(N-1)n}\sigma_y^2(1-\rho_{yx}^2), \qquad (1)$$

where σ_y^2 = variance of y, ρ_{yx} = correlation coefficient between y and x.

[1] First author is a professor of Department of Statistics, Utkal University, Bhubaneswar 751004, India. email: lnsahoostatuu@rediffmail.com. Second author is a lecturer of Department of Statistics, R.D. Women's College, Bhubaneswar 751022, India. Third author is a professor of Department of Statistics, Orissa University of Agriculture and Technology, Bhubaneswar 751003, India.

II. THE SUGGESTED REGRESSION ESTIMATOR

Agrawal and Jain (1989) introduced a product method of estimation using harmonic means of $x-$values in U and s being defined by $\bar{X}_h = N/\sum_{k\in U} x_k^{-1}$ and $\bar{x}_h = n/\sum_{k\in s} x_k^{-1}$ respectively. But, defining $z = x^{-1}$ as a transformed auxiliary variable with ρ_{yx} and ρ_{yz} (correlation coefficient between y and z) are of opposite signs, we have $\bar{Z} = N^{-1}\sum_{k\in U} z_k = \bar{X}_h^{-1}$ and $\bar{z} = n^{-1}\sum_{k\in s} z_k = \bar{x}_h^{-1}$, the arithmetic means of $z-$values for U and s respectively. Thus, the use of harmonic means of the main auxiliary variable-values leads to the use of arithmetic means of the transformed auxiliary variable-values. This philosophy of Agrawal and Jain (1989) encouraged Sahoo and Dalabehera (1995, 2005) to consider some ratio- and product-type estimators. However, in this paper, an attempt has been made to develop a new regression estimator.

The new regression estimator can be arrived at by introducing a difference estimator

$$\ell_d = \bar{y} - d\left(\bar{x}_h^{-1} - \bar{X}_h^{-1}\right),$$

which is evidently a consistent and unbiased estimator of $\bar{Y}$ for a fixed value of d. Theoretically, the optimum value of d is β_{yz}, the regression coefficient of y on z, being determined by minimizing variance of ℓ_d in the usual way. Since β_{yz} is unknown, a practicable alternative is its consistent estimator

$$b_{yz} = \frac{\sum_{k\in s}(y_k - \bar{y})(z_k - \bar{z})}{\sum_{k\in s}(z_k - \bar{z})^2}.$$

Hence, our new regression estimator is defined by

$$\ell_{RG} = \bar{y} - b_{yz}\left(\bar{x}_h^{-1} - \bar{X}_h^{-1}\right),$$

with an approximate mean square error to $O(n^{-1})$ is given by

$$M(\ell_{RG}) = \frac{N-n}{(N-1)n}\sigma_y^2\left(1-\rho_{yz}^2\right). \tag{2}$$

From (1) and (2), it follows that $M(\bar{y}_{RG}) >$ or $< M(\ell_{RG})$ according as

$$\rho_{yx}^2 < \text{or} > \rho_{yz}^2. \tag{3}$$

To convince the readers that situations do arise where any one of these two conditions are fulfilled, we consider the following common regression models:

(i) Let $y = \alpha + \beta x$ so that $\rho_{yx}^2 = 1 \geq \rho_{yz}^2$. Then, ℓ_{RG} has a lower precision than $\bar{y}_{RG}$.

(ii) Let $y = \alpha + \beta x^{-1}$. Then, $\rho_{yx}^2 \leq 1 = \rho_{yz}^2$, which implies that ℓ_{RG} has a higher precision than $\bar{y}_{RG}$.

The above results clearly indicate that ℓ_{RG} has a tendency of achieving more accuracy than $\bar{y}_{RG}$ in estimation of $\bar{Y}$ when $\rho_{yx} < 0$, and regression line of y on z is approximately linear without touching the origin. Inspecting a scatter diagram for at least some pairs (y_k, x_k) of sample values may help in this respect. The condition $x_k > 0$, $k \in U$, is not unrealistic in a sample survey situation and $\bar{X}_h$ can be calculated from the known x – values of the population units.

III. EMPIRICAL STUDY

To illustrate the relative performance of $\bar{y}_{RG}$ and ℓ_{RG}, an empirical study was carried out by considering all the ${}^{N}C_{n}$ possible samples, for $n = 2$, 3 and 4 drawn from 10 populations described in Table 1. The following performance measures were taken into consideration:

(i) Relative Bias (RB) = $100 \times |bias(t)| / \bar{Y}$,
where the bias of an estimator t (either $\bar{y}_{RG}$ or ℓ_{RG}) is defined by $bias(t) = E(t) - \bar{Y}$.

(ii) Relative Efficiency (RE) = $100 \times V(\bar{y}) / M(t)$,
where $V(\bar{y})$ is the variance of $\bar{y}$ and $M(t)$ is the mean square error of the estimator t.

(iii) Coverage Rate (CR) based on $100(1-\alpha)\%$ (95% or 99%) confidence interval

$$t \pm u_{1-\frac{\alpha}{2}} \sqrt{v(t)},$$

where $u_{1-\frac{\alpha}{2}}$ is exceeded with probability $\alpha/2$ by the unit normal variate under the assumption that the sampling distribution of the estimator t is approximately normal and $v(t)$ is it's approximate variance estimator. The approximate variance estimators for $\bar{y}_{RG}$ and ℓ_{RG} used to construct confidence intervals are as follows:

$$v(\bar{y}_{RG}) = \frac{N-n}{Nn(n-1)} \sum_{k \in s} \left[(y_k - \bar{y}) - b_{yx}(x_k - \bar{x}) \right]^2$$

$$v(\ell_{RG}) = \frac{N-n}{Nn(n-1)} \sum_{k \in s} \left[(y_k - \bar{y}) - b_{yz}(z_k - \bar{z}) \right]^2.$$

This performance measure gives us an idea about which percentage of the so constructed confidence intervals covers the true value of $\bar{Y}$ under repeated draws of samples from a population.

For each selected sample, the estimators $\bar{y}$, $\bar{y}_{RG}$ and ℓ_{RG} were calculated and their average behaviors through different performance measures are presented in Tables 2 to 4. The findings of the study are discussed in subsections 3.1 to 3.3.

3.1 Results Based on the RB

The numerical values with regard to RB shown in Table 2 reveal that ℓ_{RG} outperforms $\bar{y}_{RG}$ in three cases and $\bar{y}_{RG}$ outperforms ℓ_{RG} in one case for all values of n. The bias of $\bar{y}_{RG}$ diminishes gradually with enlargement of sample size. But, this tendency is not strictly observed for ℓ_{RG} in the sense that, for some populations, its RB increases with the increase of sample size. However, in respect of RB, both the estimators behave very much erratically and there is no clear indication that which one would have a decidedly better overall performance than other.

Table 1. Description of Populations

Pop. No.	Source	y	x	N	ρ_{yx}
1	Gujarati(1978) p. 59	quit rate	unemployment rate	13	-0.808
2	Maddala(1977) p. 96	consumption of veal	price of veal	16	-0.682
3	Maddala(1977) p.96	consumption of chicken	price of chicken	16	-0.976
4	Maddala(1977) p. 96	consumption of beef	price of beef	16	-0.844
5	Maddala(1977) p. 96	consumption of lamb	price of lamb	16	-0.751
6	Dobson(1990) p. 47	survival time of leukemia patient	initial white blood cell count	17	-0.685
7	Dobson(1990) p. 69	carbohydrate for male insulin dependent diabetics	Age	20	-0.059
8	Dobson(1990) p. 69	carbohydrate for male insulin dependent diabetics	Weight	20	-0.407
9	Pandey and Dubey(1988)	artificial	Artificial	20	-0.919
10	Srivenkataramana and Tracy(1980)	artificial	Artificial	20	-0.914

Table 2. Features of the RB

Pop No.	$n=2$		$n=3$		$n=4$	
	$\bar{y}_{RG}$	ℓ_{RG}	$\bar{y}_{RG}$	ℓ_{RG}	$\bar{y}_{RG}$	ℓ_{RG}
1	2.57	1.51	1.78	1.45	1.45	1.36
2	1.17	1.65	1.10	1.08	1.01	0.45
3	1.85	0.45	1.42	1.91	0.25	2.16
4	3.52	3.45	2.83	4.86	0.95	5.37
5	3.10	2.86	1.33	1.95	1.11	1.73
6	4.65	3.47	3.19	2.95	2.65	2.76
7	5.89	3.36	4.76	2.45	1.58	1.29
8	2.83	2.06	2.22	1.73	2.01	1.11
9	0.63	0.86	0.47	1.35	0.09	4.89
10	1.01	0.37	0.76	1.08	0.15	1.99

3.2 Results Based on the RE

The relative efficiencies of the comparable estimators are displayed in Table 3. A close scrutiny of the entries in the table clearly indicates that ℓ_{RG} outperforms $\bar{y}_{RG}$ in 8 cases. It is inferior to $\bar{y}_{RG}$ for population 8. But, its performance for population 7 is not satisfactory where the value of ρ_{yx} is very small *i.e.*, -0.059.

Table 3. Features of the RE

Pop No.	$n=2$		$n=3$		$n=4$	
	$\bar{y}_{RG}$	ℓ_{RG}	$\bar{y}_{RG}$	ℓ_{RG}	$\bar{y}_{RG}$	ℓ_{RG}
1	171	201	188	236	195	273
2	125	138	137	146	152	159
3	110	183	114	205	117	214
4	142	187	169	201	181	221
5	169	175	174	199	189	218
6	108	127	128	142	130	151
7	101	104	103	105	105	106
8	106	103	110	108	115	111
9	157	213	186	244	189	254
10	149	174	159	191	179	203

Table 4. Features of the CR of Nominal 95% Confidence Interval

Pop No.	$n=2$		$n=3$		$n=4$	
	$\bar{y}_{RG}$	ℓ_{RG}	$\bar{y}_{RG}$	ℓ_{RG}	$\bar{y}_{RG}$	ℓ_{RG}
1	73.49	75.33	79.02	81.41	85.45	90.30
2	68.38	69.90	74.11	76.81	80.65	82.46
3	79.98	65.88	82.14	66.45	84.35	71.24
4	62.34	70.31	73.75	79.57	81.75	85.52
5	70.54	68.19	76.78	72.08	82.58	77.98
6	48.43	55.26	54.17	58.21	61.80	80.11
7	31.17	48.48	35.82	55.55	45.19	69.97
8	41.41	50.44	53.21	68.36	64.48	81.73
9	50.23	65.12	56.43	78.95	62.23	80.13
10	49.74	49.85	55.00	56.27	61.34	64.80

3.3 Results Based on the CR

Table 4 shows the coverage rates of the nominal 95% confidence interval for $\bar{Y}$ using the competing estimators. Results for 99% are not shown as they confirm more or less the tendencies found in the case of 95%. Table 4 gives an indication of improvement in the performance of an estimator as the sample size increases. However, on the ground of the achieved CR, ℓ_{RG} is better than $\bar{y}_{RG}$ for 8 populations (except populations 3 and 5).

The empirical performance of ℓ_{RG} compared to $\bar{y}_{RG}$ (results of which are not presented here) is also examined for a number of populations with $\rho_{yx} > 0$ found in books

and papers on survey sampling. But, in most of the cases it is observed that ℓ_{RG} is inferior to $\bar{y}_{RG}$.

Our empirical study (although has a limited scope), suggests that the overall performance of ℓ_{RG} compared to $\bar{y}_{RG}$ on the basis of two performance measures *viz.*, RE and CR is very much satisfactory. On consideration of RB, although ℓ_{RG} seems to be inferior to $\bar{y}_{RG}$, the bias of ℓ_{RG} is not a factor of concern usually for $n = 2$. It means that, for small samples, effect of bias of ℓ_{RG} may not be more severe than that of $\bar{y}_{RG}$.

IV. CONCLUSION

Construction of ℓ_{RG}, of course, seems to be a mathematical exercise of considering z in place of x. But, analytical as well as empirical results reported here lead to an interesting conclusion that the use of harmonic mean of the auxiliary variable in regression method of estimation can achieve better performance than the classical regression method when ρ_{yx} has a high negative value.

Acknowledgement

The authors are grateful to the referee whose constructive comments led to an improvement in the paper.

References

AGRAWAL, M.C. and JAIN, N. (1989). A new predictive product estimator. *Biometrika*, **76**, 822-823.

DOBSON, A.J. (1990). *An Introduction to Generalized Linear Models*. Chapman and Hall.

GUJARATI, D. (1978). *Basic Econometrics*. Mc-Graw Hill Book Co.

MADDALA, G.S. (1977). *Econometrics*. Mc-Graw Hill Kogakusha Limited.

PANDEY, B.N. and DUBEY, V. (1988). Modified product estimator using coefficient of variation of auxiliary variable. *Assam Statistical Review*, **2**, 64-66.

SAHOO, L.N. and DALABEHERA, M. (1995). Some product-type strategies using harmonic mean of an auxiliary variable. *Statistics in Transition*, **2**, 839-845.

SAHOO, L.N. and DALABEHERA, M. (2005). Unbiased estimators using harmonic mean of the auxiliary variable. *International Journal of Agricultural and Statistical Sciences*, **1**, 11-16.

SRIVENKATARAMANA, T. and TRACY, D.S. (1980). An alternative to ratio method in sample surveys. *Annals of the Institute of Statistical Mathematics*, **32**, 111-120.