

MODIFIED REGRESSION APPROACH IN PREDICTION OF FINITE POPULATION MEAN USING KNOWN COEFFICIENT OF VARIATION

Sheela Misra*, R. K. Singh, and Ashish Kumar Shukla

Department of Statistics, University of Lucknow, Lucknow-226007 (INDIA)
Email: * drsheelamisra@gmail.com

(Received July 13, 2012)

Abstract

In this paper, we are utilizing the modified regression approach for the prediction of finite population mean, with known coefficient of variation of study variable y , under simple random sampling without replacement. The bias and mean square error of the proposed estimator are obtained and compared with the usual regression estimator of the population mean and comes out to be more efficient in the sense of having lesser mean square error. The optimum class of estimators is obtained and for the greater practical utility proposed optimum estimator based on estimated optimum value of the characterizing scalar is also obtained and is shown to retain the same efficiency to the first order of approximation as the former one. A numerical illustration is also given to support the theoretical conclusions.

Key words: Regression estimator, Prediction, Predictor, Coefficient of variation, Bias, Mean square error, Auxiliary variable, Relative efficiency, Simple random sampling, Effective sample size.

1. Introduction

Let y be the study variable taking real value Y_i for the i^{th} unit $i = 1, 2, \dots, N$ of the finite population U of size N . Consider the problem of estimating the population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

on the basis of observed values of y on units in a sample which is an ordered subset of the finite population U of size N . Let S denote the collection of all possible samples from U . For any given $s \in S$, let $v(s)$ denote its effective sample size and \bar{s} denote the set of all those units of U which are not in s . Further let

$$\bar{y}_s = \frac{1}{v(s)} \sum_{i \in s} y_i$$

$$\bar{y}_{\bar{s}} = \frac{1}{N - v(s)} \sum_{i \in \bar{s}} y_i \tag{1.1}$$

For any given $s \in S$, we have

$$\bar{Y} = \frac{\nu(s)}{N} \bar{y} + \frac{N - \nu(s)}{N} \bar{y}_{\bar{s}}$$

Basu (1971) argued that in the representation of \bar{Y} above, the sample mean $\bar{y}_{\bar{s}}$ being based on the observed y values on units in the sample s is known, therefore the statistician should attempt a prediction of the mean $\bar{y}_{\bar{s}}$ of the unobserved units of the population on the basis of observed units in s . For any given $s \in S$ using simple random sampling without replacement and effective sample size $\nu(s) = n$ and $\bar{y}_s = \bar{y}$, the population mean \bar{Y} is given by ,

$$\bar{Y} = \frac{n}{N} \bar{y} + \frac{N - n}{N} \bar{y}_{\bar{s}} \tag{1.2}$$

Considering T as a predictor of $\bar{y}_{\bar{s}}$, an estimator \bar{y} of \bar{Y} can be written as

$$\bar{Y} = \frac{n}{N} \bar{y} + \frac{N - n}{N} T \tag{1.3}$$

where T is a predictor of $\bar{y}_{\bar{s}}$.

The literature describes a great variety of techniques for using auxiliary information by means of ratio, and product and regression methods for estimating population mean which most of the time leads to the gain in efficiency of the estimator. Some efforts in this direction are due to Srivastava (1983), Srivastava & Jhaji (1995), Singh. & Espejo (2003), Kadilar & Cingi (2006). For x being the auxiliary variable correlated with study variable y and X_i being the value of x on the i^{th} unit

$i = 1, 2, \dots, N$ of the population U , let $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ and further, for x_i being the

value of x on the i^{th} unit $i = 1, 2, \dots, n$ of the sample s , let $\bar{x} = \frac{1}{n} \sum_{i \in s} x_i$. Using

regression type estimator $T_1 = \{ \bar{y} + b(\bar{X}_{\bar{s}} - \bar{x}) \} \left(\frac{C_y^2}{s_y^2} \bar{y}^2 \right)^k$, where $\bar{X}_{\bar{s}} =$

$\frac{1}{N - n} \sum_{i \in s} x_i = \frac{N \bar{X} - n \bar{x}}{N - n}$ as predictor for T in (1.3), we have the proposed

estimator of the population mean \bar{Y} under prediction approach to be

$$\tilde{y}_k = \frac{n}{N} \bar{y} + \frac{N - n}{N} \{ \bar{y} + b(\bar{X}_{\bar{s}} - \bar{x}) \} \left(\frac{C_y^2}{s_y^2} \bar{y}^2 \right)^k$$

$$\begin{aligned}
 &= \frac{n}{N} \bar{y} + \left(1 - \frac{n}{N}\right) \left\{ \bar{y} + b \left(\frac{N \bar{X} - n \bar{x}}{N - n} - \bar{x} \right) \right\} \left(\frac{C_y^2}{s_y^2} \bar{y}^{-2} \right)^k \\
 &= \frac{n}{N} \bar{y} + \left(1 - \frac{n}{N}\right) \left\{ \bar{y} + bN \left(\frac{\bar{X} - \bar{x}}{N - n} \right) \right\} \left(\frac{C_y^2}{s_y^2} \bar{y}^{-2} \right)^k \\
 &= \frac{n}{N} \bar{y} + \left(1 - \frac{n}{N}\right) \left[\bar{y} + b \left\{ \frac{\bar{X} - \bar{x}}{\left(1 - \frac{n}{N}\right)} \right\} \right] \left(\frac{C_y^2}{s_y^2} \bar{y}^{-2} \right)^k \\
 &= \frac{n}{N} \bar{y} + \left\{ \left(1 - \frac{n}{N}\right) \bar{y} + b(\bar{X} - \bar{x}) \right\} \left(\frac{C_y^2}{s_y^2} \bar{y}^{-2} \right)^k
 \end{aligned}$$

Let $\frac{n}{N} = \theta$ then

$$\tilde{y}_k = \theta \bar{y} + \left\{ (1 - \theta) \bar{y} + b(\bar{X} - \bar{x}) \right\} \left(\frac{C_y^2}{s_y^2} \bar{y}^{-2} \right)^k \tag{1.4}$$

2. Bias and Mean Square Error of \tilde{y}_k

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 terms may be ignored.

We define

$$\begin{aligned}
 \bar{y} &= \bar{Y}(1 + e_0), & \bar{x} &= \bar{X}(1 + e_1), & s_y^2 &= S_y^2(1 + e_2), & s_x^2 &= S_x^2(1 + e_3), \\
 s_{xy} &= \sigma_{xy}(1 + e_4)
 \end{aligned}$$

It can be verified that $E(e_i) = 0$; $i = 0,1,2,3,4$. Also up to the first order of approximation, we have obtained the following expectations on the lines of Sukhatme *et al.* (1984):

$$E(e_0^2) = \frac{\sigma_y^2}{n\bar{Y}^2}, E(e_1^2) = \frac{\sigma_x^2}{n\bar{X}^2}, E(e_2^2) = \frac{\beta_2 - 1}{n}, E(e_1e_3) = \frac{\mu_3(x)}{n\bar{X}\sigma_x^2},$$

$$E(e_0e_2) = \frac{\mu_3}{n\bar{Y}\sigma_y^2}, E(e_0e_1) = \frac{\sigma_{xy}}{n\bar{X}\bar{Y}}, E(e_1e_2) = \frac{\mu_{12}}{n\bar{X}\sigma_y^2}, E(e_1e_4) = \frac{\mu_{21}}{n\bar{X}\mu_{11}}$$

Also

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sigma_{xy}(1 + e_4)}{\sigma_x^2(1 + e_3)} = B(1 + e_4)(1 + e_3)^{-1} = B(1 + e_4)(1 - e_3 + e_3^2 - \dots)$$

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

where $B = \frac{\sigma_{xy}}{\sigma_x^2}$, is the population regression coefficient of y on x . From (1.4), the

proposed predictive regression type estimator \tilde{y}_k of population mean is

$$\tilde{y}_k = \left\{ \bar{y} + b(\bar{X} - \bar{x}) \right\} \left\{ \frac{C_y^2}{s_y^2} \bar{y}^2 \right\}^k \tag{2.1}$$

which, when written in the terms of e_i 's becomes

$$\begin{aligned} &= \left[\bar{Y}(1 + e_0) + B(1 - e_3 + e_4 + e_3^2 - e_3e_4 + \dots) \{ \bar{X} - \bar{X}(1 + e_1) \} \right] \\ &\quad \left\{ \frac{\sigma_y^2}{\sigma_y^2(1 + e_2)} \bar{Y}^2 (1 + e_0)^2 \right\}^k \\ &= \left\{ \bar{Y}(1 + e_0) + B(1 - e_3 + e_4 + e_3^2 - e_3e_4 + \dots)(-\bar{X}e_1) \right\} \left\{ (1 + e_2)^{-k} (1 + e_0)^{2k} \right\} \\ &= \left\{ \bar{Y} + \bar{Y}e_0 + B(e_1e_3 - e_1e_4 - e_1) \right\} \\ &\quad \left\{ 1 + 2ke_0 - ke_2 - 2k^2e_0e_2 + k(2k - 1)e_0^2 + \dots \right\} \\ &\quad \left\{ 1 - ke_2 + \frac{k(k + 1)}{2}e_2^2 + \dots \right\} \\ \tilde{y}_k - \bar{Y} &= \bar{Y} \left\{ (2k + 1)e_0 + k(2k + 1)e_0^2 - k(2k + 1)e_0e_2 - ke_2 + \frac{k(k + 1)}{2}e_2^2 \right\} \\ &\quad + B\bar{X} \{ e_1e_3 - e_1e_4 - e_1 + ke_1e_2 - 2ke_0e_1 \} + \dots \tag{2.2} \end{aligned}$$

Taking expectation on both sides of (2.2), the bias of \tilde{y}_k to the first degree of approximation, is

$$\begin{aligned} \text{Bias}(\tilde{y}_k) &= E(\tilde{y}_k - \bar{Y}) \\ &= E \left[\bar{Y} \left\{ (2k + 1)e_0 + k(2k + 1)e_0^2 - k(2k + 1)e_0e_2 - ke_2 + \frac{k(k + 1)}{2}e_2^2 \right\} + \right. \\ &\quad \left. B\bar{X} \{ e_1e_3 - e_1e_4 - e_1 + ke_1e_2 - 2ke_0e_1 \} \right] \\ &= \bar{Y} \left\{ (2k + 1)E(e_0) + k(2k + 1)E(e_0^2) - k(2k + 1)E(e_0e_2) - kE(e_2) \right\} \\ &\quad + \frac{k(k + 1)}{2}E(e_2^2) \\ &\quad + B\bar{X} \{ E(e_1e_3) - E(e_1e_4) - E(e_1) + kE(e_1e_2) - 2kE(e_0e_1) \} \tag{2.3} \end{aligned}$$

Squaring both sides of (2.2), taking expectation, we have mean square error of to the first degree of approximation i.e., upto of $O\left(\frac{1}{n}\right)$ to be

$$\begin{aligned}
 MSE(\tilde{y}_k) &= E(\tilde{y}_k - \bar{Y})^2 \\
 &= \bar{Y}^2 \left[\{(2k + 1)e_0 - ke_2\} - B\bar{X}e_1 \right]^2 \\
 &= k^2 \bar{Y}^2 \left\{ 4E(e_0^2) + E(e_2^2) - 4E(e_0e_2) \right\} + \\
 &\quad k \left\{ 4\bar{Y}^2 E(e_0^2) - 2\bar{Y}^2 E(e_0e_2) - \right. \\
 &\quad \left. 4\bar{X}\bar{Y}BE(e_0e_1) + 2\bar{X}\bar{Y}BE(e_1e_2) \right\} \\
 &\quad + \left\{ \bar{Y}^2 E(e_0^2) + B^2 \bar{X}^2 E(e_1^2) - 2\bar{X}\bar{Y}BE(e_0e_2) \right\} \\
 &= \frac{k^2 \bar{Y}^2}{n} \left\{ 4C_y^2 + (\beta_2 - 1) - 4\gamma_1 C_y \right\} + \frac{2k\bar{Y}^2}{n} \left\{ 2C_y^2(1 - \rho^2) - \gamma_1 C_y + \right. \\
 &\quad \left. \rho \frac{\mu_{12}}{\sigma_x \sigma_y \bar{Y}} \right\} + \\
 &\quad \left\{ (1 - \rho^2) \frac{\sigma_y^2}{n} \right\} \tag{2.4}
 \end{aligned}$$

The optimum value of k minimizing the mean square error of \tilde{y}_k is

$$k_0 = - \frac{\left\{ 2C_y^2(1 - \rho^2) - \gamma_1 C_y + \rho \frac{\mu_{12}}{\sigma_x \sigma_y \bar{Y}} \right\}}{\left\{ (\beta_2 - 1) + 4C_y^2 - 4\gamma_1 C_y \right\}} \tag{2.5}$$

and the minimum mean square error of \tilde{y}_k is given by

$$MSE(\tilde{y}_{k_0}) = (1 - \rho^2) \frac{\sigma_y^2}{n} - \frac{\bar{Y}^2 \left\{ 2C_y^2(1 - \rho^2) - \gamma_1 C_y + \rho \frac{\mu_{12}}{\sigma_x \sigma_y \bar{Y}} \right\}^2}{n \left\{ (\beta_2 - 1) + 4C_y^2 - 4\gamma_1 C_y \right\}} \tag{2.6}$$

The proposed estimator \tilde{y}_k is more efficient than the usual regression estimator under the condition $\gamma_1 < \left\{ \frac{(\beta_2 - 1)}{4C_y} \right\} + C_y$.

3. Estimator Based on Estimated Optimum \hat{k}

The optimum value k_0 or its good guessed value may not be available always in practice; hence the alternative is to replace the parameters in the optimum value by their unbiased or consistent estimators based on sample observations. We can replace these $\beta_2, \rho, \gamma_1, \bar{Y}$ and μ_{12} by their sample estimates $\hat{\beta}_2, \hat{\rho}, \hat{\gamma}_1, \bar{y}$ and μ_{12} respectively in (2.4) and get the estimated optimum value of k_0 denoted by \hat{k} as-

$$\hat{k} = - \frac{\left\{ 2C_y^2 (1 - \hat{\rho}^2) - \hat{\gamma}_1 \bar{C}_y + \hat{\rho} \frac{\hat{\mu}_{12}}{\hat{\sigma}_x \hat{\sigma}_y} \right\}}{\left\{ (\hat{\beta}_2 - 1) + 4C_y^2 - 4\hat{\gamma}_1 \bar{C}_y \right\}}$$

$$= - \frac{\left\{ 2C_y^2 (1 - \hat{\rho}^2) - \frac{\hat{\mu}_3}{s_y} \hat{C}_y + \hat{\rho} \frac{\hat{\mu}_{12}}{\hat{\sigma}_x \hat{\sigma}_y} \right\}}{\left\{ \left(\frac{\hat{\mu}_4}{s_y^4} - 1 \right) + 4C_y^2 - 4 \frac{\hat{\mu}_3}{s_y^3} \bar{C}_y \right\}} \tag{3.1}$$

Thus, from (1.4) and (3.1), we get the predicted regression type estimator \tilde{y}_e depending on estimated optimum \hat{k} to be

$$\tilde{y}_e = \left\{ \bar{y} + b(\bar{X} - \bar{x}) \right\} \left(\frac{C_y^2}{s_y^2} \bar{y}^{-2} \right)^{\hat{k}} \tag{3.2}$$

Defining $\hat{\mu}_3 = \mu_3(1 + e_5)$, $\hat{\mu}_4 = \mu_4(1 + e_6)$, $\hat{\mu}_{12} = \mu_{12}(1 + e_7)$, we have

$$\hat{k} = - \frac{2C_y^2 \left\{ 1 - \frac{\sigma_{xy}^2(1+e_4)^2}{\sigma_x^2(1+e_3)\sigma_y^2(1+e_2)} \right\} - \frac{\mu_3(1+e_5)}{\sigma_y^3(1+e_2)^{\frac{3}{2}}} C_y + \frac{\sigma_{xy}(1+e_4)\mu_{12}(1+e_7)}{\sigma_y \sigma_x (1+e_2)^{\frac{1}{2}} (1+e_3)^{\frac{1}{2}} \sigma_y (1+e_2)^{\frac{1}{2}} \sigma_x (1+e_3)^{\frac{1}{2}} \bar{Y} (1+e_0)}}{\left(\frac{\mu_4(1+e_6)}{\sigma_y^4(1+e_2)^2} - 1 \right) + 4C_y^2 - 4 \frac{\mu_3(1+e_5)}{\sigma_y^3(1+e_2)^2} C_y}$$

$$= - \frac{2C_y^2 \left\{ 1 - \rho^2(1 - e_3 + 2e_4 - e_2 + e_2^2 \dots) \right\} - \gamma_1 C_y \left(1 - \frac{3}{2}e_2 + e_5 + \frac{15}{8}e_2^2 \dots \right) + \frac{\rho \mu_{12}}{\sigma_x \sigma_y} \bar{Y} (1 - e_2 - e_3 - e_0 + e_4 + e_7 + e_2^2 \dots)}{\left\{ \beta_2(1 - 2e_2 + e_0 + \dots) - 1 \right\} + 4C_y^2 - 4\gamma_1 \left(1 - \frac{3}{2}e_2 + e_5 + \dots \right) C_y}$$

(3.3)

Using (3.3) in (3.2), we have

$$\begin{aligned} \tilde{y}_e - \bar{Y} = \bar{Y} \left\{ (2\hat{k} + 1)e_0 + \hat{k}(2\hat{k} + 1)e_0^2 - \hat{k}(2\hat{k} + 1)e_0e_2 - \hat{k}e_2 + \frac{\hat{k}(\hat{k} + 1)}{2}e_2^2 \right\} \\ + B\bar{X}(e_1e_3 - e_1e_4 - e_1 + ke_1e_2 - 2\hat{k}e_0e_1) \end{aligned} \tag{3.4}$$

Now squaring both sides of (3.4), taking expectation, retaining terms upto $O\left(\frac{1}{n}\right)$ we have mean square error of \tilde{y}_e given by

$$MSE(\tilde{y}_e) = (1 - \rho^2) \frac{\sigma_y^2}{n} - \frac{\bar{Y}^2 \left\{ 2C_y^2(1 - \rho^2) - \gamma_1C_y + \rho \frac{\mu_{12}}{\sigma_x\sigma_y\bar{Y}} \right\}^2}{n \left\{ (\beta_2 - 1) + 4C_y^2 - 4\gamma_1C_y \right\}} \tag{3.5}$$

which is the same expression as that of the mean square error $MSE(\tilde{y}_{k_0})$ in (2.6) giving the result that the estimator based on estimated optimum k attains the same minimum mean square error of \tilde{y}_{k_0} depending on optimum value and under the

condition $\gamma_1 < \left\{ \frac{(\beta_2 - 1)}{4C_y} \right\} + C_y$.

4. Concluding Remarks

a) The Bias of \tilde{y}_{k_0} obtained for the optimum value of k is

$$\begin{aligned} Bias(\tilde{y}_{k_0}) = \frac{\bar{Y}}{2n} \left[\frac{2C_y^2(1 - \rho^2) - \gamma_1C_y + \rho \frac{\mu_{12}}{\sigma_x\sigma_y\bar{Y}}}{\left\{ (\beta_2 - 1) + 4C_y^2 - 4\gamma_1C_y \right\}} \right] \\ + \left\{ \gamma_1C_y + 2\rho^2C_y^2 - \frac{\rho\mu_{12}}{n\sigma_x\sigma_y} - (\beta_2 - 1) \right\} + \\ \left(\frac{\rho\mu_{12}}{n\sigma_x\sigma_y} - \frac{\rho\sigma_y\mu_{21}}{n\sigma_x\mu_{11}} \right) \end{aligned} \tag{4.1}$$

b) For the optimum value of k , it is clear in (2.6), that the estimator \tilde{y}_k attains the minimum mean square error

$$MSE(\tilde{y}_{k_0}) = (1 - \rho^2) \frac{\sigma_y^2}{n} - \frac{\bar{Y}^2 \left\{ 2C_y^2(1 - \rho^2) - \gamma_1 C_y + \rho \frac{\mu_{12}}{\sigma_x \sigma_y \bar{Y}} \right\}^2}{n \{ (\beta_2 - 1) + 4C_y^2 - 4\gamma_1 C_y \}} \quad (4.2)$$

c) From(3.5), the estimator \tilde{y}_e based on estimated optimum \hat{k} has the mean square error

$$MSE(\tilde{y}_e) = (1 - \rho^2) \frac{\sigma_y^2}{n} - \frac{\bar{Y}^2 \left\{ 2C_y^2(1 - \rho^2) - \gamma_1 C_y + \rho \frac{\mu_{12}}{\sigma_x \sigma_y \bar{Y}} \right\}^2}{n \{ (\beta_2 - 1) + 4C_y^2 - 4\gamma_1 C_y \}} \quad (4.3)$$

d) From(4.3), we see that the estimator depending on estimated optimum value is always more efficient than the usual linear regression estimator in the sense of having lesser mean square error under the condition, $\gamma_1 < \left\{ \frac{(\beta_2 - 1)}{4C_y} \right\} + C_y$.

5. An Illustration

Considering the data given in Walpole R.E., Myers R.H., Myers 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). Computation of required values have been done and we have the following

$$\bar{Y} = 47.37581, \bar{X} = 10.58613, \sigma_y^2 = 27.46392, \sigma_x^2 = 1.86282, \mu_4 = 2523.46629,$$

$$\mu_3 = 59.71969, \beta_2 = 3.34559, \mu_{12} = -2.35772, C_y = 0.11, \gamma_1 = 0.041493,$$

$$\rho = -0.86219, n = 31$$

Using the required values, we have

$$MSE(\bar{y}_{lr}) = 0.22735 \quad (5.1)$$

$$MSE(\tilde{y}_e) = 0.19033 \quad (5.2)$$

From (5.1) and (5.2), the percent relative efficiency (PRE) of the predictive regression type estimator over the usual regression mean per unit estimator is **119%**.

References

1. Basu, D (1971). An easy on the logical foundations of survey sampling, Part I. Foundations of Statistical Inference, Ed. By V.P. Godambe and D. A. Sprott. New York, p. 203-233.
2. Kadilar, C. and Cingi, H (2006). An improvement in estimating the population mean by using the correlation coefficient, Hacettepe Journal of Mathematics and Statistics, 35(1), p. 103-109.
3. Singh, H.P. and Espejo, M R. (2003). On linear regression and ratio-product estimation of a finite population mean, The Statistician 52, Part 1, p. 59-67.
4. Srivastava, S.K. (1983). Predictive estimation of finite population mean using product estimator, Metrika, 30, p. 93-99.
5. Srivastava, S. K. and Jhajj, H. S. (1995). Classes of estimators of finite population mean and variance using auxiliary information, Journal of the Indian Society of Agricultural Statistics, 47, p. 119–128.
6. Sukhatme, P.V., Sukhatme, B. V., Sukhatme, S. and Ashok, C. (1984). Sampling Theory of Surveys with Applications, IOWA State University press (AMES), IOWA (U.S.A.) and Indian Society of Agricultural Statistics New Delhi-110012 (India).
7. Walpole R E, Myers R H, Myers S L and Ye K. (2005). Probability & Statistics for Engineers & Scientists, Pearson Education (Singapore) Pte. Ltd., Indian Branch, 482 F .I. E. Patparganj , Delhi-110092, (India).