International Journal of Data Science and Analysis
2018; 4(4): 53-57
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20180404.12
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Robust Estimation of Finite Population Totals Using a Model Based Approach in the Presence of Two Auxiliary Variables

Damaris Felistus Mulwa*, George Otieno Orwa, Romanus Odhiambo

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
mulwadamaris01@gmail.com (D. F. Mulwa)

*Corresponding author

To cite this article:
Damaris Felistus Mulwa, George Otieno Orwa, Romanus Odhiambo. Robust Estimation of Finite Population Totals Using a Model Based Approach in the Presence of Two Auxiliary Variables. *International Journal of Data Science and Analysis*. Vol. 4, No. 4, 2018, pp. 53-57.
doi: 10.11648/j.ijdsa.20180404.12

Received: October 16, 2018; **Accepted**: October 31, 2018; **Published**: November 14, 2018

Abstract: The utilization of auxiliary information during surveys increases the accuracy of estimators, thereby giving more reliable estimates of the population parameters of interest. It has been established that the presence of more than one auxiliary variables, some more robust estimators can be formed by combining different estimators like product, ratio or even regression estimators and in each case the individual estimators uses its own random variable. One of the most commonly used methods is the ratio method of estimating finite totals which is the foundation of all the other methods that use auxiliary information. In this paper, an estimator of the ratio-exponential class that uses two auxiliary variables has been proposed and its variance derived. After deriving the proposed estimator the coverage probabilities were estimated. Results showed that the interval length of the proposed estimator was narrower and tighter than that of the known Horwitz-Thompson's estimator. Two datasets from the agricultural and environmental sectors were used in order to investigate the properties of the estimator and they gave satisfactory results. Mean squared error criteria was used to investigate the performance of the proposed estimator and in both cases it had the minimum squared error values. The analysis in these paper is of very great importance in understanding environmental and agricultural data.

Keywords: Auxiliary Variable, Coverage Probabilities, Precision, Predictive Approach

1. Introduction

The main purpose of surveys which are conducted at Local, National and International levels is to gather information and aid public and private sectors in effective policy making [1]. Information regarding a study variable is obtained only for the sampled elements, the way how auxiliary information relates with the study variable across the sample allows inferences on the non-sampled portion of the population. Auxiliary information in the ratio type of estimator was first used in 1820 [2]. Since then, and as may be seen from studies that followed, it has emerged that the more robust estimators are the ones using auxiliary information [1]. In the recent years, researches have proposed estimators that are more efficient in estimating finite population totals using two auxiliary variables [4-5].

It has been established that the presence of multivariate auxiliary variables, some more robust estimators can be formed by combining up different estimators such as ratio, product or even regression estimators and in each case the individual estimators uses its own random variable [6]. Several researchers who have used auxiliary information in the estimation stage of parametric super population models include, Chambers and Danstan [7], Wang and Dorfman [8], Rao *et al* [9].

The use of auxiliary information in double sampling found out that the proposed estimators did perform better than the mean per unit estimator and also compared to the other estimators that don't utilize the auxiliary information and they are not asymptotically optimum with two auxiliary variables [10].

The use of local polynomial regression with two auxiliary

information to estimate population total has been investigated as well [11]. A super population approach was used where they assumed a working model and simulation was done. In all the models used, when the model is specified incorrectly, the local linear regression dominated the linear regression.

In this paper a new estimator is proposed and its variance derived, Mean Squared Error and the proposed estimator compared with the Horvitz Thompson estimator under the design based approach. The proposed estimator has been developed in relation to the motivation behind Cem Kadilar, Hulya Cingi [4-5]. Two datasets from the agricultural and environmental sectors were used in order to investigate the properties of the estimator and they gave satisfactory results.

2. The Proposed Estimator

2.1. Some Useful Information

Consider an auxiliary variable W_i, correlated with the variable of interest Z_i is obtained for every unit in the sample that has been drawn by simple random sampling from the study population and in addition the population mean of the W_i is known. The estimate of the mean of Z, the population mean of theZ_i, is:

$$\overline{Z_J} = \bar{z} + t(\overline{W} - \overline{w}) \quad (1)$$

Where t is an estimate of the change in Z when W is increased by one unit, $\bar{z}$, $\overline{w}$ are the means of w_i and z_i respectively. MSE of the estimate is given by:

$$MSE(\overline{Z_J}) = \frac{1-f}{n} s_z^2 (1 - \rho_{wz}^2) \quad (2)$$

Where $f = n/N$, n is size of the sample, N is size of the population, s_z^2 is the population variance of z_i, ρ_{wz} is the correlation coefficient between w and$z$.

Auxiliary information can be used at the sample stage as well as at the estimation stage. The regression estimate of the population mean when there are two auxiliary variables W_1 and W_2, will be:

$$\overline{zp} = \bar{z} + t1(\overline{W_1} - \overline{w_1}) + t2(\overline{W_2} - \overline{w_2}) \quad (3)$$

Where $t_1 = \frac{s_{zw1}}{s_{w1}^2}$ and $t_2 = \frac{s_{zw2}}{s_{w2}^2}$, s_{zw1} and s_{zw2} are the sample covariance between the study variable and the auxiliary information.

The MSE of the estimator is given by:

$$MSE(\overline{zp}) = \frac{1-f}{n} s_z^2 (1 - \rho_{zw1}^2 - \rho_{zw2}^2) + 2\rho_{zw1}\rho_{zw2}\rho_{w1w2}` \quad (4)$$

Survey variables are often estimated by the auxiliary variables, a super population approach is used whereby a model which is working relating the two auxiliary variables is used.

2.2. Sampling with One Auxiliary Variable

The use of auxiliary information in finite population increases the precision of the estimators of the population mean, Total or population distribution function. If a researcher in his sampling comes across an auxiliary variable, the first thing is always to think of how to utilize it in a more efficient manner. The auxiliary information may be correlated with the study character and may be put into use either at the design stage or estimation stage or sometimes at both stages. When sampling using one auxiliary variable, the regression estimate of $\bar{Z}$ the population mean of z_i is given by:

$$\bar{z}_{reg} = \bar{z} + t(\overline{W} - \overline{w}) \quad (5)$$

where t is an estimate of the change in z when w is increased by only one unit. From the foregoing, the estimate of the population total Z is given by

$$\bar{z}_{reg} = N\big(\bar{z} + t(\overline{W} - \overline{w})\big) \quad (6)$$

The MSE of the regression estimate is also given by:

$$MSE\big(\bar{z}_{reg}\big) = \frac{1-f}{n} s_z^2 (1 - \sigma_{zw}^2) \quad (7)$$

where $= \frac{n}{N}$, s_z^2 is the population variance of z and $\sigma_{zw} = \frac{s_{ZW}}{s_Z s_W}$ the population correlation coefficient between z and w.

2.3. Estimators in Literature Using Two Auxiliary Variables

An estimator for the population mean that relies on the assumption that the means of the two auxiliary variables are known was proposed by Abu-Dayyeh [12]. The proposed estimator is given by:

$$\bar{y} = \bar{y}\left(\frac{\overline{X_1}}{\overline{x_1}}\right)^{\alpha_1}\left(\frac{\overline{X_2}}{\overline{x_2}}\right)^{\alpha_2} \quad (8)$$

Where α_1 and α_2 are real numbers.

Motivated by his work, [4] proposed the estimator given by:

$$\bar{y} = \bar{y}\left(\frac{\overline{X_1}}{\overline{x_1}}\right)^{\alpha_1}\left(\frac{\overline{X_2}}{\overline{x_2}}\right)^{\alpha_2} + b_1(\overline{X_1} - \overline{x_1}) + b_2(\overline{X_2} - \overline{x_2}) \quad (9)$$

An exponential-ratio estimator which was proposed by [13] for estimating finite population mean is given by:

$$\overline{yBT} = \bar{y}_{exp}\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right) \quad (10)$$

Utilizing a linear combination of two auxiliary variables, Jinglu Lu [14] proposed an exponential ratio type estimator given by:

$$\overline{yclr} = \bar{y}_{exp}\left(\frac{\overline{Xlc} - \overline{xlc}}{\overline{Xlc} + \overline{xlc}}\right) \quad (11)$$

2.4. Proposed Estimator

On the lines of [4], an estimator which belongs to the exponential ratio regression class using two auxiliary variables for estimating population totals is given by:

$$\hat{Y}_{pr} = N\left[\bar{y}_{exp}\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right) + b_1(\overline{X_1} - \overline{x_1}) + b_2(\overline{X_2} - \overline{x_2})\right] \tag{12}$$

Further, this estimator is unbiased since

$$\begin{aligned} E\left[\hat{Y}_{pr}\right] &= E\left[N\left[\bar{y}_{exp}\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right) + b_1(\overline{X_1} - \overline{x_1}) + b_2(\overline{X_2} - \overline{x_2})\right]\right] \\ &= NE\left[\bar{y}_{exp}\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right) + b_1(\overline{X_1} - \overline{x_1}) + b_2(\overline{X_2} - \overline{x_2})\right] \\ &= N\bar{Y}_{exp}E\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right) + b_1E(\overline{X_1} - \overline{x_1}) + b_2E(\overline{X_2} - \overline{x_2}) = N\bar{Y}_{exp}(0) \\ &= Y \end{aligned} \tag{13}$$

In a similar manner, the variance of the estimator is derived as follows;

$$Var\left[\hat{Y}_{pr}\right] = E\left[N\left[\bar{y}_{exp}\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right) + b_1(\overline{X_1} - \overline{x_1}) + b_2(\overline{X_2} - \overline{x_2})\right]\right]^2 - E\left[\hat{Y}_{pr}\right]^2$$

But,

$$E\left[\hat{Y}_{pr}\right] = N\bar{Y} = E\left[N^2\bar{y}^2 exp\left\{2\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right)\right\} + \left(b_1(\overline{X_1} - \overline{x_1})\right)^2 + \left(b_2(\overline{X_2} - \overline{x_2})\right)^2\right]$$

$$+b_1(\overline{X_1} - \overline{x_1})b_2(\overline{X_2} - \overline{x_2}) = N\bar{Y}_{exp}\left(E\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right)\right) + b_1(\overline{X_1} - \overline{x_1}) + b_2(\overline{X_2} - \overline{x_2})$$

$$= N\bar{Y}_{exp}(0) = Y = N^2E\left(\bar{y}^2 exp\left(2\left(\frac{\overline{X_1}}{\overline{x_1}} - \frac{\overline{x_1}}{\overline{X_1}}\right)\left(\frac{\overline{X_2}}{\overline{x_2}} - \frac{\overline{x_2}}{\overline{X_2}}\right)\right)\right) + N^2\left(b_1^2\left(\frac{\sigma_1^2}{n} + \bar{X}_1^2\right) + b_1^2\left(\frac{\sigma_1^2}{n} + \bar{X}_1^2\right)\right)$$

$$N^2\left(b_1^2\left(\frac{\sigma_1^2}{n} + \bar{X}_1^2\right) + b_1^2\left(\frac{\sigma_1^2}{n} + \bar{X}_1^2\right) - 2b_1^2\bar{X}_1^2 + b_2^2\left(\frac{\sigma_2^2}{n} + \bar{X}_2^2\right) + b_2^2\left(\frac{\sigma_1^2}{n} + \bar{X}_2^2\right)\right)$$

$$= N^2\left(\frac{\sigma_y^2}{n} + \bar{Y}^2 + 2b_1^2\frac{\sigma_1^2}{n} + 2b_2^2\frac{\sigma_2^2}{n} - \bar{Y}^2\right) = N^2\frac{\sigma_y^2}{n} + 2N^2b_1^2\frac{\sigma_1^2}{n} + 2N^2b_2^2\frac{\sigma_2^2}{n}$$

Thus

$$Var\left[\hat{Y}_{pr}\right] = \frac{1}{f}\left(N\sigma_y^2 + 2Nb_1^2\sigma_{X1}^2 + 2Nb_2^2\sigma_{X2}^2\right) \tag{14}$$

3. Description of the Datasets and Studied Variables

3.1. The Datasets

To demonstrate the performance of the estimator give in equation (12) over the Horwitz-Thompson's estimator for finite population total, two datasets were used.

3.2. Description of Variables

Dataset I: Cost of milk in the US. Source (www.data.gov/food)

Y: Retail Cost
X_1: Farm Value
X_2: Farm Value Share

$N = 18$
$n = 10$
$\bar{X}_1 = 141.8465$
$\bar{X}_2 = 31.28$
$\bar{Y} = 126.2067$
$S_{x1}^2 = 421.7146$
$S_{x2}^2 = 15.04273$
$S_y^2 = 39.26279$
$\rho_{yx1} = 0.5834$
$\rho_{yx2} = 0.2846$
$\rho_{x1x2} = 0.9428$

Dataset II: The Trees Data

Y: Girth in inches
X_1: Height
X_2: Volume in Cubic feet

$N = 31$
$n = 18$
$\bar{X}_1 = 75$
$\bar{X}_2 = 19.577$
$\bar{Y} = 11.05$
$S_{x1}^2 = 44.70$
$S_{x2}^2 = 35.17$
$S_y^2 = 1.92$
$\rho_{yx1} = 0.88350$
$\rho_{yx2} = 0.6516$
$\rho_{x1x2} = 0.809158$

3.3. Graphical Relationships Between Variables

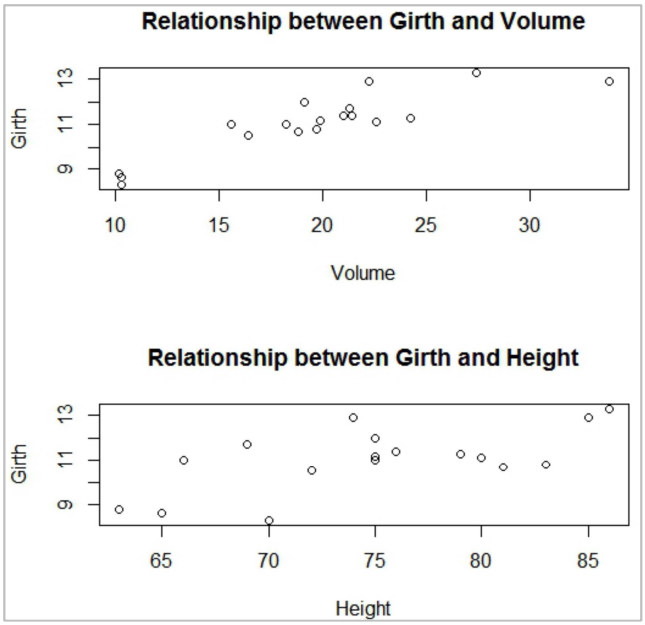

Figure 1. *Scatter plot exploring the relationships for dataset 1.*

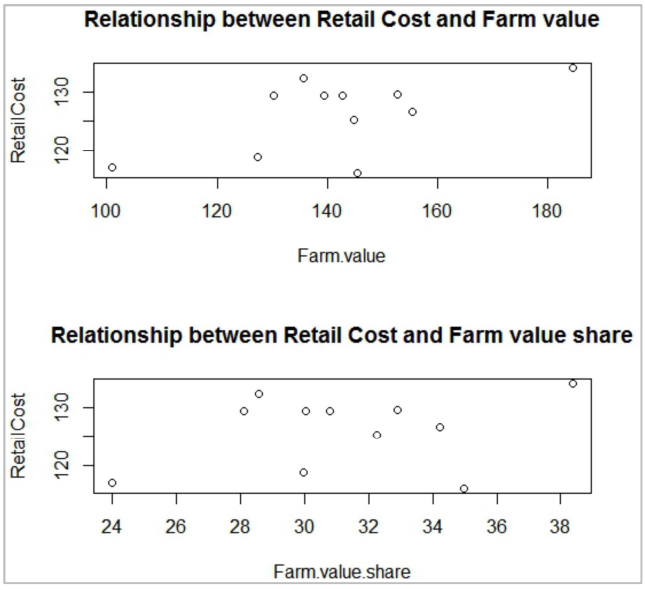

Figure 2. *Scatter plot exploring the relationships for dataset 1.*

The scatter plots to explore linear relationship of the different variables in the data was drawn. The relationship between each of the variables is linear and positive as indicated in figures 1 and 2. This observation concurs with the existing literature in that the auxiliary variables should be positively correlated with the study variable in question.

4. Results and Discussions

In order to use two auxiliary variables, the variables have to be positively correlated. As indicated in Figure 1, there is a positive relationship between the girth of the trees and the volume and height. This supports the requirement that auxiliary variable should be positively correlated with the study variable. The same conclusion of positive correlation between the study variables and auxiliary information can be deduced from Figure 2.

In order to compare the performance of the proposed estimator with the Horvitz Thompson estimator the mean squared error was investigated. As seen in Table 1, the proposed estimator gives the minimum variance as compared to the existing estimator under the two populations.

Table 1. *MSE of the existing estimator and proposed estimator.*

Estimator	Population 1	Population 2
Horvitz Thompson	92924.86	411.9107
Proposed Estimator	2138.294	350.0661

The confidence interval of the proposed estimator was calculated and the results tabulated in Table 2. According to researchers, a wider confidence width expresses a high level of uncertainty. As seen from Table 1, the confidence interval of the proposed estimator are narrower and tighter than the one compared to the Horvitz Thompson estimator at 95% coverage rate.

Table 2. *Confidence interval for the estimates.*

Population 1		
Estimator	Estimate	Interval Length
Horvitz-Thompson	127.8283	377.8784
Proposed estimator	401.5215	181.2675
Population 2		
Horvitz Thompson	11.20314	79.55868
Proposed Estimator	2252.198	73.6710

5. Conclusion

In this study, a ratio-regression type estimator using two auxiliary variables has been developed. The confidence intervals of the proposed estimator were tighter and narrower than those of the Horvitz Thompson estimator. The proposed estimator was found to be more efficient than the traditional design based Horvitz Thompson estimator. The comparison was made in terms of MSE and it was established that the MSE of the proposed estimator was smaller than that of the design based Horvitz Thompson estimator.

References

[1] S. C and L. S, Estimation in surveys with nonresponse, John Wiley & Sons, 2005.

[2] L. P. S, "Theorie analtytique des probabilities," *Courcier,* 1820.

[3] K. Cem and C. Hulya, "A new estimator using two auxilary variables," *Applied Mathematics and Computation,* vol. 162, no. 2, pp. 901-908, 2005.

[4] M. Sachin and J. Singh, "An improved estimator using two auxiliary attributes," *Applied Mathematical and Computation,* vol. 219, no. 23, pp. 10983-10986, 2013.

[5] K. M, A. O and I. A, "Use of auxiliary variables and asymptotically optimum estimators in double sampling," *International Journal of Statistics and Probability,* vol. 5, no. 3, 2016.

[6] R. E-H and Z. D, "Estimation of population tota; using local polynomial regression withtwo auxiliary variabkles," *Journal of Statistics Application and Probbility,* vol. 3, no. 2, 2014.

[7] D. Robson, "Applications of multivariate polykays to the theory of unbiased ratio-type estimation," *Journal of the American Statistical Association,* vol. 52, no. 280, pp. 511-522, 1957.

[8] M. H. e. a. Hansen, "Some history and reminiscences on survey sampling," *Statistical science,* vol. 2, no. 2, pp. 180-190, 1987.

[9] S. Bahl and R. Tuteja, "Ratio and Product Type exponential estimators," *Journal of Information and Optimization Sciences,* vol. 12, no. 1, pp. 159-164, 1991.

[10] J. Lu, Efficient estimator of a finite population mean using two auxiliary variables, biomedical, and power engineering, Mathematical problems in engineering, 2017.

[11] A. H. a. H. P. Dorfman, Estimators of the finite population distribution function using nonparametric regression., The Annals of Statistics, 1993.

[12] R. L. a. D. R. Chambers, "Estimating distribution functions from survey data," *Biometrika,* vol. 73, no. 3, pp. 597-604, 1986.

[13] W. A. A. M. A. R. a. M. H. A. Abu-Dayyeh, "Some estimators of a finite population mean using auxiliary information.," *Applied Mathematics and computation,* vol. 139, no. 2, pp. 287-298, 2003.

[14] J. K. J. a. M. H. Rao, "On estimating distribution functions and quantiles from survey data using auxiliary information.," *Biometrika,* pp. 365-375.

[15] H. P. a. E. M. R. Singh, "Double sampling ratio-product estimator of a finite population mean in sample surveys.," *Journal of Applied Statistics,* vol. 34, no. 1, pp. 71-85, 2007.