International Journal of Mathematics and Computational Science
Vol. 4, No. 3, 2018, pp. 112-117
http://www.aiscience.org/journal/ijmcs
ISSN: 2381-7011 (Print); ISSN: 2381-702X (Online)

Estimating Bounded Population Total Using Linear Regression in the Presence of Supporting Information

Lamin Kabareh[1, *], Thomas Mageto[2]

[1]Department of Mathematics and Statistics, Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI), Juja, Kenya
[2]Department of Mathematics and Statistical Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Abstract

Estimation of finite population total using linear regression in the presence of auxiliary information is considered. Model based on linear regression model is proposed. Like the existing estimators, this estimation technique deals with QR decomposition approach based on normalized yearly population totals in order to best fit in a model within a given period of time in this study. The proposed linear regression model technique has shown to be efficient. The empirical study indicated that the linear regression model is efficient and can estimate properly when the QR decomposition approach is applied even in the presence of outliers.

Keywords

Linear, Model, Estimation, Population Total, Supporting Information, QR Decomposition, Regression, Outliers

Received: April 2, 2018 / Accepted: May 21, 2018 / Published online: August 20, 2018

1. Introduction

Sample surveys are widely used as a cost effective tools of data acquisition and for making tangible conclusion about population parameters. Government offices and organizations use such approaches to get the recent information. The primary reason of a statistician in a sample survey is to get information about the population by deriving reliable estimates of unknown population parameters.

The use of information on a supporting variable holds a key idea in estimating population parameters such as mean or total. Ratio, product, and regression technique of estimation are common. The idea of supporting variable having different estimators along with their properties in simple random sampling procedure proposed by many found in the papers by [1-4]. However, in some day-today usage, the technique of systematic sampling has the merit of choosing the whole sample beginning with just random. Many writers have concentrated towards estimating population mean using ideas on single supporting variable in systematic sampling found in the papers by [5, 6].

This study is using estimation technique to estimate the finite population total with linear regression model that uses the technique of the QR decomposition approach to cater for ill conditioned models. The linear regression model is use for best fitting.

The sampling properties (e.g., bias and variance) of these estimators are well known and can be found in any textbook on sampling techniques. The idea of considering finite populations as a realization of super population represented by a model and invoked the classical least squares theory to discuss the bias of articles have appeared in the literature expounding on either the sampling design approach or the model-based approach, or a synthesis of the two approaches,

* Corresponding author
E-mail address: lkabareh@yahoo.com (L. Kabareh)

to investigate these and other similar estimators of $\bar{Y}$ can be found in the paper by [7]. An account of the underlying controversy can be found in the paper by [8] and its discussion by others.

In this article, we introduce the idea of a finite population total using the technique of the QR decomposition based on a regression fit to the population values and then make a prediction of a population total considering the errors of each estimator from the sampling design viewpoint. We show, by fitting a regression line of y on x to the finite population, that the leading term of the bias accounted for in terms of the intercept of the regression line.

Under simple random sampling (SRS) without replacement design, an exactly unbiased estimator for θ_{yx} was proposed by [9]. The proposed estimator is given by

$$\hat{\theta}_{HR} = \bar{r}_s + \frac{n(N-1)}{N(n-1)\bar{x}_u}(\bar{y}_s - \bar{r}_s\bar{x}_s) \quad (1)$$

where, $\bar{y}_s = \sum_{i\in s}\frac{y_i}{n}, \bar{r}_s = \sum_{i\in s}\frac{r_i}{n}, r_i = \frac{y_i}{x_i}, \bar{x}_s = \sum_{i\in s}\frac{x_i}{n}, \bar{x}_u = \frac{t_x}{N}, the\ population\ ratio$

$\theta_{yx} = \frac{t_y}{t_x}$, where $t_y = \sum_{i\in U} y_i$ be the population total for the variable Y, $t_x = \sum_{i\in U} x_i$ be the population total for the variable X and U of Nunits indexed by the set $\{1,2,\dots,N\}$ a finite population. This estimator can be rewritten under general sampling design p (·). In this case, this estimator is no longer unbiased but still with negligible bias according to [10].

Under general sampling design, an estimator for estimating the population ratio θ_{yx} was proposed by [11]. This estimator, has negligible relative bias especially for small sample sizes n and approaches zero with increasing n. Under SRS, and based on simulation results, the performance of this estimator is better than that of (1.1). Their estimator is defined by

$$\hat{\theta}_{JM} = \bar{r}_s + \frac{1}{\bar{x}_s}(\bar{y}_s - \bar{r}_s\bar{x}_s) \quad (2)$$

Define π_i, the first order inclusion probability, by

$$\pi_i = P_r(i^{th} element\ \epsilon\ s) = \sum_{i,j\epsilon s} P(s) \quad (3)$$

For $i \neq j$, the second order inclusion probability is defined by

$$\pi_{ij} = P_r(i^{th}\ and\ j^{th}\ elements\ \epsilon\ s) = \sum_{i,j\epsilon s} P(s) \quad (4)$$

The estimator for the population total $t_y = \sum_{i\in U} y_i$ is defined by

$$\hat{t}_{y\pi} = \sum_{i\epsilon U} y_i \frac{I_{\{i\epsilon s\}}}{\pi_i}, \quad (5)$$

where $I_{\{i\in s\}}$ is one if $i \in s$ and zero otherwise can be found on the paper by [12]. Further,

$$\bar{y}_s = \frac{1}{N}\hat{t}_{y\pi} \quad (6)$$

can be used to estimate the population mean $\bar{y}_u = \frac{1}{N}t_y$. It can be noted that $\hat{t}_{y\pi}$ and $\bar{y}_s$ are unbiased estimators for t_y, and $\bar{y}_u$ respectively. However, $\hat{t}_{y\pi}$ and $\bar{y}_s$ do not use the availability of auxiliary variables in the study. In similar way,

$$\bar{x}_s = \frac{1}{N}\hat{t}_{x\pi}, \text{ and } \bar{r}_s = \frac{1}{N}\hat{t}_{r\pi} \quad (7)$$

are unbiased estimators for $\bar{x}_u$ and $\bar{r}_u$ respectively. Where $\bar{x}_s$ is the sample mean of the inclusion probability of the auxiliary variable.

The availability of more than one auxiliary variable is used in literature for estimating the finite population total t_y, or finite population mean y_u.

Under SRS, estimating the population mean using more than one auxiliary variables was first dealt with by [13]. His estimator is given by

$$\hat{\bar{y}}_u = \sum_{i=1}^{P} w_i\bar{x}_{iu}\hat{\theta}_{yx} \quad (8)$$

where p is the number of the auxiliary variables, $\hat{\theta}_{yx_i} = \frac{\bar{y}_s}{\bar{x}_{is}} w_i$ is the weight of the ith auxiliary variable such that $\sum_{i=1}^{P} w_i = 1$ $\bar{y}_s$ is the sample mean of Y and $\bar{x}_{iu}, \bar{x}_{is}$ are the population mean and the sample mean of X_i, respectively, for $i = 1,\dots, p$. The following estimator

$$\hat{\bar{y}}_u = \bar{y}_s\left(w_1\frac{\bar{x}_{1u}}{\bar{x}_{1s}} + w_2\frac{\bar{x}_{2u}}{\bar{x}_{2s}}\right) \quad (9)$$

for estimating the population mean y_u, $w_1 + w_2 = 1$ was proposed by [14].

The general form of (9) can be traced in the paper by [15]. They proposed two classes of estimators using two auxiliary variables to estimate the population mean for the variable of interest Y. A new multivariate ratio estimator using the regression estimator instead of $\bar{y}_s$ which is used in (9) can be found in the paper by [16]. Their estimator is given by

$$\bar{y}_{pr} = \sum_{i=1}^{2} w_i\frac{\bar{y}_s + b_i(\bar{x}_{iu} - \bar{x}_{is})}{\bar{x}_{is}}\bar{x}_{iu} \quad (10)$$

where b_i, $i = 1,2$ are the regression coefficients. Based on the mean squares error (MSE), they found that their estimator is more efficient than (9) when

$$MSE\ (\bar{y}_{pr}) < MSE\ (\bar{y}_u),$$

where $MSE\ (\bar{y}_{pr})$, and $MSE\ (\bar{y}_u)$ are defined by Equations (2.4), and (1.2) of [17], respectively.

Subsection 2.1 introduced a general population model, while subsection 2.2 talked about the asymptotic properties and section 3.0 talked about the main results of the study. Finally, sections 4.0 and 5.0 crown it all with a discussion and

conclusion on the study respectively.

2. Estimation of Finite Population Total

This section is purposely considering an estimator, that is the linear regression model in the presence of auxiliary information.

2.1. Proposed Linear Regression Model

Let us consider a finite population U=$(U_1, U_2, \ldots\ldots, U_N)$ of size N units. A sample of size n units is drawn from the population U using the technique of successive sampling as $(x_1, y_1)\ldots.(x_n, y_n)$, "best" fit $y = a_0 + a_1 x + \cdots + a_m x^m$ to the data $(m \leq n-1)$ where x_i represents census year and y_i the corresponding population total. Suppose $y = a_0 + a_1 x$ to the data is considered which means $(n \geq 2)$, then;

$$S_r = \sum_{i=1}^{n}(y_i - (a_0 + a_1 x_i))^2 \quad (11)$$

$$= \sum_{i=1}^{n}(y_i - a_0 + a_1 x_i)^2 \quad (12)$$

Minimizing the residuals gives:

$$\frac{\partial S_r}{\partial a_0} = 0; \quad (13)$$

$$2\sum_{i=1}^{n}(y_i - a_0 + a_1 x_i)(-1) = 0 \quad (14)$$

$$na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (15)$$

$$\frac{\partial S_r}{\partial a_1} = 0; \quad (16)$$

$$2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i)(-x_i) = 0 \quad (17)$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad (18)$$

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \quad (19)$$

$$\hat{a} = (X'X)^{-1}X'Y \quad (20)$$

For numerical stability and efficiency, the QR decomposition should be used to calculate the estimate $\hat{a}$.

Definition:

For X, the QR decomposition is X=QR, where Q is the orthogonal matrix and R is the upper triangular matrix.

The quantities of interest are:

$$V_a = (X'X)^{-1} = ([Q\ R]'QR)^{-1} = (R'Q'QR)^{-1} = (R'R)^{-1} = R^{-1}[R']^{-1} \quad (21)$$

$$\hat{a} = (X'X)^{-1}X'Y = R'[R']^{-1}R'Q'Y = R^{-1}Q'Y \quad (22)$$

Table 1. Census Results.

#	YEAR	POPULATION TOTAL
0	1969	10,942,705
1	1979	15,327,061
2	1989	21,448,774
3	1999	28,686,607
4	2009	38,610,097

The five census years obtained from a sample frame is shown in Table 1 above. However, the aim is to normalize the values of the year and population from 1969 to 2009. The normalized sample sizes will be used to estimate the population total in 2019 census using the proposed technique.

We now use the normalized values to obtain our linear regression.

n= 5; $\sum_{i=1}^{n} x_i$=37.9768; $\sum_{i=1}^{n} x_i^2$=288.4477; $\sum_{i=1}^{n} y_i$=84.2754; $\sum_{i=1}^{n} x_i y_i$=640.1178

Substituting these values in matrix (19), the QR decomposition matrices can be obtained as follows in MATLAB:

[Q R]=qr(X)

$$Q = \begin{pmatrix} -0.1305 & -0.9914 \\ -0.9914 & 0.1305 \end{pmatrix}$$
$$R = \begin{pmatrix} 84.2754 \\ 640.1178 \end{pmatrix} \quad (23)$$

â can be calculated as follows:

$$\hat{a} = \text{inv}(R) * Q' * Y$$

$$\hat{a} = \begin{pmatrix} -499.5614 \\ 67.9910 \end{pmatrix} \quad (24)$$

Regression model:

Estimated Log(Y)=-499.5614+67.9910*Log(X)

With this model, the values of column two in Table 2 below can be used to obtain the values of column 6 on the same table. The estimated log(Y) values can be used to obtain the corresponding population totals in MATLAB as follows:

Estimated population total = exp(Estimated Log Y)
$= e^{\text{Estimated Log Y}}$

2.2. Asymptotic Properties

Theorem: Central Limit

Let $Y_1, Y_2, \ldots\ldots, Y_n$ be independent and identically distributed random variables with

$$E(Y_i) = \mu \text{ and } \text{Var}(Y_i) = \sigma^2 < \infty$$

$U_n = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then as $n \to \infty$, U_n converges in distribution to the standard normal.

Proof of Theorem:

Assuming the Moment Generating Function (MGF) of Y_i exists. Assuming $\mu = 0$, $\sigma = 1$ and consider $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Y_i-\mu}{\sigma}$. Let $S_n = \sum_{i=1}^{n} Y_i$, show MGF of $\frac{S_n}{\sqrt{n}}$ goes to $N(0,1)$

By definition, MGF is;

$$E\left(e^{tS_n/\sqrt{n}}\right) = E\left(e^{tY_1/\sqrt{n}}\right) \dots\dots\dots\dots E(e^{tY_n/\sqrt{n}}) \quad (25)$$

$$= (M(\tfrac{t}{\sqrt{n}}))^n \quad (26)$$

Taking the log of (26) then the limit gives:

$$\lim_{n\to\infty} nlogM(\tfrac{t}{\sqrt{n}}) = \lim_{n\to\infty} \frac{logM(\frac{t}{\sqrt{n}})}{\frac{1}{n}} \quad (27)$$

We let $x = \frac{1}{\sqrt{n}}$ then let x be real

$$= \lim_{x\to 0} \frac{logM(xt)}{x^2} \quad (28)$$

By using Hospital's rule results to;

$$= \lim_{x\to 0} \frac{tM\prime(xt)}{2xM(xt)} = \frac{t}{2}\lim_{x\to 0}\frac{M\prime(xt)}{x} = \frac{t^2}{2}\lim_{x\to 0}\frac{M\prime\prime(xt)}{1} = \frac{t^2}{2} \quad (29)$$

But the log of $e^{\frac{t^2}{2}}$ is exactly the $N(0,1)$ MGF

Note:

$M(t) = E(e^{tY_1})$, M(0)=1, $M'(0) = 0\ (mean), M''(0) = 1$ (variance)

3. Main Results

Table 2. Normalized data and estimated population.

Year(X)	Log(X)=x_i	Actual Population(Y)	Log(Y)= y_i	Estimated Population(Y)	Estimated Log(Y)
1969	7.5853	10,942,705	16.2082	10,540,000	16.1707
1979	7.5903	15,327,061	16.5451	14,808,000	16.5107
1989	7.5954	21,448,774	16.8812	20,945,000	16.8574
1999	7.6004	28,686,607	17.1719	29,426,000	17.1974
2009	7.6054	38,610,097	17.4690	41,342,000	17.5374
2019	7.6104			58,078,000	17.8773

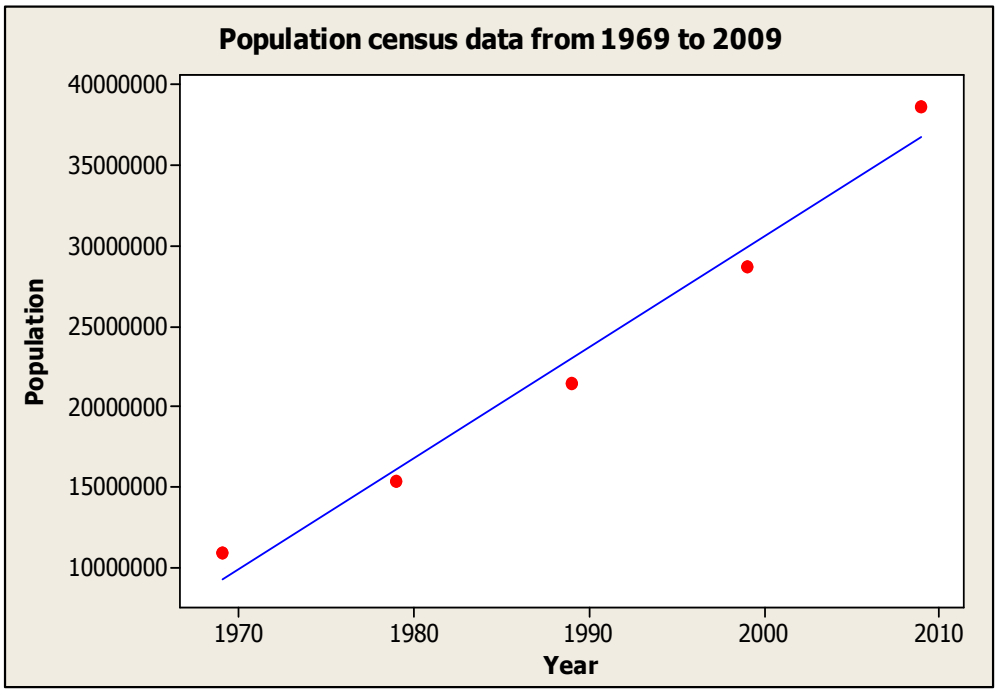

Figure 1. Best fitting line for the original census data.

Population=-1.34e+09+686943Year R^2=97.8%

Table 3. Actual and Estimated populations.

Year	Actual Population	Estimated Population
1969	10,942,705	12,590,767
1979	15,327,061	19,460,197
1989	21,448,774	26,329,627

Year	Actual Population	Estimated Population
1999	28,686,607	33,199,057
2009	38,610,097	40,068,487
2019	Pop=-1.34e+09+686943Year	46,937,917

Table 3 showed the actual and estimated population totals obtained from the regression model of the figure presented using MINITAB. This has projected the population total to be at 46,937,917 in the coming 2019.

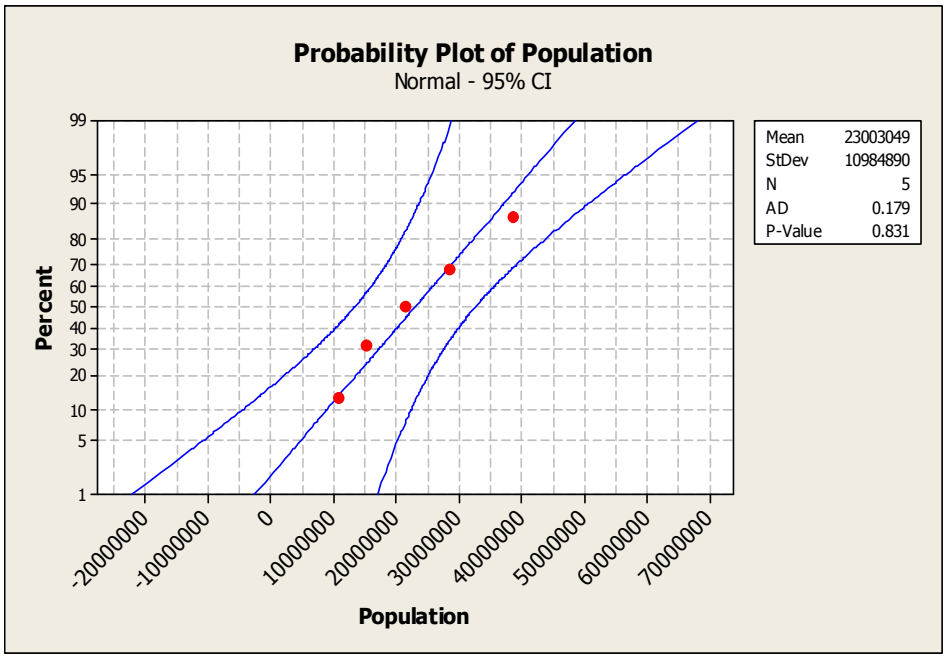

Figure 2. Test for normality plot.

Figure 2 above showed the test for normality plot with three of the points slightly deviating from the center line. Although the figure showed normality in the data but not a perfect normality which means there is room for improving the data in order to get a good regression. Therefore, a normalization technique will be applied to see the difference.

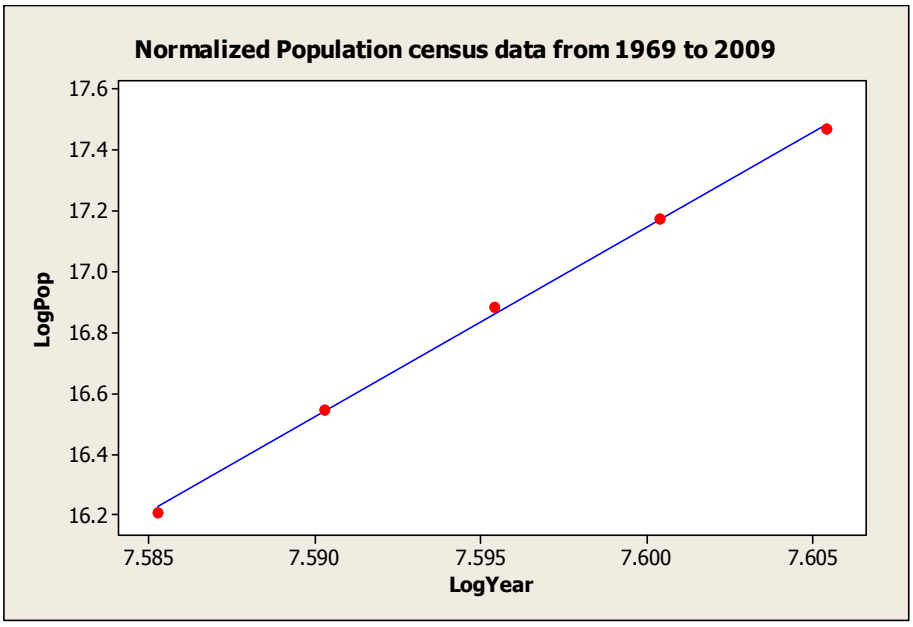

Figure 3. Best fitting line for the normalized census data.

LogPop=-458.6+62.60LogYear R^2=99.9%

Table 4. Actual and estimated populations obtained from normalized data.

YEAR	LOGYEAR	LOGPOP	ACTUAL POPULATION	ESTIMATED POPULATION
1969	7.5853	16.2398	10,942,705	11,294,000
1979	7.5903	16.5528	15,327,061	15,445,000
1989	7.5954	16.8720	21,448,774	21,253,000
1999	7.6004	17.1850	28,686,607	29,064,000
2009	7.6054	17.4980	38,610,097	39,745,000
2019	7.6104	17.8110		54,352,000

Table 4 above presented the normalized data that has been used to estimate the population totals using the regression model obtained from MINITAB as shown in figure 2 above.

4. Discussion

Previous studies used regression techniques to estimate population parameters but were faced with two tradeoffs, bias and variance. These two are a great concern especially in sample survey which is mostly attributed to the sample collected or technique used. Most of the regression techniques used do suffer from weak estimates. This distinction can be observed in figure 1 and table 3 compared to figure 2 and table 4 above with regard to the error margins.

5. Conclusion

In this work, the technique of normalizing the data prior to using it is very effective especially with the presence of outliers in trying to maintain precision. This has been manifested in figure 2 with an R^2=99.9% while the none normalized data has an R^2=97.8%. The QR decomposition can be more efficient in prediction especially where a regression model is ill conditioned.

Conflicts of Interest

The authors declare that they have no competing interests.

References

[1] Khan, M., Singh, R. 2015, Estimation of population mean in chain ratio-type estimator under systematic sampling. *Jour. Prob. Statist.* 2015, 1-5. DOI: 10.1155/2015/248374.

[2] Singh, H. P., Pal, S. K., Solanki, R. S., 2017, A new class of estimators of finite population mean in sample surveys. *Commun. Statist. Theo. Meth.,* 46 (6), 2630-2637.

[3] Singh, H. P., Pal, S. K., 2017a, A class of exponential type estimator of a general parameter. *Commun. Statist. Theo. Meth.* 46 (8), 3957-3984.

[4] Sharma, M. K., Barar, S. S., Kaur, H., 2017, Most efficient estimators of population mean using known information of population mode of auxiliary variable. *Inter. Jour. Appl. Math. Statist.* 56 (5), 108-117.

[5] Singh, H. P., Pal, S. K., 2015, A new chain ratio-ratio-type exponential estimator using auxiliary information in sample surveys. *Inter. Jour. Math. Appl.* 3 (4-B), 37-46.

[6] Singh, H. P., Pal, S. K., 2017b, A new family of estimators of the population variance using information on population variance of auxiliary variable in sample surveys. *Statist. Transit.-New Series,* 17 (4), 126.

[7] Royall, R. M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," *Biometrika,* 57, 377-387.

[8] Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). "An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys" (with discussion), *Journal* of *the American Statistical Association,* 78, 776-807.

[9] C. Henry Edwards, David E. Penney, 2008. Differential equations: Computing and modeling, 4th edition, 79-92.

[10] Hartley H, Ross A (1954). "Unbiased Ratio Estimates." *Nature*, 174, 270–271.

[11] Al-Jararha J (2012). *Unbiased Ratio Estimation for Finite Populations*. LAMBERT Academic Publishing, Germany.

[12] Al-Jararha J, Al-Haj Ebrahem M (2012). "A Ratio Estimator Under General Sampling Design." *Austrian Journal of Statistics*, 41, 105–115.

[13] Horvitz D, Thompson D (1952). "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association*, 47, 663–685.

[14] Olkin I (1958). "Multivariate Ratio Estimation for the Finite Populations." *Biometrika*, 45, 154–165.

[15] Singh D, Chaudhary F (1986). *Theory and Analysis of Sample Survey Design*. New Age Publication, New Delhi, India.

[16] Abu-Dayyeh W, Ahmad M, Ahmad R, Hassen A (2003). "Some Estimators of a Finite Population Mean Using Auxiliary Information." *Applied Mathematics and Computations*, 139, 287–298.

[17] Kadilar C, Cingi H (2004). "Estimator of a Population Mean Using Two Auxiliary Variables in Simple Random Sampling." *International Mathematical Journal*, 5, 357–367.