Uniport Journal of Engineering and Scientific Research (UJESR)

Vol. 6, Issue 1, 2021, Page 85-89
ISSN: 2616-1192

(www.ujesr.org)

A Modified Estimator for Population Parameters in the Presence of Outliers

Olayiwola, O.M*[1], Apantaku, F.S[2], Imarhia, F.O[3], Yusuf, K.M[4], Ogunsola, I.A[5] and Olawoore, S.A[6]
[1,2,3,4,5]Department of Statistics, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria
[6]Department of Statistics, Oyo State College of Agriculture and Technology, Igboora, Oyo State, Nigeria
*Corresponding author's email: olayiwolaom@funaab.edu.ng

Abstract
An outlier is one of the factors affecting socio-economic data, thus any wrong inference made on such data misleads and affects the decision of policy makers. This work proposed a modified regression estimator that made provision for outliers in order to improve the validity of the population parameters. Two data sets were considered in validating the efficiency of the proposed estimator, by using double sampling technique. Variance of the proposed estimator was derived up to first order approximation. The estimator was shown to be unbiased and efficient for large and small samples. The proposed estimator was theoretically and numerically shown to perform better than existing estimators.
Keywords: *Double sampling, Regression estimator, Outliers, Variance, Unbiased estimator*
Received: 21st April, 2021 *Accepted: 3rd September, 2021*

1. Introduction

Outliers are of frequent concern in surveys with quantitative variables like household budget surveys or business surveys on production or turnover. In some situations, a relatively small fraction of the data has extreme values in one or several variables. Often time, these extreme values occur when the bulk of the data has already a markedly skewed distribution. The traditional approach is to detect these extreme values on the basis of fixed univariate or bivariate limits, to review these observations manually and either "correct" the outliers, dismiss them from analysis or to leave them unchanged. These extreme values are usually detected as outliers.

Today the outlier detection or nomination methods use robust estimators. In these methods, imputations based on robust models are used to replace these values (outliers) and, finally, traditional non-robust estimators may be used on the data treated beforehand. The alternative to this editing and imputation approach is to use robust estimators directly on the raw data. The double sampling technique helps to provide information about the auxiliary variable which in turn helps to improve the estimation of the population parameter for the study variable.

Various authors have studied ways of estimating the population mean of the study variables using double sampling technique. Hansen and Hurwitz (1943) firstly incorporated auxiliary information in probability proportional to size sampling. Chand (1975) suggested two chain ratio-type estimators to estimate the population mean using two auxiliary variables. Bahl and Tuteja (1991) proposed exponential ratio and product-type estimators to estimate the population mean using single auxiliary variable. Singh and Espejo (2003) proposed a class of ratio-product estimators for estimating the unknown population parameter in double sampling. Khan and Shabbir (2013) proposed a ratio-type estimator for the estimation of population variance using the knowledge of quartiles and their functions as auxiliary information. They proposed different modified estimators for the estimation of finite population mean using maximum and minimum values. Recently Al-Hossain and Khan (2014) worked on the estimation of population mean using maximum and minimum values under simple random sampling by incorporating the knowledge of two auxiliary variables. Khan and Shabbir (2013), Khan (2015), Nasir*et al.* (2018), Cekim and Cingi (2016), Abid*et al.* (2016), Khan*et al.* (2015) and Darez*et al.* (2018) proposed some estimators for the estimation of finite population mean under minimum and maximum values using known information of the auxiliary variable. Sarndal (1972) suggested the following unbiased estimator for the estimation of finite population mean, which takes into account the presence of

extreme values in the samples. Khan (2015) extended the work of Sarndal (1972) and proposed the ratio type estimator to estimate the finite population mean of the study variable. This work proposed a modified regression estimator that made provision for outliers in order to improve the validity of the population parameters.

2. Materials and methods

The modified double sampling regression estimator that makes provision for the presence of outliers in a data set is given by:

$$\bar{y}_{m2} = \bar{y}_{C_{11}} + \beta\left(\bar{x}'_{C_{21}} - \bar{x}_{C_{21}}\right),$$

$$\bar{y}_{m2} = \begin{cases} (\bar{y}+c_1)+\beta\left[(\bar{x}'+c_2)-(\bar{x}+c_2)\right], & \text{if sample contains } y_{\min} \text{ and } x_{\min}, \\ (\bar{y}-c_1)+\beta\left[(\bar{x}'-c_2)-(\bar{x}-c_2)\right], & \text{if sample contains } y_{\max} \text{ and } x_{\max}, \\ \bar{y}+\beta(\bar{x}'-\bar{x}), & \text{for all other samples.} \end{cases} \quad (1)$$

The variance of $\bar{y}_{m2}$ up to first order of approximation by neglecting terms of e_j's having power greater than 2 is:

$$V\left(\bar{y}_{m2}\right) = \left\{\theta S_y^2 + \theta_2\beta^2 S_x^2 - 2\beta\theta_2 S_{yx}\right\} + \frac{2}{N-1}\left\{\begin{array}{l} C_1\left[n\theta\Delta y - \beta\Delta x\left(n\theta - n'\theta'\right)\right] \\ +C_2\left[\left(\beta^2\Delta x - \beta\Delta y\right)\left(n'\theta' - n\theta\right)\right] \\ +n^2\theta C_1^2 + \left(R^2C_2^2 - 2RC_1C_2\right)\left(n^2\theta - n'^2\theta'\right) \end{array}\right\}. \quad (2)$$

Let

$$e_0 = \frac{\bar{y}}{\bar{Y}} - 1,\ e_1 = \frac{\bar{x}}{\bar{X}} - 1,\ e_2 = \frac{\bar{x}'}{\bar{X}} - 1, \quad (3)$$

Expressing $\bar{y}_{m2}$ in terms of e_j's, yields

$$\bar{y}_{m2} = \bar{Y}\left(1+e_0\right) + \beta\left(e_2 - e_1\right)\bar{X}, \quad (4)$$

Since β is constant in repeated sampling then

$$E\left(\bar{y}_{m2}\right) = E\left(\bar{Y}(1+e_0)+\beta(e_2-e_1)\bar{X}\right) = \bar{Y}, \quad (5)$$

This implies that the proposed estimator is unbiased.

$$E\left(\bar{y}_{m2} - \bar{Y}\right)^2 = E\left[\bar{Y}^2e_0^2 + \beta^2\bar{X}^2\left(e_2^2 + e_1^2 - 2e_1e_2\right) - 2\beta\bar{Y}\bar{X}\left(e_0e_1 - e_0e_2\right)\right]. \quad (6)$$

Then,

$$Var\left(\bar{y}_{m2}\right) = \bar{Y}^2\left[\frac{\theta}{\bar{Y}^2}\left(S_y^2 - \frac{2nC_1}{N-1}\left(\Delta y - nC_1\right)\right)\right] + \beta^2\bar{X}^2\left[\begin{array}{l} \frac{\theta}{\bar{X}^2}\left(S_x^2 - \frac{2nC_2}{N-1}\left(\Delta x - nC_2\right)\right) \\ +\frac{\theta'}{\bar{X}^2}\left(S_x^2 - \frac{2n'C_2}{N-1}\left(\Delta x - n'C_2\right)\right) \\ -\frac{2\theta'}{\bar{X}^2}\left(S_x^2 - \frac{2n'C_2}{N-1}\left(\Delta x - n'C_2\right)\right) \end{array}\right] \quad (7)$$

$$-2\beta\bar{Y}\bar{X}\left[\begin{array}{l} \frac{2\theta'}{\bar{X}\bar{Y}}\left(S_{yx} - \frac{n'}{N-1}\left(C_2\Delta y + C_1\Delta x - 2n'C_1C_2\right)\right) \\ -\frac{2\theta'}{\bar{X}\bar{Y}}\left(S_{yx} - \frac{n'}{N-1}\left(C_2\Delta y + C_1\Delta x - 2n'C_1C_2\right)\right) \end{array}\right]$$

Therefore,

$$Var\left(\bar{y}_{m2}\right)=\left\{\theta S_y^2+\theta_2\beta^2 S_x^2-2\beta\theta_2 S_{yx}\right\}+\frac{2}{N-1}\begin{Bmatrix} C_1\left[n\theta\Delta y-\beta\Delta x\left(n\theta-n'\theta'\right)\right] \\ +C_2\left[\left(\beta^2\Delta x-\beta\Delta y\right)\left(n'\theta'-n\theta\right)\right] \\ +n^2\theta C_1^2+\left(R^2C_2^2-2RC_1C_2\right)\left(n^2\theta-n'^2\theta'\right)\end{Bmatrix}. \quad (8)$$

Differentiating the variance with respect to C_1 and C_2 gives the optimum values of C_1 and C_2. When Equation (8) is differentiated with respect to C_1 and C_2 , it becomes zero.

Hence,

$$Var\left(\bar{y}_{m2}\right)_{\min}=\left\{\theta S_y^2+\theta_2\beta^2 S_x^2-2\beta\theta_2 S_{yx}\right\}-\frac{1}{2n'\left(N-1\right)N}\begin{Bmatrix}\Delta y\left\{\left(N-n\right)\Delta y+2\beta\left(n-n'\right)\Delta x\right\} \\ +\dfrac{\left(n-n'\right)\left[\left(N-n\right)\Delta y-n'\beta\Delta x\right]^2}{n'\left(N-n'-n\right)}\end{Bmatrix}. \quad (9)$$

where

$$C_{1opt}=\frac{\Delta y}{2n'} \quad \text{and} \quad C_{2opt}=\frac{\left(N-n\right)\Delta y-n'\beta\Delta x\right)}{2n'\beta\left(N-n'-n\right)}. \quad (10)$$

3. Results and discussion

3.1 Comparison of estimators

Let the conventional unbiased estimator of the population mean be denoted by $\bar{y}$. Thus, $Var\left(\bar{y}_{m2}\right)_{\min}<Var\left(\bar{y}\right)$ if

$$\rho_{yx}>\frac{C_x}{2C_y}-\frac{1}{4n'\theta_2 N\left(N-1\right)\bar{Y}^2C_xC_y}\begin{Bmatrix}\Delta y\left[\left(N-n\right)\Delta y+2\beta\left(n-n'\right)\Delta x\right] \\ +\dfrac{\left(n-n'\right)\left[\left(N-n\right)\Delta y-n'\beta\Delta x\right]^2}{n'\left(N-n'-n\right)}\end{Bmatrix}. \quad (11)$$

Let $\bar{y}_{lr}$ denote the usual double sampling regression estimator of the population mean. Then:

$$Var\left(\bar{y}_{lr}\right)=\theta S_y^2+\theta_2\beta^2 S_x^2-2\beta\theta_2 S_{yx} \quad (12)$$

Therefore,

$Var\left(\bar{y}_{m2}\right)_{\min}<Var\left(\bar{y}_{lr}\right)$ if

$$\left[\left(y_{\max}-y_{\min}\right)\left(\left(N-n\right)\left(y_{\max}-y_{\min}\right)-2\beta\left(n'-n\right)\left(x_{\max}-x_{\min}\right)\right)-\frac{\left(n'-n\right)\left(\left(N-n\right)\left(y_{\max}-y_{\min}\right)-n'\beta\left(x_{\max}-x_{\min}\right)\right)^2}{n'\left(N-n'-n\right)}\right]>0 \quad (13)$$

Relative efficiency is given by:

$$\frac{Var\left(\bar{y}_{lr}\right)-Var\left(\bar{y}_{m2}\right)_{\min}}{Var\left(\bar{y}_{lr}\right)}\times 100\%. \quad (14)$$

3.2 Numerical illustration

In this section, the performance of the proposed estimator is presented, using two data sets. It is seen to perform better than various other estimators. The description and the necessary data statistics of the populations are given as follows:

Data set 1: The first set of data is on the enrolment of students into secondary schools with the corresponding number of staff in each school in Ogun state. The enrolment of students is the variable of interest while the number of staff is the auxiliary variable. The schools were divided into

four zones namely Egba, Yewa, Ijebu and Remo zones. There are 89, 91, 69, 53 secondary Schools at Egba, Yewa, Ijebu and Remo zones respectively, giving a total of 302 secondary schools under consideration. Table 1 presents the summary statistics for the enrolment of students into secondary schools in Ogun State.

Table 1: Summary statistics on the enrolment of students into secondary school

Parameters	**Y**	**X**
N	302	302
Population mean	1284.721	49.00
n' (First Phase sample)	200	200
n (Second phase sample)	124	124
Minimum	74	1
Maximum	9884	200
$\Delta(Range)$	9810	199

$$\rho_{xy} = 0.7981, \quad \beta = 28.56, \quad R = 26.22$$
$$S_y^2 = 2095373.531 \; S_x^2 = 1335.34$$
$$S_{yx} = 42216.494$$

Data set 2: The second data set is the inflation rate and external reserves of Nigeria from the year 1981-2015. Table 2 describes the relationship between the external Reserves and inflation rate.

Table 2: Correlation between inflation rate and external reserves of Nigeria

	External reserves	**Inflation**
External reserves	1	-0.36
Inflation	-0.36	1

The mean variances and relative efficiency of the proposed and the existing estimators is shown in Table 3.

Table 3: Variances of estimators for data set 1 and data set 2

Estimators	**Variances data set 1**	**Variances data set 2**
$\bar{y}_{lr}$	5907(100%)	10.9 (100%)
$\bar{y}_{m1}$	5497(6.9%)	
$\bar{y}_{(m2)_{\min}}$	5493(7.0%)	6.142(43.5%)

Relative efficiencies in parenthesis

4. Conclusions

In this work, a double sampling regression estimator that considers the presence of minimum and maximum values (outliers) in the data sets (large and small data sets) is proposed. The variances of the proposed estimator (5493 and 6.142 for first and second data sets, respectively) were lesser than the mean square error of the existing Khan (2015) ratio estimator (5904 and 10.876 for first and second data sets, respectively). Thus, the proposed estimator is found to be more efficient, both in the cases of large data set and small data set.

References

Abid, M., Sherwani, R.A.K., Abbas N. and Nawaz T. (2016) Some improved modified ratio estimators based on decile mean of an auxiliary variable. Pakistan Journal of Statistics and Operation Research, 12(4): 787-797.

Al-Hossain, Y. and Khan, M. (2014) Efficiency of ratio, product and regression estimators under maximum and minimum values using two auxiliary variables. Journal of Applied Mathematics, Vol. 2014: 1-6.

Bahl, S. and Tuteja, R.K. (1991) Ratio and product type exponential estimator. Information and Optimization Sciences, 12: 159-163.

Cekim, H.O. and Cingi, H. (2016) Some estimator types for population mean using linear transformation with the help of the minimum and maximum values of the auxiliary variable. Hacettepe Journal of Mathematics and Statistics,46(4): 1-10.

Chand, L. (1975) Some ratio type estimator based on two or more auxiliary variables. Ph.D Thesis, Lowa State University, Ames, Lowa.

Darez, U., Shabbir, J. and Khan, H. (2018) Estimation of finite population mean by using minimum and maximum values in stratified random sampling. Journal of Modern Applied Statistical Methods: 17(1): 7-19.

Hansen, M.H. and Hurwitz, W.N. (1943) On the theory of sampling from finite populations. The Annals of Mathematical Statistics,14: 333-362.

Khan, M. (2015) Improvement in estimating the finite population mean under maximum and minimum values in double sampling scheme. Journal of Statistics Applications & Probability Letters, 2(2): 115-121.

Khan, M. and Shabbir, J. (2013) A Ratio Type Estimator for the Estimation of Population Variance Using Quartiles and its Functions of an Auxiliary Variable, Journal of Statistics Applications and Probability, 2(3): 157-162.

Khan, M., Ullah, S., Al-Hossain, A.Y. and Bashir, N. (2015) Improved ratio-type estimators using

maximum and minimum values under simple random sampling scheme. Hacettepe Journal of Mathematics and Statistics, 44(4): 923-931.

Nasir, A., Abid, M., Tahir, M., Abbas, N. and Hussain, Z (2018) Enhancing ratio estimators for estimating population mean using maximum value of auxiliary variable. Journal of National Science Foundation Sri Lanka, 46(3): 453-463.

Singh, H.P. and Espejo, M.R. (2003) On linear regression and ratio-product estimation of a finite population mean. The Statistician, 1: 59-67.