Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online): 2229-5666
Vol. 12, Issue 2 (2019): 01-10

REGRESSION-IN-RATIO ESTIMATORS IN THE PRESENCE OF OUTLIERS BASED ON REDESCENDINGM-ESTIMATOR

[1]Aamir Raza, *[2]Muhmmad Noor-ul-Amin and [3]Muhammad Hanif
[1,3]National College of Business Administration & Economics, Lahore, Pakistan
[2]COMSATS University Islamabad-Lahore Campus, Pakistan
E Mail: *[2]nooramin.stats@gmail.com

Received March 09, 2019
Modified July 03, 2019
Accepted September 02, 2019

Abstract

In this paper, a robust redescending M-estimator is used to construct the regression-in-ratio estimators to estimate population when data contain outliers. The expression of mean square error of proposed estimators is derived using Taylor series approximation up to order one. Extensive simulation study is conducted for the comparison between the proposed and existing class of ratio estimators. It is revealed form the results that proposed regression-in-ratio estimators have high relative efficiency (R.E) as compared to previously developed estimators. Practical examples are also cited to validate the performance of proposed estimators.

Key Words: Redescending, Ratio Estimator, Robust Regression, Outliers, Auxiliary Information.
Mathematics Subject Classification: 62D05

1. Introduction

The ratio or regression method of estimation is used to increase the efficiency of estimator when auxiliary information is used in estimation procedure. Kadilar and Cingi (2004) used coefficient of kurtosis and coefficient of variation of auxiliary variable to increase the performance of regression-in-ratio estimators to estimate the population mean.

It is known phenomena that ordinary least square (OLS) estimator does not perform well in the presence of outliers. In this situation, redescending M-estimators are used to reduce the effect of outliers. Various authors have discussed the M-estimators such as Huber (1964), Andrew et al. (1972, 1974), Roisseeuw and Leroy (1987), Hample et al. (1986), Beaton and Tukey (1974), Qadir (1996),Insha et al. (2006), Khalil et al. (2016) and Noor-ul-Amin et al. (2018). Some authors have suggested ratio type estimators using the M-estimators in the presence of outliers such as Noor-ul-Amin (2016), Subzar et al. (2019) and Zaman (2019).

We used a function (1.1) in terms of r, where r is the error term obtained from OLS regression line i.e. $y = a_o + b_o x + r$. This $\rho_1(r)$ function has following form

$$\rho_1(r)=\frac{v^2}{2c}\left[1-\left\{1+\left(\frac{r}{v}\right)^2\right\}^{-c}\right] \qquad |r|\geq 0 \tag{1.1}$$

where v and c are tuning constants which control the robustness of estimator and for current study, optimum values are $c=2.5$ and $v=8$. Above $\rho_1(r)$ function is used to develop robust regression-in-ratio estimators in SRS design to estimate the population mean. This study deals with the regression-in-ratio estimators that are useful when data contain some outliers. In this regards redescending M-estimator is used to improve the performance of ratio type estimators. In second section, Kadilar and Cingi (2004) estimators are discussed. Section 3 consists in regression-in-ratio estimators based on Huber (1964) M-estimator. In section 4, proposed regression-in-ratio estimators based on (1.1) are discussed. Comparative study of proposed estimators based on theoretical illustrations of real data and simulation study is presented in section 5. Conclusions drawn based on results obtained in section 5 are presented in section 6.

2. Regression-in-ratio Estimators based on OLS Method

In simple random sampling (SRS), Kadilar and Cingi (2004) developed following regression-in-ratio estimators for population mean using information of supporting variable. It was concluded that suggested estimators are more efficient than OLS estimators.

$$\bar{y}_{KCi}=\frac{\bar{y}+b\left(\bar{X}-\bar{x}\right)}{\left(\alpha_i\bar{x}+\beta_i\right)}\left(\alpha_i\bar{X}+\beta_i\right) \tag{2.1}$$

where b is the OLS estimator of regression coefficient and $i=1,2,3,4,5$,

$$\alpha_1=1\&\beta_1=0,\ \alpha_2=1\&\beta_2=C_x,\ \alpha_3=1\&\beta_3=B_2(x),\ \alpha_4=B_2(x)\&\beta_4=C_x \text{ and}$$
$$\alpha_5=C_x\&\beta_5=B_2(x) \tag{2.2}$$

where C_x and $B_2(x)$ are the co-efficient of variation and coefficient of kurtosis of auxiliary variable respectively , sample mean of study variable is $\bar{y}$ and sample mean of supporting variable is and $\bar{x}$ .The $b=\frac{s_{yx}}{s_x^2}$ is calculated by OLS method, where s_x^2 is sample variance of x and s_{yx} is sample covariance between x and y. The mean square error (*MSE)* of estimator given in (2.1) can be obtained by using Taylor series approximation up-to order one, and is given as

$$MSE\left(\bar{y}_{KCi}\right)\cong\frac{1-f}{n}\left(R_{KCi}^2S_x^2+2BR_{KCi}S_x^2+B^2S_x^2-2R_{KCi}S_{yx}-2BS_{yx}+S_y^2\right) \tag{2.3}$$

for further details see Kadilar and Cingi (2004), where $f=\frac{n}{N}$, *n and N* are sample size and the population size respectively and

$$R_{KC1}=R=\frac{\bar{Y}}{\bar{X}} \quad , \qquad R_{KC2}=\frac{\bar{Y}}{\bar{X}+C_x}, \quad R_{KC3}=\frac{\bar{Y}}{\bar{X}+B_2(x)}$$

$$R_{KC4} = \frac{\bar{Y}\, B_2(x)}{\bar{X}\, B_2(x) + C_x} \text{ and } R_{KC5} = \frac{\bar{Y}\, C_x}{\bar{X}\, C_x + B_2(x)}.$$

It is important to note that $E(b) = B$. Kadilar and Cingi (2004) showed that estimators in (2.1) are more efficient than traditional estimators given by Sisodia and Dwivedi (1981) and Upadhayaya and Singh (1999).

3. Ratio Estimators based on Huber M-Estimators

Kadilar et al. (2007) suggested following class of robust regression-in-ratio estimators for the population mean using robust regression instead of OLS method.

$$\bar{y}_{robi} = \frac{\bar{y} + b_{rob}(\bar{X} - \bar{x})}{(\alpha_i \bar{x} + \beta_i)}\left[\alpha_i \bar{X} + \beta_i\right] \tag{3.1}$$

where b_{rob} is calculated using Huber M-estimator of robust regression and $i = 1,2,3,4,5$ The estimator in (3.1) is more efficient than estimator in (2.1) when data consist of some outliers. Huber (1964) discussed following $\rho_2(r)$ function, where r is the error term of OLS model.

$$\rho_2(r) = \begin{cases} r^2 & -v \le r \le v \\ 2v|r| - r^2 & r < -v \; or \; v < r \end{cases}$$

where v is tuning constant. Huber (1964) advised $v = 1.5s$, where s is the estimate of population standard deviation of error term is. Estimate of b_{rob} is found by minimizing the following expression with respect to b.

$$\sum_{i=1}^{n} \rho_2(y_i - a - bx_i)$$

The *MSE* of $\bar{y}_{robi}$ is obtained by replacing B by B_{rob} in (2.3) and is given as

$$MSE(\bar{y}_{robi}) \cong \frac{1-f}{n}\left(R_{KCi}^2 S_x^2 + 2B_{rob}R_{KCi}S_x^2 + B_{rob}^2 S_x^2 - 2R_{KCi}S_{yx} - 2B_{rob}S_{yx} + S_y^2\right) \tag{3.2}$$

4. Proposed Regression-in-ratio estimators

Huber (1964) M-estimatoris are not flexible to weight the larger residuals. To overcome this deficiency, the redescending M-estimator discussed in (1.1) is used. On the basis of $\rho_1(r)$ given in (1.1), following robust regression-in-ratio are proposed estimator to estimate the population mean when data contain outliers.

$$\bar{y}_{proi} = \frac{\bar{y} + b_{pro}(\bar{X} - \bar{x})}{(\alpha_i \bar{x} + \beta_i)}\left[\alpha_i \bar{X} + \beta_i\right] \tag{4.1}$$

where b_{pro} is obtained by minimizing $\sum_{i=1}^{n} \rho_1(y_i - a - bx_i)$.

The *MSE* of $\bar{y}_{proi}$, where $i = 1,2..,5$ is found by replacing B_{rob} by B_{pro} in(3.2) and is given by

$$MSE(\bar{y}_{proi}) \cong \frac{1-f}{n}\left(R_{KCi}^2 S_x^2 + 2B_{pro}R_{KCi}S_x^2 + B_{pro}^2 S_x^2 - 2R_{KCi}S_{yx} - 2B_{pro}S_{yx} + S_y^2\right) \tag{4.2}$$

5. Comparative Study

A comparative study of proposed regression-in-ratio estimators with considered estimators is performed in terms of *R.E* using *MSE*. The comparisons are made on the basis of two practical examples. A simulation study is included to validate the performance of proposed estimators.

5.1 Example 1

We have used the data from Kadilar et al. (2007)about the production of apple '*y*' in tons as study variable and numbers of apple trees (x, 1 unit = 100 trees) as auxiliary variable in 204 villages of the Karadeniz Region in Turkey. Scatter plot of collected data is dawn to see the presence of outliers and shown in Figure 1.

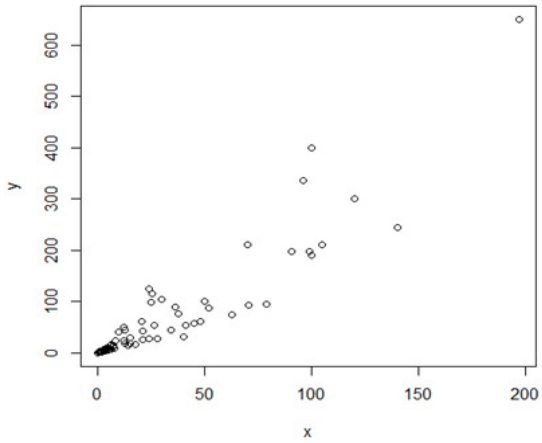

Figure 1: Production of Apples and Numbers of Apple Trees

It is examined from Figure 1 that data contain outliers and one can expect to obtain efficient estimates of population mean using proposed estimators. A random sample of size 30is selected from the population to calculate estimates. Note that sample size does not affect the efficiency comparison as it is not involved in efficiency expression. It is noted that there is a high correlation between apple production and number of apple trees i.e. 0.713. Statistics regarding the example 1 are given in Table 1.

$N = 204$ $n = 30$ $\rho = 0.713$ $B_{rob} = 3.547$ $B = 3.753$ $B_{pro} = 2.484$	$S_{yx} = 773727.8$ $\bar{X} = 264.42$ $\bar{Y} = 966.96$ $S_x = 454.03$ $S_y = 2389.77$ $B_2(x) = 29.77$	$C_x = 1.72$ $R_{KC1} = 3.6333$ $R_{KC2} = 3.6333$ $R_{KC3} = 3.2868$ $R_{KC4} = 3.6561$ $R_{KC5} = 3.4319$

Table 1: Statistics Regarding Example 1

The *MSE* of estimator $\bar{y}_{KCi}$, $\bar{y}_{robi}$ and $\bar{y}_{proi}$ are obtained by using (2.3), (3.2) and (4.2). The *R.Es* of proposed regression-in-ratio estimators to estimators given in (2.1) and (3.1) are presented in Table 2 and 3 respectively using (5.1)

$$R.E\left(\bar{y}_{proi}\right)=\frac{MSE\left(\bar{y}_{ti}\right)}{MSE\left(\bar{y}_{proi}\right)}, i=1,2,..5 \; and \; t=1,2 \qquad (5.1)$$

where $\bar{y}_{1i}=\bar{y}_{KCi}$ *and* $\bar{y}_{2i}=\bar{y}_{robi}$

R.E	$\bar{y}_{KC1}$	$\bar{y}_{KC2}$	$\bar{y}_{KC3}$	$\bar{y}_{KC4}$	$\bar{y}_{KC5}$
$\bar{y}_{pro1}$	1.406	1.397	1.273	1.406	1.323
$\bar{y}_{pro2}$	1.415	1.406	1.280	1.414	1.331
$\bar{y}_{pro3}$	1.535	1.526	1.389	1.535	1.445
$\bar{y}_{pro4}$	1.407	1.398	1.273	1.406	1.324
$\bar{y}_{pro5}$	1.485	1.475	1.343	1.484	1.397

Table 2: *R.E* of Proposed Estimator w.r.t. Kadilar & Cingi (2004)

It is revealed from Table 2 that proposed estimators are superior to estimators suggested by Kadilar and Cingi (2004) as outliers highly affected the performance of these estimators and proposed estimators reduced the influence of outliers up to a significant level to enhance the efficiency.

R.E	$\bar{y}_{rob1}$	$\bar{y}_{rob2}$	$\bar{y}_{rob3}$	$\bar{y}_{rob4}$	$\bar{y}_{rob5}$
$\bar{y}_{pro1}$	1.330	1.322	1.204	1.330	1.252
$\bar{y}_{pro2}$	1.338	1.329	1.211	1.337	1.259
$\bar{y}_{pro3}$	1.452	1.443	1.314	1.452	1.366
$\bar{y}_{pro4}$	1.330	1.322	1.204	1.330	1.252
$\bar{y}_{pro5}$	1.404	1.395	1.271	1.404	1.321

Table 3: *R.E* of Proposed Estimator w.r.t. to Kadilar et al. (2007)

The drawback of Huber robust M-estimator can be seen in Table 3as it failed to reduce the influence of large residuals and presence of outliers reduced the efficiency of estimators based on robust M-estimator suggested by Kadilar et al. (2007). A close look at Table 2 and 3 showed that performance of $\bar{y}_{proi}$ is better than $\bar{y}_{KCi}$ and $\bar{y}_{robi}$ considering the *MSE* of the estimators for all values of *i.*

5.2 Example 2

The data regarding the U.S. State Public-School Expenditures is used from Fox (2008). This data consists of 51 observations indicating the per-capita income in dollars and per-capita education expenditure in dollars for the U. S. states in 1970. The Per-capita income is taken as independent variable and per-capita education expenditures is taken as dependent variable. The original data was free from extreme values so 6% outliers are injected in original data to validate the performance of estimators. The scatter plot of income and expenditures with outliers is given in Figure 2.Statistics regarding the population used in Example 2 after inclusion of outliers are given in Table 4. The *R.E* of proposed regression-in-ratio estimators to the estimators given in (2.1) is given in table 5.

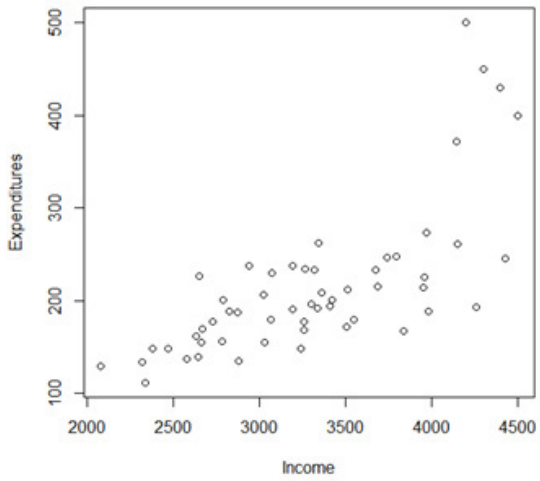

Figure 2: Income and Expenditures with 6% Outliers

$N = 54$ $n = 5$ $\rho = 0.7179$ $B_{rob} = 0.06539$ $B = 0.08223$ $B_{pro} = 0.05383$	$S_{yx} = 30129.52$ $\bar{X} = 3288.7040$ $\bar{Y} = 208.9259$ $S_x = 605.3097$ $S_y = 69.3353$ $B_2(x) = 2.2168$	$C_x = 0.1840$ $R_{KC1} = 0.06353$ $R_{KC2} = 0.06353$ $R_{KC3} = 0.06349$ $R_{KC4} = .063526$ $R_{KC5} = 0.06330$

Table 4: Statistics of data for example 2 with 6% outliers

R.E	$\bar{y}_{KC1}$	$\bar{y}_{KC2}$	$\bar{y}_{KC3}$	$\bar{y}_{KC4}$	$\bar{y}_{KC5}$
$\bar{y}_{pro1}$	1.369	1.368	1.368	1.368	1.365
$\bar{y}_{pro2}$	1.369	1.369	1.368	1.369	1.365
$\bar{y}_{pro3}$	1.369	1.369	1.368	1.369	1.365
$\bar{y}_{pro4}$	1.369	1.368	1.368	1.369	1.365
$\bar{y}_{pro5}$	1.371	1.371	1.371	1.371	1.368

Table 5: *R.E* of Proposed Estimator w.r.t. Kadilar & Cingi (2004)

Table 5 shows that suggested estimators out fit the estimators given by Kadilar and Cingi (2004)for population means under SRS for the data represented in Table 4. For example, estimator $\bar{y}_{pro1}$ is 137% efficient than the estimator $\bar{y}_{KC4}$ as its efficiency is 1.37 relative to estimator given in (2.1).

R.E	$\bar{y}_{rob1}$	$\bar{y}_{rob2}$	$\bar{y}_{rob3}$	$\bar{y}_{rob4}$	$\bar{y}_{rob5}$
$\bar{y}_{pro1}$	1.124	1.124	1.124	1.124	1.121
$\bar{y}_{pro2}$	1.124	1.124	1.124	1.124	1.121
$\bar{y}_{pro3}$	1.125	1.124	1.124	1.125	1.122
$\bar{y}_{pro4}$	1.124	1.124	1.124	1.124	1.122
$\bar{y}_{pro5}$	1.126	1.126	1.126	1.126	1.124

Table 6: *R.E* of Proposed Estimators w.r.t. Kadilar et al. (2007)

It is depicted from Table 6, when contamination of outliers is 6%, previously exiting ratio-in-regression estimators given by Kadilar et al. (2007) produced efficient results as compare to the estimators given by Kadilar and Cingi (2004) but their efficiencies are still significantly lower than proposed robust estimators given in (4.1). For example *R.E* of $\bar{y}_{pro1}$ is 112% relative to $\bar{y}_{rob4}$. It can be concluded that proposed estimators worked efficiently as compare to all considered estimators.

5.3 Simulation Study

To conduct simulation study, simple random samples of size 20, 30, 40 and 50 are selected using SRS and *R.E* of proposed regression-in-ratio estimators with respect to considered estimators are obtained for each sample and presented in Table 4 using (5.1).The R-programming is used to take samples of different sizes using SRS without replacement from the population defined in example 1 and for each sample, 50000 iterations are carried out to obtained the *MSE* of $\hat{\bar{Y}}$ using following formula

$$MSE(\hat{\bar{Y}}) = \frac{1}{50000}\sum_{k=1}^{50000}\left(\hat{\bar{Y}}_k - \bar{Y}\right)^2 \tag{5.1}$$

where $\hat{\bar{Y}} = \bar{y}_{KCi}, \bar{y}_{robi}, \bar{y}_{proi}$ and the population mean of study variable is represented by $\bar{Y}$.

It is revealed from above table that suggested regression-in-ratio estimators have high *R.E* than Kadilar and Cingi (2004) for each sample size. These results also verify the theoretical results obtained in section 5 shown in Table 2-3. As expected, by increasing the sample size the performance of traditional estimators is slightly improved but still remains lower than proposed estimators. From Table 7, it is also revealed that efficiency of proposed estimators is better than the Kadilar et al. (2007).

n	Estimator	$\bar{y}_{rob1}$	$\bar{y}_{rob2}$	$\bar{y}_{rob3}$	$\bar{y}_{rob4}$	$\bar{y}_{rob5}$
20	$\bar{y}_{pro1}$	1.602	1.582	1.334	1.602	1.427
	$\bar{y}_{pro2}$	1.620	1.599	1.348	1.619	1.442
	$\bar{y}_{pro3}$	1.842	1.819	1.538	1.841	1.644
	$\bar{y}_{pro4}$	1.603	1.583	1.334	1.602	1.427
	$\bar{y}_{pro5}$	1.760	1.737	1.465	1.759	1.567
30	$\bar{y}_{pro1}$	1.531	1.515	1.317	1.531	1.393
	$\bar{y}_{pro2}$	1.544	1.528	1.328	1.544	1.405
	$\bar{y}_{pro3}$	1.735	1.717	1.489	1.735	1.577
	$\bar{y}_{pro4}$	1.532	1.516	1.317	1.531	1.394
	$\bar{y}_{pro5}$	1.653	1.636	1.422	1.653	1.505
40	$\bar{y}_{pro1}$	1.491	1.478	1.301	1.491	1.370
	$\bar{y}_{pro2}$	1.503	1.489	1.311	1.502	1.381
	$\bar{y}_{pro3}$	1.667	1.652	1.454	1.667	1.531
	$\bar{y}_{pro4}$	1.492	1.478	1.301	1.491	1.370
	$\bar{y}_{pro5}$	1.600	1.585	1.396	1.600	1.470
50	$\bar{y}_{pro1}$	1.473	1.460	1.292	1.472	1.358
	$\bar{y}_{pro2}$	1.484	1.471	1.302	1.483	1.368
	$\bar{y}_{pro3}$	1.638	1.624	1.438	1.638	1.511
	$\bar{y}_{pro4}$	1.473	1.460	1.292	1.472	1.359
	$\bar{y}_{pro5}$	1.576	1.563	1.383	1.576	1.454

Table 7: *R.E* of Proposed Estimators w.r.t. Kadilar and Cingi (2004)

6 Conclusion

Form the theoretical and simulation results shown in Section 5, it is concluded that proposed robust regression-in-ratio estimators are more efficient to estimate population mean than estimators developed by Kadilar and Cingi (2004) and Kadilar et al. (2007) in SRS when data have outliers. The results showed that population mean can be estimated more efficiently by using proposed regression-in-ratio estimators. In future studies, one can propose regression-in-ratio estimators in various other sampling designs such as stratified sampling and two phase sampling.

Acknowledgements

The authors are very obliged to the reviewers for their valuable suggestions to enhance the quality of this manuscript. The authors are also thankful to prof. Cem Kadilar for providing the data of example 1.

References

1. Andrews, D.F.(1974). A robust method for multiple linear regressions, Technometrics, 16, p. 523-531.
2. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972).Robust Estimates of Location.Survey and Advances, Princeton University Press.
3. Beaton, A.E. and Tukey, J.W.(1974). The fitting of power series, meaning polynomials, illustrated on banned-spectroscopic data, Technometrics, 16, p. 147-185.
4. Fox, J. (2008). Applied Regression Analysis and Generalized Linear Models, Second Edition. Sage.
5. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W. A.1986). Robust Statistics, The Approach Based on Influence Functions, New York: John Wiley and Sons.
6. Huber, P.J.(1964). Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35(1), p. 73-10.
7. InshaUllah, Qadir, M.F. and Ali, A. (2006). Insha's redescending M-estimator for robust regression: A comparative study, Pakistan Journal of Statistics and Operation Research, 2, p. 135-144.
8. Kadilar, C. and Cingi, H.(2004). Ratio estimators in simple random sampling, Applied Mathematics and Computation, 151, p. 893-902.
9. Kadilar, C., Candan, M. and Cingi, H.(2007). Ratio estimation using robust regression, Hacettepe Journal of Mathematics and Statistics, 36, p. 181-188.
10. Khalil, U., Alamgir, Amjad, A. and Khan, D.M., (2016). Efficient Uk's re-descending M-estimator for robust regression, Pakistan Journal of Statistics, 32(2), p. 125-138.
11. Noor-ul-Amin, M., Asghar, S.U.D., Sanaullah, A., and Shahzad, M.A. (2018). Redescending M-estimator for robust regression, Journal of Reliability and Statistical Studies, 11(2), p. 69-80
12. Noor-Ul-Amin, M., Shahbaz, M.Q. and Kadilar, C. (2016). Ratio estimators for population mean using robust regression in double sampling, Gazi University Journal of Science, 29 (4), p. 793-798.

13. Qadir, M.F.(1996). Robust method of detection of single and multiple outliers, Scientific Khyber, 9(2**)**, p. 135-144.
14. Rousseeuw, P.J. and Leroy, A.M.(1987).Robust Regression and Outlier Detection, Wiley-Interscience, New York.
15. Sisodia, B.V.S. and Dwivedi, V.K.(1981). A modified ratio estimator using coefficient of variation of auxiliary variable, Journal of Indian Society of Agricultural Statistics, 33, p. 13-18.
16. Subzar, M., Bouza, C. N., Maqbool, S., Raja, T.A. and Para, B.A.(2019). Robust ratio type estimators in simple random sampling using Huber M estimation, Revista Investigacion Operacional, l40(2)**,**p. 201-209.
17. Upadhyaya, L.N. and Singh, H.P. (1999). Use of transformed auxiliary variable in estimating the finite populations mean, Biometrical Journal, 41, p. 627-636.
18. Zaman, T. (2019). Improvement of modified ratio estimators using robust regression methods, Applied Mathematics and Computation, 348, p. 627-631.