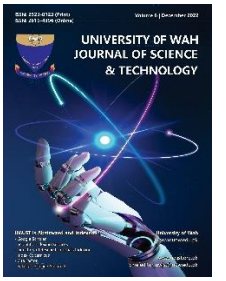
UW Journal of Science and Technology Vol. 7, Issue 1 (2023) 11-18
ISSN: 2523-0123 (Print) 2616-4396 (Online)

University of Wah
Journal of Science and Technology

www.uow.edu.pk



An improved regression type mean estimator using redescending M-estimator

Mubeen Mukhtar[1], Nasir Ali[1,2], Usman Shahzad[1]

[1] Department of Mathematics and Statistics, PMAS-Arid Agriculture University, Rawalpindi 46300, Pakistan.

[2] Department of Mathematics and Statistics, International Islamic University, Islamabad 44000, Pakistan.

Corresponding author email: usman.stat@yahoo.com

Abstract
In this article, a novel technique is presented for calculating the mean of a countable population using simple random sampling (SRS) in situations where there are outliers in the data. The proposed approach employs a robust regression type estimation method called re-descending M-estimation. To assess the effectiveness of the new method, the mean square error (MSE) equation is derived using a first-order approximation and compared against existing estimation methods. Furthermore, the percentage relative efficiency (PRE) of the proposed estimator is calculated in comparison to other estimators. Real-life data sets are employed to demonstrate the efficacy of the suggested approach. The results indicate that the proposed estimator outperforms other estimators in the literature.

Keywords
Robust Regression, Outliers, Redescending M-estimator.

1. Introduction

When utilizing supplementary information in estimation methods, the ratio or regression approach is employed to enhance effectiveness. [1]aimed to improve mean estimators by incorporating the traditional measures of descriptive statistics related to additional information.

In survey sampling, there are several ways we can improve our estimate by using supplementary data. It should be noted that the ratio, regression, and product type estimation techniques are useful when the supplementary data are available [2]. However, situations can also occur where many authors develop various estimators using auxiliary data, improving the performance of the estimation methods. In this situation, numerous writers, including [3-4], constructed a number of improved and modified estimators using auxiliary data.

The term "outlier" refers to an observation in a dataset that deviates significantly from the norm and frequently has an important statistical impact. In sample studies, it is not always feasible to access

the entire population, and therefore it can be challenging to identify the presence of outliers. If a sample includes an outlier observation, the efficiency of the estimation method can be reduced, especially in situations involving small sample sizes. To mitigate the impact of outlier(s) on the actual data, robust regression techniques are commonly employed. In the field of statistics, M-estimators serve as a robust alternative to traditional estimators. The redescending M-estimation method is suggested [5] as a viable solution to the problem of outliers in comparison to other robust estimation methodologies.

It is generally known that the presence of outliers degrades the performance of the ordinary least squares (OLS) estimation method. Re-descending M-estimation techniques are employed in this case. Many researchers have discussed about the M-estimators including [6-9]. Some writers, including [10-11] developed ratio based mean estimators employing the M-estimation methods in the presence of outliers.

This article introduces a fresh robust regression mean estimator utilizing redescending M-estimators to enhance efficiency. The structure of the article is as follows: a review of existing estimators from the literature is presented in Section 2, followed by a detailed description of the proposed robust regression mean estimator in Section 3. Section 4 provides a report on a numerical study conducted to assess the estimator's performance. Finally, the article concludes with Section 5.

2. Existing estimators in the literature

2.1 **Kadilar and Cingi (2004) estimators**

In simple random sampling (SRS), [1] suggested the estimators $\bar{y}_{KC_l}$. According to their analysis, the suggested estimators outperform then OLS estimators in terms of effectiveness.

$$\bar{y}_{KC_l} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{\gamma_l \bar{x} + \psi_l}(\gamma_l \bar{X} + \psi_l) \qquad l = 1,2,3,4,5 \qquad (1)$$

where $\bar{y}$ and $\bar{x}$ are the study's sample mean and supplemental variable, respectively

$$\gamma_1 = 1 \;\&\; \psi_1 = 0, \gamma_2 = 1 \;\&\; \psi_2 = C_x, \gamma_3 = 1 \;\&\; \psi_3 = \beta_2(x), \gamma_4 = \beta_2(x) \;\&\; \psi_4 = C_x,$$
$$\gamma_5 = C_x \quad \& \quad \psi_5 = \beta_2(x)$$

Where $\beta_2(x)$ and C_x are the coefficient of kurtosis of population and CV of supplementary variable respectively. Then $b = \frac{s_{yx}}{s_x^2}$ is calculated by OLS method, where s_x^2 is sample variance of x, s_{yx} is sample covariance between x and y and $\bar{Y}$ and $\bar{X}$ is the population mean of the study and supplementary variable. The mean square error (MSE) of the estimator can be obtained by utilizing the Taylor series approximation up to the first order given in (1) as follows, which is given as

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

For more details follow Cingi and Kadilar (2004), where, $f = \frac{n}{N}$, n & N are sample size and the population size respectively and

$$R_{KC_1} = R = \frac{\bar{Y}}{\bar{X}}, R_{KC_2} = \frac{\bar{Y}}{\bar{X} + C_x}, R_{KC_3} = \frac{\bar{Y}}{\bar{X} + \beta_2(x)}, R_{KC_4} = \frac{\bar{Y}\beta_2(x)}{\bar{X}\beta_2(x) + C_x}, R_{KC_5} = \frac{\bar{Y}C_x}{\bar{X}C_x + \beta_2(x)}$$

It's crucial to keep in mind that E (b) = B. For gaining more knowledge interested readers may refer to Shabbir and Onyango (2022).

2.2. Kadilar et al., (2007) suggested estimators

The following robust estimators were proposed by [12] to estimate the population mean in the presence of outliers by utilising data on the auxiliary variable's parameter.

$$\bar{y}_{rob_l} = \frac{\bar{y} + b_{rob}(\bar{X} - \bar{x})}{\gamma_l \bar{x} + \psi_l}(\gamma_l \bar{X} + \psi_l) \quad (3)$$

Where b_{rob} is generated using the robust regression Huber M-estimator with l =1, 2, 3, 4, 5. When data contain some outliers, the estimate in (3) is more effective than the estimator in (1). Huber (1964) discussed following $\rho_2(r)$ function, where r is the error term of OLS model.

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

The tuning constant v governs the robustness of the M-estimator, and its value can be adjusted accordingly. Huber (1964) suggested setting $v = 1.5s$. The b_{rob} is obtained by reducing $\sum_{i=1}^{n} \rho_2(y_i - a - bx_i)$.

MSE of $\bar{y}_{rob_l}$ is obtaining by replace B by B_{rob} in (2) and is given as

$$MSE(\bar{y}_{rob_l}) = \frac{1-f}{n}\left(R_{KC_l}^2 S_x^2 + 2B_{rob}R_{KC_l}S_x^2 + B_{rob}^2 S_x^2 - 2R_{KC_l}S_{yx} - 2B_{rob}S_{yx} + S_y^2\right) \quad (4)$$

For more details see kadilar et al. (2007).

2.3. Raza et al. (2019) estimators

[5] developed ratio estimators based on a newly proposed robust redescending M-estimator. The Re-decending M-estimators (RM) are stated by

$$\bar{y}_{RM_l} = \frac{\bar{y} + b_{RM}(\bar{X} - \bar{x})}{\gamma_l \bar{x} + \psi_l}(\gamma_l \bar{X} + \psi_l) \quad (5)$$

where b_{RM} is provided by [5]. The Raza's objective function's $\rho_1(r_l)$ design is explained as

$$\rho_1(r_l) = \frac{v^2}{2c}\left\{1 - \left[1 + \left(\frac{r}{v}\right)^2\right]^{-c}\right\} \; for \; |r| \ge 0$$

where c and v are tuning parameters. The optimal tuning constant values for the current investigation are c = 2.5 and v = 8. The b_{RM} the redescending M-estimator is used in the MSE equation of the ratio estimators in equation (5).

$$MSE(\bar{y}_{RM_l}) = \frac{1-f}{n}\left(R_{KC_l}^2 S_x^2 + 2B_{RM}R_{KC_l}S_x^2 + B_{RM}^2 S_x^2 - 2R_{KC_l}S_{yx} - 2B_{RM}S_{yx} + S_y^2\right) \quad (6)$$

3. Proposed robust regression mean estimator

By extending the idea of existing estimator, we propose the robust regression mean estimator as

$$\bar{y}_{N_i} = a\bar{y} + b_{RM}(\bar{X} - \bar{x}) \quad (7)$$

where b_{RM} is the slope coefficient of [5] redescending M-estimators.

Using Taylor series of expansion, the MSE of the proposed estimator is described as

$$h(\bar{x},\bar{y}) - h(\bar{X},\bar{Y}) = \left[\frac{\delta h\left(a\bar{y} + b_{RM}(\bar{X} - \bar{x})\right)}{\delta \bar{x}}\right]_{\bar{X},\bar{Y}} (\bar{x} - \bar{X}) + \left[\frac{\delta h\left(a\bar{y} + b_{RM}(\bar{X} - \bar{x})\right)}{\delta \bar{y}}\right]_{\bar{X},\bar{Y}} (\bar{y} - \bar{Y}) \quad (8)$$

Now, by partially differentiating the first and second terms of equation (8) w.r.t. $\bar{x}$ and $\bar{y}$, respectively, we obtain.

$$h(\bar{x},\bar{y}) - h(\bar{X},\bar{Y}) = -b_{RM}(\bar{x} - \bar{X}) + a(\bar{y} - \bar{Y}) \quad (9)$$

We obtain the MSE of the proposed estimator by squaring and taking expectation to both sides as.

$$MSE\left(\bar{y}_{N_i}\right) = \frac{1-f}{n}\left[B_{RM}^2 S_x^2 + a^2 S_y^2 - 2B_{RM} a S_{xy}\right] \quad (10)$$

Partially differentiating the equation (10) to obtain the optimum value of a that minimizes the $MSE$of $\bar{y}_{N_i}$, we get

$$a^{opt} = B_{RM}\frac{S_{xy}}{S_y^2} \quad (11)$$

Substituting the value of a in equation (10), the minimum MSE of the proposed estimator is given by

$$MSE_{min}\left(\bar{y}_{N_i}\right) = \frac{1-f}{n} B_{RM}^2 S_x^2 (1-\rho^2) \quad (12)$$

4. Numerical illustration

This section compares the performance of the proposed estimators to the existing estimators in terms of MSE and (PRE) using two real-world datasets. PRE of an estimator can be computed through the following expressions:

$$PRE\left(\bar{y}_{KC_l}, \bar{y}_Q\right) = \frac{\bar{y}_{KC_l}}{\bar{y}_Q} \times 100$$

where $Q = \bar{y}_{rob_l},\ \bar{y}_{RM_l}\ and\ \bar{y}_{N_i} \quad l = 1,2,3,4,5$

4.1. Population I

We have used the data from [12] about the production of apple y in tons as Numbers of apple trees (x, 1 unit = 100 trees) were used as a research variable and a supplementary variable in 204 villages in Turkey's Karadeniz Region. Figure 1 displays a scatter plot of the gathered data that is used to look for outliers.

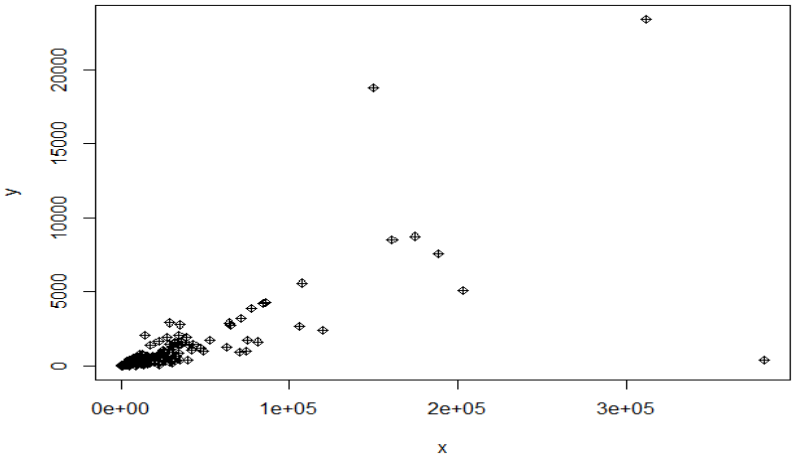


Figure 1: Numbers of Apple Trees and production of Apple

Table 1: Statistics Regarding Population 1

$N = 204$	$S_x = 454.03$	$B_{RM} = 2.484$
$n = 30$	$S_y = 2389.77$	$R_{KC_1} = 3.6569$
$\bar{X} = 264.42$	$\beta_2(x) = 29.77$	$R_{KC_2} = 3.6333$
$\bar{Y} = 966.96$	$\rho = 0.713$	$R_{KC_3} = 3.2868$
$C_x = 1.72$	$B = 3.753$	$R_{KC_4} = 3.6561$
$S_{yx} = 773727.8$	$B_{rob} = 3.547$	$R_{KC_5} = 3.4319$

4.2. Population II

The information about state public-school spending, in the United States is taken from Fox (2008). This information, which comprises of 51 observations, shows the per-capita income and per-capita education spending for the American states in 1970. The independent variable is the per-capita income, while the dependent variable is the per-capita expenditure on education. Although the original data didn't contain any outliers, we added 6% of them to test how well the suggested estimators performed. Figure 2 shows the income and expense graph with outliers.

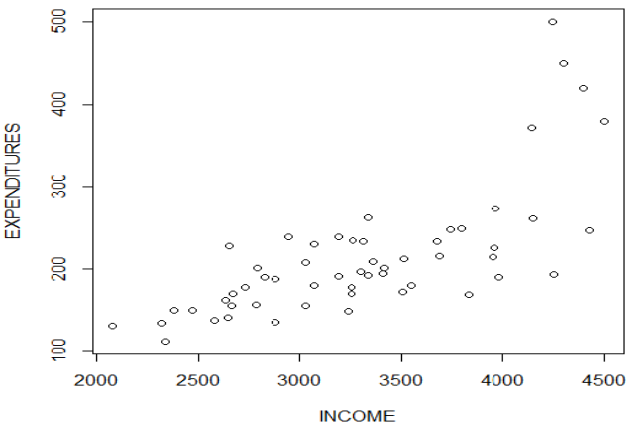


Figure 2: Expenditures and Income with 6% outliers

Table 2: Statistics Regarding Population 2

$N = 54$	$S_x = 605.3097$	$B_{RM} = 0.05383$
$n = 5$	$S_y = 69.3353$	$R_{KC_1} = 0.06353$
$\bar{X} = 3288.7040$	$\beta_2(x) = 2.2168$	$R_{KC_2} = 0.06353$
$\bar{Y} = 208.9259$	$\rho = 0.7179$	$R_{KC_3} = 0.06349$
$C_x = 0.1840$	$B = 0.08223$	$R_{KC_4} = 0.063526$
$S_{yx} = 30129.52$	$B_{rob} = 0.06539$	$R_{KC_5} = 0.06330$

Table 3: The MSE of proposed and existing estimators

Estimators	Population I	Population II
$\bar{y}_{KC_1}$	158167.6	691.1811
$\bar{y}_{KC_2}$	157159.4	691.1811
$\bar{y}_{KC_3}$	143107.4	690.8432
$\bar{y}_{KC_4}$	158133.4	691.1473
$\bar{y}_{KC_5}$	148820.5	689.2414
$\bar{y}_{rob_1}$	149586.9	567.7626
$\bar{y}_{rob_2}$	148635.6	567.7626
$\bar{y}_{rob_3}$	135420.3	567.5144
$\bar{y}_{rob_4}$	149554.5	567.7378
$\bar{y}_{rob_5}$	140783.0	566.338
$\bar{y}_{RM_1}$	113214.5	504.8712
$\bar{y}_{RM_2}$	112557.3	504.8712
$\bar{y}_{RM_3}$	103659.5	504.6844
$\bar{y}_{RM_4}$	113192.1	504.8525
$\bar{y}_{RM_5}$	107214.2	503.8002
$\bar{y}_{N_i}$	17779.08	93.37644

Table 4: PRE for population 1

$\frac{\bar{y}_{KC_1}}{\bar{y}_{rob_1}} \times 100$ $= 105.7363$	$\frac{\bar{y}_{KC_2}}{\bar{y}_{rob_2}} \times 100$ $= 105.7347$	$\frac{\bar{y}_{KC_3}}{\bar{y}_{rob_3}} \times 100$ $= 105.6765$	$\frac{\bar{y}_{KC_4}}{\bar{y}_{rob_4}} \times 100$ $= 105.7363$	$\frac{\bar{y}_{KC_5}}{\bar{y}_{rob_5}} \times 100$ $= 105.7091$
$\frac{\bar{y}_{KC_1}}{\bar{y}_{RM_1}} \times 100$ $= 139.7062$	$\frac{\bar{y}_{KC_2}}{\bar{y}_{RM_2}} \times 100$ $= 139.6261$	$\frac{\bar{y}_{KC_3}}{\bar{y}_{RM_3}} \times 100$ $= 138.0553$	$\frac{\bar{y}_{KC_4}}{\bar{y}_{RM_4}} \times 100$ $= 139.7035$	$\frac{\bar{y}_{KC_5}}{\bar{y}_{RM_5}} \times 100$ $= 138.8067$
$\frac{\bar{y}_{KC_1}}{\bar{y}_{N_i}} \times 100$ $= 889.6276$	$\frac{\bar{y}_{KC_2}}{\bar{y}_{N_i}} \times 100$ $= 883.9565$	$\frac{\bar{y}_{KC_3}}{\bar{y}_{N_i}} \times 100$ $= 804.9198$	$\frac{\bar{y}_{KC_4}}{\bar{y}_{N_i}} \times 100$ $= 889.4348$	$\frac{\bar{y}_{KC_5}}{\bar{y}_{N_i}} \times 100$ $= 837.0538$
$\frac{\bar{y}_{rob_1}}{\bar{y}_{N_i}} \times 100$ $= 841.3643$	$\frac{\bar{y}_{rob_2}}{\bar{y}_{N_i}} \times 100$ $= 836.0137$	$\frac{\bar{y}_{rob_3}}{\bar{y}_{N_i}} \times 100$ $= 761.683$	$\frac{\bar{y}_{rob_4}}{\bar{y}_{N_i}} \times 100$ $= 841.1823$	$\frac{\bar{y}_{rob_5}}{\bar{y}_{N_i}} \times 100$ $= 791.8463$
$\frac{\bar{y}_{RM_1}}{\bar{y}_{N_i}} \times 100$ $= 636.7847$	$\frac{\bar{y}_{RM_2}}{\bar{y}_{N_i}} \times 100$ $= 633.0881$	$\frac{\bar{y}_{RM_3}}{\bar{y}_{N_i}} \times 100$ $= 583.0416$	$\frac{\bar{y}_{RM_4}}{\bar{y}_{N_i}} \times 100$ $= 636.6588$	$\frac{\bar{y}_{RM_5}}{\bar{y}_{N_i}} \times 100$ $= 603.0357$

Table 5: PRE for population 2

$\frac{\bar{y}_{KC_1}}{\bar{y}_{rob_1}} \times 100 = 121.7377$	$\frac{\bar{y}_{KC_2}}{\bar{y}_{rob_2}} \times 100 = 121.7377$	$\frac{\bar{y}_{KC_3}}{\bar{y}_{rob_3}} \times 100 = 121.7314$	$\frac{\bar{y}_{KC_4}}{\bar{y}_{rob_4}} \times 100 = 121.7371$	$\frac{\bar{y}_{KC_5}}{\bar{y}_{rob_5}} \times 100 = 121.7014$
$\frac{\bar{y}_{KC_1}}{\bar{y}_{RM_1}} \times 100 = 136.9025$	$\frac{\bar{y}_{KC_2}}{\bar{y}_{RM_2}} \times 100 = 136.9025$	$\frac{\bar{y}_{KC_3}}{\bar{y}_{RM_3}} \times 100 = 136.8862$	$\frac{\bar{y}_{KC_4}}{\bar{y}_{RM_4}} \times 100 = 136.9008$	$\frac{\bar{y}_{KC_5}}{\bar{y}_{RM_5}} \times 100 = 136.8085$
$\frac{\bar{y}_{KC_1}}{\bar{y}_{N_i}} \times 100 = 740.2093$	$\frac{\bar{y}_{KC_2}}{\bar{y}_{N_i}} \times 100 = 740.2093$	$\frac{\bar{y}_{KC_3}}{\bar{y}_{N_i}} \times 100 = 739.8475$	$\frac{\bar{y}_{KC_4}}{\bar{y}_{N_i}} \times 100 = 740.1731$	$\frac{\bar{y}_{KC_5}}{\bar{y}_{N_i}} \times 100 = 738.132$
$\frac{\bar{y}_{rob_1}}{\bar{y}_{N_i}} \times 100 = 608.0363$	$\frac{\bar{y}_{rob_2}}{\bar{y}_{N_i}} \times 100 = 608.0363$	$\frac{\bar{y}_{rob_3}}{\bar{y}_{N_i}} \times 100 = 607.7704$	$\frac{\bar{y}_{rob_4}}{\bar{y}_{N_i}} \times 100 = 608.0097$	$\frac{\bar{y}_{rob_5}}{\bar{y}_{N_i}} \times 100 = 606.5106$
$\frac{\bar{y}_{RM_1}}{\bar{y}_{N_i}} \times 100 = 540.6837$	$\frac{\bar{y}_{RM_2}}{\bar{y}_{N_i}} \times 100 = 540.6837$	$\frac{\bar{y}_{RM_3}}{\bar{y}_{N_i}} \times 100 = 540.4837$	$\frac{\bar{y}_{RM_4}}{\bar{y}_{N_i}} \times 100 = 540.6637$	$\frac{\bar{y}_{RM_5}}{\bar{y}_{N_i}} \times 100 = 539.5367$

4.3. Interpretation.

The interpretation is provided in the following upcoming points:

1. Estimators $(\bar{y}_{rob_1}, \dots, \bar{y}_{rob_5})$ are performing better than $(\bar{y}_{KC_1}, \dots, \bar{y}_{KC_5})$.
2. Estimators $(\bar{y}_{RM_1}, \dots, \bar{y}_{RM_5})$ is performing better than $(\bar{y}_{KC_1}, \dots, \bar{y}_{KC_5})$.
3. Estimator $\bar{y}_{N_i}$ is performing better than $(\bar{y}_{KC_1}, \dots, \bar{y}_{KC_5})$, $(\bar{y}_{rob_1}, \dots, \bar{y}_{rob_5})$ and $(\bar{y}_{RM_1}, \dots, \bar{y}_{RM_5})$.

Thus, it can be seen from Tables 3, 4, and 5 that the proposed estimators are more effective than the current estimators under the SRS for the provided datasets.

5. Conclusion

In SRS, under the determined conditions, using the redescending M-estimator, a new robust regression type estimator has been proposed. Utilizing the numerical illustration it has also been demonstrated that the proposed estimator produces smallest MSE value as compared to the existing estimators. When data contain outliers, it is found that the proposed recommended robust mean estimator performs better than the reviewed estimators.

Acknowledgements

Authors are thankful to the referees for the valuable suggestions.

References

[1] C. Kadilar, and H. Cingi. Ratio estimators in simple random sampling, Applied Mathematics and Computation, Vol.151, pp. 893-902, 2004.
[2] H. Bulut, and T. Zaman. An improved class of robust ratio estimators by using the minimum covariance determinant estimation. Communications in Statistics-Simulation and Computation, Vol. 51(5), pp. 2457-2463, 2022.
[3] T. Zaman, and C. Kadilar. On estimating the population mean using auxiliary character in stratified random sampling. Journal of Statistics and Management Systems, Vol. 23(8), pp.1415-1426, 2020.
[4] T. Zaman, and C. Kadilar. Exponential ratio and product type estimators of the mean in stratified two-phase sampling. AIMS Mathematics, Vol. 6(5), pp. 4265-4279, 2021.
[5] Raza, M. Noor-ul-Amin, and M. Hanif. Regression-in-ratio estimators in the presence of outliers based on redescendingm-estimator. Journal of Reliability and Statistical Studies, pp.01-10, 2019.
[6] P. J. Huber. Robust estimation of a location parameter, The Annals of Mathematical Statistics, Vol. 35(1), pp. 73-10, 1964.
[7] U. Khalil, Alamgir, A. Amjad, and D. M. Khan. Efficient Uk's redescending M-estimator for robust regression, Pakistan Journal of Statistics, Vol. 32(2), pp. 125-138, 2016.
[8] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. Robust Statistics, The Approach Based on Influence Functions, New York: John Wiley and Sons. 1986
[9] M. Noor-Ul-Amin, M. Q. Shahbaz, and C. Kadilar. Ratio estimators for population mean using robust regression in double sampling, Gazi University Journal of Science, Vol. 29 (4), pp. 793-798, 2016
[10] M. Subzar, C. N. Bouza, S. Maqbool, T.A. Raja, and B.A. Para. Robust ratio type estimators in simple random sampling using Huber M estimation, Revista Investigacion Operacional, Vol. l40(2),pp. 201-209, 2019
[11] T. Zaman. Improvement of modified ratio estimators using robust regression methods, Applied Mathematics and Computation,Vol. 348, pp. 627-631, 2019.
[12] M. Kadilar, Candan. and H. Cingi. Ratio estimation using robust regression, Hacettepe Journal of Mathematics and Statistics, Vol. 36, pp. 181-188, 2007.