The Robust Regression Methods for Estimating of Finite Population Mean Based on SRSWOR in Case of Outliers

Mir Subzar[1], Amer Ibrahim Al-Omari[2] and Ayed R. A. Alanzi[3, *]

Abstract: The ordinary least square (OLS) method is commonly used in regression analysis. But in the presence of outlier in the data, its results are unreliable. Hence, the robust regression methods have been suggested for a long time as alternatives to the OLS to solve the outliers problem. In the present study, new ratio type estimators of finite population mean are suggested using simple random sampling without replacement (SRSWOR) utilizing the supplementary information in Bowley's coefficient of skewness with quartiles. For these proposed estimators, we have used the OLS, Huber-M, Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate methods for estimating the population parameters. Theoretically, the mean square error (MSE) equations of various estimators are obtained and compared with the OLS competitor. Simulations for skewed distributions as the Gamma distribution support the results, and an application of real data set containing outliers is considered for illustration.

Keywords: Efficiency, GM-estimates, Huber-M, ordinary least square, ratio type estimators.

1 Introduction

The OLS scheme is widely used in estimating the parameter of a linear regression model, which has a wide range of applications in real-life provided that the OLS assumptions are satisfied. In many cases, these assumptions may be violated due to the nature of the data under consideration, especially of the occurrence of an outlier. Therefore, several robust regression methods are suggested to overcome this problem.

Some of the commonly known robust regression methods are the least absolute deviations method, where the Least Absolute Deviations (LAD) regression is the first step for robust regression methods [Nadia and Mohammad (2013)]. The least median squares method is suggested and improved by Rousseeuw et al. [Rousseeuw and Leroy (1987)]. The least trimmed squares method, the Huber-M plan, is introduced by Huber

[1] Division of Agricultural Statistics SKUAST-K, Shalimar, 190025, India.

[2] Department of Mathematics, Faculty of Science, Al al-Bayt University, Mafraq, 25113, Jordan.

[3] Department of Mathematics, College of Science and Human Studies at HotatSudair, Majmaah University, Majmaah, 11952, Saudi Arabia.

[*]Corresponding Author: Ayed R. A. Alanzi. Email: a.alanzi@mu.edu.sa.

Received: 18 February 2020; Accepted: 03 June 2020.

[Huber (1973)]. The Hampel-M method is suggested by Hampel [Hampel (1971)], the Tukey-M method is proposed by Tukey [Tukey (1977)], and the Huber-MM method by Yohai [Yohai (1987)].

In this study, we considered the following generalized M methods. The Mallows GM-estimator which was proposed by Mallows [Mallows (1975)], the Schweppes GM-estimate method that was introduced by Handschin et al. [Handschin, Kohlas, Fiechter et al. (1975)], the SIS GM-estimate method which was submitted by Coakley et al. [Coakley and Hettmansperger (1993)], with illustrations given in the next section. However, the Huber-M was adopted by Subzar et al. [Subzar, Bouza, Maqbool et al. (2019a)] in the case of outliers, and was compared with the OLS method. It was shown that the Huber-M estimation performs better than the OLS method. In the current study, we have adopted the generalized case of M-estimation methods and compared it with the OLS and Huber-M estimation. Suppose that Y is a study variable, and X is an auxiliary variable that is correlated with Y. Also, let the population means of Y and X, respectively, are $\bar{Y}$ and $\bar{X}$, with variances σ_Y^2 and σ_X^2, and let the correlation coefficient between Y and X is $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$, where σ_{XY} is the covariance between X and Y. The usual simple random sampling (SRS) ratio estimator of the population mean of Y is given by $\hat{\bar{y}} = \frac{\bar{y}}{\bar{X}}\bar{Y}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means of X and Y, respectively. Based on Cochran [Cochran (1977)], the mean squared error of $\hat{\bar{y}}$ is given by $MSE\left(\hat{\bar{y}}\right) = \frac{1-f}{n}\left[\sigma_Y^2 + R^2\sigma_X^2\left(1-2D\right)\right]$, where $f = \frac{n}{N}$, N is the population size, and n is the sample size, where $R = \frac{\bar{Y}}{\bar{X}}$, $D = \rho\frac{CV_Y}{CV_X}$, $CV_X = \frac{\sigma_X}{\bar{X}}$, $CV_Y = \frac{\sigma_Y}{\bar{Y}}$ and $Cov\left(X,Y\right) = E\left[\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)\right]$. Al-Omari et al. [Al-Omari, Ibrahim and Jemain (2009)] have suggested ratio-type estimators of the population mean using SRS as $\hat{\bar{y}}_1 = \frac{\bar{y}}{\bar{x}+Q_1}\left(\bar{X}+Q_1\right)$, and $\hat{\bar{y}}_2 = \frac{\bar{y}}{\bar{x}+Q_3}\left(\bar{X}+Q_3\right)$, where, Q_1 and Q_3 are the first and third quartiles of the auxiliary variable X, respectively. Al-Omari et al. [Al-Omari, Jaber, Ibrahim (2008)] have suggested ratio-type estimators of the mean using an extreme ranked set sampling technique. Additionally, Al-Omari [Al-Omari (2012)] proposed ratio estimators of population mean using auxiliary information based on simple random sampling and median ranked set sampling. Al-Omari et al. [Al-Omari and Bouza (2015)] investigated some ratio estimators of the population mean with missing values using ranked set sampling. Al-Omari et al. [Al-Omari and Al-Nasser (2018)] considered the problem of ratio estimation using multistage median ranked set sampling. Zeinalova et al. [Zeinalova, Huseynov and Sharghi (2018)] considered A Z-Number valued regression model. Subzar et al. [Subzar, Maqbool, Raja et al. (2019)] introduced a new ratio estimator as an alternative to the regression estimator using auxiliary information. Moreover, Subzar et al. [Subzar, Maqbool, Raja et al. (2018)] introduced ratio estimators for the population mean in simple random sampling using supplemental information. For more details about ratio and regression estimators, see Jemain et al. [Jemain, Al-Omari and Ibrahim (2008); Krasker (1980); Krasker and Welsch (1982); Subzar, Bouza and Al-Omari (2019b); Bouza, Al-Omari, Santiago et al. (2017); Yu and Yao (2017)].

The rest of this paper is prepared in seven sections and subsections. The robust regression techniques are illustrated in Section 2, while the suggested ratio estimators are presented with their main properties in Section 3. In Section 4, efficiency comparisons of the OLS method with the robust regression techniques are presented. Numerical illustrations are provided in Section 5, and in Section 6, an application of real data is supported. Finally, the paper is concluded in Section 7.

2 Robust regression techniques

In this section, we summarized the main robust regression methods considered in this study.

2.1-Huber-M estimation function

The M-Estimator is a well-known estimator advocated by Huber [Huber (1973)]. The M-Estimator is given by

$$\delta H(Z) = \left\{ \frac{1}{2} Z^2 \text{ for } Z \le Q \text{ and } |Z| - \frac{1}{2} Q^2 \text{ for } Z > Q, Q \text{ is small} \right\}. \tag{1}$$

The influence function is determined by taking the derivative of this function as

$$\psi H(Z) = \{ q \text{ for } Z > Q, Z \text{ for } Z \le Q \text{ and} - Q \ for\ Z < -Q \}, \tag{2}$$

where the tuning constant Q defines the center and tails.

2.2 Generalized M estimation function

The generalized M-Estimate (GM-estimate) is proposed to provide reliable results. The general GM class of estimators is defined by

$$\sum_{i=1}^{n} w_i (X_i) \psi \left\{ \frac{e_i(\hat{\beta})}{V(X_i)\hat{\sigma}} \right\} X_i = 0, \tag{3}$$

where ψ is the certain function, as in the case of M-estimate.

2.3 Mallows GM estimation function

Mallows [Mallows (1975)] proposed Mallows GM-estimate to M-estimate resistant to high leverage outliers. The Mallows GM-estimate is defined by

$$\sum_{i=1}^{n} w_i \Re \left\{ \frac{r_i(\hat{\beta})}{\hat{\sigma}} \right\} x_i = 0 ,$$

(4)

where $\Re(e) = \rho'(e)$ and $w_i = \sqrt{1 - l_i}$ with l_i being the leverage of the ith observation.

The weight w_i ensures that the observations with high leverage receive less weight than observations with small leverage.

2.4 Schweppe GM estimation function

The Schweppe GM-estimate is suggested by Handschin et al. [Handschin, Kohlas, Fiechter et al. (1975)] to be the solution of the equation

$$\sum_{i=1}^{n} w_i \Re\left\{\frac{r_i\hat{\beta}}{w_i\hat{\sigma}}\right\} x_i = 0, \tag{5}$$

which adjusts the leverage weights according to the size of the residual r_i.

2.5 SIS GM estimation function

Coakley et al. [Coakley and Hettmansperger (1993)] proposed Schweppe one step (SIS) estimate, which extended from the original Schweppe estimator. The SIS estimator is defined as

$$\hat{\beta} = \hat{\beta}_0 + \left[\sum_{i=1}^{n} \Re\left(\frac{r_i\hat{\beta}_0}{\hat{\sigma}w_i}\right) x_i x_i'\right]^{-1} \times \sum_{i=1}^{n} \hat{\sigma} w_i \Re\left(\frac{r_i\hat{\beta}_0}{\hat{\sigma}w_i}\right) x_i, \tag{6}$$

where the weight w_i is defined in the same way as Schweppe's GM-estimate.

3 Suggested estimators

In this section, the proposed ratio estimators are presented. The suggested estimators are suggested based on the supplementary information of Bowley's coefficient of skewness with quartiles. For estimating the parameters, we considered the OLS method, Huber M-estimate, Mallows GM-estimate, Schweppe's GM-estimate and SIS GM-estimate method. The proposed estimators are as follows.

3.1 Using the OLS method

$$\hat{\bar{Y}}_1 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}S_k + Q_1)}\left(\bar{X}S_k + Q_1\right) = \hat{R}_1\left(\bar{X}S_k + Q_1\right), \tag{7}$$

$$\hat{\bar{Y}}_2 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}S_k + Q_2)}(\bar{X}S_k + Q_2) = \hat{R}_2(\bar{X}S_k + Q_2), \tag{8}$$

$$\hat{\bar{Y}}_3 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}S_k + Q_3)}\left(\bar{X}S_k + Q_3\right) = \hat{R}_3\left(\bar{X}S_k + Q_3\right), \tag{9}$$

where the Bowley's coefficient of skewness is defined as $S_k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$ and Q_i is the *i*th quartile. The mean squared error expressions for the above estimators can be derived as follows. For the estimator given in Eq. (7), the mean squared error equation can be obtained as

$$h(\bar{x},\bar{y}) \cong h(\bar{X},\bar{Y}) + \frac{\partial h(c,d)}{\partial c}|_{\bar{X},\bar{Y}} (\bar{x}-\bar{X}) + \frac{\partial h(c,d)}{\partial d}|_{\bar{X},\bar{Y}} (\bar{y}-\bar{Y}), \tag{10}$$

where $h(\bar{x},\bar{y}) = \hat{R}_1$ and $h(\bar{X},\bar{Y}) = R$. As shown in Wolter [Wolter (1985)], Eq. (10) can be applied to the proposed estimator to obtain its MSE as follows:

$$\hat{R}_1 - R \cong \frac{\partial((\bar{y}+b(\bar{X}-\bar{x}))/(\bar{x}S_K+Q_1)}{\partial\left(\bar{x}S_K+Q_1\right)}|_{\bar{X},\bar{Y}} (\bar{x}-\bar{X})$$

$$+\frac{\partial((\bar{y}+b(\bar{X}-\bar{x}))/(\bar{x}S_K+Q_1)}{\partial\left(\bar{y}S_K+Q_1\right)}|_{\bar{X},\bar{Y}} (\bar{y}-\bar{Y})$$

$$\cong -\left(\frac{\bar{y}S_K}{(\bar{x}S_K+Q_1)^2} + \frac{b(\bar{X}S_K+Q_1)}{(\bar{x}S_K+Q_1)^2}\right)|_{\bar{X},\bar{Y}} (\bar{x}-\bar{X}) + \frac{1}{\bar{x}S_K+Q_1}|_{\bar{X},\bar{Y}} (\bar{y}-\bar{Y}).$$

Squaring both sides of the last equation and taking the expectation to obtain

$$E(\hat{R}_1-R)^2 \cong \frac{(\bar{Y}+B(\bar{X}S_K+Q_1))^2}{(\bar{X}S_K+Q_1)^4}V(\bar{x}) - \frac{2(\bar{Y}+B(\bar{X}S_K+Q_1))}{(\bar{X}S_K+Q_1)^3}Cov(\bar{x},\bar{y}) + \frac{1}{(\bar{X}S_K+Q_1)^2}V(\bar{y})$$

$$\cong \frac{1}{(\bar{X}S_K+Q_1)^2}\left\{\frac{(\bar{Y}S_K+B(\bar{X}S_K+Q_1))^2}{(\bar{X}S_K+Q_1)^2}V(\bar{x}) - \frac{2(\bar{Y}S_K+B(\bar{X}S_K+Q_1)}{\bar{X}S_K+Q_1}Cov(\bar{x},\bar{y}) + V(\bar{y})\right\}$$

where $B = \frac{s_{xy}}{s_x^2} = \frac{\rho s_x s_y}{s_x^2} = \frac{\rho s_y}{s_x}$. Note that we omit the difference of $(E(b)-B)$. Hence,

$$MSE(\bar{y}_1) = (\bar{X}S_K+Q_1)^2 E(\hat{R}_1-R)^2$$

$$\cong \frac{(\bar{Y}S_K+B(\bar{X}S_K+Q_1))^2}{(\bar{X}S_K+Q_1)^2}V(\bar{x}) - \frac{2(\bar{Y}S_K+B(\bar{X}S_K+Q_1))}{\bar{X}S_K+Q_1}Cov(\bar{x},\bar{y}) + V(\bar{y})$$

$$\cong \frac{\bar{Y}^2S_K + 2B(\bar{X}S_K+Q_1)\bar{Y} + B^2(\bar{X}S_K+Q_1)^2}{(\bar{X}S_K+Q_1)^2}V(\bar{x}) - \frac{2\bar{Y}S_K+2B(\bar{X}S_K+Q_1)}{\bar{X}S_K+Q_1}Cov(\bar{x},\bar{y}) + V(\bar{y})$$

$$\cong \frac{1-f}{n}\left\{\left(\frac{\bar{Y}^2S_K}{(\bar{X}S_K+Q_1)^2} + \frac{2B\bar{Y}S_K}{\bar{X}S_K+Q_1} + B^2\right)S_x^2 - \left(\frac{2\bar{Y}S_K}{\bar{X}S_K+Q_1} + 2B\right)S_{xy} + S_y^2\right\}.$$

Therefore,

$$MSE(\hat{\bar{Y}}_1) \cong \frac{1-f}{n}\left(R_1^2S_x^2 + 2BR_1S_x^2 + B^2S_x^2 - 2R_1S_{xy} - 2BS_{xy} + S_y^2\right). \tag{11}$$

Similarly, the MSEs of Eqs. (8)-(9) are

$$MSE(\hat{\bar{Y}}_2) \cong \frac{1-f}{n}\left(R_2^2S_x^2 + 2BR_2S_x^2 + B^2S_x^2 - 2R_2S_{xy} - 2BS_{xy} + S_y^2\right), \tag{12}$$

$$MSE(\hat{\bar{Y}}_3) \cong \frac{1-f}{n}\left(R_3^2S_x^2 + 2BR_3S_x^2 + B^2S_x^2 - 2R_3S_{xy} - 2BS_{xy} + S_y^2\right), \tag{13}$$

where $R_1 = \dfrac{\overline{Y}S_k}{\overline{X}S_k + Q_1}$, $R_2 = \dfrac{\overline{Y}S_k}{\overline{X}S_k + Q_2}$ and $R_3 = \dfrac{\overline{Y}S_k}{\overline{X}S_k + Q_3}$.

3.1 Using the Huber M-estimation

The suggested estimators based on the Huber M-estimation are given by

$$\hat{\overline{Y}}_4 = \frac{\overline{y} + b_{rob(Huber)}(\overline{X} - \overline{x})}{\overline{x}S_k + Q_2}(\overline{X}S_k + Q_2), \tag{14}$$

$$\hat{\overline{Y}}_5 = \frac{\overline{y} + b_{rob(Huber)}(\overline{X} - \overline{x})}{\overline{x}S_k + Q_2}(\overline{X}S_k + Q_2), \tag{15}$$

$$\hat{\overline{Y}}_6 = \frac{\overline{y} + b_{rob(Huber)}(\overline{X} - \overline{x})}{\overline{x}S_k + Q_3}(\overline{X}S_k + Q_3), \tag{16}$$

with respective MSE's defined as

$$MSE(\hat{\overline{Y}}_4) \cong \frac{1-f}{n}\left(R_2^2 S_x^2 + 2B_{rob(Huber)}R_2 S_x^2 + B_{rob(Huber)}^2 S_x^2 - 2R_2 S_{xy} - 2B_{rob(Huber)}S_{xy} + S_y^2\right), \tag{17}$$

$$MSE(\hat{\overline{Y}}_5) \cong \frac{1-f}{n}\left(R_2^2 S_x^2 + 2B_{rob(Huber)}R_2 S_x^2 + B_{rob(Huber)}^2 S_x^2 - 2R_2 S_{xy} - 2B_{rob(Huber)}S_{xy} + S_y^2\right), \tag{18}$$

$$MSE(\hat{\overline{Y}}_6) \cong \frac{1-f}{n}\left(R_3^2 S_x^2 + 2B_{rob(Huber)}R_3 S_x^2 + B_{rob(Huber)}^2 S_x^2 - 2R_3 S_{xy} - 2B_{rob(Huber)}S_{xy} + S_y^2\right). \tag{19}$$

3.3 Using the Mallows GM-estimate

The suggested estimators based on the Mallows GM-estimate method with their mean squared error expressions are provided here as

$$\hat{\overline{Y}}_7 = \frac{\overline{y} + b_{rob(Mallows)}(\overline{X} - \overline{x})}{\overline{x}S_k + Q_1}(\overline{X}S_k + Q_1), \tag{20}$$

$$\hat{\overline{Y}}_8 = \frac{\overline{y} + b_{rob(Mallows)}(\overline{X} - \overline{x})}{\overline{x}S_k + Q_2}(\overline{X}S_k + Q_2), \tag{21}$$

$$\hat{\overline{Y}}_9 = \frac{\overline{y} + b_{rob(Mallows)}(\overline{X} - \overline{x})}{\overline{x}S_k + Q_3}(\overline{X}S_k + Q_3), \tag{22}$$

with respective MSE's given by

$$MSE(\hat{\overline{Y}}_7) \cong \frac{1-f}{n}\begin{pmatrix} R_1^2 S_x^2 + 2B_{rob(Mallows)}R_1 S_x^2 + B_{rob(Mallows)}^2 S_x^2 - 2R_1 S_{xy} - 2B_{rob(Mallows)}S_{xy} \\ +S_y^2 \end{pmatrix}, \tag{23}$$

$$MSE(\hat{\overline{Y}}_8) \cong \frac{1-f}{n}\begin{pmatrix} R_2^2 S_x^2 + 2B_{rob(Mallows)}R_2 S_x^2 + B_{rob(Mallows)}^2 S_x^2 - 2R_2 S_{xy} - 2B_{rob(Mallows)}S_{xy} \\ +S_y^2 \end{pmatrix}, \tag{24}$$

$$MSE(\hat{\bar{Y}}_9) \cong \frac{1-f}{n}\begin{pmatrix} R_3^2 S_x^2 + 2B_{rob(Mallows)} R_3 S_x^2 + B_{rob(Mallows)}^2 S_x^2 - 2R_3 S_{xy} - 2B_{rob(Mallows)} S_{xy} \\ + S_y^2 \end{pmatrix}. \quad (25)$$

3.4 Using the Schweppe GM-estimate

The Schweppe GM-estimate is used to suggest the following estimators as

$$\hat{\bar{Y}}_{10} = \frac{\bar{y} + b_{rob(Schweppe)}(\bar{X} - \bar{x})}{\bar{x}S_k + Q_1}(\bar{X}S_k + Q_1), \quad (26)$$

$$\hat{\bar{Y}}_{11} = \frac{\bar{y} + b_{rob(Schweppe)}(\bar{X} - \bar{x})}{\bar{x}S_k + Q_2}(\bar{X}S_k + Q_2), \quad (27)$$

$$\hat{\bar{Y}}_{12} = \frac{\bar{y} + b_{rob(Schweppe)}(\bar{X} - \bar{x})}{\bar{x}S_k + Q_3}(\bar{X}S_k + Q_3). \quad (28)$$

The MSE for the Eqs. (26)-(28), respectively, are

$$MSE(\hat{\bar{Y}}_{10}) \cong \frac{1-f}{n}\begin{pmatrix} R_1^2 S_x^2 + 2B_{rob(Schweppe)} R_1 S_x^2 + B_{rob(Schweppe)}^2 S_x^2 - 2R_1 S_{xy} \\ -2B_{rob(Schweppe)} S_{xy} + S_y^2 \end{pmatrix}, \quad (29)$$

$$MSE(\hat{\bar{Y}}_{11}) \cong \frac{1-f}{n}\begin{pmatrix} R_2^2 S_x^2 + 2B_{rob(Schweppe)} R_2 S_x^2 + B_{rob(Schweppe)}^2 S_x^2 - 2R_2 S_{xy} \\ -2B_{rob(Schweppe)} S_{xy} + S_y^2 \end{pmatrix}, \quad (30)$$

$$MSE(\hat{\bar{Y}}_{12}) \cong \frac{1-f}{n}\begin{pmatrix} R_3^2 S_x^2 + 2B_{rob(Schweppe)} R_3 S_x^2 + B_{rob(Schweppe)}^2 S_x^2 - 2R_3 S_{xy} \\ -2B_{rob(Schweppe)} S_{xy} + S_y^2 \end{pmatrix}. \quad (31)$$

3.5 Using the SIS GM-estimate

The suggested ratio estimators using the SIS GM-estimate are given by

$$\hat{\bar{Y}}_{13} = \frac{\bar{y} + b_{rob(SIS\,GM)}(\bar{X} - \bar{x})}{\bar{x}S_k + Q_1}(\bar{X}S_k + Q_1), \quad (32)$$

$$\hat{\bar{Y}}_{14} = \frac{\bar{y} + b_{rob(SIS\,GM)}(\bar{X} - \bar{x})}{\bar{x}S_k + Q_2}(\bar{X}S_k + Q_2), \quad (33)$$

$$\hat{\bar{Y}}_{15} = \frac{\bar{y} + b_{rob(SIS\,GM)}(\bar{X} - \bar{x})}{\bar{x}S_k + Q_3}(\bar{X}S_k + Q_3), \quad (34)$$

with the following MSE equations defined as

$$MSE(\hat{\bar{Y}}_{13}) \cong \frac{1-f}{n}\begin{pmatrix} R_1^2 S_x^2 + 2B_{rob(SIS\,GM)} R_1 S_x^2 + B_{rob(SIS\,GM)}^2 S_x^2 - 2R_1 S_{xy} - 2B_{rob(SIS\,GM)} S_{xy} \\ + S_y^2 \end{pmatrix}, \quad (35)$$

$$MSE(\hat{\bar{Y}}_{14}) \cong \frac{1-f}{n}\begin{pmatrix} R_2^2 S_x^2 + 2B_{rob(SIS\,GM)} R_2 S_x^2 + B_{rob(SIS\,GM)}^2 S_x^2 - 2R_2 S_{xy} - 2B_{rob(SIS\,GM)} S_{xy} \\ +S_y^2 \end{pmatrix}, \quad (36)$$

$$MSE(\hat{\bar{Y}}_{15}) \cong \frac{1-f}{n}\begin{pmatrix} R_3^2 S_x^2 + 2B_{rob(SIS\,GM)} R_3 S_x^2 + B_{rob(SIS\,GM)}^2 S_x^2 - 2R_3 S_{xy} - 2B_{rob(SIS\,GM)} S_{xy} \\ +S_y^2 \end{pmatrix}. \quad (37)$$

4 Efficiency comparison of the OLS method with robust regression techniques

In this section, a theoretical comparison between the OLS method with the robust regression methods is presented for the estimators considered in this study. Let

$MSE(\hat{\bar{Y}}_k) < MSE(\hat{\bar{Y}}_i), \quad i = 1,2,3,\ k = 3,4,...,15,$

$(2B_{(rob)l} R_j S_x^2 + B_{(rob)l} S_x^2 - 2B_{(rob)l} S_{xy}) < (2BR_j S_x^2 + BS_x^2 - 2BS_{xy}),\ j = 1,2,3,$

where $B_{(rob)l}$ indicates the robust regression techniques (Huber M-estimate, Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate) used to the ratio estimators proposed in the present study. Let

$2R_j S_x^2 (B_{(rob)l} - B) - 2S_{xy}(B_{(rob)l} - B) + S_x^2 (B_{(rob)l}^2 - B^2) < 0.$

Hence,

$(B_{(rob)l} - B)[2R_j S_x^2 - 2S_{xy} + S_x^2 (B_{(rob)l} + B)] < 0.$

For $B_{(rob)l} - B > 0$, that is, and so

$2R_j S_x^2 - 2S_{xy} + S_x^2 (B_{(rob)l} + B) < 0.$

Therefore, $\left(B_{(rob)l} + B\right) < -2R_j + 2B$, where $B = \left(2S_{xy} \big/ S_x^2\right)$.

$$\begin{aligned} B_{(rob)l} &< -2R_j + 2B - B \\ &< B - 2R_j. \end{aligned}$$

Similarly, for $B_{(rob)l} - B < 0,$ that is $B_{(rob)l} < B$:, and hence $B_{(rob)l} > B - 2R_j$.

Consequently, we have the following conditions:

$$0 < B_{(rob)l} - B < 2R_j \quad (38)$$

or

$$-2R_j < B_{(rob)l} - B < 0. \quad (39)$$

If one of the conditions (38) or (39) is satisfied, the proposed estimators using the mentioned robust regression methods are more efficient than the usual ratio estimators based on the OLS method.

5 Numerical illustration

For numerical illustration, a real data set is selected from Division of Agricultural

Statistics, Faculty of Horticulture Shalimar in which the data of apple production amount (as an interest of variate) and the number of apple trees (as an auxiliary variate) in 499 villages of District Baramulla of Jammu and Kashmir from 2010 to 2011. (Source: RCM project, pilot survey for estimation of cultivation and production of apple in District Baramulla, RCM approved project). First, we have stratified the data by area wise and from each stratum (region), and the samples (villages) have been selected randomly. Here, we have taken the sample size to 170. We joined two areas, then chose four strata where each one contains three blocks (as 1: Zaniger, Boniyar, Tangmarg; 2: Wagoora, Sopore, Baramulla; 3: Uri, Pattan, Rohama; 4: Rafiabad, Kunzer, Singapore) for this data. However, in the present study, we have used only the data of Uri, Pattan, Rohama of district Baramulla of Jammu and Kashmir, due to the interest in simple random sampling. We have applied our proposed ratio estimators on the data of apple production amount and number of apple trees in 117 villages of Uri, Pattan, Rohama of district Baramulla of Jammu and Kashmir, in which the apple production (in tons) is denoted by Y (study variable), and the number of apple trees is denoted by X (auxiliary variable, 1 unit = 100 trees). The characteristics of the data set are given in the Tab. 1, and the statistical analysis of the suggested estimators is carried in Tab. 2. Tab. 3 provides the % RE of the proposed estimators using the OLS method with the suggested estimators using Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate. Similarly, Tab. 4 presents the % RE of suggested estimators using Huber M-estimates with the proposed estimators using Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate.

Table 1: Description of apple trees data set

Population (Apple Data 2010-2011)		
$N = 117$	$C_y = 0.9728$	$Q_3 = 701.0$
$n = 40$	$S_x = 235.5$	$B = 3.19$
$\bar{Y} = 1263$	$C_x = 0.7395$	$B_{rob(Huber)} = 2.16$
$\bar{X} = 560.0$	$S_k = 0.6654$	$B_{rob(Mallows)} = 1.01$
$\rho = 0.987$	$Q_1 = 300.1$	$B_{rob(Schweppe)} = 0.96$
$S_y = 862$	$Q_2 = 600.6$	$B_{rob(SIS\,GM)} = 0.85$

Table 2: Statistical analysis of the suggested estimators

Estimators	Constant	OLS method	Huber M estimate	Mallows GM estimate	Schweppe GM estimate	SIS GM estimate
1	2.155	37014.83	27755.01	20278.93	20022.39	19478.11
2	1.490	30756.43	23061.56	17332.77	17152.20	16775.05
3	1.350	29567.41	22202.09	16841.23	16676.66	16334.70

Table 3: The % RE of the estimators using OLS with the estimators using Mallows GM, Schweppe GM and SIS GM estimates

Estimators	OLS/Mallows GM estimate	OLS/Schweppe GM estimate	OLS/SIS GM estimate
1	182.53	184.87	190.03
2	177.45	179.31	183.35
3	175.57	177.30	181.01

Table 4: The % RE of the estimators using Huber M estimation with the estimators using Mallows GM, Schweppe GM, and SIS GM estimates

Estimators	Huber M/Mallows GM estimate	Huber M /Schweppe GM estimate	Huber M /SIS GM estimate
1	136.87	138.62	142.49
2	133.05	134.45	137.48
3	131.83	133.13	135.92

The empirical results in Tab. 2, show that the proposed estimators using Huber M-estimate, Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate are more efficient than the suggested estimators based on the OLS methods due to the smallest mean square error values. However, from Tab. 3, one can conclude that, the suggested estimators based on the SIS GM-estimate are more efficient than other robust methods considered in this study. Similarly, Tab. 4 indicates that the suggested estimators using the SIS GM-estimate are superior to other estimators in the current study.

6 Simulation study

To investigate the usefulness of the proposed estimators relative to the OLS, Huber-M, Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate in the case of skewed distributions, the Gamma distribution as an example, a simulation study is carried out. Five thousand samples are generated using the SRSWOR method. Then, the values of $\hat{\bar{Y}}_i$ are calculated for 5000 times from the 5000 samples. The mean squared error values are obtained as follows

$$MSE = \frac{1}{5000}\sum_{i=1}^{5000}\left(\hat{\bar{Y}}_i - \bar{Y}\right)^2, \tag{40}$$

where $\hat{\bar{Y}}_i$ represents the estimated mean squared error for $i = 1,2,\cdots,5000$ and $\bar{Y}$ is the population mean. Different sample sizes such as $n = 20,30,40,50,60$ are considered in this study to investigate performance of the suggested estimators using Mallows GM-estimate, Schweppe GM-estimate and SIS GM-estimate compared to the estimators using the OLS and Huber-M. The relative efficiency is defined by

$$RE\left(\hat{\bar{Y}}_{p(MAllows,Schweppe,SIS)}\right)=\frac{MSE\left(\hat{\bar{Y}}_{p(OLS)}\right)}{MSE\left(\hat{\bar{Y}}_{p(MAllows,Schweppe,SIS)}\right)}\times 100 \tag{41}$$

and

$$RE\left(\hat{\bar{Y}}_{p(MAllows,Schweppe,SIS)}\right)=\frac{MSE\left(\hat{\bar{Y}}_{p(Huber\,M)}\right)}{MSE\left(\hat{\bar{Y}}_{p(MAllows,Schweppe,SIS)}\right)}\times 100 \tag{42}$$

The results are summarized in Tab. 5. It turns out that while using the OLS method, the estimators do not rely on the precise results in case of outliers. Then, by adopting the above-mentioned robust regression techniques, the suggested estimators perform better, and as the sample size increasing, these estimators seem to be much better. Also the Mallows GM-estimate, Schweppe GM-estimate, and SIS GM-estimate are better than the Huber-M estimate. Moreover, as the sample size increases, these techniques give precise results in the presence of outliers.

Table 5: The % RE of the estimators (Est.) using OLS, Huber-M estimation with the estimators using Mallows GM, Schweppe GM, and SIS GM-estimates

N	Est.	OLS/ Mallows GM	OLS/ Schweppe GM	OLS/ SIS GM	Huber M/ Mallows GM	Huber M/ Schweppe GM	Huber M/ SIS GM
20	1	121.78	123.67	130.67	115.08	117.89	120.08
	2	118.92	121.89	127.03	113.67	115.78	118.76
	3	115.01	119.98	125.98	111.76	112.03	116.08
30	1	126.67	130.76	134.45	118.56	121.06	125.90
	2	123.47	128.07	132.33	117.01	119.67	123.56
	3	119.07	126.05	130.97	115.67	117.74	121.82
	1	131.78	135.98	139.41	120.54	124.31	128.90
40	2	128.90	132.69	137.97	119.07	122.76	127.32
	3	125.67	130.56	136.01	117.89	120.09	125.92
	1	133.68	139.99	144.78	123.09	127.89	130.54
50	2	132.45	137.87	142.37	121.67	125.04	129.05
	3	131.09	134.87	140.90	120.79	123.21	128.21
60	1	141.09	143.89	148.78	125.67	129.06	133.89
	2	139.56	141.69	145.87	123.89	128.42	131.34
	3	136.08	139.99	144.03	122.03	127.08	129.67

7 Conclusion

The results of this study revealed that by adopting the robust methods, Mallows GM-estimates, Schweppe GM-estimates, and SIS GM-estimates, the proposed estimators of the population mean perform better than their competitors based on the OLS and Huber-M methods. Hence, we strongly recommend considering the suggested estimators using Mallows GM-estimates, Schweppe GM-estimates, and SIS GM-estimate to estimate the population parameters as compared to the OLS and Huber-M estimation methods in the presence of outliers. The suggested estimators in this paper can be modified using other sampling methods as ranked set sampling and median ranked set sampling methods. See for illustration [Haq, Brown, Moltchanova et al. (2016a, 2016b); Al-Omari and Haq (2019); Haq, Brown, Moltchanova et al. (2015); Al-Nasser and Al-Omari (2018); Jemain, Al-Omari and Ibrahim (2007); Zamanzade and Al-Omari (2016)].

Acknowledgment: The authors would like to thank the anonymous reviewers for their insightful comments and suggestions on the manuscript.

Funding Statement: A. R. A. Alanzi would like to thank the Deanship of Scientific Research at Majmaah University for financial support and encouragement.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Al-Nasser, D. A.; Al-Omari, A. I. (2018): MiniMax ranked set sampling. *Revista Investigacion Operacional*, vol. 39, no. 4, pp. 560-570.

Al-Omari, A. I. (2012): Ratio estimation of population mean using auxiliary information in simple random sampling and median ranked set sampling. *Statistics and Probability Letters*, vol. 82, no. 11, pp. 1883-1990.

Al-Omari, A. I.; Al-Nasser, A. D. (2018): Ratio estimation using multistage median ranked set sampling approach. *Journal of Statistical Theory and Practice*, vol. 12, no.3, pp. 512-529.

Al-Omari, A. I.; Bouza, C. N. (2015): Ratio estimators of the population mean with missing values using ranked set sampling. *Environmetrics*, vol. 26, no. 2, pp. 67-76.

Al-Omari, A. I.; Haq, A. (2019): A new sampling method for estimating the population mean. *Journal of Statistical Computation and Simulation*, vol. 89, no. 11, pp. 1973-1985.

Al-Omari, A. I.; Ibrahim, K.; Jemain, A. A. (2009): New ratio estimators of the mean using simple random sampling and ranked set sampling methods. *Revista Investigacion Operacional*, vol. 30, no. 2, pp. 97-108.

Al-Omari, A. I.; Jaber; Ibrahim, K. (2008): Modified ratio-type estimators of the mean using extreme ranked set sampling, *Journal of Mathematics and Statistics,* vol. 4, no. 3, pp. 150-155.

Bouza, C. N.; Al-Omari, A. I.; Santiago, A.; Sautto, J. M. (2017): Ratio type estimation using the knowledge of the auxiliary variable for ranking and estimating. *International Journal of Statistics and Probability*, vol. 6, no. 2, pp. 21-30.

Coakley, C. W.; Hettmansperger, T. P. (1993): A bounded influence, high breakdown, efficient regression estimator. *Journal of American Statistical Association*, vol. 88, pp. 872-880.

Cochran, W. G. (1977): *Sampling Techniques, 3rd Edition*, John Wiley and Sons, New York.

Hampel, F. R. (1971): A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1887-1896.

Handschin, E.; Kohlas, J.; Fiechter, A.; Schweppe, F. (1975): Bad data analysis for power system state estimation. *IEEE Transactions on Power Apparatus and Systems*, vol. 2, pp. 329-337.

Haq, A.; Brown, J.; Moltchanova, E.; Al-Omari, A. I. (2015): Varied L ranked set sampling scheme. *Journal of Statistical Theory and Practice*, vol. 9, pp. 741-767.

Haq, A.; Brown, J.; Moltchanova, E.; Al-Omari, A. I. (2016a): Best linear unbiased and invariant estimation in location-scale families based on double ranked set sampling. *Communications in Statistics-Theory and Methods*, vol. 45, no. 1, pp. 225-248.

Haq, A.; Brown, J.; Moltchanova, E.; Al-Omari, A. I. (2016b): Paired double ranked set sampling. *Communications in Statistics-Theory and Methods*, vol. 45, no. 1, pp. 2873-2889.

Huber, P. J. (1973): Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pp. 799-821.

Jemain, A. A.; Al-Omari, A. I.; Ibrahim, K. (2007): Multistage extreme ranked set samples for estimating the population mean. *Journal of Statistical Theory and Applications*, vol. 6, no. 4, pp. 456-471.

Jemain, A. A.; Al-Omari, A. I.; Ibrahim, K. (2008): Modified ratio estimator for the population mean using double median ranked set sampling. *Pakistan Journal of Statistics*, vol. 24, no. 3, pp. 217-226.

Krasker, W. S. (1980): Estimation in linear regression models with disparate data points. *Econometrica*, vol. 48, pp. 1333-1346.

Krasker, W. S.; Welsch, R. E. (1982): Efficient bounded-influence regression estimation. *Journal of American Statistical Association*, vol. 77, pp. 595-604.

Mallows, C. L. (1975): On some topics in robustness. Unpublished memorandum, Bell Tel. Laboratories, Murray Hill.

Nadia, H.; Mohammad, A. A. (2013): Model of robust regression with parametric and nonparametric methods. *Mathematical Theory and Modeling*, vol. 3, pp. 27-39.

Rousseeuw, P. J.; Leroy, A. M. (1987): Robust regression and outlier detection. *Wiley Series in Probability and Mathematical Statistics*, New York.

Subzar, M.; Bouza, C. N.; Al-Omari, A. I. (2019b): Utilization of different robust regression techniques for estimation of finite population mean in SRSWOR in case of presence of outliers through ratio method of estimation. *Revista Investigacion Operacional*, vol. 40, no. 5, pp. 600-609.

Subzar, M.; Bouza, C. N.; Maqbool, S.; Raja, T. A.; Para, B. A. (2019a): Robust ratio type estimators in simple random sampling using Huber M estimation. *Revista Investigacion Operacional*, vol. 40, no. 2, pp. 201-209.

Subzar, M.; Maqbool, S.; Raja, T. A.; Muhammad, A. (2018): Ratio estimators for estimating population mean in simple random sampling using auxiliary information. *Applied Mathematics & Information Sciences Letters*, vol. 6, no. 3, pp. 123-130.

Subzar, M.; Maqbool, S.; Raja, T. A.; Sharma; P. (2019): A new ratio estimator: an alternative to regression estimator in survey sampling using auxiliary information. *Statistics in Transition New Series*, vol. 20, no. 4, pp. 181-189.

Tukey, J. W. (1977): *Exploratory Data Analysis*. MA: Addison-Wesley.

Wolter K. M. (1985): *Introduction to Variance Estimation*, Springer-Verlag.

Yohai, V. J. (1987): High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, vol. 15, no. 2, pp. 642-656.

Yu, C.; Yao, W. (2017): Robust linear regression: a review and comparison. *Communications in Statistics-Simulation and Computation*, vol. 46, no. 8, pp. 6261-6282.

Zamanzade, E.; Al-Omari, A. I. (2016): New ranked set sampling for estimating the population mean and variance. *Hacettepe Journal of Mathematics and Statistics*, vol. 45, no. 6, pp. 1891-1905.

ZeinalovaLala, M.; Huseynov, O. H.; Sharghi, P. (2018): A Z-Number valued regression model and its application. *Intelligent Automation and Soft Computing*, vol. 24, no. 1, pp. 187-191.